STRUCTURE AND ENERGY-BASED ANALYSES OF FGFR2 KINASE MUTATIONS REVEALING DIFFERENCES IN CANCER AND SYNDROME MUTATIONS AND INCONCLUSIVE NATURE OF ENERGY ANALYSIS FOR THE MUTANTS

by

Snehal Vilas Sambare

A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Bioinformatics and Computational Biology

Committee:

_____	Dr. Amarda Shehu, Thesis Director
_____	Dr. Donald Seto, Committee Member
_____	Dr. Dmitri Klimov Committee Member
_____	Dr. Iosif Vaisman, Director, School of Systems Biology
_____	Dr. Donna Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: _____	Summer Semester 2019 George Mason University Fairfax, VA

Structure and Energy-based Analyses of FGFR2 Kinase Mutations Revealing Differences in Cancer and Syndrome Mutations and Inconclusive Nature of Energy Analysis for the Mutants

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Snehal Vilas Sambare
Bachelor of Engineering
Shri Ramdeobaba College of Engineering and Management

Director: Amarda Shehu, Professor
Computer Science Department

Summer Semester 2019
George Mason University
Fairfax, VA

DEDICATION

This thesis is dedicated to my grandparents, parents and all my friends for their endless love, support and encouragement.

ACKNOWLEDGEMENTS

There are many people who walked alongside me during my master's journey. They have guided me, placed opportunities in front of me and guided me to overcome challenges. I would like to thank each one of them. I would like to thank my advisor, Dr. Amarda Shehu for guiding me in this work and steering me in the right direction whenever needed. I would also like to thank Yiyan Lian for all the insightful discussions we had that helped in this research. I would also like to thank my committee member Dr. Donald Seto and Dr Dimitri Klimov for being supportive throughout this work.

Finally, I would like to thank my grandparents and parents, whose love and guidance are with me in whatever I pursue. A very big thanks goes to my friends for being there throughout this journey.

Thank you, Lord, for always being there.

This thesis is only a beginning of my journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Fibroblast Growth Factor Receptor 2 ..... FGFR2
Fibroblast Growth Factor ..... FGF
Familial Scaphocephaly ..... FSPC
Crouzon Syndrome ..... CS
Pfeiffer Syndrome ..... PS
Lacrimo-auriculo-dento-digital syndrome ..... LADDS
Blocks Substitution Matrix ..... BLOSUM
Single nucleotide polymorphisms ..... SNP
Universal Protein Resource ..... UniProt
Research Collaboratory for Structural ..... RCSB
Bioinformatics-Protein Data Bank ..... PDB
Catalogue of Somatic Mutations in Cancer ..... COSMIC
National Center for Biotechnology Information ..... NCBI
Hidden Markov Model ..... HMM
Protein Variability Server ..... PVS
Yet Another Scientific Artificial Reality Application ..... YASARA
Wild Type ..... WT

ABSTRACT

STRUCTURE AND ENERGY-BASED ANALYSES OF FGFR2 KINASE MUTATIONS REVEALING DIFFERENCES IN CANCER AND SYNDROME MUTATIONS AND INCONCLUSIVE NATURE OF ENERGY ANALYSIS FOR THE MUTANTS

Snehal Vilas Sambare, M.S.

George Mason University, 2019

Thesis Director: Dr. Amarda Shehu

Fibroblast growth factor receptor 2 (FGFR2) is a protein in humans encoded by gene FGFR2. It plays an important role in the regulation of cell proliferation, differentiation, migration and apoptosis, and in the regulation of embryonic development. Mutations in FGFR2 gene are associated with numerous medical conditions that include craniosynostosis syndromes (abnormal bone development) and various cancers. In fact, FGFR2 is shown to be activated in many cancers through the mechanisms of gene amplification, translocations, and point mutations. There remains many FGFR2 mutations whose effects are unknown. In this work we have investigated point mutations in FGFR2 kinase. We have performed region-based analysis wherein we mapped mutations to various domains of protein and performed Shannon entropy analysis on the mutant positions. BLOSUM matrix values were also obtained for the mutations to get insights

about differences in amino acids substitution for cancer and syndromes. Structure energy-based analysis was performed using FoldX, a protein design algorithm. Statistical analysis like Normality tests, T-tests, Mann-Whitney-Wilcoxon Tests were performed on the energy values obtained from FoldX, and histograms were generated. This analysis makes the following contributions. The region-based analysis shows that cancer causing mutations are distributed across all regions, whereas syndrome causing mutations are not uniformly distributed across all domains. The BLOSUM analysis reveals that in cancer causing mutations substitution takes place between amino acids with similar physicochemical properties, whereas in syndrome causing mutations all types of amino acids can be substituted. The structure-energy based, and statistical analysis shows that cancer causing and syndrome causing mutations have identical energy distributions, indicating that energy cannot be used as predictor for differentiating cancer causing and syndromes causing mutants in FGFR2. The results of histogram analysis are inconclusive. In summary, this study has provided interesting insights that can be helpful for further research of FGFR2 kinase mutations.

INTRODUCTION

Fibroblast growth factor 2 (FGFR2) is a protein in humans encoded by gene FGFR2 [1]. The amino acid sequence for this protein is highly conserved between members and throughout evolution [2]. The entire protein consists of three immunoglobulin domains, a single hydrophobic membrane spanning segment and cytoplasmic tyrosine kinase domain (FIGURE 1) [2]. The extracellular portion of this protein interacts with fibroblast growth factors resulting in dimerization and cascade of downstream signals which influence mitogenesis and differentiation [2]. Tyrosine kinases occupy a central role in cellular regulation, acting as intermediaries in relaying signals from extracellular ligands to major signaling pathways in the cell [3]. FGFR2 has two natural isoforms which are created by splicing of third immunoglobulin domain. FGFR2lllb is found in skin and internal organs, whereas FGFR2lllc is found in mesenchyme which includes craniofacial bone [2]. These two isoforms differ in their bindings to the ligand [4]. FGFR2 plays an important role in the regulation of cell proliferation, differentiation, migration and apoptosis, and in the regulation of embryonic development. It is required for normal embryonic patterning, trophoblast function, limb bud development, lung morphogenesis, osteogenesis, skin development and normal skeletal development [5].

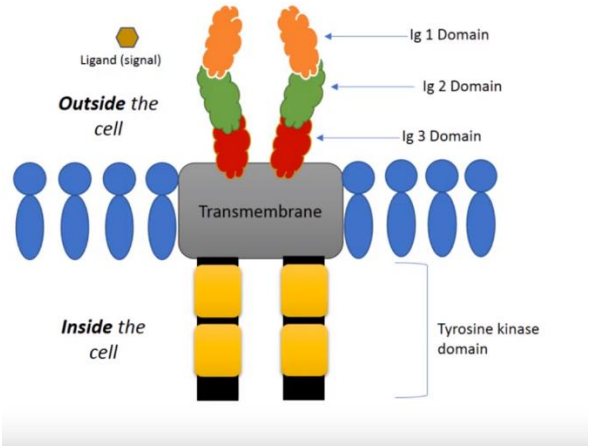

Figure 1. Structure of FGFR2.Three extracellular immunoglobulin domains Ig1, Ig2, Ig3; transmembrane and cytoplasmic tyrosine kinase domain. (https://www.youtube.com/watch?v=1Y1el3_vQ3Y)

Mutations in FGFR2 are associated with numerous medical conditions that includes Craniosynostosis syndromes and cancer. Craniosynostosis syndromes are skull malformations which are caused by premature fusion of cranial structures and other features which are dependent on the mutation. The syndromes are Apert syndrome, Jackson-Weiss syndrome, Beare-Stevenson cutis gyrata syndrome, Saethre-Chotzen syndrome, and syndromic craniosynostosis. Gain of function mutations in FGFR2 kinase cause diseases like Familial scaphocephaly syndrome (FSCP), Crouzon Syndrome (CS), Pfeiffer Syndrome (PS) [6][7]. Lacrimo-auriculo-dento-digital syndrome (LADDS) is cause by reduced tyrosine kinase activity in FGFR2 mutants [8]. Missense mutations of FGFR2 have been found in endometrial cancers [9], cervical cancer [10] breast cancer [11] and melanoma [12] resulting in loss of function in some cases.

There are different types of mutations. Missense mutations result in change of an amino acid, nonsense mutations result in change of an amino acid to STOP codon which cause

premature termination of translation, silent mutations have no effect and frameshift mutations are caused by deletion or insertion of multiple bases resulting in lots of amino acid changes. Missense mutations can result in loss of function or gain of function in proteins. Mapping of various mutations onto different areas of protein structures has led to valuable insights into molecular mechanism underlying a disorder [13]. Potential energy of a protein molecule is the sum of many different components which are related to the internal structure and its interaction with other molecules. Backbone configuration, side chain interactions, Van der Waals clashes, electrostatics interactions, bond length, bond angles, torsion angles contribute to determine the potential energy of a protein molecule [14]. Mutations in protein can lead to changes in any of these, resulting in energy changes. Studying these energy changes can give us new insights about the mutants. Analyzing the energy changes of the protein mutants has shown that there could be differences between the cancer causing mutants and syndrome causing mutants [15].

There remains many FGFR2 mutations whose effects are unknown. In this study we are specifically interested in studying mutations in tyrosine kinase domain of the protein which is involved in regulation of catalytic activity (Figure 1). FGFR2 kinase activation loop toggles between two states, active and inactive, so the mutations in this region can perturb the balance between these states, resulting in various medical conditions like syndromes and cancers [3]. Performing structure-energy based analyses of these structures can help us in determining the effect mutations have on the structures and can help in determining the relationship between the structure and disorder type (cancer or syndrome), if any. In this study we are mapping the mutations to all domains of the

protein, calculating the Shannon entropy values, BLOSUM matrix values and the changes in the potential energies of mutants. With the help of FoldX, which is a protein design algorithm, we wish to understand the effect of mutations on the protein. This research can give us new insights about the differences in the cancer causing and syndrome causing mutants.

SPECIFIC AIMS

1. Region based analysis

1.1 Mapping mutations to domains of protein

Proteins are composed of one or more functional regions called as domains. Different combination of these domains produces diverse range of proteins. Mapping various disease missense mutations can shed light on molecular mechanism of the disease [17]. Previously, differences in distribution of missense mutations in various domains were observed in different diseases after such mapping [18]. So, in this study we decided to map mutations onto the domains of FGFR2 protein to understand whether cancer causing and syndrome causing mutants fall into distinct groups.

1.2 To perform Shannon Entropy Analysis

Shannon entropy gives a measure for sequence conservation [19]. It is observed that mutated amino acid positions are highly conserved across species. Here we studied evolutionary sequence conservation of the positions where the mutation has taken place.

2. Substitution diverseness analysis of mutations

BLOSUM matrix values give information about substitution diverseness of amino acids [20]. BLOSUM matrix values for the mutants were obtained to understand the substitution diverseness for cancer causing and syndrome causing mutants.

3. Structure and energy-based analysis

3.1 FoldX Analysis

FoldX is a protein design algorithm that uses an empirical force field. It predicts the effect of point mutations or human SNPs on protein stability or protein complexes [16]. The mutations were performed in FoldX and the energies of the mutant structures were obtained. The values were plotted to find any differences in the cancer causing and syndrome causing mutants.

3.2 Statistical analysis of the energy values obtained from FoldX

Statistical analysis can give us new insights about the differences in cancer causing and syndrome causing mutants. In this research we have performed T-tests, Mann-Whitney-Wilcoxon Test and Histogram analysis on the energy values obtained from FoldX to understand the differences in cancer causing and syndrome causing mutants.

MATERIALS AND METHODS

This section summarizes databases and scripts used in this study. Then we describe the techniques employed to conduct the region-based analysis, substitution diverseness analysis and structure energy-based analysis.

1.Databases & Scripts

UniProt (Universal Protein Resource) is a database of proteins which contains information about the sequences and functions (https://www.uniprot.org/). It also has information about the mutations, structures, expression and diseases associated with the proteins [21].

RCSB-PDB (Protein data bank) is database of structures of large biological molecules like proteins and nucleic acid [22] (https://www.rcsb.org/).

COSMIC (Catalogue of Somatic Mutations in Cancer) is largest manually created database consisting of information about human cancers. It catalogues information about the type of cancers, mutations and cell lines [23]. (https://cancer.sanger.ac.uk/cosmic)

NCBI (National Center for Biotechnology Information) consists of different databases related to biotechnology and biomedicine which are very useful in bioinformatics analyses. (https://www.ncbi.nlm.nih.gov/) In this study NCBI HomoloGene was used to obtain sequences. It is tool which gives homologs among genes of various eukaryotic genomes [24]. (https://www.ncbi.nlm.nih.gov/homologene)

Clustal Omega is a tool is which is used for multiple sequence alignment of three or more sequences. It uses seeded guide trees and HMM (Hidden Markov Model) profile-profile technique for sequence alignment [25]. (https://www.ebi.ac.uk/Tools/msa/clustalo/)

Protein Variability Server (PVS) is a web server which is used for studying protein sequence variability. It uses several variability metrics to compute the absolute site variability in multiple protein-sequence alignments [26]. (http://imed.med.ucm.es/PVS/)

The scripts were written in R language [27] (Appendix III). Bio3D package in R has utilities for performing analysis on protein structures and sequences [28]. The protein chains were separated using methods from this package. Ggplot2 [29], reshape2 [30], plotrix[31] and xlsx were the packages used for getting histograms and plots. Shapiro-Wilk Normality Test, T-test and Mann-Whitney-Wilcoxon Test were performed in R on the datasets.

In this study we use Isoform 1 of FGFR2 human gene (Appendix I) consisting of 821 amino acids from UniProt database (P21802) as the reference. Specifically, our focus is on tyrosine kinase domain which ranges from 399 to 821 amino acids of the sequence. All the driver mutations in this region and their types were obtained from UniProt, RCSB-PDB and COSMIC.

2.Region based analysis

2.1 Mapping mutations to protein domains

For domains analysis the domains mentioned in Family & Domains section of UniProt database were used for mapping. The Pathology & Biotech section gives information about the mutations. From this section, number of mutations in each domain of the gene for the syndromes were obtained. COSMIC database has mapping of cancer causing mutations on the gene sequence. From this data, number of mutations in each domain were calculated.

2.2 Shannon Entropy Analysis

For a multiple protein sequence alignment, the Shannon entropy (H) for every position is as follow:

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i$$

Equation 1. Shannon Entropy Formula

Where Pi is the fraction of residues of amino acid type i, and M is the number of amino acid types (20). H ranges from 0 (only one residue in present at that position) to 4.322 (all 20 residues are equally represented in that position). Typically, positions with H $\geq$2.0 are considered variable, whereas those with H $\leq$ 2 are considered as conserved. Highly conserved positions are those with H $\leq$1.0. A minimum number of sequences is however required (~100) for H to describe the diversity of a protein family.

For calculation of Shannon entropy, protein sequences were obtained from UniProt and NCBI. The list of organisms with FGFR2 gene were obtained from NCBI-HomoloGene. The sequences for these organisms were obtained from UniProt. For every organisms UniProt gives a list of sequences. Only those sequences were considered whose length was close to 821. If there were multiple sequences with length close to 821, the sequence length mentioned in results obtained from NCBI- HomoloGene was considered. Organisms whose sequences were not in UniProt were obtained from NCBI-HomoloGene. Considering this, the sequences considered were as follows: HomoSapiens: P21802, Pan troglodytes: H2Q2P3, Canis lupus familiaris: F1PPD8, Bos Taurus: F1MNW2, Mus musculus: E9QK53, Rattus norvegicus: F1LNW0, Gallus gallus: F1NEE9, Danio rerio: Q8JG38, Xenopus tropicalis: A4IHW8. These sequences were aligned using Clustal Omega with alignment to be outputted according to the input sequences. Protein Variability Server PVS was used to calculate the Shannon Entropy for the aligned sequences. In sequence variability options, reference sequence was set to first sequence in alignment file. First sequence is human FGFR2 with 821 amino acid sequences. Shannon entropy values for the mutant positions were listed.

3. Substitution diverseness analysis

BLOcks SUbstitution Matrix (BLOSUM) is a substitution matrix based on local alignments which is used for sequence alignment of proteins. BLOCKS databases are scanned for very conserved region of protein families and then relative frequencies of amino acids and their substitution probabilities are calculated. BLOSUM matrix has log-odds score for each of the 201 possible substitution pairs of the 20 amino acids. In this

study, BLOSUM62 matrix was used [20]. A score of zero indicates that the frequency with which given two amino acids were found aligned in the database was as expected by chance, while a positive score indicates that the alignment was found more often than by chance, and negative score indicates that the alignment was found less often than by chance. For each mutation, by considering original amino acid as row entry and substituted amino acid pair as column entry, BLOSUM62 score was obtained.

4. Structure and energy-based analysis

4.1 FoldX Analysis

FoldX is protein design algorithm empirical force field. It is used to predict the effect of point mutations on the stability of proteins. The energy function includes terms that have been found to be very important for protein stability, where the energy of unfolding (ΔG) of a target protein is given by equation:

$$\begin{aligned}\Delta G = {} & \mathrm{Wvdw} * \Delta\mathrm{Gvdw} + \mathrm{WsolvH} * \Delta\mathrm{GsolvH} + \mathrm{WsolvP} * \Delta\mathrm{GsolvP} \\ & + \Delta\mathrm{Gwb} + \Delta\mathrm{Ghbond} + \Delta\mathrm{Gel} + \Delta\mathrm{GKon} + \mathrm{Wmc} * \mathrm{T} * \Delta\mathrm{Smc} \\ & + \mathrm{Wsc} * \mathrm{T} * \Delta\mathrm{Ssc}\end{aligned}$$

Equation 2. Stability of Protein given by FoldX in kcal/mol.

with:

- ΔGvdw as the sum of the van der Waals contributions of all atoms with respect to the same interactions with the solvent.

- ΔGsolvH and ΔGsolvP as the differences in solvation energy for apolar and polar groups, respectively, when these change from the unfolded to the folded state.
- ΔGhbond as the free energy difference between the formation of an intra-molecular hydrogen bond and intermolecular hydrogen bond.
- ΔGwb as the extra stabilizing free energy provided by a water molecule making more than one hydrogen bond to the protein (water bridges) that cannot be considered with non-explicit solvent approximations.
- ΔGel as the electrostatic contribution of charged groups, including the helix dipole.
- T * ΔSsc as the entropy cost of fixing the backbone in the folded state.
- ΔSsc as the entropic cost of fixing a side chain in a conformation.

ΔΔG is the energy difference between the stability of the mutant and the stability of original structure.

For studying mutations in FoldX, wild type structures of FGFR2 kinase domain were obtained from RCSB-PDB. Structure 2PSQ is inactive form of FGFR2 kinase with amino acids from 468-765. It has two chain A and B. The pdb files for two chains was separated using Bio3D package in R. Mutation on both chains were studied. Structure 2PVF is active from of FGR2 kinase with amino acids from 458-778. It has two chains Chain A and Chain B; Chain B has 15 residues. So, mutations were studied only on Chain A.

Yet Another Scientific Artificial Reality Application (YASARA) is a computer program for molecular visualizing, modelling, and dynamics. FoldX plugin is available for YASARA which gives all the functionalities of FoldX. So, all the mutations were studied in YASARA. The original structures were repaired in FoldX using "RepairObject". "RepairObject" identifies those residues which have bad torsion angles, or Van Der Waals' clashes, or total energy, and repairs them. It self-mutates residues with high energy to low energy. FoldX command "Stability" gives the stability of protein in kcal/mol. FoldX "Mutate Residue" option, mutates a single residue on the main chain of the wild type protein structure. This outputs the energy difference (ΔΔG) between mutant structure and Wild Type structure, called as FoldX energy. A total of 28 mutations were performed on each of the Wild Type structures. For every mutation, ΔG (stability of the mutant), ΔΔG (FoldX energy) and manually calculated ΔΔG (stability of mutant-stability of Wild Type referred as mutant-WT) were noted.

4.2 Statistical analysis

Gold Sets are subsets of the entire datasets which are created based on different conditions for statistical analysis. Considering the standard deviations of 0.8 [13] and 0.46 [14] in the ΔΔG of the mutants, the Gold Sets were created. For every structure, four Gold sets were created, mutants having absolute values of mutant-WT (stability of mutant-stability of Wild Type) greater than 0.8 and other set with absolute values of mutant-WT greater than 0.46; mutants having absolute values of FoldX energy greater than 0.8 and other set with absolute values of FoldX energy greater than 0.46. Each set has two categories, syndromes and cancers. Normality Tests were performed on stability,

mutant-WT, FoldX energy of cancer causing and syndrome causing mutants of each set. Based on the results, T-tests and Mann-Whitney-Wilcoxon Tests were performed on cancer causing and syndrome causing mutants of each dataset. Histograms were generated for plotting the number of mutations based on energy values.

RESULTS AND DISCUSSION

In this section, the results of the region-based analysis, substitution diverseness analysis and structure energy-based analysis are discussed in detail.

1.Region based analysis

FGFR2 protein has three immunoglobulin domains; Ig C2 type 1, Ig C2 type 2 and Ig C2 type 3, along with single protein kinase domain. On mapping the number on mutation to these domains following results were obtained, see Figure 2.

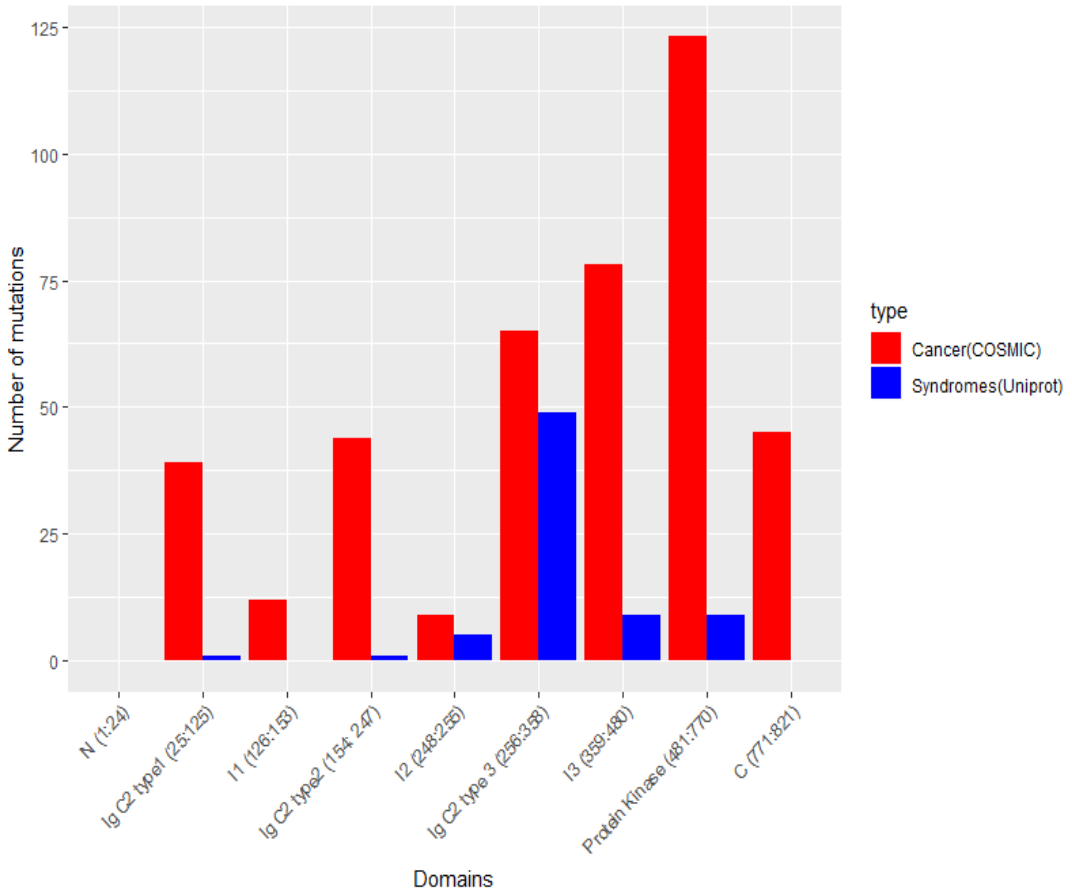

Figure 2. Mapping of mutations to protein domains. N and C are terminal regions, I1, I2, I3 are inter structural regions, Ig is Immunoglobulin domain. Cancer mutations are present in all regions except N terminal unlike syndrome mutations.

In the plot, N and C are N terminal and C terminal regions and I1, I2, I3 are inter structural regions. Cancer mutations include driver as well as passenger mutations in the protein. All the regions have cancer mutations except the N-terminal regions with highest number of mutations in protein kinase region. Syndrome mutations are concentrated from Immunoglobulin domain 3 and kinase domain as compared to the remaining part of the protein.

The Shannon entropy values for 399 to 821 amino acids are listed in Table 1.

Table 1. Shannon Entropy values for mutants. Shannon Entropy values < 1, so all mutant positions are highly conserved.

Position	Mutation	Shannon Entropy value	Phenotype
474	W474X	0	Cancer
475	E475K	0	Cancer
526	K526E	0	Syndrome
530	D530N	0	Cancer
547	I547V	0	Cancer
549	N549H	0	Syndrome
549	N549T	0	Syndrome
549	N549K	0	Cancer
565	E565G	0	Syndrome
565	E565A	0	Syndrome
574	E574K	0.53	Cancer
628	A628T	0	Syndrome
636	E636K	0	Cancer
640	M640I	0	Cancer
641	K641R	0	Syndrome
642	I642V	0	Cancer
648	A648T	0	Cancer,Syndrome
649	RD649-650S	0	Syndrome
650	D650V	0	Syndrome
659	K659Q	0	Syndrome
659	K659T	0	Syndrome
659	K659E	0	Cancer
659	K659M	0	Cancer
659	K659N	0	Cancer,Syndrome
663	G663E	0	Syndrome
678	R678G	0	Syndrome
688	S688F	0	Cancer
701	G701S	0	Cancer
708	P708S	0	Cancer
759	R759X	0	Cancer
759	R759Q	0	Cancer
770	L770V	0	Cancer

Table 1 provides documentation for cancer causing as well as syndrome causing mutations. In the table X represents a translation termination codon. Positions that have values < 1.0 are considered as highly conserved. From Table 1 it is evident that all the positions where the mutations are present are highly conserved across species for cancer causing as well as syndrome causing mutants.

2.Substitution diverseness analysis

BLOSUM62 matrix that was used is shown in Figure 3.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Figure 3. BLOSUM62 matrix.

BLOSUM matrix values can be obtained for point amino acid mutation where the new amino acid is known. Considering this, the BLOSUM matrix values for the mutants are in Table 2.

Table 2. BLOSUM matrix values for the mutants.

Mutant	BLOSUM matrix value	Phenotype
K526E	1	Syndrome
N549H	1	Syndrome
N549T	0	Syndrome
E565G	-2	Syndrome
E565A	-1	Syndrome
A628T	0	Syndrome
K641R	2	Syndrome
D650V	-3	Syndrome
K659Q	1	Syndrome
K659T	-1	Syndrome
G663E	-2	Syndrome
R678G	-2	Syndrome
E475K	1	Cancer
D530N	1	Cancer
I547V	3	Cancer
N549K	0	Cancer
E574K	1	Cancer
E636K	1	Cancer
M640I	1	Cancer
I642V	3	Cancer
K659E	1	Cancer
K659M	-1	Cancer
S688F	-2	Cancer
G701S	0	Cancer
P708S	-1	Cancer
R759Q	1	Cancer
L770V	1	Cancer

On plotting the frequency distribution of cancer and syndrome mutations across the BLOSUM matrix values results obtained are shown in Figure 4.

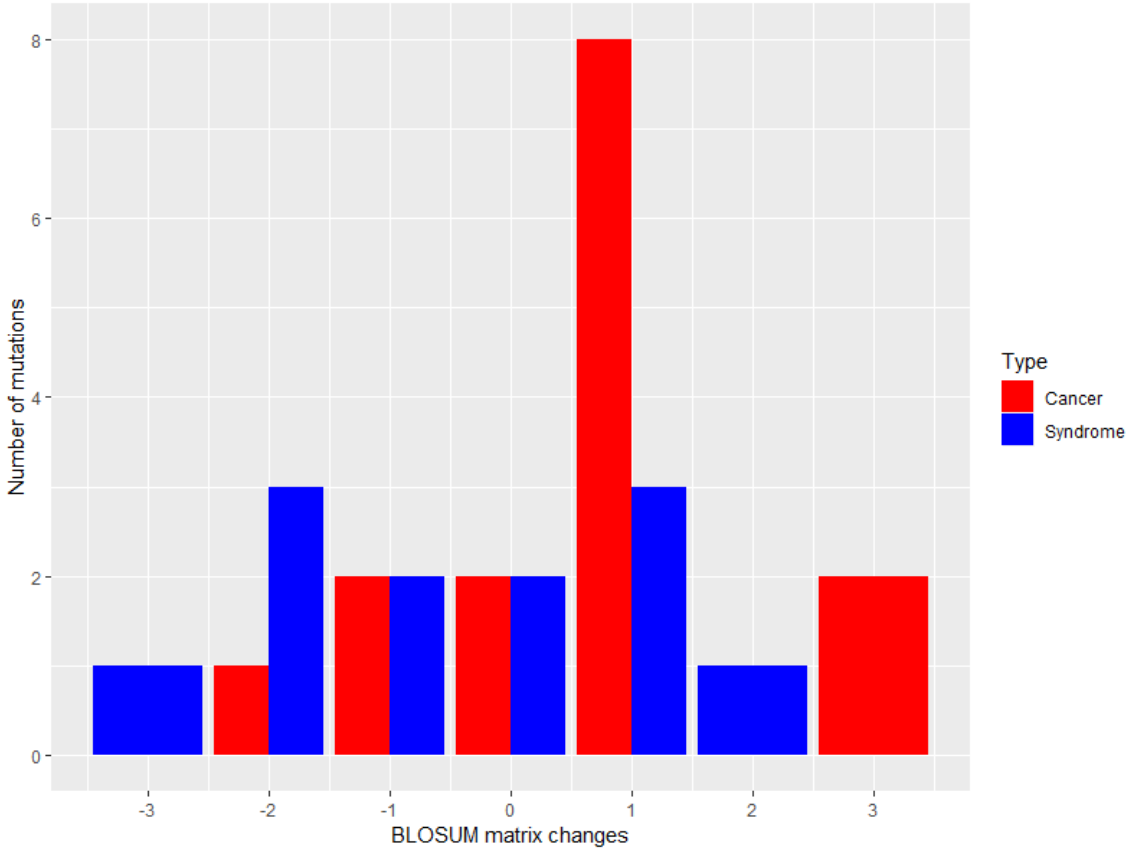

Figure 4. Number of mutations VS BLOSUM matrix values. More number of cancer causing mutants have positive values, indicating that substitution takes place between amino acids of similar physicochemical properties, whereas syndrome mutants have amino acid substitutions between any amino acid.

Higher values indicate substitution with higher similar physicochemical properties of amino acids. There are a greater number of cancer causing mutants with higher values indicating that in cancer, mutations take place between amino acids of similar physicochemical properties as compared to syndromes causing mutants which have even distribution of amino acid substitutions based on physicochemical properties.

3. Structure and energy-based analysis

3.1 FoldX Analysis

The results of FoldX analysis on all structures are mentioned in Appendix II.

Original energy of the 2PSQ Chain A structure was 116.97 kcal/mol and after repairing the energy of the structure was -11.53 kcal/mol. The stabilities of the mutant structures were plotted.

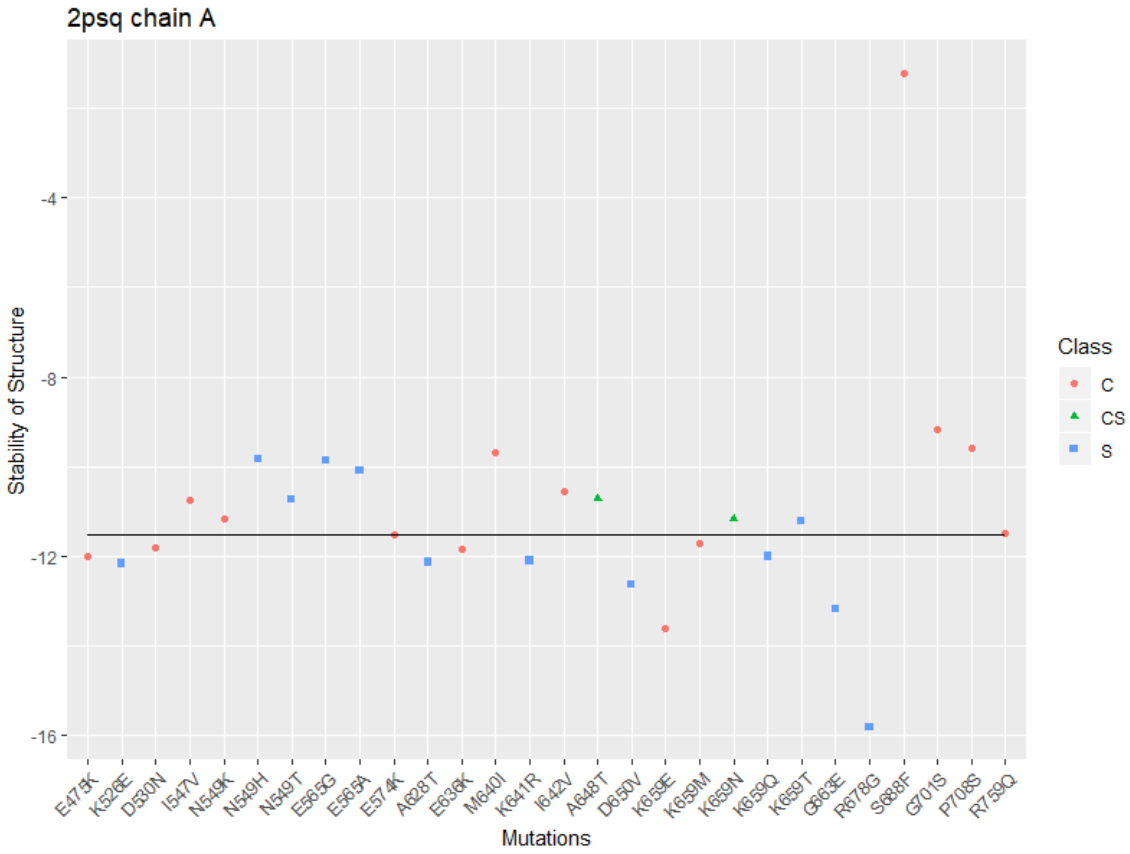

Figure 5. Stabilities (kcal/mol) of mutants for 2PSQ Chain A. C-Cancer, S-Syndrome, CS-Cancer&Syndrome. No differences between cancer causing and syndrome causing mutants with respect to stability of structures. The line in the graph represents the stability of the Wild Type with value of -11.53 kcal/mol.

The line in the graph represents the stability of the WT with value of -11.53 kcal/mol. It can be seen from the graph that cancer causing and syndrome causing mutants are randomly distributed, and all the mutants have stabilities within -8 kcal/mol

and -14 kcal/mol. R678G syndrome causing mutant has very high stability close to 16 kcal/mol and S688F cancer causing mutant is highly stable with stability of close to -1 kcal/mol.

On plotting the mutant-WT and FoldX energy values of the mutants we get the plot as shown in Figure 6. Orange lines in the plot represent standard deviation values of -0.46 and 0.46 and black lines represent standard deviation values of -0.8 and 0.8.

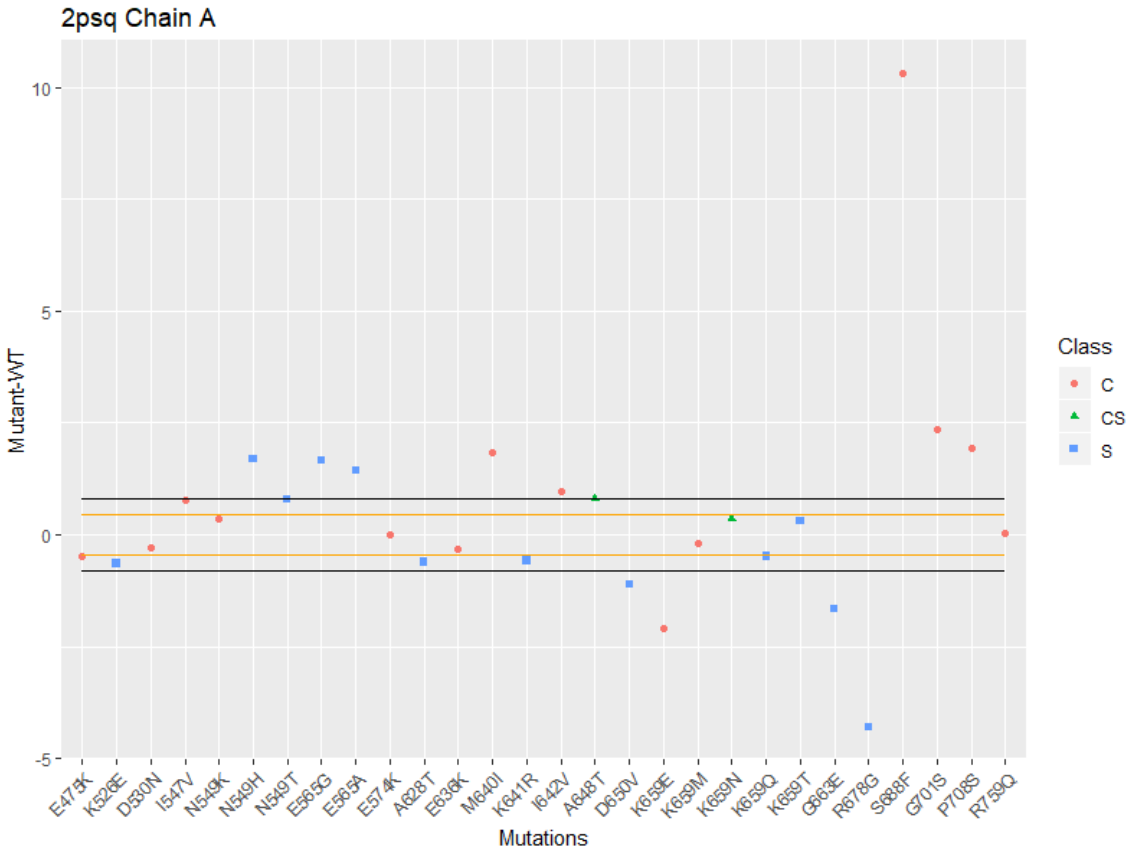

Figure 6. Mutant-WT (kcal/mol) energies of mutants for 2PSQ Chain A. C-Cancer, S-Syndrome, CS-Cancer&Syndrome. Orange lines in the plot represent standard deviation values of -0.46 and 0.46 and black lines represent standard deviation values of -0.8 and 0.8. No differences between cancer causing and syndrome causing mutants with respect to mutant-WT of structures.

In this Figure 6, among all the mutations having values greater than 0.46 and less than -0.46, there is only one cancer causing mutant K659E which is stable rest all cancer causing mutants are unstable.

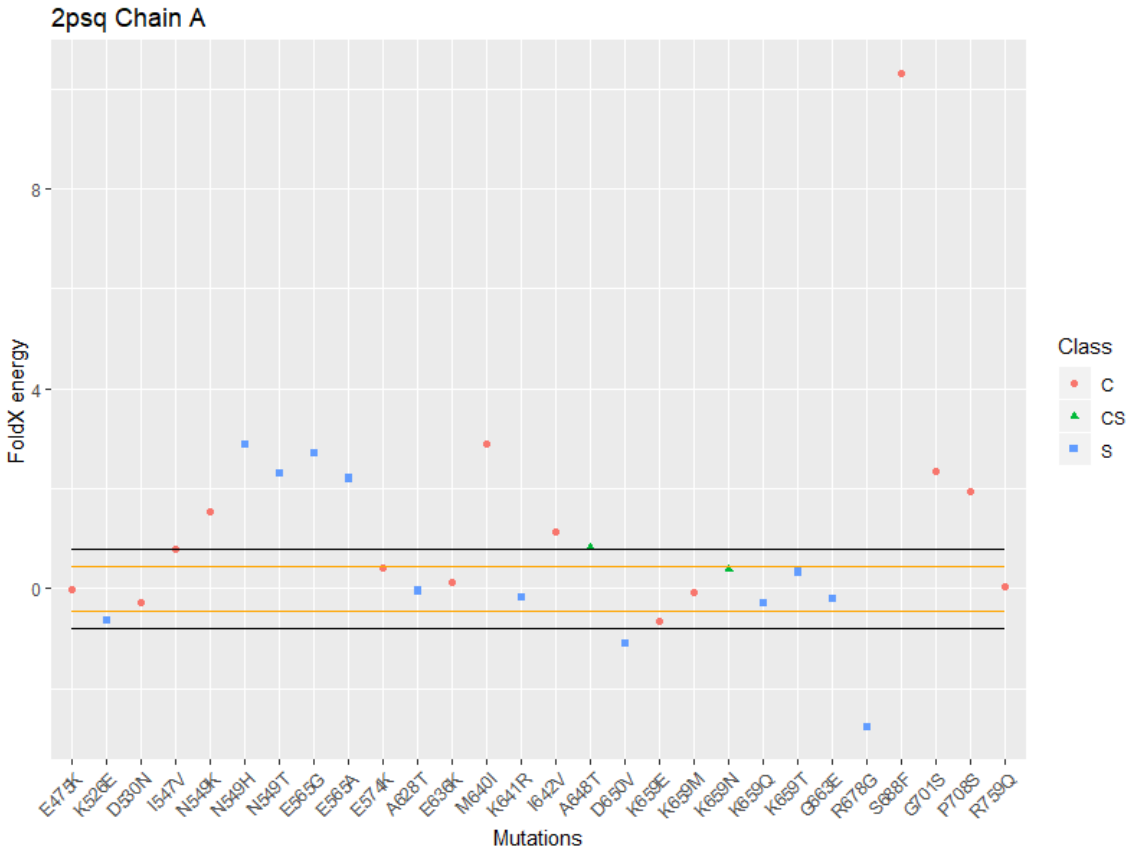

Figure 7. FoldX energies (kcal/mol) of mutants for 2PSQ Chain A. C-Cancer, S-Syndrome, CS-Cancer&Syndrome. Orange lines in the plot represent standard deviation values of -0.46 and 0.46 and black lines represent standard deviation values of -0.8 and 0.8. No differences between cancer causing and syndrome causing mutants with respect to FoldX energy of structures.

According to Figure 7, cancer causing and syndrome causing mutants are randomly distributed having positive as well as negative energy changes for both cancer and syndrome.

Similar analysis was performed on 2PSQ Chain B. The energy of the structure before repairing was 103.08 kcal/mol and after repairing it was -6.04 kcal/mol. Stability, mutant-WT and FoldX energy values were plotted.

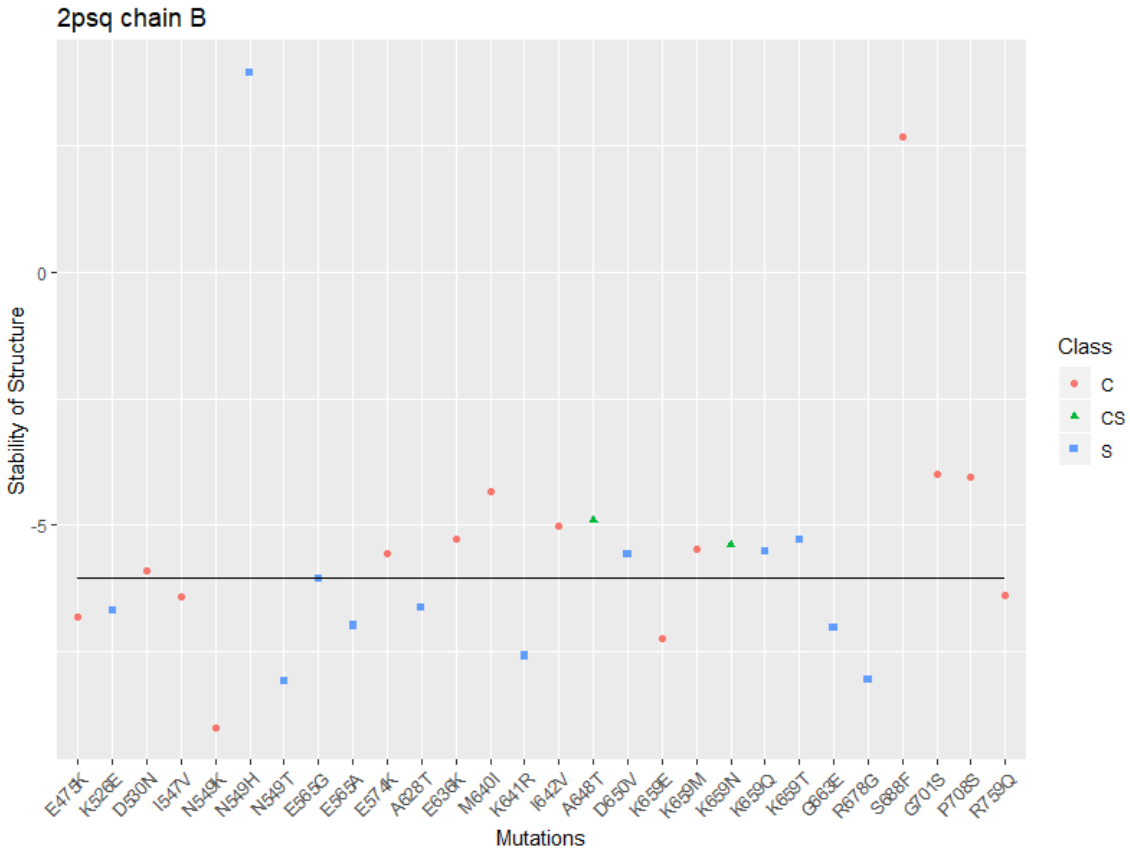

Figure 8. Stabilities (kcal/mol) of mutants for 2PSQ Chain B. C-Cancer, S-Syndrome, CS-Cancer&Syndrome. No differences between cancer causing and syndrome causing mutants with respect to stability of structures. The line in the graph represents the stability of the Wild Type with value of -6.04 kcal/mol.

Cancer causing and causing syndrome mutants are randomly distributed within range of -4 kcal/mol and -8 kcal/mol which is close to the stability of the original

structure. N549K cancer causing mutation is highly stable and N549H syndrome causing mutation and S688F cancer causing mutation are highly unstable.

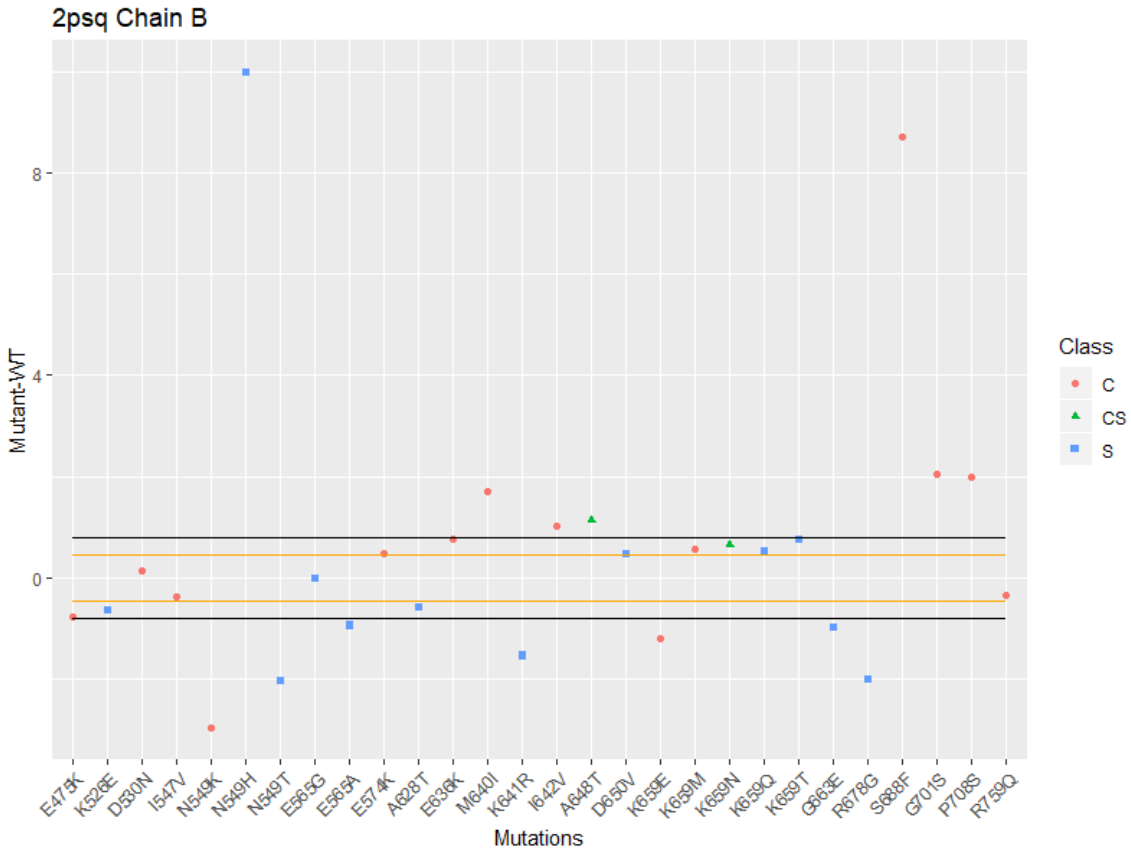

Figure 9. Mutant-WT (kcal/mol) energies of mutants for 2PSQ Chain B. C-Cancer, S-Syndrome, CS-Cancer&Syndrome. Orange lines in the plot represent standard deviation values of -0.46 and 0.46 and black lines represent standard deviation values of -0.8 and 0.8. No differences between cancer causing and syndrome causing mutants with respect to mutant-WT of structures.

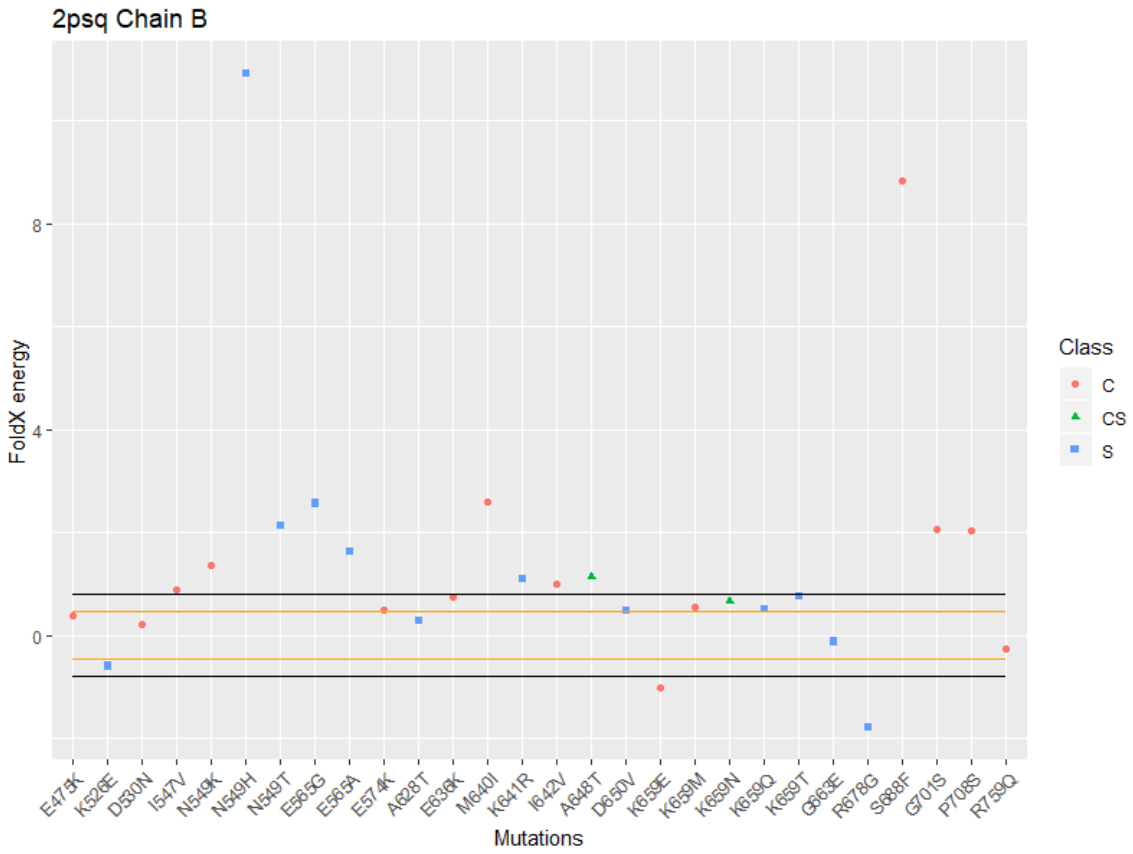

Figure 10. FoldX energies (kcal/mol) of mutants for 2PSQ Chain B. C-Cancer, S-Syndrome, CS-Cancer&Syndrome. Orange lines in the plot represent standard deviation values of -0.46 and 0.46 and black lines represent standard deviation values of -0.8 and 0.8. No differences between cancer causing and syndrome causing mutants with respect to FoldX energy of structures.

Maximum mutations have absolute energy difference values greater than 0.46 indicating they all are significant with respect to FoldX standard deviation of 0.46.

FoldX energy plots have maximum mutations with positive energy differences indicating their instability with respect to the original structure.

Similar FoldX analysis was done on 2PVF Chain A. The energy of the structure before repairing was 107.02 kcal/mol and after repairing it was -4.85 kcal/mol.

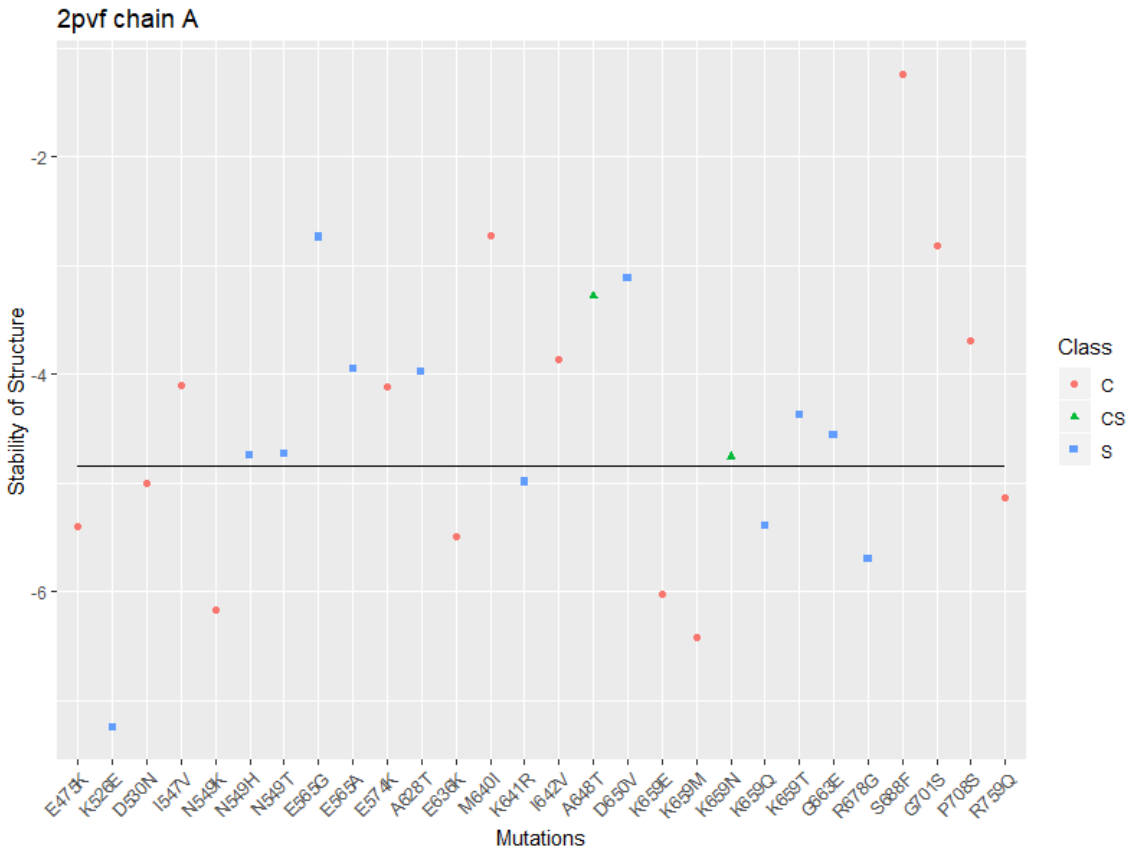

Figure 11. Stabilities (kcal/mol) of mutants for 2PVF Chain A. C-Cancer, S-Syndrome, CS-Cancer&Syndrome. No differences between cancer causing and syndrome causing mutants with respect to stability of structures. The line in the graph represents the stability of the Wild Type with value of -4.85 kcal/mol.

For 2PVF Chain A the stabilities of the mutants have random distribution with all mutations having stabilities within range of -3.5 kcal/mol to -6.5 kcal/mol. There are no significant outliers for this distribution.

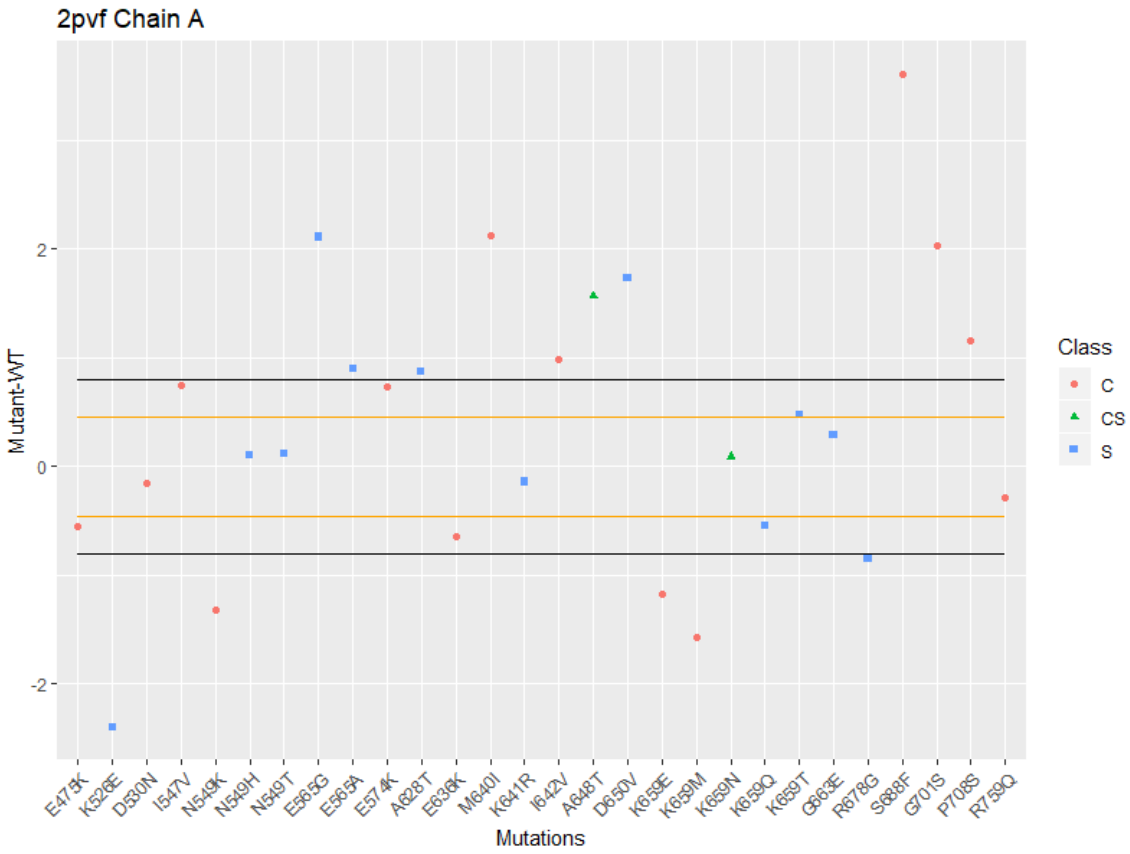

Figure 12. Mutant-WT (kcal/mol) energies of mutants for 2PVF Chain A. C-Cancer, S-Syndrome, CS-Cancer&Syndrome. Orange lines in the plot represent standard deviation values of -0.46 and 0.46 and black lines represent standard deviation values of -0.8 and 0.8. No differences between cancer causing and syndrome causing mutants with respect to mutant-WT of structures.

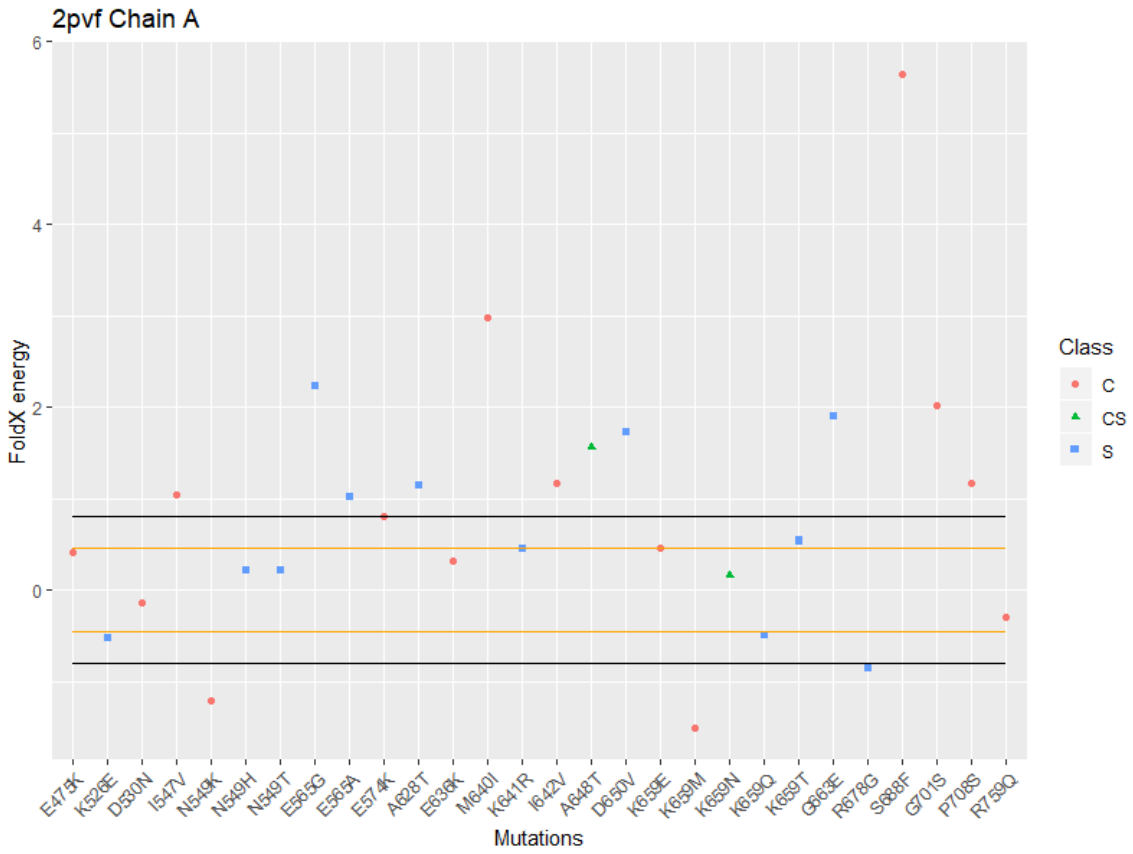

Figure 13. FoldX energies (kcal/mol) of mutants for 2PVF Chain A. C-Cancer, S-Syndrome, CS-Cancer&Syndrome. Orange lines in the plot represent standard deviation values of -0.46 and 0.46 and black lines represent standard deviation values of -0.8 and 0.8. No differences between cancer causing and syndrome causing mutants with respect to FoldX energy of structures.

FoldX energy values shows maximum mutations with positive values which are not seen in mutant-WT plot of 2PVF Chain A structure.

3.2 Statistical analysis

In order to understand the differences in cancer causing and syndrome causing mutants based on potential energies, statistical analyses were performed. Mutants which cause both cancer and syndrome were not taken into consideration for performing tests.

Based on the Normal distribution of the data, T-tests and Mann-Whitney-Wilcoxon Tests were performed. For each structure, there were four Gold Sets; mutants having absolute mutant-WT values and FoldX values greater than 0.46 (Gold Set - 0.46) and other set with values greater than 0.8 (Gold Set - 0.8). Statistical tests were performed on entire dataset and Gold Set.

a) 2PSQ Chain A

Entire dataset

Entire dataset used for statistical analysis consisted of 26 mutations:

Syndrome: K526E, N549H, N549T, E565G, E565A, A628T, K641R, D650V, K659Q, K659T, G663E, R678G

Cancer: E475K, D530N, I547V, N549K, E574K, E636K, M640I, I642V, K659E, K659M, S688F, G701S, P708S, R759Q

Table 3. Results of statistical analysis on 2PSQ Chain A entire dataset. The values are p-values.

	Stability		Mutant-WT		FoldX	
	Syndrome	Cancer	Syndrome	Cancer	Syndrome	Cancer
Normality	0.1554	0.0002261	0.1554	0.0002261	0.2372	0.0001207
Wilcoxon	0.1108		0.1108		0.252	

Table 3 shows that for Stability, Mutant-WT and FoldX values, the p-value results of Normality Test indicate that cancer dataset does not follow Normal distributions (p-value < 0.05). So, for these sets, Wilcoxon Test results were considered. Based on p-values, the two datasets are identical (p-value > 0.05).

The plots for number of mutations based on mutant-WT and FoldX energy values are shown in Figure 14.

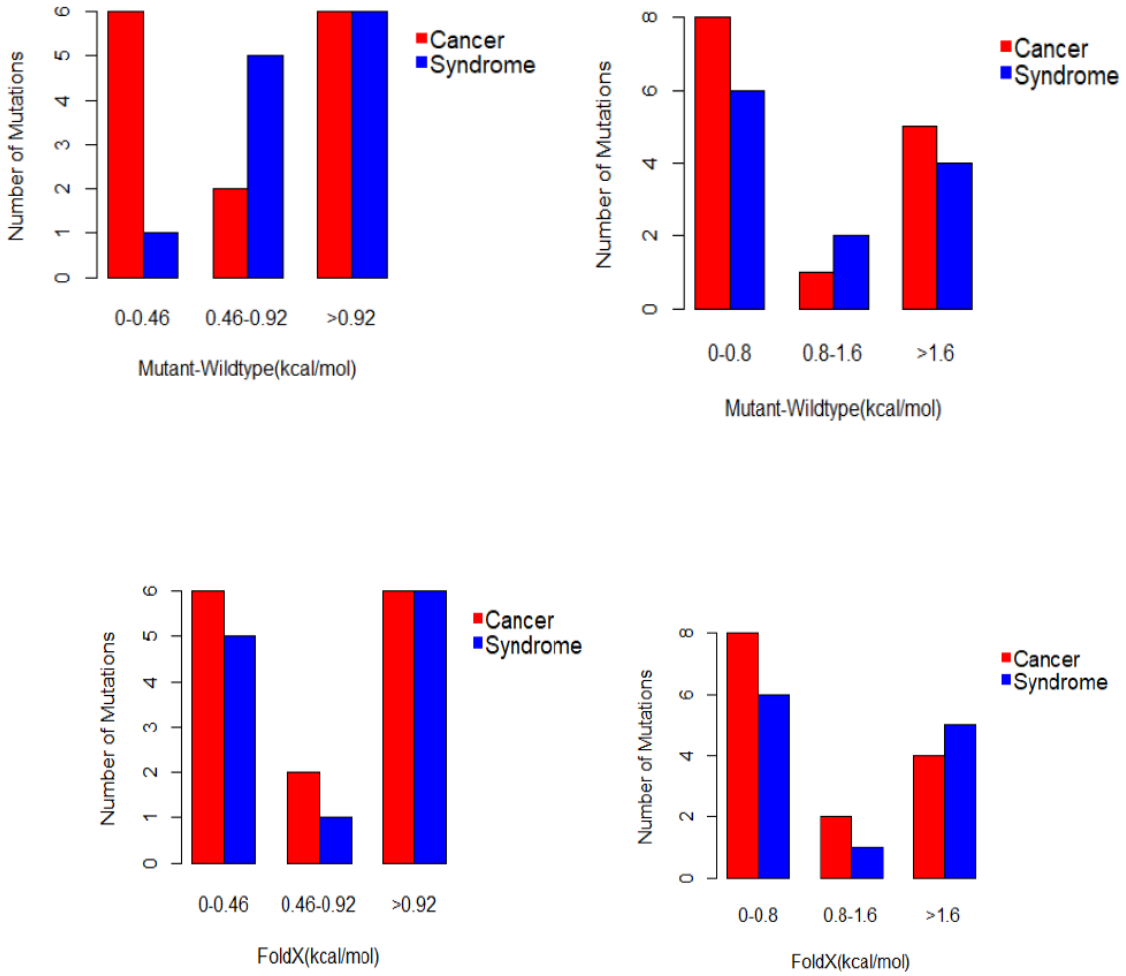

Figure 14. Frequency histograms of mutations based on Mutant-WT and FoldX energies (kcal/mol) for 2PSQ Chain A entire dataset.

In histograms of Figure 14, there are a greater number of cancer causing mutations than syndrome causing mutations that have absolute values less than 0.8.

Gold Set (0.8)

Based on mutant-WT values following mutations were considered:

Syndrome: N549H, E565G, E565A, D650V, G663E, R678G

Cancer: M640I, I642V, K659E, S688F, G701S, P708S

Table 4. Results of statistical analyses on 2PSQ Chain A Gold Set (0.8) based on Mutant-WT energy (kcal/mol). The values are p-values.

	Stability		Mutant-WT	
	Syndrome	Cancer	Syndrome	Cancer
Normality	0.1819	0.09194	0.1819	0.09194
T-Test	0.1721		0.1721	

From Table 4 as all p-values are greater than 0.05, the cancer and syndrome dataset follow Normal distribution both datasets are identical.

Based on FoldX values following mutations were considered:

Syndrome: N549H, N549T, E565G, E565A, D650V, R678G

Cancer: N549K, M640I, I642V, S688F, G701S, P708S

Table 5. Results of statistical analyses on 2PSQ Chain A Gold Set (0.8) based on FoldX energy (kcal/mol). The values are p-values.

	Stability		FoldX	
	Syndrome	Cancer	Syndrome	Cancer
Normality	0.04292	0.00425	0.04415	0.002812
Wilcoxon	0.09307		0.5887	

Table 5 shows that both syndrome dataset and cancer dataset do not follow normal distribution (p-value < 0.05). So, Wilcoxon Test was considered for understanding the distribution. From the results (p-value > 0.05), the distribution is identical.

The plots for number of mutations based on mutant-WT and FoldX values are shown in Figure 15.

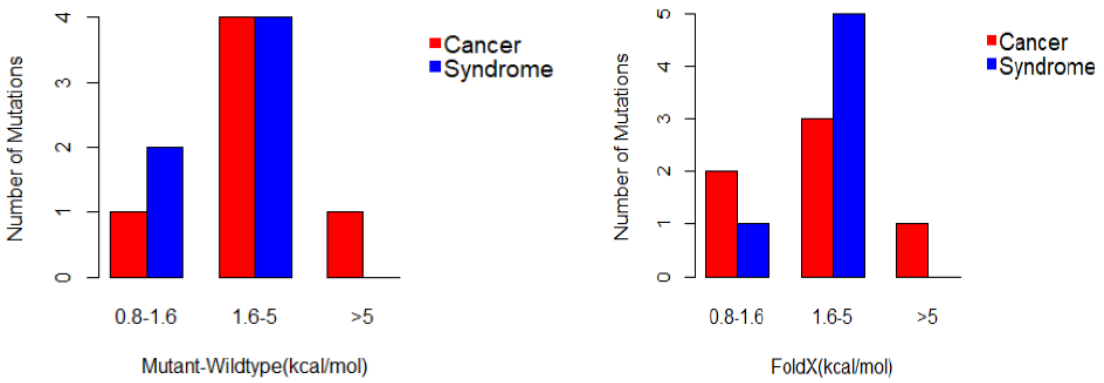

Figure 15. Frequency histograms of mutations based on Mutant-WT and FoldX energies (kcal/mol) for 2PSQ Chain A Gold Set (0.8).

Figure 15 shows that only cancer mutants have absolute energy values greater than 5.

Gold Set (0.46)

Based on mutant-WT values following mutations were considered:

Syndrome: K526E, N549H, N549T, E565G, E565A, A628T, K641R, D650V, K659Q, G663E, R678G

Cancer: E475K, I547V, M640I, I642V, K659E, S688F, G701S, P708S

Table 6. Results of statistical analyses on 2PSQ Chain A Gold Set (0.46) based on Mutant - WT energy (kcal/mol). The values are p-values.

	Stability		Mutant-WT	
	Syndrome	Cancer	Syndrome	Cancer
Normality	0.1528	0.02277	0.1528	0.02277
Wilcoxon	0.06916		0.06916	

Table 6 shows that the cancer dataset does not follow Normal distribution (p-value < 0.05), so Wilcoxon Test results were considered. Since p-value is greater than 0.05, the data distribution is identical for cancer and syndrome dataset.

Based on FoldX values following mutations were considered:

Syndrome: K526E, N549H, N549T, E565G, E565A, D650V, R678G

Cancer: I547V, N549K, M640I, I642V, K659E, S688F, G701S, P708S

Table 7. Results of statistical analyses on 2PSQ Chain A Gold Set (0.46) based on FoldX energy (kcal/mol). The values are p-values.

	Stability		FoldX	
	Syndrome	Cancer	Syndrome	Cancer
Normality	0.08788	0.01401	0.1118	0.005688
Wilcoxon	0.2026		0.6126	

Table 7 shows that for this dataset, cancer dataset does not have Normal distribution (p-value < 0.05). So based on Wilcoxon Test results, it can be concluded that both cancer and syndrome dataset have no differences based on distribution (p-value > 0.05).

The plots for number of mutations based on mutant-WT and FoldX values are shown in Figure 16.

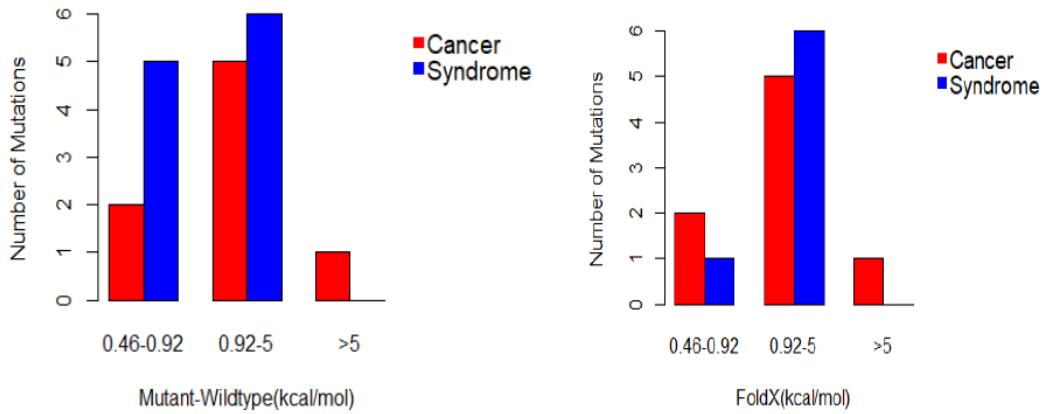

Figure 16. Frequency histograms of mutations based on Mutant-WT and FoldX energies (kcal/mol) for 2PSQ Chain A Gold Set (0.46).

Figure 16 shows that, for this dataset, the number of mutants having absolute values greater than 0.92 are identical.

b) 2PSQ Chain B

Entire Dataset

Entire dataset used for statistical analysis consisted on 26 mutations:

Syndrome mutants: K526E, N549H, N549T, E565G, E565A, A628T, K641R, D650V, K659Q, K659T, G663E, R678G

Cancer mutants: E475K, D530N, I547V, N549K, E574K, E636K, M640I, I642V, K659E, K659M, S688F, G701S, P708S, R759Q

Table 8.Results of statistical analyses on 2PSQ Chain B entire dataset. The values are p-values.

	Stability		Mutant-WT		FoldX	
	Syndrome	Cancer	Syndrome	Cancer	Syndrome	Cancer
Normality	0.0001275	0.03553	0.0001275	0.004293	0.0007004	0.0003248
Wilcoxon	0.3275		0.1228		0.8201	

According to Table 8, for the Stability, Mutant-WT and FoldX, the p-value results of Normality Test (p-value < 0.05) indicate that syndrome and cancer dataset do not follow Normal distributions. So, for these datasets, Wilcoxon Test p-value results were considered. Based on p-values (p-value > 0.05), the population distribution looks identical.

The plots for number of mutations based on mutant-WT and FoldX energy values are shown in Figure 17.

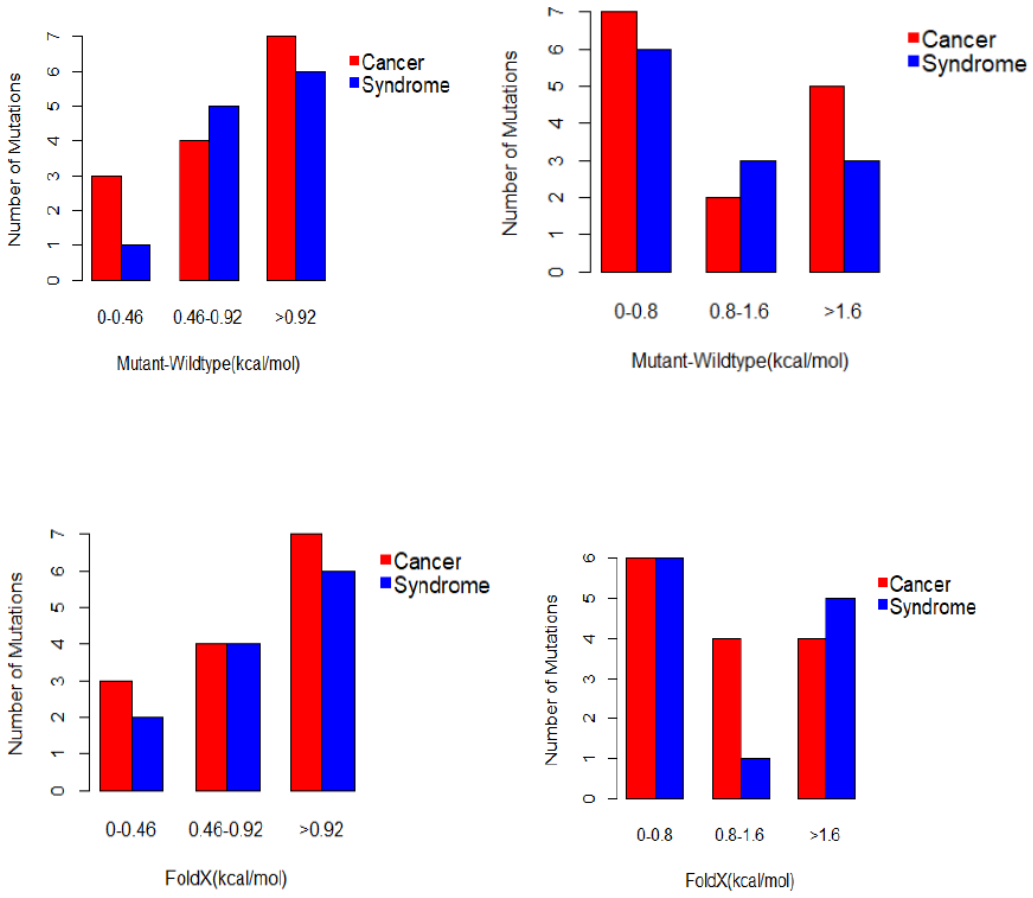

Figure 17. Frequency histograms of mutations based on Mutant-WT and FoldX energies (kcal/mol) for 2PSQ Chain B entire dataset.

According to Figure 17, the distribution of cancer causing and syndrome causing mutants is random.

Gold Set (0.8)

Based on mutant-WT values following mutations were considered:

Syndrome: N549H, N549T, E565A, K641R, G663E, R678G

Cancer: N549K, M640I, I642V, K659E, S688F, G701S, P708S

Table 9. Results of statistical analyses on 2PSQ Chain B Gold Set (0.8) based on Mutant-WT energy (kcal/mol). The values are p-values.

	Stability		Mutant-WT	
	Syndrome	Cancer	Syndrome	Cancer
Normality	0.0003248	0.2336	0.0003248	0.2336
Wilcoxon	0.366		0.366	

Table 9 shows that, for this dataset, syndrome dataset does not follow Normal distribution (p-value < 0.05). So, based on Wilcoxon Test p-value results, the two datasets are identical (p-value > 0.05).

Based on FoldX values following mutations were considered:

Syndrome: N549H, N549T, E565G, E565A, K641R, R678G

Cancer: I547V, N549K, M640I, I642V, K659E, S688F, G701S, P708S

Table 10. Results of statistical analyses on 2PSQ Chain B Gold Set (0.8) based on FoldX energy (kcal/mol). The values are p-values.

	Stability		FoldX	
	Syndrome	Cancer	Syndrome	Cancer
Normality	0.001262	0.1681	0.07421	0.01515
Wilcoxon	0.345		0.7546	

Table 10 shows that, for this dataset, the stability of syndrome and FoldX values of cancer do not follow Normal distribution (p-value < 0.05). So, by considering p-value results of Wilcoxon Test (p-value > 0.05), it can be said that the two datasets are identical

The plots for number of mutations based on mutant-WT and FoldX values are shown in Figure 18.

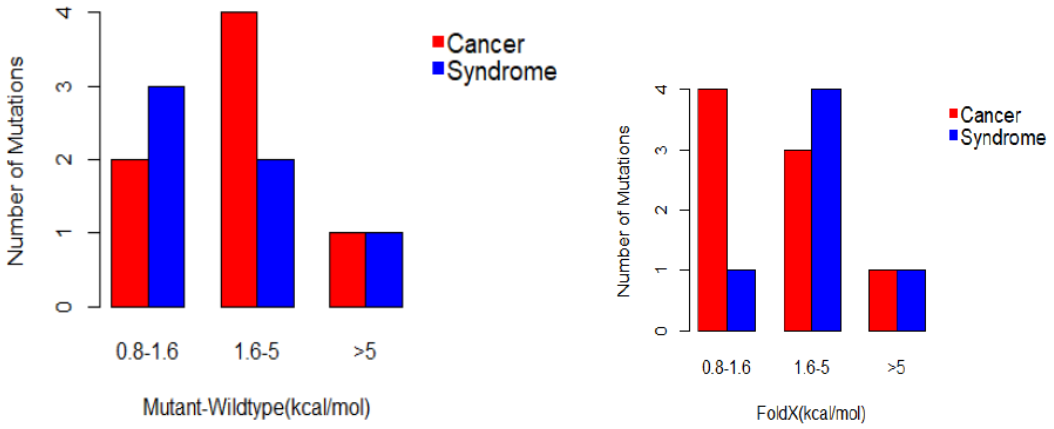

Figure 18. Frequency histograms of mutations based on Mutant-WT and FoldX energies (kcal/mol) for 2PSQ Chain B Gold Set (0.8).

Histograms in Figure 18 show same number of mutations having absolute values greater than 5 kcal/mol.

Gold Set (0.46)

Based on mutant-WT values following mutations were considered:

Syndrome: K526E, N549H, N549T, E565A, A628T, K641R, D650V, K659Q, K659T, G663E, R678G

Cancer: E475K, N549K, E574K, E636K, M640I, I642V, K659E, K659M, S688F, G701S, P708S

Table 11. Results of statistical analysis on 2PSQ Chain B Gold Set (0.46) based on Mutant - WT energy (kcal/mol). The values are p-values.

	Stability		Mutant-WT	
	Syndrome	Cancer	Syndrome	Cancer
Normality	4.036e-05	0.02179	4.036e-05	0.02179
Wilcoxon	0.1227		0.1227	

Table 11 shows that the cancer and syndrome dataset do not follow normal distribution (p-value < 0.05), so Wilcoxon p-value results were considered. Since p-value is greater than 0.05, the data distribution is identical for cancer and syndrome.

Based on FoldX values following mutations were considered:

Syndrome: K526E, N549H, N549T, E565G, E565A, K641R, D650V, K659Q, K659T, R678G

Cancer: I547V, N549K, E574K, E636K, M640I, I642V, K659E, K659M, S688F, G701S, P708S

Table 12. Results of statistical analyses on 2PSQ Chain B Gold Set (0.46) based on FoldX energy (kcal/mol). The values are p-values.

	Stability		FoldX	
	Syndrome	Cancer	Syndrome	Cancer
Normality	0.0001938	0.01905	0.00182	0.0007262
Wilcoxon	0.1391		0.7564	

Table 12 shows that, both datasets do not follow normal distribution (p-value < 0.05). So based on Wilcoxon test results (p-value > 0.05), it can be concluded that both cancer and syndrome dataset are identical based on distribution. The plots for number of mutations based on mutant-WT and FoldX values are shown in Figure 19.

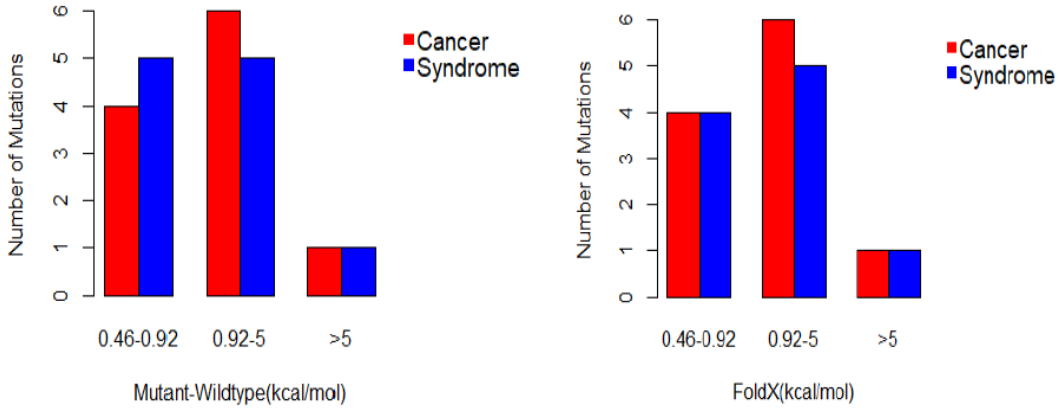

Figure 19. Frequency histograms of mutations based on Mutant-WT and FoldX energies (kcal/mol) for 2PSQ Chain B Gold Set (0.46).

According to Figure 19, most of mutations have absolute values less than 5 kcal/mol.

c) 2PVF Chain A

Entire dataset

Entire dataset used for statistical analysis consisted on 26 mutations:

Syndrome: K526E, N549H, N549T, E565G, E565A, A628T, K641R, D650V, K659Q, K659T, G663E, R678G

Cancer: E475K, D530N, I547V, N549K, E574K, E636K, M640I, I642V, K659E, K659M, S688F, G701S, P708S, R759Q

Table 13. Results of statistical analyses on 2PVF Chain A entire dataset. The values are p-values.

	Stability		Mutant-WT		FoldX	
	Syndrome	Cancer	Syndrome	Cancer	Syndrome	Cancer
Normality	0.7969	0.5513	0.7969	0.5513	0.7287	0.08996
T-Test	0.729		0.729		0.6088	

Table 13 shows that, based on p-values of Normality Test cancer and syndrome dataset (p-value > 0.05) are Normally distributed, so based on T-Test values we can conclude that the distribution is identical for cancer and syndrome (p-value > 0.05).

The plots for number of mutations based on mutant-WT and FoldX energy values are shown in Figure 20.

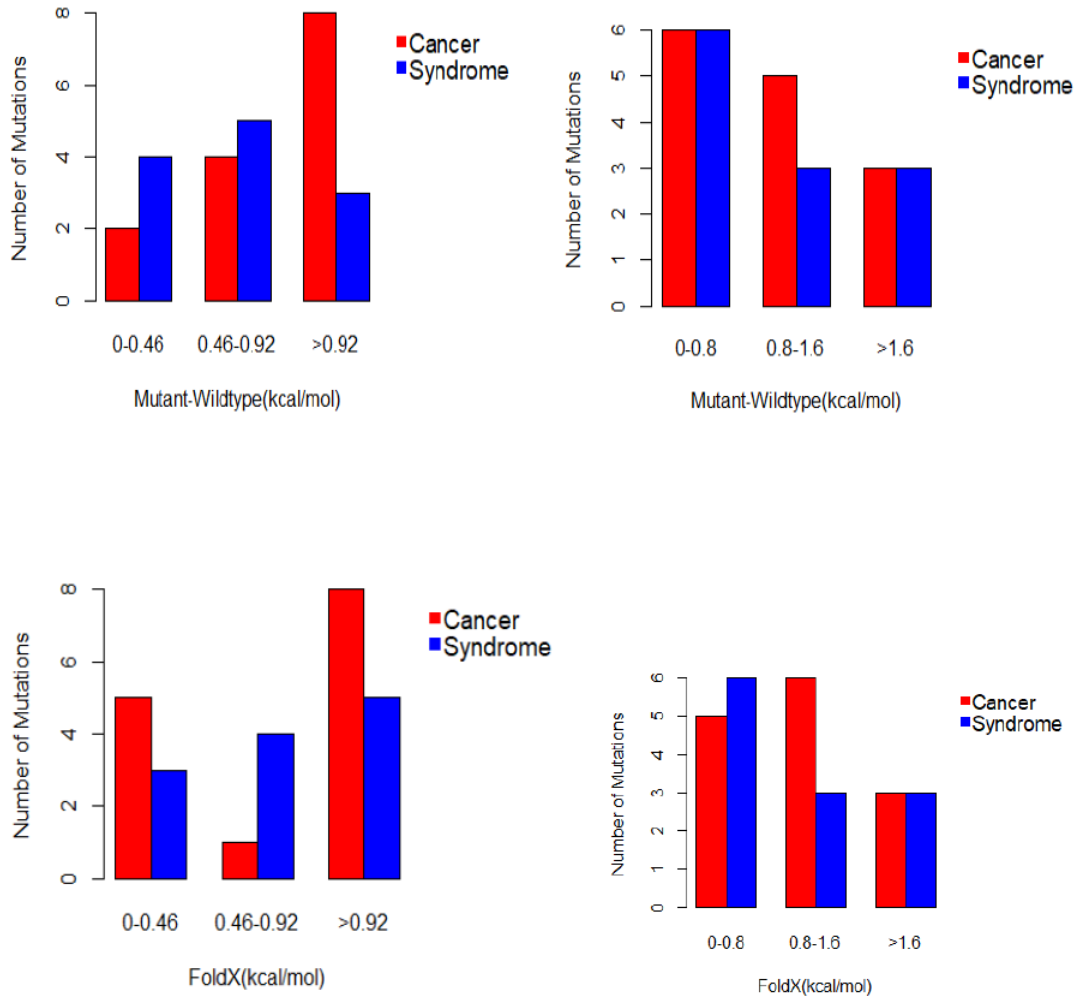

Figure 20. Frequency histograms of mutations based on Mutant-WT and FoldX energies (kcal/mol) for 2PVF Chain A entire dataset.

According to Figure 20 histograms, there are a greater number of cancer mutations with values greater than 0.92 whereas syndrome mutations are evenly distributed.

Gold Set (0.8)

Based on mutant-WT values following mutations were considered:

Syndrome: K526E, E565G, E565A, A628T, D650V, R678G

Cancer: N549K, M640I, I642V, K659E, K659M, S688F, G701S, P708S

Table 14. Results of statistical analyses on 2PVF Chain A Gold Set (0.8) based on Mutant-WT energy (kcal/mol). The values are p-values.

	Stability		Mutant-WT	
	Syndrome	Cancer	Syndrome	Cancer
Normality	0.3918	0.3454	0.3918	0.3454
T-Test		0.7333		0.7333

The results of Normality Test in Table 14 show that the two datasets have normal distribution (p-value > 0.05). Based on T-test p-values, we can conclude that the two datasets are identical (p-value > 0.05).

Based on FoldX values following mutations were considered:

Syndrome: E565G, E565A, A628T, D650V, G663E, R678G

Cancer: I547V, N549K, E574K, M640I, I642V, K659M, S688F, G701S, P708S

Table 15.. Results of statistical analyses on 2PVF Chain A Gold Set (0.8) based on FoldX energy (kcal/mol). The values are p-values.

	Stability		FoldX	
	Syndrome	Cancer	Syndrome	Cancer
Normality	0.8073	0.5506	0.1741	0.419
T-Test		0.8857		0.8587

The results for this set as seen in Table 15 also indicate that the two datasets are identical (p-value > 0.05).

The plots for number of mutations based on mutant-WT and FoldX values are shown in Figure 21.

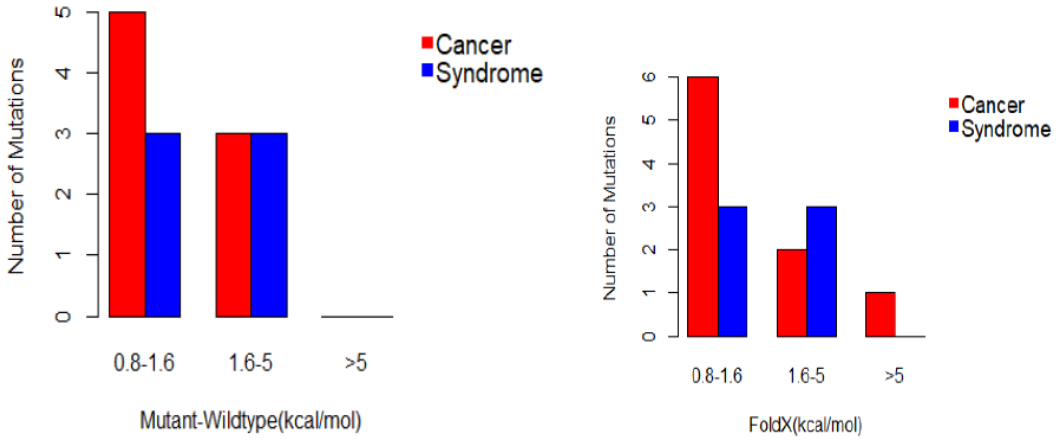

Figure 21. Frequency histograms of mutations based on Mutant-WT and FoldX energies (kcal/mol) for 2PVF Chain A Gold Set (0.8)

Histograms in Figure 21 show that, for this dataset, most of the cancer causing mutants have values within range 0.8-1.6 kcal/mol.

Gold Set (0.46)

Based on mutant-WT values following mutations were considered:

Syndrome: K526E, E565G, E565A, A628T, D650V, K659Q, K659T, R678G

Cancer: E475K, I547V, N549K, E574K, E636K, M640I, I642V, K659E, K659M, S688F, G701S, P708S

Table 16. Results of statistical analyses on 2PVF Chain A Gold Set (0.46) based on Mutant - WT energy (kcal/mol). The values are p-values.

	Stability		Mutant-WT	
	Syndrome	Cancer	Syndrome	Cancer
Normality	0.7135	0.58	0.7135	0.5408
T-Test		0.8116		0.7516

Since both cancer and syndrome dataset have normal distribution (p-value > 0.05) as seen from Table 16, T-Test p-value results are valid. Based on p-value of T-tests (p-value > 0.05), we can say that both the datasets are identical.

Based on FoldX values following mutations were considered:

Syndrome: K526E, E565G, E565A, A628T, D650V, K659Q, K659T, G663E, R678G

Cancer: I547V, N549K, E574K, M640I, I642V, K659M, S688F, G701S, P708S

Table 17. Results of statistical analyses on 2PVF Chain A Gold Set (0.46) based on FoldX energy (kcal/mol). The values are p-values.

	Stability		FoldX	
	Syndrome	Cancer	Syndrome	Cancer
Normality	0.7613	0.5506	0.3584	0.419
T-Test		0.3704		0.4689

Based on T-test p-value results in Table 17, the two datasets considered here are identical (p-value > 0.05).

The plots for number of mutations based on mutant-WT and FoldX values are shown in Figure 22.

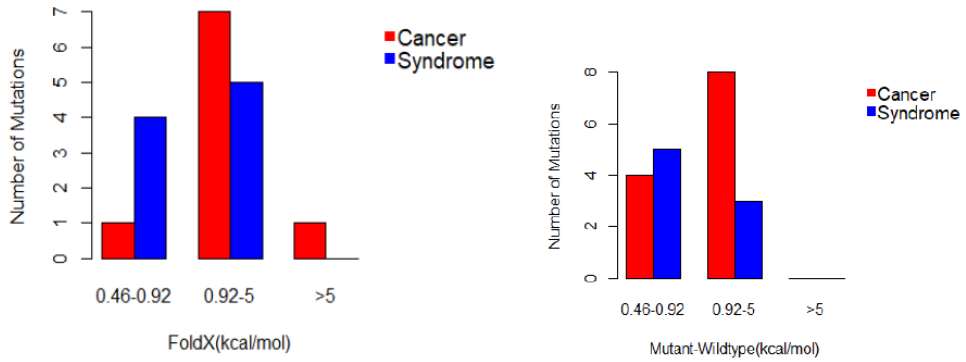

Figure 22. Frequency histograms of mutations based on Mutant-WT and FoldX energies (kcal/mol) for 2PVF Chain A Gold Set (0.46).

In the above analysis, for 2PSQ the mutations on two chains give different results which based on energies cannot be distinguished for correctness. Also, the data used for statistical analysis, did not follow Normal distribution in almost all the cases. Based on Mann-Whitney-Wilcoxon Test results, the cancer and syndrome datasets show identical distribution in all the cases. The histograms do not reveal significant differences in cancer causing and syndrome causing mutants' distribution across different energy values.

For 2PVF, the cancer and syndrome datasets follow normal distribution for all the Gold Sets. So, results of T-tests show that the both the datasets are identical in distribution which can be seen from the histograms.

CONCLUSION

In this work, we investigated FGFR2 cancer and syndrome point mutations in cytoplasmic tyrosine kinase domain by performing region-based analysis, substitution diverseness and structure energy-based analysis. Our analysis shows that cancer mutations are present in all the regions of the protein unlike syndrome mutations. The mutant positions are highly conserved across various species based on Shannon Entropy Analysis. There are differences in substitution diverseness of cancer and syndrome mutants. Cancer mutations have substitutions with amino acids having similar physicochemical properties, whereas in syndromes all types of amino acids can be substituted. Structure and energy-based analysis has revealed that based on energy of the mutants cancer and syndrome cannot be distinguished. Statistical and histogram analysis showed that the two classes of disorder, namely cancer and syndrome, have identical distribution based on energy. Thus, this analysis is inconclusive, and energy cannot be used as predictor for cancer and syndrome mutations in FGFR2 kinase domain. Such analysis also has reliance on the specific energy model used. Here FoldX energy field was used for analysis. Instead other detailed minimization techniques and molecular dynamics simulations could be useful, but they also include great computational costs and cannot currently be practical for screening mutations. However, this thesis has

revealed interesting characteristics about FGFR2 cancer and syndrome mutations that open the way for further investigation of these mutations.

REFERENCES

[1] Houssaint, E., Blanquet, P.R., Champion-Arnaud, P., Gesnel, M.C., Torriglia, A., Courtois, Y. and Breathnach, R., 1990. Related fibroblast growth factor receptor genes exist in the human genome. *Proceedings of the National Academy of Sciences*, *87*(20), pp.8180-8184.

[2] "Entrez Gene: FGFR2 fibroblast growth factor receptor 2 (bacteria-expressed kinase, keratinocyte growth factor receptor, craniofacial dysostosis 1, Crouzon syndrome, Pfeiffer syndrome, Jackson–Weiss syndrome)"

[3] Karp, J.M., Sparks, S. and Cowburn, D., 2017. Effects of FGFR2 kinase activation loop dynamics on catalytic activity. *PLoS computational biology*, *13*(2), p.e1005360.

[4] Stauber, D.J., DiGabriele, A.D. and Hendrickson, W.A., 2000. Structural interactions of fibroblast growth factor receptor with its ligands. *Proceedings of the National Academy of Sciences*, *97*(1), pp.49-54.

[5] Lu, Y., Pan, Z.Z., Devaux, Y. and Ray, P., 2003. p21-activated protein kinase 4 (PAK4) interacts with the keratinocyte growth factor receptor and participates in keratinocyte growth factor-mediated inhibition of oxidant-induced cell death. *Journal of Biological Chemistry*, *278*(12), pp.10374-10380.

[6] Chen, H., Ma, J., Li, W., Eliseenkova, A.V., Xu, C., Neubert, T.A., Miller, W.T. and Mohammadi, M., 2007. A molecular brake in the kinase hinge region regulates the activity of receptor tyrosine kinases. *Molecular cell*, *27*(5), pp.717-730.

[7] Chen, H., Marsiglia, W.M., Cho, M.K., Huang, Z., Deng, J., Blais, S.P., Gai, W., Bhattacharya, S., Neubert, T.A., Traaseth, N.J. and Mohammadi, M., 2017. Elucidation of a four-site allosteric network in fibroblast growth factor receptor tyrosine kinases. *Elife*, *6*, p.e21137.

[8] Lew, E.D., Bae, J.H., Rohmann, E., Wollnik, B. and Schlessinger, J., 2007. Structural basis for reduced FGFR2 activity in LADD syndrome: Implications for FGFR autoinhibition and activation. *Proceedings of the National Academy of Sciences*, *104*(50), pp.19802-19807.

[9] Pollock, P.M., Gartside, M.G., Dejeza, L.C., Powell, M.A., Mallon, M.A., Davies, H., Mohammadi, M., Futreal, P.A., Stratton, M.R., Trent, J.M. and Goodfellow, P.J., 2007. Frequent activating FGFR2 mutations in endometrial carcinomas parallel germline mutations associated with craniosynostosis and skeletal dysplasia syndromes. *Oncogene*, *26*(50), p.7158.

[10] Chen, H., Huang, Z., Dutta, K., Blais, S., Neubert, T.A., Li, X., Cowburn, D., Traaseth, N.J. and Mohammadi, M., 2013. Cracking the molecular origin of intrinsic tyrosine kinase activity through analysis of pathogenic gain-of-function mutations. *Cell reports*, *4*(2), pp.376-384.

[11] Reintjes, N., Li, Y., Becker, A., Rohmann, E., Schmutzler, R. and Wollnik, B., 2013. Activating somatic FGFR2 mutations in breast cancer. *PLoS One*, *8*(3), p.e60264.

[12] Gartside, M.G., Chen, H., Ibrahimi, O.A., Byron, S.A., Curtis, A.V., Wellens, C.L., Bengston, A., Yudt, L.M., Eliseenkova, A.V., Ma, J. and Curtin, J.A., 2009. Loss-of-function fibroblast growth factor receptor-2 mutations in melanoma. *Molecular Cancer Research*, *7*(1), pp.41-54.

[13] Gureasko, J., Galush, W.J., Boykevisch, S., Sondermann, H., Bar-Sagi, D., Groves, J.T. and Kuriyan, J., 2008. Membrane-dependent signal integration by the Ras activator Son of sevenless. *Nature structural & molecular biology*, *15*(5), p.452.

[14] Zhijun, W., 2008. *Lecture notes on computational structural biology*. World Scientific.

[15] Kiel, C. and Serrano, L., 2014. Structure-energy-based predictions and network modelling of RASopathy and cancer missense mutations. *Molecular systems biology*, *10*(5), p.727.

[16] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. and Serrano, L., 2005. The FoldX web server: an online force field. *Nucleic acids research*, *33*(suppl_2), pp.W382-W388.

[17] Gureasko, J., Galush, W.J., Boykevisch, S., Sondermann, H., Bar-Sagi, D., Groves, J.T. and Kuriyan, J., 2008. Membrane-dependent signal integration by the Ras activator Son of sevenless. *Nature structural & molecular biology*, *15*(5), p.452.

[18] Zhong, Q., Simonis, N., Li, Q.R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D. and Swearingen, V., 2009. Edgetic perturbation models of human inherited disorders. *Molecular systems biology*, *5*(1), p.321.

[19] Strait, B.J. and Dewey, T.G., 1996. The Shannon information entropy of protein sequences. *Biophysical journal*, *71*(1), pp.148-155.

[20] Henikoff, S. and Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), pp.10915-10919.

[21] UniProt Consortium, 2018. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, *47*(D1), pp.D506-D515.

[22] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., 2000. The protein data bank nucleic acids research, 28: 235-242. *URL: www. rcsb. org Citation*.

[23] Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A. and Stratton, M.R., The catalogue of somatic mutations in cancer (COSMIC) Curr Protoc Hum Genet. 2008.

[24] Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. and Wagner, L., 2003. Database resources of the National Center for Biotechnology. *Nucleic acids research*, *31*(1), pp.28-33.

[25] Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J. and Thompson, J.D., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, *7*(1), p.539.

[26] Garcia-Boronat, M., Diez-Rivero, C.M., Reinherz, E.L. and Reche, P.A., 2008. PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery. *Nucleic acids research*, *36*(suppl_2), pp.W35-W41.

[27] Team, R.C., 2013. R: A language and environment for statistical computing.

[28] Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A. and Caves, L.S., 2006. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, *22*(21), pp.2695-2696.

[29] Wickham, H., 2016. *ggplot2: elegant graphics for data analysis*. Springer.

[30] Wickham, H., 2012. reshape2: Flexibly reshape data: a reboot of the reshape package. *R package version*, *1*(2).

[31] Niedballa, J., Sollmann, R., Courtiol, A. and Wilting, A., 2016. camtrapR: an R package for efficient camera trap data management. *Methods in Ecology and Evolution*, *7*(12), pp.1457-1462.

APPENDIX I

All the sequences that were used for Shannon Entropy Analysis are mentioned here

>sp|P21802|FGFR2_HUMAN Fibroblast growth factor receptor 2 OS=Homo sapiens OX=9606 GN=FGFR2 PE=1 SV=1
MVSWGRFICLVVVTMATLSLARPSFSLVEDTTLEPEEPPTKYQISQPEVYVAAPG
ESLEVRCLLKDAAVISWTKDGVHLGPNNRTVLIGEYLQIKGATPRDSGLYACTAS
RTVDSETWYFMVNVTDAISSGDDEDDTDGAEDFVSENSNNKRAPYWTNTEKME
KRLHAVPAANTVKFRCPAGGNPMPTMRWLKNGKEFKQEHRIGGYKVRNQHWS
LIMESVVPSDKGNYTCVVENEYGSINHTYHLDVVERSPHRPILQAGLPANASTVV
GGDVEFVCKVYSDAQPHIQWIKHVEKNGSKYGPDGLPYLKVLKAAGVNTTDKE
IEVLYIRNVTFEDAGEYTCLAGNSIGISFHSAWLTVLPAPGREKEITASPDYLEIAI
YCIGVFLIACMVVTVILCRMKNTTKKPDFSSQPAVHKLTKRIPLRRQVTVSAESSS
SMNSNTPLVRITTRLSSTADTPMLAGVSEYELPEDPKWEFPRDKLTLGKPLGEGC
FGQVVMAEAVGIDKDKPKEAVTVAVKMLKDDATEKDLSDLVSEMEMMKMIGK
HKNIINLLGACTQDGPLYVIVEYASKGNLREYLRARRPPGMEYSYDINRVPEEQM
TFKDLVSCTYQLARGMEYLASQKCIHRDLAARNVLVTENNVMKIADFGLARDIN
NIDYYKKTTNGRLPVKWMAPEALFDRVYTHQSDVWSFGVLMWEIFTLGGSPYP
GIPVEELFKLLKEGHRMDKPANCTNELYMMMRDCWHAVPSQRPTFKQLVEDLD
RILTLTTNEEYLDLSQPLEQYSPSYPDTRSSCSSGDDSVFSPDPMPYEPCLPQYPHI
NGSVKT
>tr|H2Q2P3|H2Q2P3_PANTR Fibroblast growth factor receptor OS=Pan troglodytes OX=9598 GN=FGFR2 PE=3 SV=2
MGLTSTWRYGRGPGIGTVTMVSWGRFICLVVVTMATLSLARPSFSLVEDTTLEP
EEPPTKYQISQPEVYVAAPGESLEVRCLLKDAAVISWTKDGVHLGPNNRTVLIGE
YLQIKGATPRDSGLYACTASRTVDSETWYFMVNVTDAISSGDDEDDTDGAEDFV
SENSNNKRAPYWTNTEKMEKRLHAVPAANTVKFRCPAGGNPTPTMRWLKNGK
EFKQEHRIGGYKVRNQHWSLIMESVVPSDKGNYTCVVENEYGSINHTYHLDVVE
RSPHRPILQAGLPANASTVVGGDVEFVCKVYSDAQPHIQWIKHVEKNGSKYGPD
GLPYLKVLKHSGINSSNAEVLALFNVTEADAGEYICKVSNYIGQANQSAWLTVL
PKQQAPGREKEITASPDYLEIAIYCIGVFLIACMVVTVILCRMKNTTKKPDFSSQP
AVHKLTKRIPLRRQVTVSAESSSSMNSNTPLVRITTRLSSTADTPMLAGVSEYELP
EDPKWEFPRDKLTLGKPLGEGCFGQVVMAEAVGIDKDKPKEAVTVAVKMLKD
DATEKDLSDLVSEMEMMKMIGKHKNIINLLGACTQDGPLYVIVEYASKGNLREY
LRARRPPGMEYSYDINRVPEEQMTFKDLVSCTYQLARGMEYLASQKCIHRDLAA

RNVLVTENNVMKIADFGLARDINNIDYYKKTTNGRLPVKWMAPEALFDRVYTH
QSDVWSFGVLMWEIFTLGGSPYPGIPVEELFKLLKEGHRMDKPANCTNELYMM
MRDCWHAVPSQRPTFKQLVEDLDRILTLTTNEEYLDLSQPLEQYSPSYPDTRSSC
SSGDDSVFSPDPMPYEPCLPQYPHINGSVKT
>tr|F1PPD8|F1PPD8_CANLF Fibroblast growth factor receptor OS=Canis lupus
familiaris OX=9615 GN=FGFR2 PE=3
SV=2MVSWARFVCLAAVTMATLSLARPSFNLVEDTTLEPEEPPTKYQISQPEVYV
AAPGESLELRCLLRDAATIIWTKDGVHLGPNNRTVLIGEYLQIKGATPRDSGLYA
CTAARPVDSEAVYFMVNVTDAISSGDDEDDTDGSEDFVSENSNNKRAPYWTNT
EKMEKRLHAVPAANTVKFRCPAGGNPTPTMRWLKNGKEFKQEHRIGGYKVRN
QHWSLIMESVVPSDKGNYTCVVENEYGSINHTYHLDVVERSPHRPILQAGLPAN
ASTVVGGDVEFVCKVYSDAQPHIQWIKHVEKNGSKYGPDGLPYLKVLKHSGINS
SNAEVLALFNVTEEDAGEYICKVSNYIGQANQSAWLTVLPKQQAPVREKEITASP
DYLEIAIYCIGVFLIACMVVTVILCRMKTTTKKPDFSSQPAVHKLTKRIPLRRQVT
VSAESSSSMNSNTPLVRITTRLSSTADTPMLAGVSEYELPEDPKWEFPRDKLTLG
KPLGEGCFGQVVMAEAVGIDKEKPKEAVTVAVKMLKDDATEKDLSDLVSEME
MMKMIGKHKNIINLLGACTQDGPLYVIVEYASKGNLREYLRARRPPGMEYSYDI
NRVPEEQMTFKDLVSCTYQLARGMEYLASQKCIHRDLAARNVLVTENNVMKIA
DFGLARDINNIDYYKKTTNGRLPVKWMAPEALFDRVYTHQSDVWSFGVLMWEI
FTLGGSPYPGIPVEELFKLLKEGHRMDKPANCTNELYMMMRDCWHAVPSQRPT
FKQLVEDLDRILTLTTNEEYLDLSQPLEQYSPSYPDTRSSCSSGDDSVFSPDPMPY
EPCLPQYPHVNGSVKT
>tr|F1MNW2|F1MNW2_BOVIN Fibroblast growth factor receptor OS=Bos taurus
OX=9913 GN=FGFR2 PE=3 SV=2
MGLTSTWRYGRGQGIGTVTMVSWGRFLCLVVVTMATLSLARPSFNLVDDTTVE
PEEPPTKYQISQPEVYVAAPRESLELRCLLRDAAMISWTKDGVHLGPNNRTVLIG
EYLQIKGATPRDSGLYACTAARNVDSETVYFMVNVTDAISSGDDEDDADGSEDF
VSENSNSKRAPYWTNTEKMEKRLHAVPAANTVKFRCPAGGNPTPTMRWLKNG
KEFKQEHRIGGYKVRNQHWSLIMESVVPSDKGNYTCVVENDYGSINHTYHLDV
VERSPHRPILQAGLPANASTVVGGDVEFVCKVYSDAQPHIQWIKHVEKNGSKYG
PDGLPYLKVLKAAGVNTTDKEIEVLYIRNVTFEDAGEYTCLAGNSIGISFHSAWL
TVLPAPVREKEIPASPDYLEIAIYCIGVFFIACMVVTVILCRMRNTTKKPDFSSQPA
VHKLTKRIPLRRQVSAESSSSMNSNTPLVRITTRLSSTADTPMLAGVSEYELPEDP
KWEFPRDKLTLGKPLGEGCFGQVVMAEAVGIDKEKPKEAVTVAVKMLKDDAT
EKDLSDLVSEMEMMKMIGKHKNIINLLGACTQDATGPLYVIVEYASKGNLREYL
RARRPPGMEYSYDINRVPEEQMAFKDLVSCTYQLARGMEYLASQKCIHRDLAA
RNVLVTENNVMKIADFGLARDINNIDYYKKTTNGRLPVKWMAPEALFDRVYTH
QSDVWSFGVLMWEIFTLGGSPYPGIPVEELFKLLKEGHRMDKPANCTNELYMM
MRDCWHAVPSQRPTFKQLVEDLDRILTLTTNEEYLDLSQLLEQYSPSYPDTRSSC
SSGDDSVFSPDPMPYEPCLPQYPHRNGSVKT
>tr|E9QK53|E9QK53_MOUSE Fibroblast growth factor receptor OS=Mus musculus
OX=10090 GN=Fgfr2 PE=1 SV=1
MGLPSTWRYGRGPGIGTVTMVSWGRFICLVLVTMATLSLARPSFSLVEDTTLEPE
EPPTKYQISQPEAYVVAPGESLELQCMLKDAAVISWTKDGVHLGPNNRTVLIGE

YLQIKGATPRDSGLYACTAARTVDSETWYFMVNVTDAISSGDDEDDTDSSEDVV
SENRSNQRAPYWTNTEKMEKRLHAVPAANTVKFRCPAGGNPTPTMRWLKNGK
EFKQEHRIGGYKVRNQHWSLIMESVVPSDKGNYTCLVENEYGSINHTYHLDVVE
RSPHRPILQAGLPANASTVVGGDVEFVCKVYSDAQPHIQWIKHVEKNGSKYGPD
GLPYLKVLKAAGVNTTDKEIEVLYIRNVTFEDAGEYTCLAGNSIGISFHSAWLTV
LPAPVREKEITASPDYLEIAIYCIGVFLIACMVVTVIFCRMKTTTKKPDFSSQPAVH
KLTKRIPLRRQVTVSAESSSSMNSNTPLVRITTRLSSTADTPMLAGVSEYELPEDP
KWEFPRDKLTLGKPLGEGCFGQVVMAEAVGIDKDKPKEAVTVAVKMLKDDAT
EKDLSDLVSEMEMMKMIGKHKNIINLLGACTQDGPLYVIVEYASKGNLREYLRA
RRPPGMEYSYDINRVPEEQMTFKDLVSCTYQLARGMEYLASQKCIHRDLAARN
VLVTENNVMKIADFGLARDINNIDYYKKTTNGRLPVKWMAPEALFDRVYTHQS
DVWSFGVLMWEIFTLGGSPYPGIPVEELFKLLKEGHRMDKPTNCTNELYMMMR
DCWHAVPSQRPTFKQLVEDLDRILTLTTNEEYLDLTQPLEQYSPSYPDTRSSCSSG
DDSVFSPDPMPYEPCLPQYPHINGSVKT
>tr|F1LNW0|F1LNW0_RAT Fibroblast growth factor receptor OS=Rattus norvegicus
OX=10116 GN=Fgfr2 PE=3 SV=1
MGLPSTWRYGTGPGIGTVTMVSWGRFICLVLVTMATLSLARPSFSLVEDTTLEPE
EPPTKYQISQPEACVVAPGESLELRCMLKDAAVISWTKDGVHLGPNNRTVLIGEY
LQIKGATPRDSGLYACAAARTVDSETLYFMVNVTDAISSGDDEDDTDSSEDFVSE
NRSNQRAPYWTNTEKMEKRLHAVPAANTVKFRCPAGGNPTPTMRWLKNGKEF
KQEHRIGGYKVRNQHWSLIMESVVPSDKGNYTCLVENEYGSINHTYHLDVVERS
PHRPILQAGLPANASTVVGGDVEFVCKVYSDAQPHIQWIKHVEKNGSKYGPDGL
PYLKVLKAAGVNTTDKEIEVLYIRNVTFEDAGEYTCLAGNSIGISFHSAWLTVLP
APVREKEITASPDYLEIAIYCIGVFLIACMVVTVIFCRMKTTTKKPDFSSQPAVHK
LTKRIPLRRQVTVSAESSSSMNSNTPLVRITTRLSSTADTPMLAGVSEYELPEDPK
WEFPRDKLTLGKPLGEGCFGQVVMAEAVGIDKDRPKEAVTVAVKMLKDDATE
KDLSDLVSEMEMMKMIGKHKNIINLLGACTQDGPLYVIVEYASKGNLREYLRAR
RPPGMEYSYDINRVPEEQMTFKDLVSCTYQLARGMEYLASQKCIHRDLAARNVL
VTENNVMKIADFGLARDINNIDYYKKTTNGRLPVKWMAPEALFDRVYTHQSDV
WSFGVLMWEIFTLGGSPYPGIPVEELFKLLKEGHRMDKPTNCTNELYMMMRDC
WHAVPSQRPTFKQLVEDLDRILTLTTNEEYLDLTQPLEQYSPSYPDTRSSCSSGD
DSVFSPDPMPYDPCLPQYPHINGSVKT
>tr|F1NEE9|F1NEE9_CHICK Fibroblast growth factor receptor OS=Gallus gallus
OX=9031 GN=FGFR2 PE=3 SV=3
MVSWDSGCLICLVVVTMAGLSLARPSFNLVVEDATLEPEEPPTKYQISQPDVHSA
LPGEPLELRCQLKDAVMISWTKDGVPLGPDNRTVIIGEYLQIKDASPRDSGLYAC
TAIRTLDSDTLYFIVNVTDALSSGDDEDDNDGSEDFVNDSNQMRAPYWTHTDK
MEKRLHAVPAANTVKFRCPAMGNPTPTMRWLKNGKEFKQEHRIGGYKVRNQH
WSLIMESVVPSDKGNYTCIVENQYGSINHTYHLDVVERSPHRPILQAGLPANASA
VVGGDVEFVCKVYSDAQPHIQWIKHVERNGSKYGPDGLPYLQVLKAAGVNTTD
KEIEVLYIRNVTFEDAGEYTCLAGNSIGISFHTAWLTVLPAPEKEKEFPTSPDYLEI
AIYCIGVFLIACMVLTVILCRMKNTTKKPDFSSQPAVHKLTKRIPLRRQVTVSADS
SSSMNSNTPLVRITTRLSSTADAPMLAGVSEYELPEDPKWEFPRDKLTLGKPLGE
GCFGQVVMAEAVGIDKDRPKEAVTVAVKMLKDDATEKDLSDLVSEMEMMKMI

GKHKNIINLLGACTQDGPLYVIVEYASKGNLREYLRARRPPGMEYSFDINRVPEE
QMTFKDLVSCTYQLARGMEYLASQKCIHRDLAARNVLVTENNVMKIADFGLAR
DINNIDYYKKTTNGRLPVKWMAPEALFDRVYTHQSDVWSFGVLMWEIFTLGGS
PYPGIPVEELFKLLKEGHRMDKPANCTNELYMMMRDCWQAVPSQRPTFKQLVE
DLDRILTLTTNEEYLDLSGPLEQYSPSYPDTRSSCSSGDDSVFSPDPMPYEPCLPK
YQHMNGSVKT
>sp|Q8JG38|FGFR2_DANRE Fibroblast growth factor receptor 2 OS=Danio rerio
OX=7955 GN=fgfr2 PE=1 SV=1
MFARGWLLGALLLMTLATVSVARPSLKIDLVNTSAPEEPPTKNQNCVPVLFSVH
PGELLKLKCPLSGADDVVWTKDSSSLRPDNRTLVARDWLQISDATPKDSGLYSC
SATGLRDCDVFSFIVNVTDAISSGDDEDDTERSDDVGADGEQMRLPYWTFPEKM
EKKLHAVPAANTVKFRCAAAGNPKPKMRWLKNAKPFRQEDRMGGYKVRLQH
WTLIMESVVPSDKGNYTCLVENQYGSIDHTYTLDVVERSPHRPILQAGLPANVTV
QVGQDAKFVCKVYSDAQPHIQWLQHYTKNGSCCGPDGLPYVRVLKTAGVNTT
DKEIEVLYLPNVTFEDAGEYTCLAGNSIGISYHTAWLTVHPAETNPIETDYPPDYV
EIAIYCIGVFLIACMVVIVVVCRMRTSAKKPDFSSQPAVHKLTKQIPLRRQVTVSS
DSSSSMSSSTPLVRITTRRSSAHDDPIPEYDLPEDPRWEFSRDKLTLGKPLGEGCF
GQVVMAEALGIDKDKPKEAVTVAVKMLKDDATEKDLSDLVSEMEMMKMIGRH
KNIINLLGACTQDGPLYVIVEYASKGNLREYLRARRPPGMEYSYDIARVSDEPLT
FKDLVSCTYQVARGMEYLASQKCIHRDLAARNVLVTESNVMKIADFGLARDVH
NIDYYKKTTNGRLPVKWMAPEALFDRVYTHQSDVWSFGVLMWEIFTLGGSPYP
GIPVEELFKLLKEGHRMDKPANCTNELYMMMKDCWHAISSHRPTFKQLVEDLD
RILTLATNEEYLDLCAPVEQYSPSFPDTRSSCPSGDDSVFSHDPLADEPCLPKYQH
INGGIKT
>tr|A4IHW8|A4IHW8_XENTR Fibroblast growth factor receptor OS=Xenopus tropicalis
OX=8364 GN=fgfr2 PE=2 SV=1
MGMSLVWRSGKAGGGGHADRMLVLVLLGLLLVSRTIARPSYHMAEDTTSEPEE
PPAKYQISKADVFPVLPGEPLDLRCPLADGPPVTWNKDGAKLEVNNRTLIVRNY
LQIKETTPRDSGLYSCSVLKNSHFFHVNVTEASSSGDDEDDNDGSEDFTNDNNNI
RAPYWTNTEKMEKKLHAVPAANTVKLRCPAGGNPTPRMRWLKNGKEFKQEHR
IGGYKVRNQHWSLIMESVVPSDKGIYTCIVENEHGSINHTYHLDVIERSSHRPILQ
AGLPANTTAMVGGDAEFVCKVYSDAQPHIRWVRYIEKNGSRFGVDGLPYIKVL
KAAGVNVTDEEIEVLYVRNVSFEDAGEYTCIAGNSIGISQHSAWLTVHPATVSPG
EDNPVPYYMEIGIYSAGIFIIFCMVVICVVCRMRQGAKKKKNFTGPPVHKLTKRIP
LHRQVSADSSSSMNSTTPLVRITTRLLSSTDAMPLPNVSEYELPHDPLWEFSRDKL
TLGKPLGEGCFGQVVMAEALGIDKDRPKESVTVAVKMLKDDATEKDLADLVSE
MEMMKIIGKHKNIINLLGACTQGGTLYVIVEYAAKGNLRQYLRARRPLEMEYSF
DVTRVPDEQMTFKDLVSCTYQIARGMEYLASQKCIHRDLAARNVLVTENNVMK
IADFGLARDVNNIDYYKKTTNGRLPVKWMAPEALFDRVYTHQSDVWSFGVLM
WEIFTLGGSPYPGIPVEELFKLLKEGHRMDKPGNCTNELYMMMRDCWHAIPSHR
PTFKQLVEDLDRILTLTTNEEYLDLSAPLEQYSPSFPDSSCSASSSSGDDSVFSPDP
MPHDPCLPKFPHVNGVVKT

APPENDIX II

The results of FoldX Modeling on the mutants are mentioned here. The Tables have energy values for all the mutants on every structure used in modeling. Stability is the potential energy of the mutant. Mutant-WT is stability of mutant structure – stability of Wild Type structure and FoldX energy is the value given by FoldX. All the values are in kacl/mol.

Sr No	Mutation	Stability	Mutant-WT	FoldX	Type
1	E475K	-12.02	-0.49	-0.0176213	Cancer
2	K526E	-12.17	-0.64	-0.644175	Syndrome
3	D530N	-11.83	-0.3	-0.27725	Cancer
4	I547V	-10.74	0.79	0.79778	Cancer
5	N549K	-11.18	0.35	1.53597	Cancer
6	N549H	-9.83	1.7	2.88411	Syndrome
7	N549T	-10.74	0.79	2.31738	Syndrome
8	E565G	-9.86	1.67	2.72189	Syndrome
9	E565A	-10.1	1.43	2.20475	Syndrome
10	E574K	-11.53	0	0.428975	Cancer
11	A628T	-12.14	-0.61	-0.0423916	Syndrome
12	E636K	-11.85	-0.32	0.142083	Cancer
13	M640I	-9.69	1.84	2.90628	Cancer
14	K641R	-12.11	-0.58	-0.166695	Syndrome
15	I642V	-10.55	0.98	1.12599	Cancer
16	A648T	-10.72	0.81	0.812361	Cancer,Syndrome
17	D650V	-12.63	-1.1	-1.09612	Syndrome
18	K659E	-13.63	-2.1	-0.648642	Cancer
19	K659M	-11.73	-0.2	-0.0665327	Cancer
20	K659N	-11.18	0.35	0.375055	Cancer,Syndrome

21	K659Q	-12.01	-0.48	-0.29223	Syndrome
22	K659T	-11.23	0.3	0.328151	Syndrome
23	G663E	-13.19	-1.66	-0.200961	Syndrome
24	R678G	-15.83	-4.3	-2.77075	Syndrome
25	S688F	-1.21	10.32	10.3186	Cancer
26	G701S	-9.18	2.35	2.35022	Cancer
27	P708S	-9.6	1.93	1.93416	Cancer
28	R759Q	-11.5	0.03	0.034757	Cancer

2PSQ Chain A FoldX modeling results.

Sr No	Mutation	Stability	Mutant-WT	FoldX	Type
1	E475K	-6.81	-0.77	0.374304	Cancer
2	K526E	-6.7	-0.66	-0.599949	Syndrome
3	D530N	-5.91	0.13	0.215462	Cancer
4	I547V	-6.42	-0.38	0.882638	Cancer
5	N549K	-9	-2.96	1.35076	Cancer
6	N549H	3.94	9.98	10.9037	Syndrome
7	N549T	-8.09	-2.05	2.11862	Syndrome
8	E565G	-6.07	-0.03	2.56068	Syndrome
9	E565A	-6.99	-0.95	1.63468	Syndrome
10	E574K	-5.55	0.49	0.487605	Cancer
11	A628T	-6.64	-0.6	0.276559	Syndrome
12	E636K	-5.29	0.75	0.750145	Cancer
13	M640I	-4.35	1.69	2.60144	Cancer
14	K641R	-7.59	-1.55	1.0828	Syndrome
15	I642V	-5.03	1.01	1.0058	Cancer
16	A648T	-4.91	1.13	1.13074	Cancer,Syndrome
17	D650V	-5.57	0.47	0.487702	Syndrome
18	K659E	-7.25	-1.21	-1.02156	Cancer
19	K659M	-5.48	0.56	0.555619	Cancer
20	K659N	-5.38	0.66	0.656886	Cancer,Syndrome
21	K659Q	-5.53	0.51	0.509913	Syndrome
22	K659T	-5.29	0.75	0.749185	Syndrome
23	G663E	-7.03	-0.99	-0.125286	Syndrome
24	R678G	-8.05	-2.01	-1.78792	Syndrome
25	S688F	2.68	8.72	8.82344	Cancer
26	G701S	-3.99	2.05	2.04672	Cancer
27	P708S	-4.06	1.98	2.03928	Cancer

28	R759Q	-6.38	-0.34	-0.276328	Cancer

2PSQ Chain B FoldX modeling results.

Sr No	Mutation	Stability	Mutant-WT	FoldX	Type
1	E475K	-5.4	-0.55	0.413131	Cancer
2	K526E	-7.25	-2.4	-0.522624	Syndrome
3	D530N	-5	-0.15	-0.140033	Cancer
4	I547V	-4.11	0.74	1.0452	Cancer
5	N549K	-6.17	-1.32	-1.20724	Cancer
6	N549H	-4.75	0.1	0.209906	Syndrome
7	N549T	-4.73	0.12	0.220434	Syndrome
8	E565G	-2.74	2.11	2.22817	Syndrome
9	E565A	-3.95	0.9	1.0186	Syndrome
10	E574K	-4.12	0.73	0.800792	Cancer
11	A628T	-3.98	0.87	1.14309	Syndrome
12	E636K	-5.5	-0.65	0.322243	Cancer
13	M640I	-2.73	2.12	2.98507	Cancer
14	K641R	-4.99	-0.14	0.450132	Syndrome
15	I642V	-3.86	0.99	1.17217	Cancer
16	A648T	-3.28	1.57	1.56564	Cancer,Syndrome
17	D650V	-3.12	1.73	1.72913	Syndrome
18	K659E	-6.02	-1.17	0.457548	Cancer
19	K659M	-6.42	-1.57	-1.5022	Cancer
20	K659N	-4.76	0.09	0.156327	Cancer,Syndrome
21	K659Q	-5.4	-0.55	-0.496762	Syndrome
22	K659T	-4.38	0.47	0.536941	Syndrome
23	G663E	-4.56	0.29	1.90395	Syndrome
24	R678G	-5.7	-0.85	-0.853792	Syndrome
25	S688F	-1.24	3.61	5.64689	Cancer
26	G701S	-2.82	2.03	2.02666	Cancer
27	P708S	-3.69	1.16	1.16715	Cancer
28	R759Q	-5.14	-0.29	-0.289966	Cancer

2PVF Chain A FoldX modeling results.

APPENDIX III

The R scripts that were used in the study are mentioned here.

Following is the script to separate the two chains of protein in different PDB files

```
library(bio3d)
pdb = read.pdb(PDB filename)
inds = atom.select(pdb,chain=”Chain name”)
newpdb = trim.pdb(pdb,inds)
write.pdb(newpdb,file="Output PDB filename")
```

Following is the script for normality test, T-test and Mann Whitney Wilcoxon test.

```
shapiro.test(dataset)
t.test(dataset1, dataset2)
wilcox.test(dataset1, dataset2)
```

BIOGRAPHY

Snehal Vilas Sambare is master's student at George Mason University in Bioinformatics and Computational Biology department. She received her Bachelor of Engineering in Computer Science from Ramdeobaba College of Engineering and Management, Nagpur, India in 2016. Her research interests are in proteins, cancer genomics, epigenetics and computer science.