# In-silico Analysis of Damaging SNPs of Human DHFR Protein and Determining their Functional and Structural Consequences

SHUBHAM RAMLE

*A dissertation submitted for the partial fulfilment of the BS-MS dual degree in Biological sciences*.


Indian Institute of Science Education and Research, Mohali

April 2020

# *Certificate of Examination*

This is to certify that the dissertation titled **"In silico Analysis of Damaging SNPs of Human DHFR protein and Determining their Functional and Structural Consequences"** submitted by **Shubham Ramle** (Reg. No. MS15163) for the fulfilment of **BS-MS dual degree programme** of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Shashi Bhushan Pandit  Dr Sabyasachi Rakshit  Dr. Monika Sharma

(Supervisor)

.

# *Declaration*

The work presented in this dissertation has been carried out by me under the guidance of Dr. Monika Sharma at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Shubham Ramle

(Candidate)

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Monika Sharma

(Supervisor)

# *Acknowledgements*

Foremost, I would like to express my deep and sincere gratitude to my supervisor Dr. Monika Sharma for her immense support, kindness, and encouragement for my thesis and research. I am thankful to her for giving me the opportunity to work under her guidance.

I would like to thank Mr. Milind Kale and Mr. Ashish Kumar Swain for their insightful comments during my research work.

Finally, I would like to thank my parents and IISER Mohali for supporting me through my thesis.

# *List of Tables*

# *List of Figures*

# *Abbreviations*

**SNPs** Single Nucleotide Polymorphisms

**DHFR** Dihydrofolate Reductase

**SIFT** Sorting Intolerant From Tolerant

**PolyPhen2** Polymorphism Phenotyping v2

**PHD-SNP** Predictor of human Deleterious Single Nucleotide Polymorphisms

**PROVEAN** Protein Variation Effect Analyzer

**YASARA** Yet Another Scientific Artificial Reality Application

**PTM Site** Post-Translational Modifications Site

# *Abstract*

DHFR is an enzyme that is ubiquitous in all organisms. DHFR's main role is to maintain tetrahydrofolate at intracellular levels, which is required for certain cofactors to biosynthesize purine, pyrimidine, and several amino acids. As, it is the primary source of THF, it is vulnerable to quickly proliferating cells, which ends in making it a preferable target for many essential anticancer and antimicrobial drugs. With a set of SNPs data accessible via dbSNP, my thesis is planned to point out functional SNPs in DHFR by applying various in silico tools such as SIFT, PolyPhen2, PROVEAN, SNP&GO, PHD-SNP, Consurf, ModPred, MutPred, Tm-Align and lastly Project HOPE was used for estimating the impact of SNPs on a protein, functionally and structurally, PTM sites and energy minimization analysis. 241 SNPs found to be non-synonymous among 7967 DHFR SNP entries out of which SIFT estimated 64 nsSNPs as non-tolerable, while PolyPhen-2 estimated 60. An aggregate result was obtained by evaluating five tools with different perceptions where twenty-five nsSNPs were considered most likely to exert deleterious impact. To evaluate mutation's functional and structural impact on DHFR, Phyre2 was used to create 3D models of mutated proteins. Results from FoldX and Project HOPE reinforced the initial findings, as they predicted, upon mutation there will be significant structural and functional instability. To determine whether the mutations lies in any protein's functional domains. Considering these analyses, my study picked up 10 most damaging nsSNPs.

# Contents

## Table of Contents

# Chapter 1

# Introduction

## 1.1 What are SNPs?

Single nucleotide polymorphisms or SNPs are substitution of a nucleotide of one base with a different one, that have a tendency to occur in more than 1% of the total population. SNP, reflects the most prevalent type of human genetic mutation. It was reported that for nearly 93% of all human genes, at least one SNP would be there[1]. Therefore they are responsible for producing most biological differences between individuals.

SNPs may fall within the coding regions or non-coding gene regions or between two genes in the intergenomic region[2], described as follows:-

1) Linked SNPs:- They are not present in the genes and that's why can't influence the role of proteins. However, they can cause a certain drug response or they can induce a risk of having a some type of disease.
2) Causative SNPs:- These SNPs alters the functioning of a protein, either it leads to a some type of disease or it may have an affect on how a person reacts to some medication. These SNPs are further divided into two forms:-
    1) Coding SNPs:- They are found inside a gene's coding region and are known to alter the protein's amino acid sequence.
    2) Non-coding SNPs:- They are found in the gene's regulatory sequences, alter the timing, location or gene expression level.

Nonsynonymous coding SNPs (nsSNPs) are known to make a major effect on phenotype by altering the amino acid sequence of a protein. Since SNPs induce modification on protein's amino acid sequence, which can leads to having deleterious effects on the protein structure, operation, stability or solubility[3][4]. Therefore, nsSNPs are found to have a significant role in the functional and structural range of proteins and sometimes are correlated with human diseases. Previous research has shown that more than 50% of mutations are linked with inherited genetic defects which are caused by nsSNPs[5][6].

# 1.2 Dihydrofolate Reductase (DHFR)

Dihydrofolate Reductase (DHFR) is a key metabolic enzyme whose role is to reduce dihydrofolate to tetrahydrofolate, which can then be integrated into the Purine and amino acid synthesis. DHFR is known as an oxidoreductase that uses NADP+ as the electron acceptor[7]. DHFR contains 159 amino acids, around 18KDa and thus it's a small protein. With eight main B strands and four helices, it has an a/b arrangement. The protein may be assumed to consist of two subdomains, separated by the active site cleft. The Adenosine binding loop made up of 38-88 residues, and the major subdomain is composed of 100 residues. Three loops can be found in the major subdomain and they make up around 50% of this domain. These are residues which vary from 9-24 (Met20 loop), residues which vary from 116-132 (F-G loop)and residues which vary from 142-150 (G-H loop). During catalysis, the Met20 loop assumes multiple conformations and accommodation of ligands is held possible by the motion of 'hinge bending' around Lys 38 and Val 88 of the Adenosine binding domain.

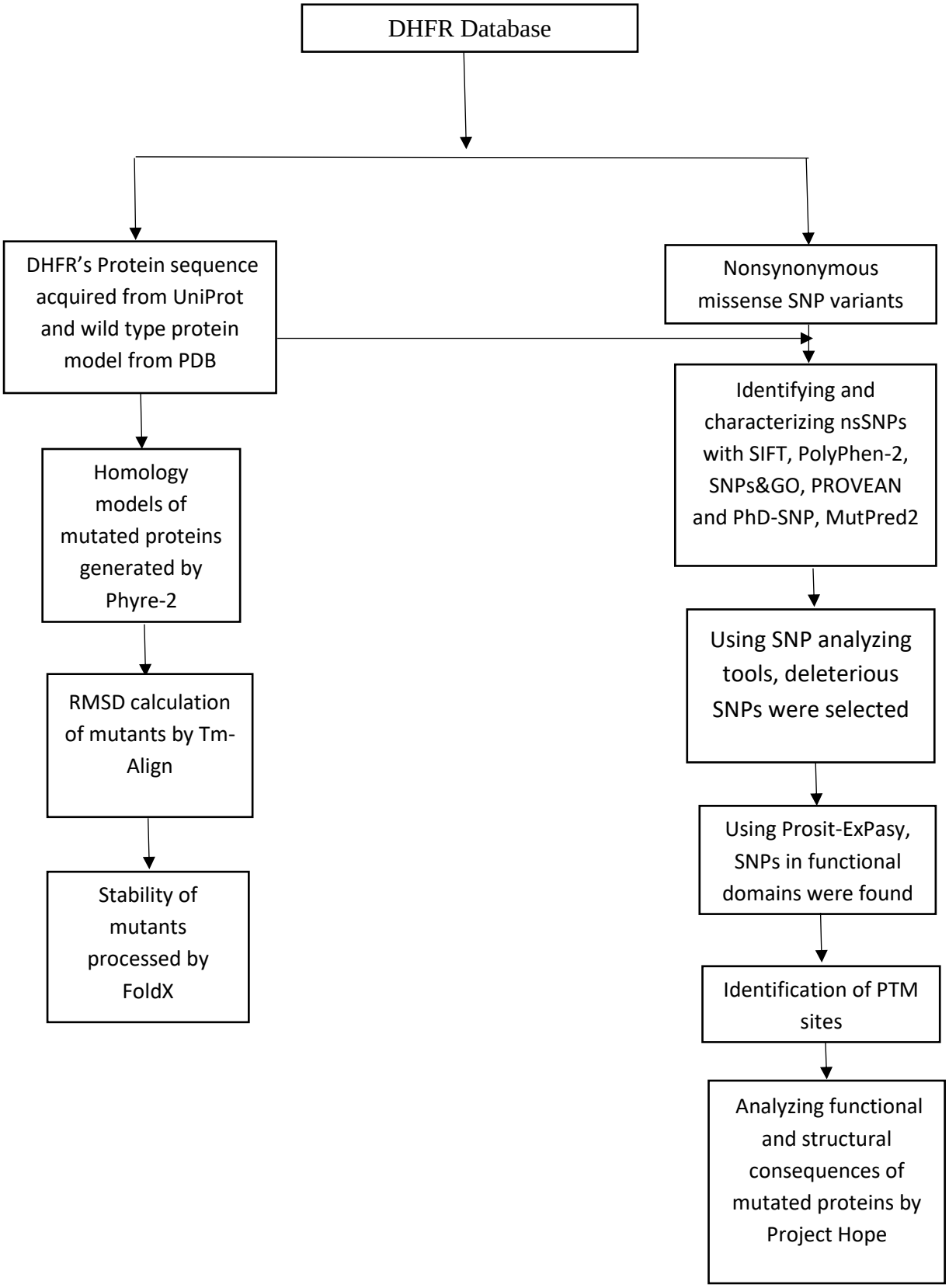
DHFR role is to catalyse the reduction of 7,8-dihydrofolate, using reduced NADPH, to 5,6,7,8-tetrahydrofolate. This method now became a crucial template for deciphering enzyme catalysis, and crystallography has identified the intermediates in the catalytic chain.

Because DHFR is so strategically located in every organism's metabolic homeostasis, it has become the goal of preference for anti-microbial and anti-cancer therapy. This enzyme's inhibitors are basically folate mimics, methotrexate that was first developed to block human DHFR and used as a medication for cancer and autoimmune disorders. Trimethoprim, another folate analogue has been established as an antibacterial agent with far more binding specificity to bacterial DHFR than its counterpart to mammals. Both drugs bind at the enzyme's active site and are bound irreversibly, blocking the activity of enzymes.

In my thesis, DHFR gene's in-silico analysis was carried out to determine deleterious mutations. My thesis involved (1) collection of nsSNPs from the accessible databases, (2) finding damaging or deleterious nsSNPs by using various in-silico tools, (3) analyzing protein evolutionary conservation sites, (4) finding PTM sites (5) estimating the changes a substitution brings on secondary structure of protein.

# Chapter 2

# Materials and Methods



Above flow chart shows generally the method for identifying and characterizing deleterious nsSNPs in DHFR, alongside with the study of the changes due to a mutation on protein's structure.

## 2.1 Retrieving nsSNPs

Data regarding DHFR's gene related single nucleotide polymorphisms (SNP IDs and protein accession number) was obtained from the NCBI dbSNP ( Single Nucleotide Polymorphism Database)[8] and DHFR's protein structure and its sequence was subsequently obtained via Uniprot[9] and the RCSB Protein database[10].

## 2.2 Determining Damaging nsSNPs by SIFT

SIFT[11] is an is silico tool which detects non-synonymous SNPs which are deleterious. During human mutagenesis study, the functionally benign and deleterious polymorphisms may be distinguished via SIFT. To locate homologous sequences, the SIFT software algorithms use the SWISSPROT, TrEMBL and nr databases. The intolerance index threshold is greater than or equal to 0.05 ($\geq$0.05). In this study, chromosome coordinates, orientation (1, -1) and alleles were given as an input, obtained from dbSNP database. The SIFT value below or equal to ($\leq$0.050) depicts the damaging nature of that nsSNP.

## 2.3 Determining Damaging nsSNPs By PolyPhen2

To recognize a protein's functional significance, analysing the damaging coding nsSNPs at the structural level is crucial. PolyPhen-2[5] uses Naïve Bayes, which is a supervised learning algorithm to measure the effect of an allele change functionally , to evaluate structural implications. Protein sequence was submitted as a query, with the wild type residue and the mutant along with mutational position. PolyPhen-2 categorises coding nsSNPs as probably damaging or possibly damaging or benign by calculating the position-specific independent count (PSIC) score. If difference in PSIC score is higher for a particular mutation, then that substitution will have a greater functional impact.

## 2.4 Determining Damaging nsSNPs

SNP&GO was used for the characterization of functional nsSNPs. It collects from various functional tools like SNPs&GO[12], PHD-SNP[13] and PANTHER[14]. Single Nucleotide Polymorphism Database ( SNPs) and PhD-SNP and Gene Ontology (GO) uses analysing based method, support vector machine ( SVM), in which PANTHER determines the role of damaging coding nsSNPs by measuring the subPSEC ( position-specific evolutionary conservation) score[12][15]. PROVEAN predicts the variant to be deleterious when it scores below -2.5 threshold and neutral if it is above this value.

## 2.5 Stability Effect of Mutation

From PDB, the human DHFR crystallographic structure was obtained (PDB Id-6DAV). YASARA was used to calculate the stability change caused by the mutations. YASARA is a modelling and simulation program which by analysing the sequence or the structure of the protein, predicts the change in structural stability. It is used to predict the stability of the mutant in comparison with the wild type in form of ΔΔG value, which is the difference between the mutant's and wild type's Gibbs free energy and is calculated in Kcal/mol.

## 2.6 Determining Damaging SNPs in Conserved Regions

Evolutionary conservation scores of amino acid was predicted by ConSurf web server[16], which uses a Bayesian algorithm to score the conservation of amino acids[17]. If conservation score is 7-9, then amino acid is conserved and if score 5-6 then it's intermediate and if the score is between 1-4, then it's variable. Amino acid sequence of DHFR protein was given as the input and the regions of conservation were predicted using colour scheme and conservation score. Also structural and functional residues were predicted too. For further analysis, amino acids which are in highly conserved regions were selected for further analysis.

## 2.7 Association of nsSNPs with Disease

MutPred server is used for predicting the connection of nsSNP with disease along with the molecular impact of that specific substitution[18]. Protein sequence is given in FASTA format

as an input along with the list of mutations and predicts the 14 separate structural and functional properties and then it gives a score which predicts the cause of disease/deleterious mutation.

## 2.8 PTM Sites

To predict the PTM sites, a ModPred web server was used[19]. PTM sites are essential for evaluating nsSNPs, since the SNPs on PTM sites can be very damaging and will affect the functioning of the protein. The output is given as modifications of a particular residue with a confidence score.

## 2.9 Mutations in HEME Binding Site

For determining any disease related to a protein, it is essential to precisely find the HEME binding residues, For this, HEMEsPred was used to accurately find the HEME binding residues. Input was given as amino acid sequence of the protein and threshold was set to 0.5.

## 2.10 Protein Modelling and RMSD Calculations

From PDB, the human DHFR crystallographic structure was obtained (PDB Id-6DAV).For a wide range of research activities, the determination of macromolecule's properties and three dimensional structure is an essential element. For modelling, Phyre2 was used to create structures of the mutated proteins[20]. Phyre2 creates the protein model by selecting the best suited template for the protein. Normal mode was used for the effective estimate of the protein modelling. For Phyre2, amino acid sequence of the DHFR protein was given as an input.

TM-Align was used to calculate the RMSD values and the TM-Scores for respective mutants[21]. These values tells that how much a mutant's protein structure is different from that of the wild type.

## 2.11 Structural Consequences of nsSNPs

To estimate the structural and functional effects of the substitution, Project HOPE was used. HOPE was utilized to determine the consequence of respective mutations on DHFR's structure via molecular dynamics simulations. It tells us about the various changes a substitution brings

with it and how those changes affect the protein structure. Amino acid sequence of the DHFR protein was given as an input and further provided with the respective mutations[22].

# Chapter 3

# Results

## 3.1 nsSNPs Retrieved from dbSNP Database

Several databases can provide you with polymorphism data, like Ensembl genome browser, NCBI dbSNP database etc. The large st of all the SNP databases is NCBI dbSNP database, but the drawback is that it has validated as well as non-validated SNPs[23]. In spite of this, SNPs data was acquired from dbSNP database, for DHFR gene, since it comprises of the biggest SNP database. There were 7967 SNPs in total for the gene DHFR in dbSNP, of which 241 were nsSNPs, 1226 occurred in 3'UTR region, 835 occurred in 5'UTR. 17 nonsense SNPs were also there, but for further analysis only missense SNPs were selected.

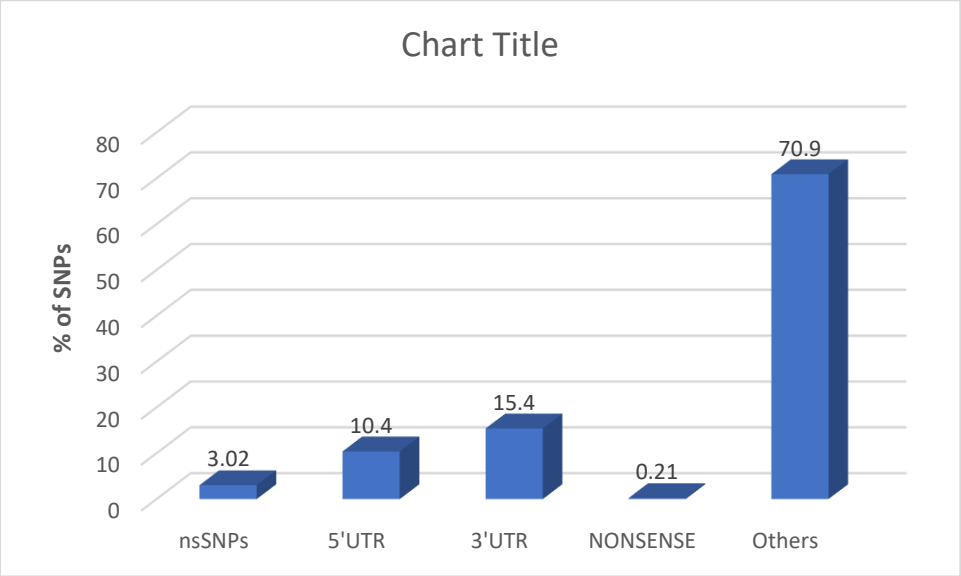



Fig.1 Data showing number of nsSNPs, SNPs found in 5'UTR, 3'UTR, nonsense, others in dbSNP database.

## 3.2 Nonsynonymous SNP Analysis

All nsSNPs were subjected to five separate in silico nsSNPs algorithm predictor to determine damaging nsSNPs. SIFT score is termed as intolerant (0.051-0.10), borderline (0.101-0.20) and tolerant (0.201-1.00)[11]. Threshold value for PROVEAN is -2.5, so if a substitution has a

score below -2.5 is deemed as deleterious and neutral if it is above it. The PolyPhen-2 findings anticipated possibly damaging, probably damaging and benign nsSNPs, with the most confident prediction being probably damaging. Those nsSNPS, which were standard in 4 of the above mentioned tools, were shortlisted.

SIFT predicted 64nsSNPs as diseased and likewise PROVEAN, PolyPhen-2, PHD-SNP and SNP&GO predicted 76,60,59 and 68 nsSNPs as diseased respectively. Furthermore, out of 112 possible nsSNPs, total 52 of them fulfilled the requirement and marked them as highly risky. Additional inquiries were held only for these 52 nsSNPs.

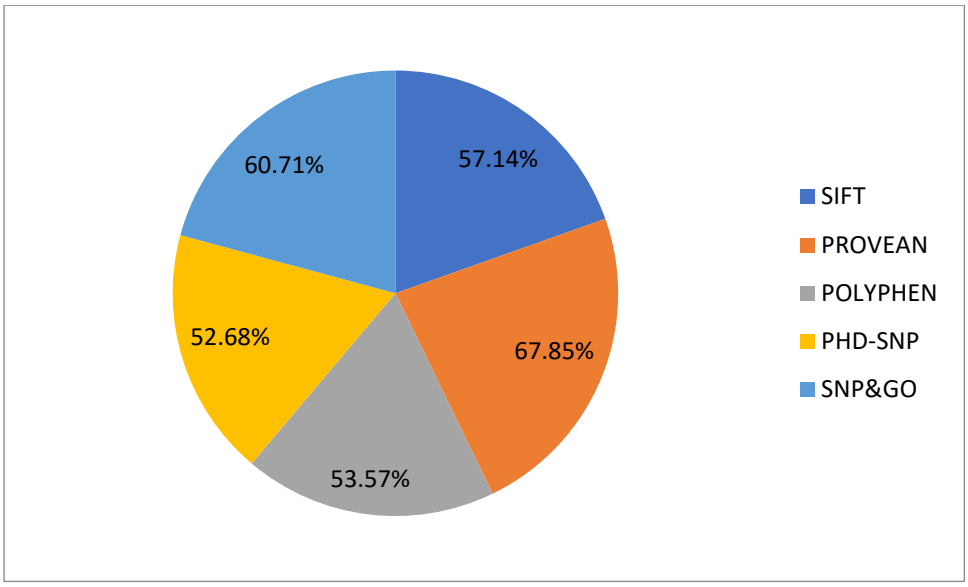



Fig.2 Data showing number of each SNPs that were found deleterious by various in-silico tools.

# 3.3 Stability Effect of Mutation

From PDB, the human DHFR crystallographic structure was obtained (PDB Id-6DAV). First, energy minimization was done on the DHFR protein structure, obtained from PDB, via FoldX tool using YASARA. Then, stability change of the 52 damaging nsSNPs was checked via again FoldX tool in YASARA. Free energy is denoted as protein's stability (ΔG), expressed in Kcal/mol. The lower the value becomes, the more stable it becomes. ΔΔG is the difference between wild-type's and mutant's free energy. Destabilization of the system occurs when mutation results in an increase of energy (ΔΔG>0 Kcal/mol), whereas when energy decreases due to mutation (ΔΔG<0 Kcal/mol), will end in stabilizing the protein structure. For my thesis I have chosen the threshold value of ΔΔG>1.5 Kcal/mol. In total, there were 25 nsSNPs which showed ΔΔG>1.5 Kcal/mol. These 25 damaging nsSNPs were selected for further analysis.

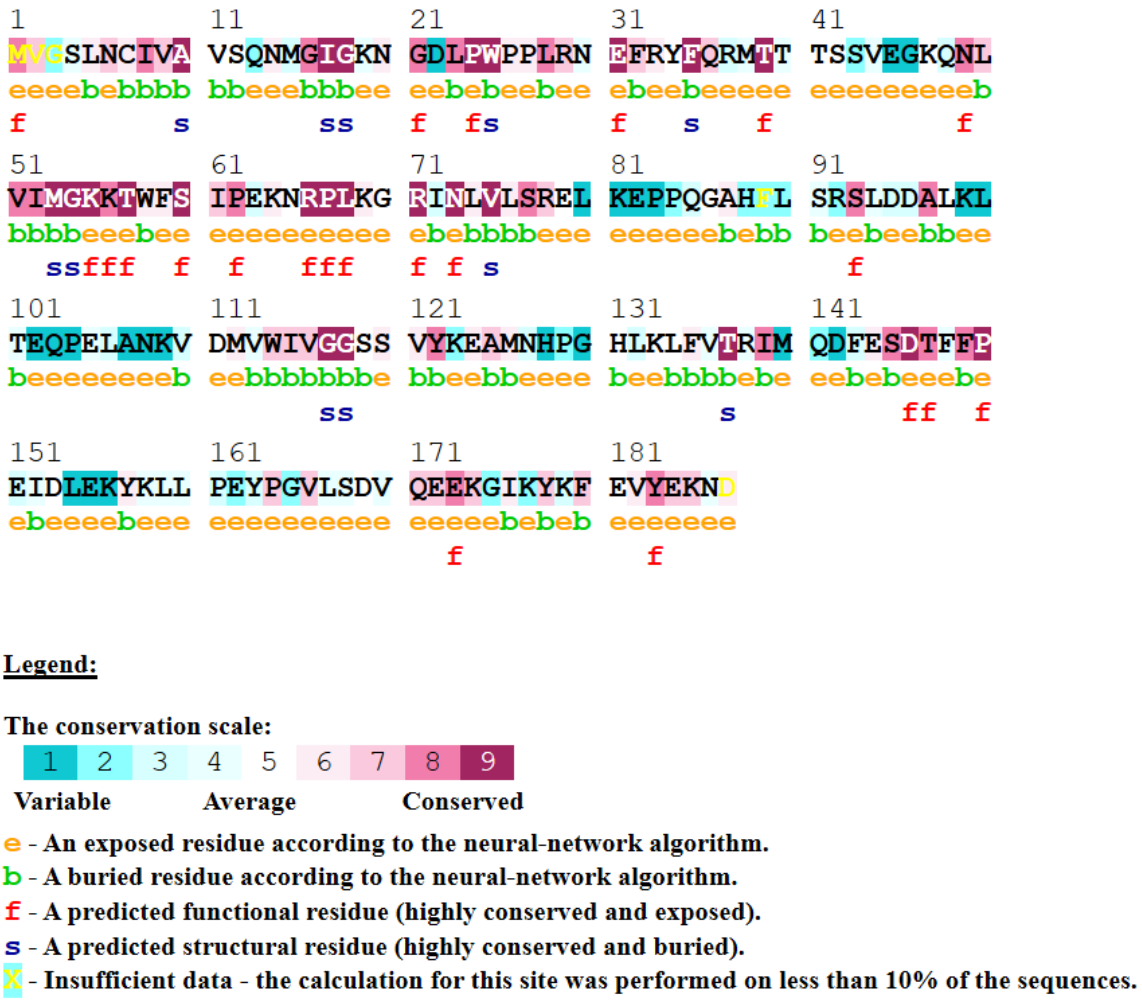
| SNP IDs | Amino Acid Change | Allele Change | ΔΔG | SNP IDs | Amino Acid Change | Allele Change | ΔΔG |
|---|---|---|---|---|---|---|---|
| **Rs371161421** | S60F | G>A | 2.09949 | **rs959468467** | Y122D | A>C | 6.01212 |
| **Rs746083870** | C7F | C>A | 8.13754 | **rs969810454** | L76P | A>G | 4.56436 |
| **Rs747824863** | T137A | T>C | 2.6678 | **rs992651401** | G118D | C>T | 1.56788 |
| **rs750437833** | G86E | C>T | 3.16115 | **rs1190920458** | D153Y<br>D153H | C>A<br>C>G | 4.30231<br>5.20077 |
| **rs755733770** | M53R<br>M53T | A>C<br>A>G | 3.76699<br>4.75206 | **rs1197915916** | Y34C | T>C | 4.3388 |
| **rs756401520** | N6K | G>C | 3.67209 | **rs1213637006** | I52T | A>G | 3.80952 |
| **rs760708940** | N73H | T>G | 15.9757 | **rs1214007828** | V113F | C>A | 9.62893 |
| **rs766341761** | G70V<br>G70D | C>A<br>C>T | 5.76554<br>3.10772 | **rs1223136219** | K179N | T>G | 1.5907 |
| **rs768327109** | F180C | A>C | 2.81047 | **rs1251977833** | N49K | A>C | 4.81038 |
| **rs954427933** | P150A | G>C | 2.3531 | **rs1301676659** | L50Q | A>T | 3.66457 |
| **rs1435266382** | R71T | C>G | 2.58139 | **rs1466161664** | D146G | T>C | 1.83404 |

Table 1: ΔΔG values of damaging nsSNPs predicted by in-silico tools.

# 3.4 Conservation Profile of Deleterious nsSNPs in DHFR

ConSurf was used for predicting residues which are in a conserved site to further investigate the possible consequences of most deleterious nsSNPs. Sites which take part in the protein function like protein-protein interaction or enzymatic sites are generally conserved sites. Loss of such sites can result in a total loss of function of the protein. So, that's why these conserved

sites are very important part for functioning of the protein. Evolutionary conservation degree of the DHFR protein at each amino acid position was calculated by ConSurf.. The output is given as shown below




**Legend:**

**The conservation scale:**

1 2 3 4 5 6 7 8 9

**Variable** **Average** **Conserved**

**e - An exposed residue according to the neural-network algorithm.**
**b - A buried residue according to the neural-network algorithm.**
**f - A predicted functional residue (highly conserved and exposed).**
**s - A predicted structural residue (highly conserved and buried).**
**X - Insufficient data - the calculation for this site was performed on less than 10% of the sequences.**

Figure 3: Output of ConSurf showing conservation scale of residues.

| Amino Acid Change | Variation Scale | Buried or Exposed Residue | Function | Amino Acid Change | Variation Scale | Buried or Exposed Residue | Function |
|---|---|---|---|---|---|---|---|
| **S60F** | 9 | Exposed | Predicted functional residue | **Y122D** | 8 | Buried | - |
| **C7F** | 6 | Buried | - | **L76P** | 7 | Buried | - |
| **T137A** | 9 | Buried | Predicted structural residue | **G118D** | 9 | Buried | Predicted structural residue |
| **G86E** | 5 | Exposed | - | **D153Y**<br>**D153H** | 5 | Exposed | - |
| **M53R**<br>**M53T** | 9 | Buried | Predicted structural residue | **Y34C** | 5 | Exposed | - |
| **N6K** | 7 | Exposed | - | **I52T** | 8 | Buried | - |
| **N73H** | 9 | Exposed | Predicted functional residue | **V113F** | 3 | Buried | - |
| **G70V**<br>**G70D** | 5 | Exposed | - | **K179N** | 3 | Exposed | - |
| **F180C** | 7 | Buried | - | **N49K** | 9 | Exposed | Predicted functional residue |
| **P150A** | 9 | Exposed | Predicted functional residue | **L50Q** | 7 | Buried | - |
| **R71T** | 9 | Exposed | Predicted functional residue | **D146G** | 9 | Exposed | Predicted functional residue |

Table 2: Conservation degree of DHFR protein at damaging mutants positions.

In the above ConSurf analysis, 9 residues (S60F, T137A, M53R, M53T, N73H, P150A, G118D, N49K, R71T, D146G) which were depicted as essential structural and functional residues. This depicts that the corresponding mutations S60F, T137A, M53R, M53T, N73H,

P150A, G118D, N49K, R71T, D146G to nsSNPs rs371161421, rs747824863, rs755733770, rs760708940, rs954427933, rs992651401, rs1251977833, rs1435266382, rs1466161664 can significantly disrupt the functional and structural properties of DHFR gene.

## 3.5 Prediction of Connection of Substitution with Disease by MutPred2

MutPred's results consists of a general score (g), which tells whether the substitution is deleterious or not and the characteristic scores (p), where P-value is the score which tells the influence of a substitution on the functional or structural properties of a protein. If general score>0.75 and P-value<0.05, then it is confident hypothesis and if general score>0.75 and P-value<0.01, then it is very confident hypothesis.

| SNPs IDs | Substitution | General Score | Effects |
|---|---|---|---|
| **rs371161421** | S60F | 0.927 | A(DNA binding) (P=0.0037)<br>G(Allosteric site at K55) (P=0.03)<br>L(Acetylation at K64) (P=0.05)<br>L(Catalytic site at K55) (P=0.02) |
| **rs747824863** | T137A | 0.834 | L( Relative solvent accessibility) (P=0.01)<br>A(Metal binding) (P=0.0067)<br>L(Strand) (P=0.03)<br>A(Transmembrane protein) (P=0.04) |
| **rs755733770** | M53R | 0.966 | L(Catalytic site at K55) (P=0.00033)<br>A(DNA binding) (P=0.00071)<br>A(Disordered interface) (P=0.0091)<br>L(Strand) (P=0.05) |

| | | | |
|---|---|---|---|
| | | | G(Allosteric site at M53 )(P=0.01)<br>A(Stability) (P=0.01) |
| **rs755733770** | M53T | 0.963 | A(Stability) (P=0.0019)<br>A(DNA binding) (P=0.0055)<br>L(Allosteric site at K55) (P=0.04) |
| **rs760708940** | N73H | 0.923 | L(Allosteric site at R71) (P=0.03)<br>A(Metal binding )(P=0.04)<br>A(Stability) (P=0.04) |
| **rs954427933** | P150A | 0.881 | A(Ordered interface) (P=0.03)<br>L(Relative solvent accessibility) (P=0.02)<br>A(Metal binding) (P=0.0092)<br>A(Transmembrane protein) (P=0.02)<br>L(Catalytic site at S145) (P=0.03) |
| **rs992651401** | G118D | 0.954 | A(Metal binding) (P=0.0001)<br>G(Catalytic site at G118) (P=0.00098)<br>G(Allosteric site at G118) (P=0.0029) |
| **rs1251977833** | N49K | 0.899 | A(Ordered interface) (P=0.008)<br>L(Relative solvent accessibility) (P=0.03)<br>L(Strand )(P=0.05)<br>G(Acetylation at K47 )(P=0.02)<br>A(Disordered interface) (P=0.04)<br>A( DNA binding) (P=0.03)<br>G(Methylation at K47) (P=0.03) |
| **rs1435266382** | R71T | 0.944 | L(Allosteric site at R71) (P=0.01)<br>A(DNA binding )(P=0.05) |

| | | | |
|---|---|---|---|
| **rs1466161664** | D146G | 0.956 | A(Metal binding) (P=0.00032)<br>A(Ordered interface) (P=0.02)<br>L(Relative solvent accessibility) (P=0.01)<br>L(Catalytic site at D146) (P=0.009)<br>A(Transmembrane protein) (P=0.03) |

Table 3: MutPred2 results showing effect of substitutions on functional and structural aspects of protein. Here A() means altered, G() means gain of and L() means Loss of.

As the general score, g>0.75 for every substitution and property score, either P<0.01 or P<0.05, so it implies that each and every substitution and its effect is either a confident hypothesis or a very confident hypothesis.

# 3.6 Prediction of Post-translational Modification Sites

PTM-SNPs are highly disease-related in comparison to other nsSNP sites, so to predict posttranslational modification sites, in DHFR protein, ModPred server was used.

| SNP IDs | Substitution | PTM sites (wild type) | Score with Confidence | PTM sites (mutants) | Score with Confidence |
|---|---|---|---|---|---|
| **rs371161421** | S60F | - | - | - | - |
| **rs747824863** | T137A | - | - | - | - |
| **rs755733770** | M53R | - | - | ADP-ribosylation | 0.83<br>Medium |
| **rs755733770** | M53T | - | - | - | - |
| **rs760708940** | N73H | Proteolytic cleavage | 0.51<br>Low | Proteolytic cleavage | 0.73<br>Medium |
| **rs954427933** | P150A | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| **rs992651401** | G118D | Proteolytic cleavage | 0.58<br>Low | Proteolytic cleavage | 0.83<br>Medium |
| **rs1251977833** | N49K | Proteolytic cleavage | 0.62<br>Low | Methylation | 0.65<br>Low |
| **rs1435266382** | R71T | ADP-ribosylation | 0.72<br>Medium | O-linked glycosylation | 0.65<br>Low |
| **rs1466161664** | D146G | Proteolytic cleavage | 0.62<br>Low | Amidation | 0.60<br>Medium |

Table 4: ModPred results showing damaging SNPs on various Post-Translational Modification sites.

# 3.7 Determining HEME Binding Residues

The output by HEMEsPred provided us with information about residues with their respective positions and predicted score. 20 such residues were predicted by HEMEsPred as shown below.

| Amino Acid | Position | Prediction Score |
|---|---|---|
| **Q** | 13 | 0.586 |
| **M** | 15 | 0.714 |
| **K** | 47 | 0.542 |
| **Q** | 48 | 0.787 |
| **P** | 67 | 0.58 |
| **E** | 79 | 0.574 |
| **E** | 82 | 0.683 |
| **P** | 83 | 0.87 |
| **H** | 88 | 0.566 |
| **E** | 102 | 0.57 |
| **Q** | 103 | 0.732 |

| | | |
|---|---|---|
| **V** | 121 | 0.576 |
| **E** | 124 | 0.765 |
| **H** | 128 | 0.936 |
| **H** | 131 | 0.596 |
| **Q** | 141 | 0.554 |
| **Q** | 171 | 0.647 |
| **E** | 173 | 0.539 |
| **K** | 174 | 0.581 |
| **E** | 181 | 0.557 |

Table 5: Predicted amino acids binding to Heme Protein.

Out of these, 7 mutants were found on the above listed residues which binds with HEME protein. They are as follows.

| **Mutants** | **SNP analysing tools Net Result** | **ΔΔG** |
|---|---|---|
| **E79K** | Neutral | - |
| **H88R** | Neutral | - |
| **V121A** | Deleterious | 1.4355 |
| **H128Q** | Neutral | - |
| **H131R** | Neutral | - |
| **Q171R** | Deleterious | 0.0649 |
| **E173K** | Neutral | - |

Table 6: Mutants Binding to Heme Protein

There were only two mutations which were found damaging by SNP analysing tool. But these two mutations are not damaging or deleterious because of their low ΔΔG values.

# 3.8 Comparative Modelling of Wild Type DHFR and its Mutants and RMSD Calculation of the Protein Models

From PDB, the human DHFR crystallographic structure was obtained (PDB Id-6DAV). 3D structures of mutants were generated by using Phyre2. For each mutant model TM-Scores and RMSD value were calculated. The greater the RMSD value, greater will be the difference between mutant structure and that of wild type.

| SNP IDs | Substitution | ΔΔG | TM-Score | RMSD |
|---|---|---|---|---|
| **rs371161421** | S60F | 2.09949 | 0.97532 | 0.76 |
| **rs747824863** | T137A | 2.66788 | 0.99547 | 0.54 |
| **rs755733770** | M53R<br>M53T | 3.76699<br>4.75206 | 0.99492<br>0.99854 | 0.55<br>0.20 |
| **rs760708940** | N73H | 15.9757 | 0.99890 | 0.17 |
| **rs954427933** | P150A | 2.3531 | 0.99886 | 0.17 |
| **rs992651401** | G118D | 1.56788 | 0.99536 | 0.54 |
| **rs1251977833** | N49K | 4.81038 | 0.99521 | 0.56 |
| **rs1435266382** | R71T | 2.58139 | 0.99780 | 0.24 |
| **rs1466161664** | D146G | 1.83404 | 0.99531 | 0.53 |

Table 7: Damaging SNPs with RMSD value and TM-Score.

TM-align showed the highest RMSD, 0.76, for S60F, which is still less, and lowest RMSD, 0.17 for N73H and P150A. Higher the RMSD, greater will be the changes in the protein structure which can result in disturbing its natural function.

# 3.9 Prediction of Structural Effects upon Mutation

Mutations in the domain of DHFR protein were found by the Prosit-ExPasy tool, for evaluating the effects of these mutations. So, Prosit-ExPasy and uniport database was used to find that the

mutation S60F was in Casein kinase II phosphorylation site and M53R, M53T were in an Amidation site. Also, R71T was in a substrate binding site and G118D was in an NADP binding site.

To evaluate the effects of amino acid substitutions on structure of the protein, Project Hope server was used. It takes protein's amino acid sequence as an input with the mutations and provides with all the changes it brings to the structure of the protein.

Mutation S60F with SNP ID- rs371161421, has an amino acid substitution from Serine to Phenylalanine at position 60, which is in Casein kinase II phosphorylation domain. The wild type residue was interacting with a ligand named as LII; NDP. It also forms a hydrogen bond with Threonine at 70 position. Because of the difference in wild type and mutant residue, such as: mutant is bigger and more hydrophobic than the wild type residue, there can easily be loss of interactions with the ligand and also due to the size difference the hydrogen bond with Threonine at 70 position is also been affected.

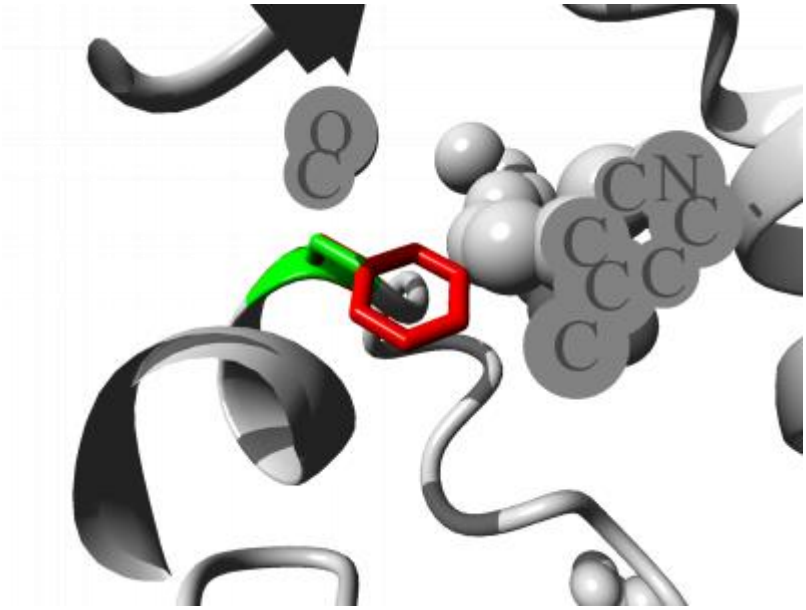



Figure 4a

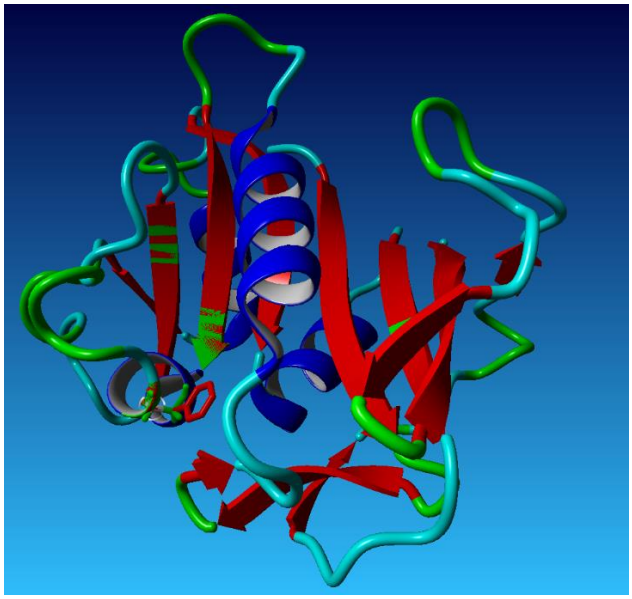


Figure 4b

Figure 4a and 4b: Mutation S60F, showing side chains of wild type residue (coloured green) and mutant (coloured red).

Similarly in the mutation T137A with SNP ID-rs747824863, the wild type residue has interactions with ligand LII and also it forms H-bond with Glutamic acid at position 31 and because of the changes which mutant residue brings, such as mutant is smaller and more hydrophobic, there will be an effect on the interaction with the ligand and the hydrogen bond formation. This residue is termed as 100% conserved by Project HOPE.

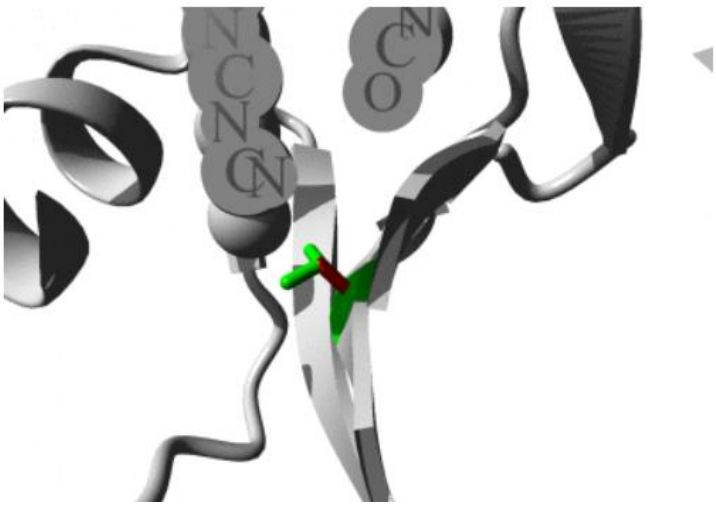


Figure 5a

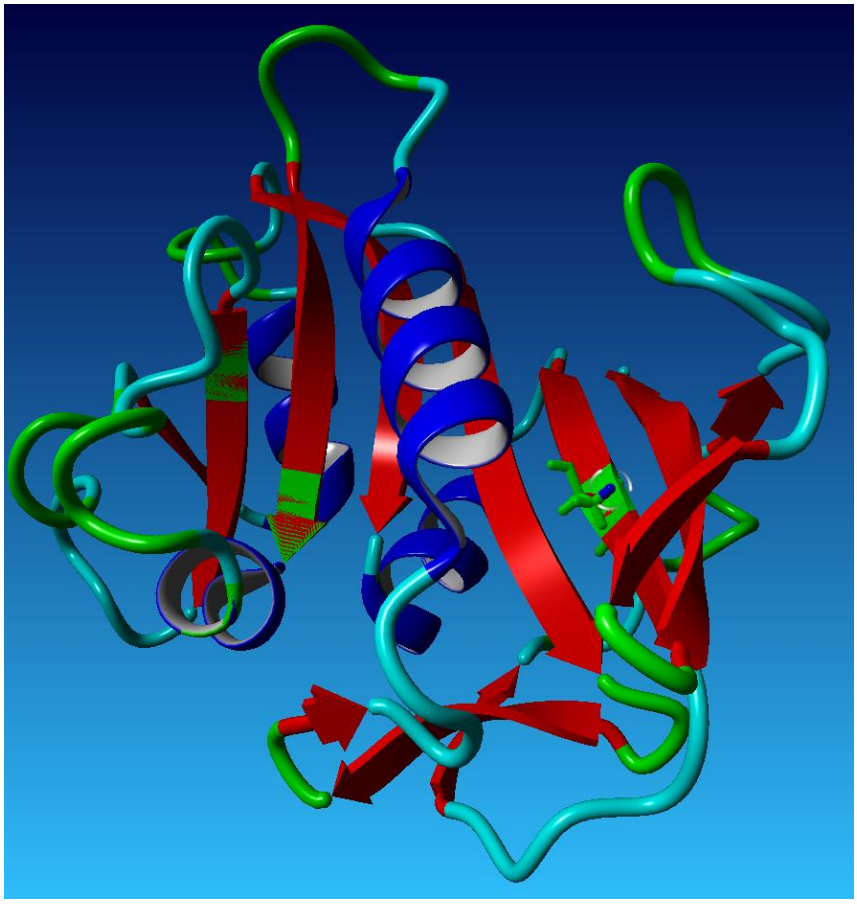


Figure 5b

Figure 5a and 5b: Mutation T137A showing side chains of wild type residue (coloured green) and mutant (coloured red).

Since M53R and M53T are in amidation site, so a change can disturb the function of the protein. The wild type residue has no contact with ligand, although due to the difference in size between the residues, protein's function may get affected. In M53R, the mutant residue is bigger and less hydrophobic than the wild type residue and it also brings a positive charge with it, since the wild type residue is neutral. Whereas, in case of M53T, the mutant is smaller and less hydrophobic than the wild type residue. And since it is smaller, the mutant residue will bring a hollow space which might interfere with the function of the protein.

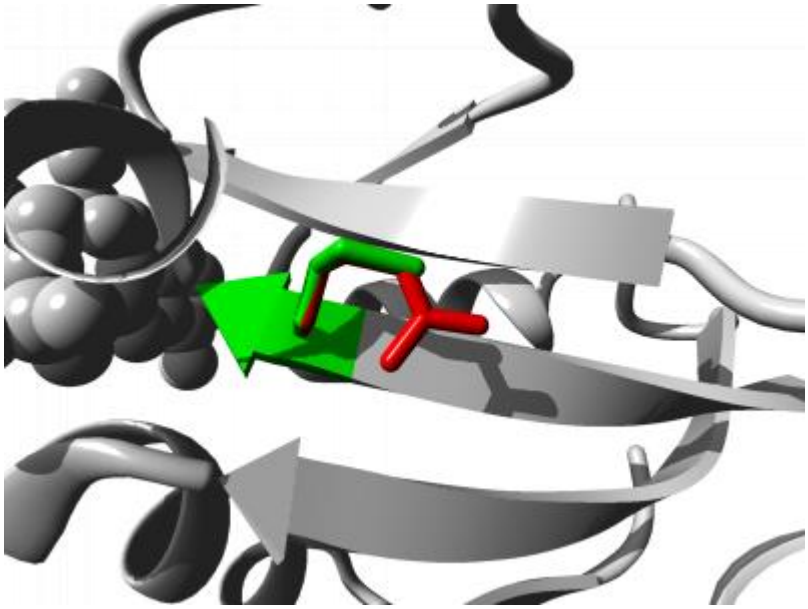


Figure 6a

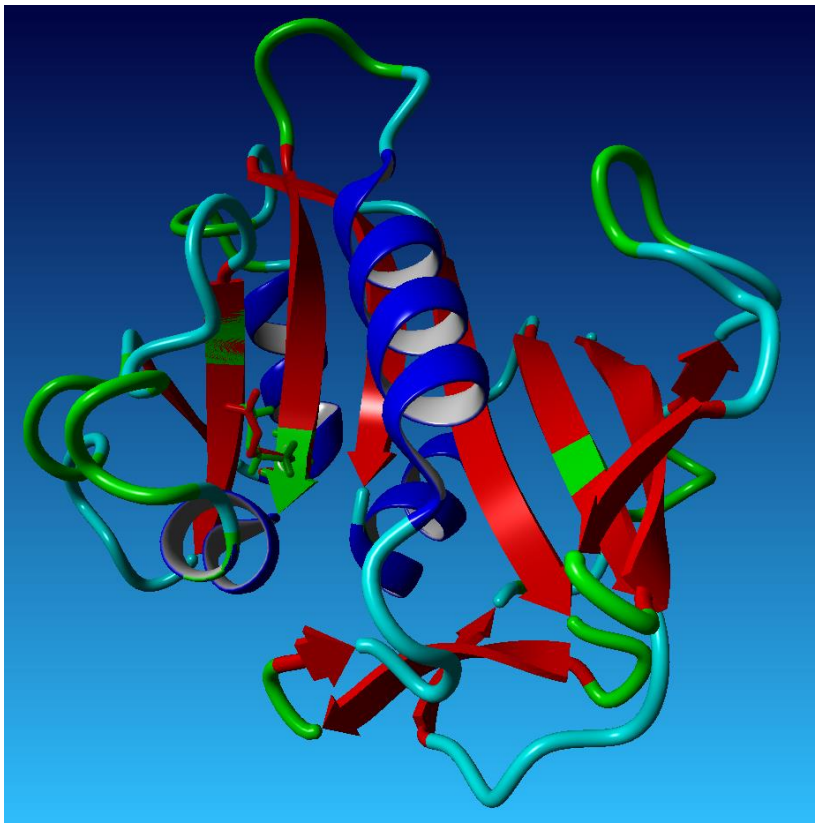


Figure 6b

Figure 6a and 6b: Mutant M53R, showing side chains of wild type residue (coloured green) and mutant (coloured red).

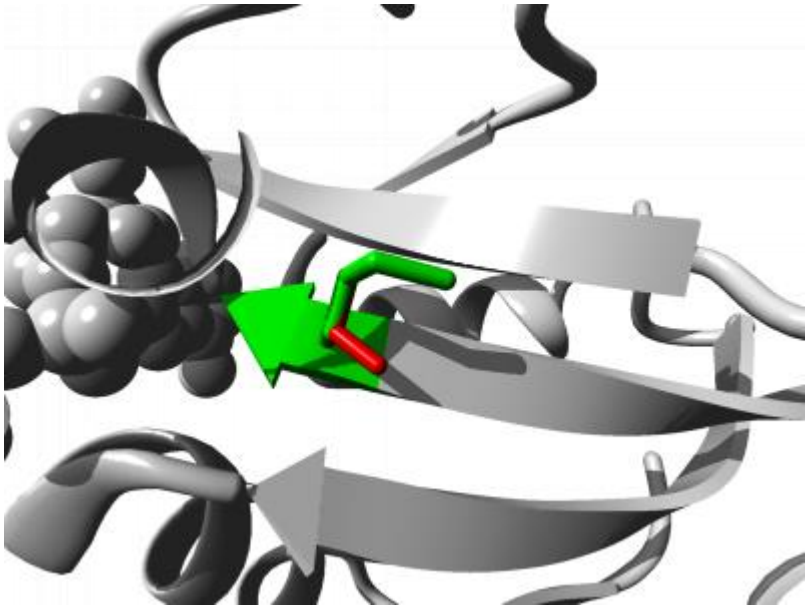


Figure 7a

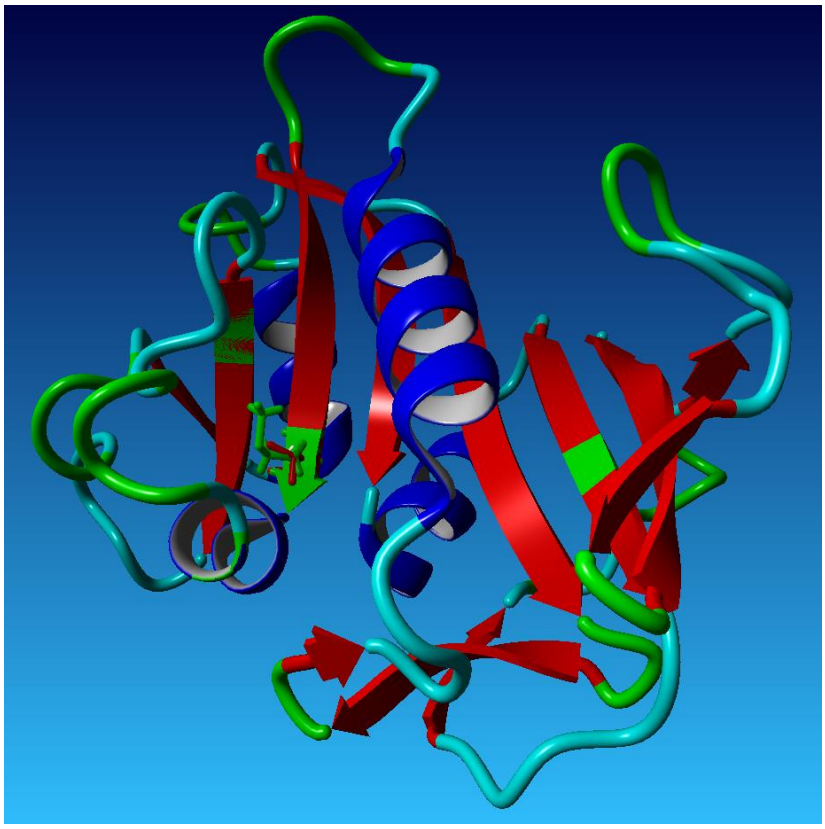


Figure 7b

Figure 7a and 7b: Mutant M53T, showing side chains of wild type residue (coloured green) and mutant (coloured red).

In case of mutation N73H, the wild type residue has no contact with ligand, but it forms hydrogen bond with Leucine at 68, Arginine at position 71 and as the mutant residue is bigger, the new residue will not be in the correct position to form the hydrogen bonds and as the wild

type residue is a conserved residue, due to the bigger size of the mutant residue, it will probably not fit.

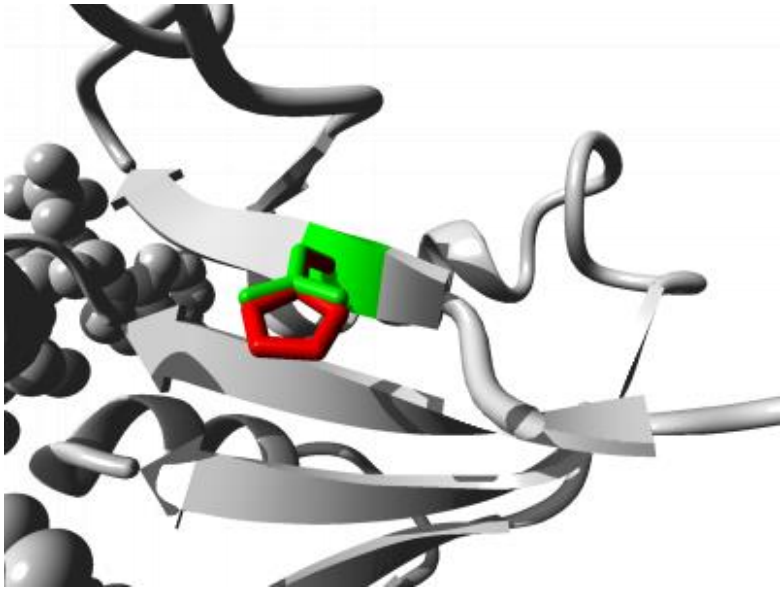


Figure 8a

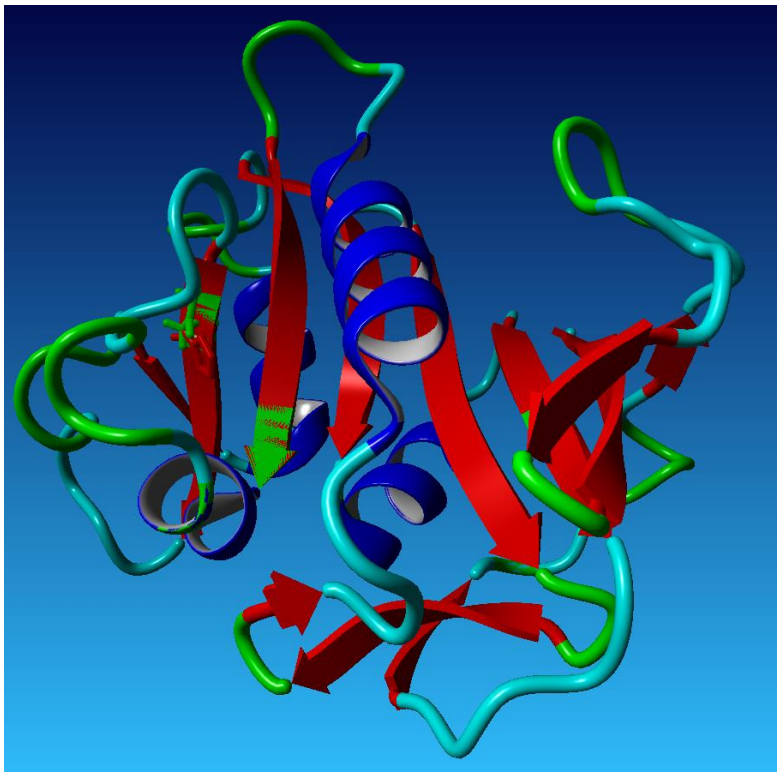


Figure 8b

Figure 8a and 8b: Mutant N73H, showing side chains of wild type residue (coloured green) and mutant (coloured red).

In case of mutation P150A, The mutant residue is smaller than the wild type residue and will eventually cause an empty space in core of the protein.

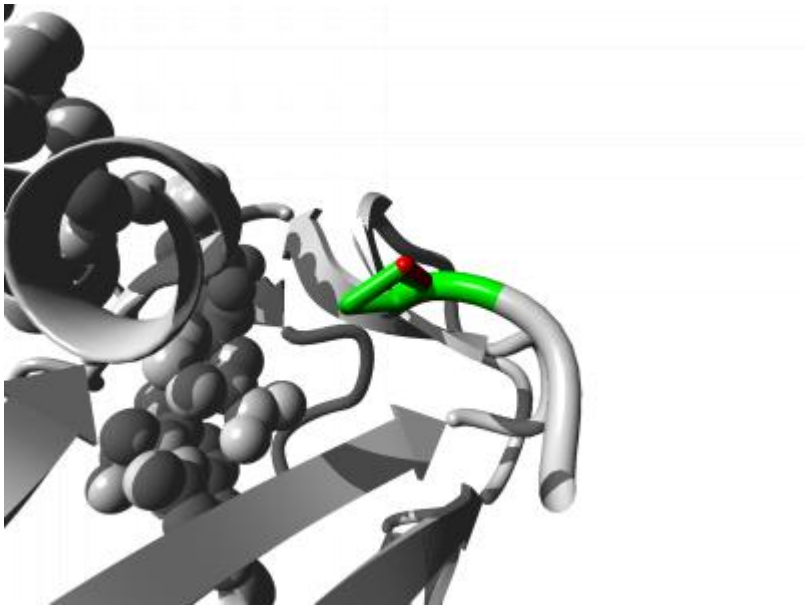


Figure 9a

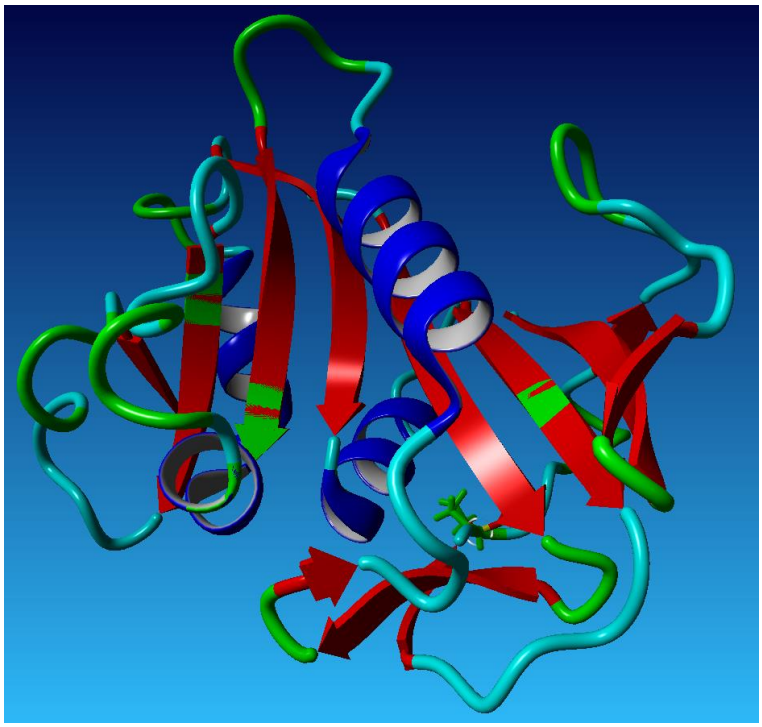


Figure 9b

Figure 9a and 9b: Mutant P150A, showing side chains of wild type residue (coloured green) and mutant (coloured red).

In mutation G118D, as the wild type residue is in NADP binding site, so it has interactions with ligand NDP. Mutant residue is bigger, has a negative charge (whereas wild residue is

neutral) and is less hydrophobic, so due to such changes, mutant residue can cause loss of interaction with the ligand. Also, torsion angles for this residue are unusual and as Glycine is the most flexible residue, mutation will disturb the local structure.

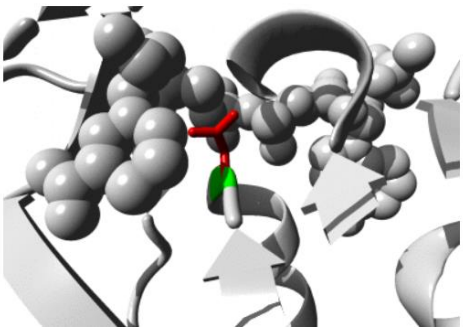


Figure 10a

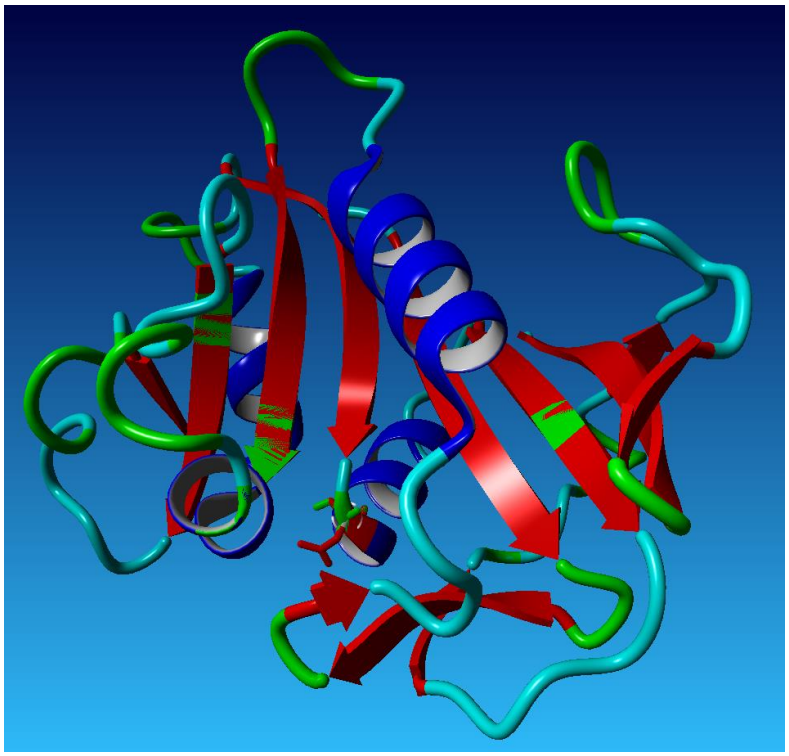


Figure 10b

Figure 10a and 10b: Mutant G118D, showing side chains of wild type residue (coloured green) and mutant (coloured red).

In mutation N49K, wild type residue form hydrogen bonds with Threonine at position 39 and 41 and with Methionine at position 112. Mutant is bigger, has positive charge (whereas wild type is neutral) and because of these changes mutant residue will disturb the hydrogen bonds. And because of the positive charge on the mutant residue, it can induce protein folding problems.

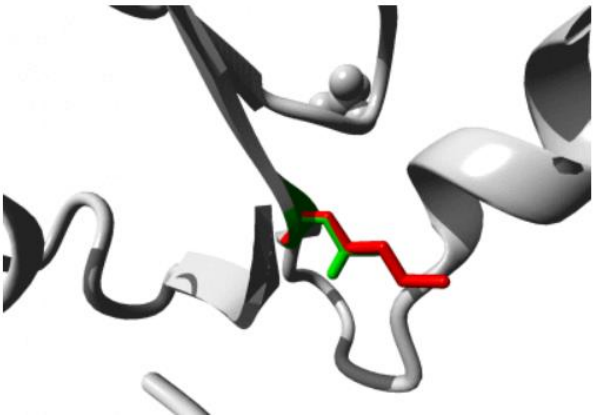


Figure 11a

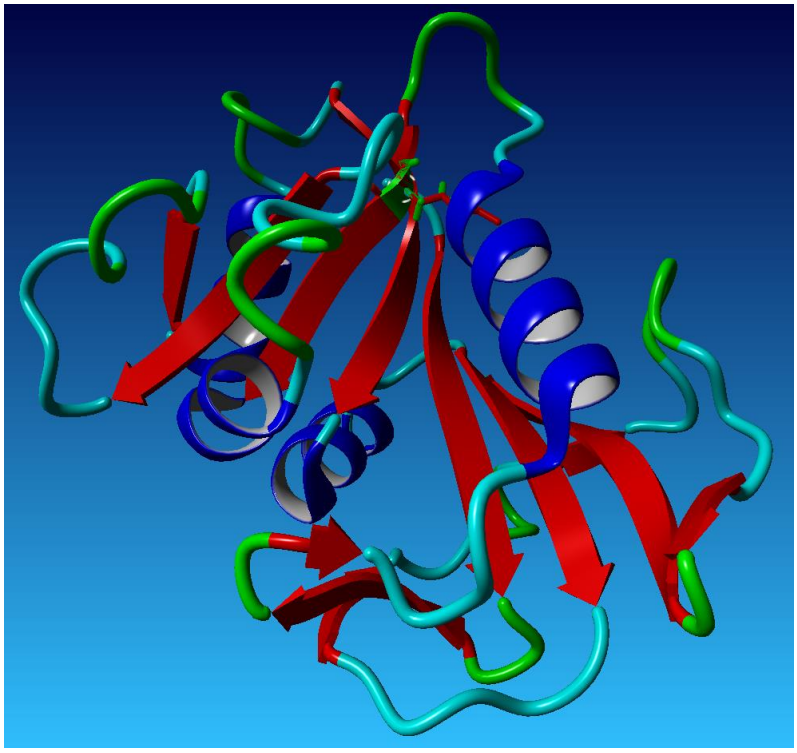


Figure 11b

Figure 11a and 11b: Mutant N49K, showing side chains of wild type residue (coloured green) and mutant (coloured red).

In mutation R71T, wild type residue has positive charge and form hydrogen bonds with Threonine at position 39, Leucine at position 69. Whereas, mutant residue is smaller, neutral and more hydrophobic, so because of the size difference, mutant residue is not in the correct position to form hydrogen bonds. Project Hope also reported that no other residue was observed at this position with similar properties.

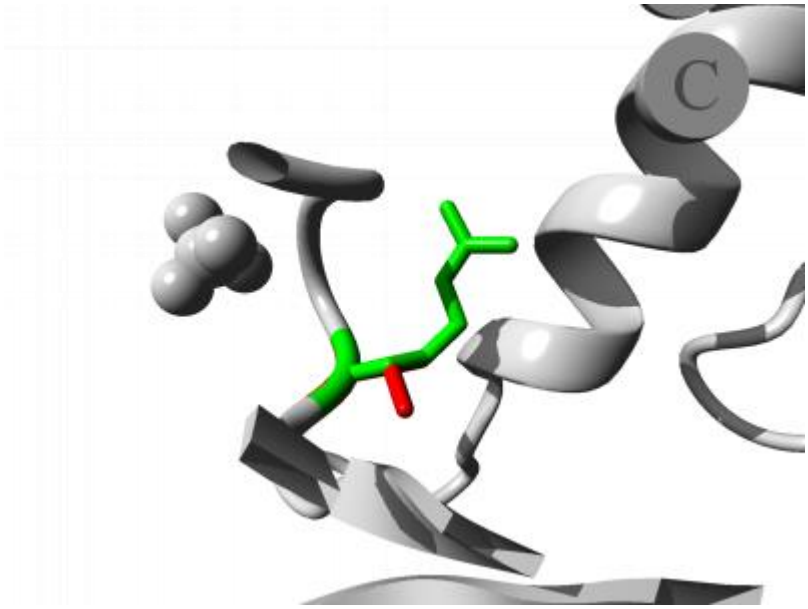



Figure 12a

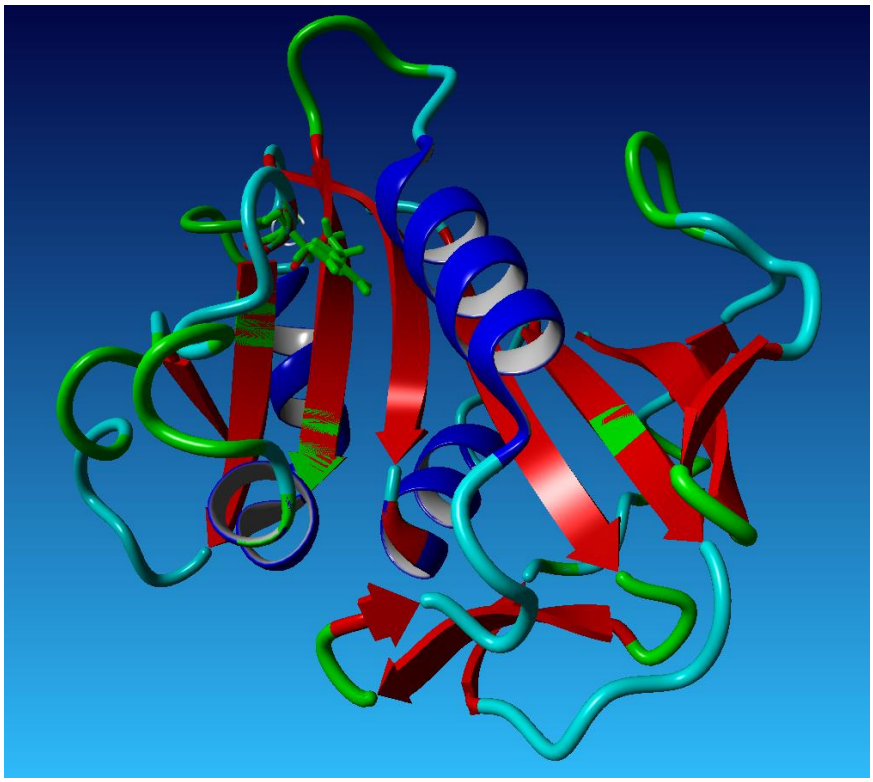


Figure 12b

Figure 12a and 12b: Mutant R71T, showing side chains of wild type residue (coloured green) and mutant (coloured red).

In case of mutation D146G, wild type residue has negative charge, forms hydrogen bond with Asparagine at position 20, Threonine at position 147, forms a salt-bridge with Lysine at position 19. Whereas, mutant residue is smaller, neutral, more hydrophobic and because of the

size difference, it will disturb the hydrogen bonds and because it is neutral, it will disturb the ionic interaction. Also, as mutant residue is Glycine and it is very flexible, so it can disturb the required rigidity of the protein.

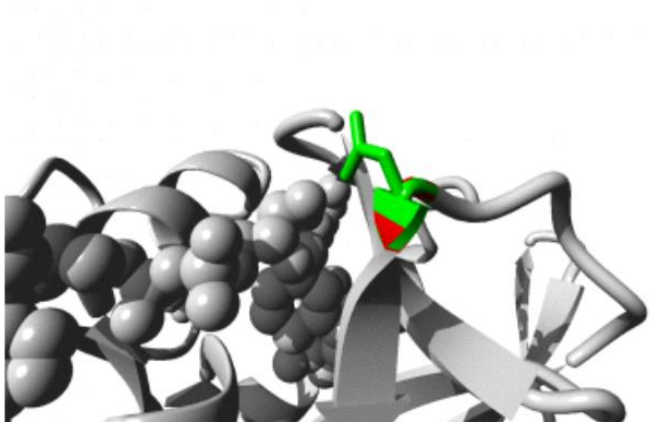


Figure 13a

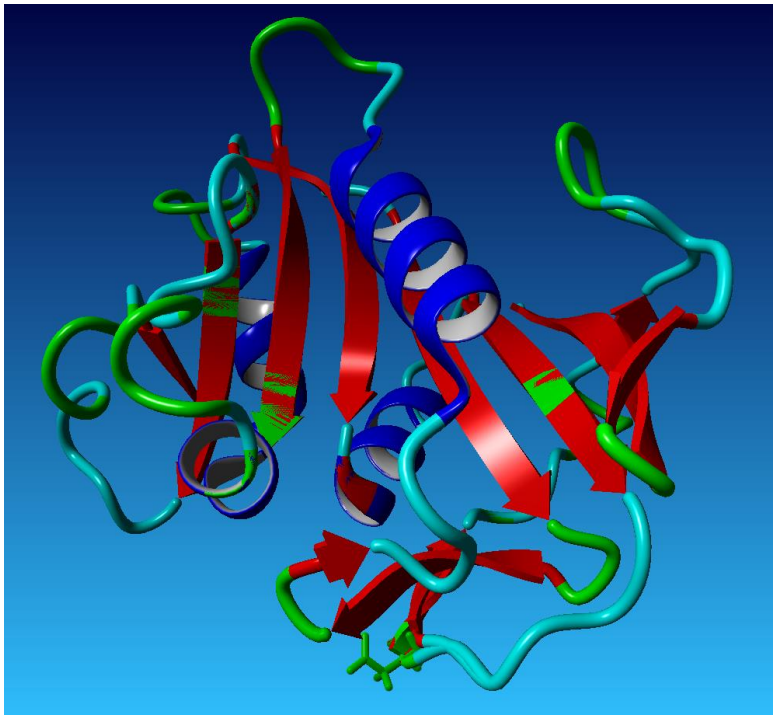


Figure 13b

Figure 13a and 13b: Mutant D146G, showing side chains of wild type residue (coloured green) and mutant (coloured red).

# 3.10 Clinically Damaging nsSNPs

There were two nsSNPs which were found clinically damaging as shown below.

| SNP IDs | Substitution | Allele Change | ΔΔG |
|---|---|---|---|
| **rs121913223** | D153V | T>A | -0.215366 |
| **rs387906619** | L80F | G>A | 0.346991 |

Table 8: Clinically damaging nsSNPs.

The mutation L80F was found in Pakistani siblings, who have deficiency of dihydrofolate reductase, causing megaloblastic anaemia[24]. The other mutation D153V was found in three European siblings[25].

The reason why they weren't picked up by my study is that they were excluded because of their very low ΔΔG value, although they were predicted damaging by four out of the five in silico nsSNPs algorithm.

# Chapter 4

# Conclusion

This study indicates that numerous nsSNPs can disturb the structure and/or the role of DHFR protein. After using various SNP analysing tools and predicting their harmful effects on the structure or the function of the DHFR protein, 10 SNPs were found most damaging and those are: rs371161421 (S60F), rs747824863 (T137A), rs755733770 (M53R,M53T), rs760708940 (N73H), rs954427933 (P150A), rs992651401 (G118D), rs1251977833 (N49K), rs1435266382 (R71T), rs1466161664 (D146G). Out of these, mutation S60F was in Casein kinase II phosphorylation site and M53R, M53T were in an Amidation site. Also, R71T was present in substrate binding site and G118D was in an NADP binding site. Two clinically nsSNPs: rs121913223 (D153V), rs387906619 (L80F) were reported in ClinVar, but they were not picked up by my study, because they got screened out in estimation of ΔΔG values. So, these nsSNPs can be strongly considered as key candidates in causing DHFR-related malfunction diseases.

# Bibliography


[1] A. Chakravarti, ...“to a future of genetic medicine,” *Nature*. 2001, doi: 10.1038/35057281.

[2] P. Carninci *et al.*, “Molecular biology: The transcriptional landscape of the mammalian genome,” *Science (80-. ).*, 2005, doi: 10.1126/science.1112014.

[3] T. P. Dryja *et al.*, “Mutations within the rhodopsin gene in patients with autosomal dominant retinitis pigmentosa,” *N. Engl. J. Med.*, 1990, doi: 10.1056/NEJM199011083231903.

[4] E. P. Smith *et al.*, “Estrogen resistance caused by a mutation in the estrogen-receptor gene in a man,” *N. Engl. J. Med.*, 1994, doi: 10.1056/NEJM199410203311604.

[5] V. Ramensky, “Human non-synonymous SNPs: server and survey,” *Nucleic Acids Res.*, 2002, doi: 10.1093/nar/gkf493.

[6] S. W. Doniger *et al.*, “A catalog of neutral and deleterious polymorphism in yeast,” *PLoS Genet.*, 2008, doi: 10.1371/journal.pgen.1000183.

[7] B. S. Askari and M. Krajinovic, “Dihydrofolate Reductase Gene Variations in Susceptibility to Disease and Treatment Outcomes,” *Curr. Genomics*, 2010, doi: 10.2174/138920210793360925.

[8] S. T. Sherry, “dbSNP: the NCBI database of genetic variation,” *Nucleic Acids Res.*, 2001, doi: 10.1093/nar/29.1.308.

[9] R. Apweiler, “The universal protein resource (UniProt) in 2010,” *Nucleic Acids Res.*, 2009, doi: 10.1093/nar/gkp846.

[10] “The Protein Data Bank.,” *Methods Biochem. Anal.*, 2003, doi: 10.4135/9781412994231.n75.

[11] P. C. Ng and S. Henikoff, “SIFT: Predicting amino acid changes that affect protein function,” *Nucleic Acids Res.*, 2003, doi: 10.1093/nar/gkg509.

[12] R. Calabrese, E. Capriotti, P. Fariselli, P. L. Martelli, and R. Casadio, “Functional

annotations improve the predictive score of human disease-related mutations in proteins," *Hum. Mutat.*, 2009, doi: 10.1002/humu.21047.

[13] E. Capriotti, R. Calabrese, and R. Casadio, "Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information," *Bioinformatics*, 2006, doi: 10.1093/bioinformatics/btl423.

[14] P. D. Thomas and A. Kejariwal, "Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects," *Proc. Natl. Acad. Sci. U. S. A.*, 2004, doi: 10.1073/pnas.0404380101.

[15] E. Capriotti and R. B. Altman, "Improving the prediction of disease-related variants using protein three-dimensional structure," *BMC Bioinformatics*, 2011, doi: 10.1186/1471-2105-12-S4-S3.

[16] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal, "ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids," *Nucleic Acids Res.*, 2010, doi: 10.1093/nar/gkq399.

[17] I. Mayrose, D. Graur, N. Ben-Tal, and T. Pupko, "Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior," *Mol. Biol. Evol.*, 2004, doi: 10.1093/molbev/msh194.

[18] O. Altindag, O. Erel, N. Aksoy, S. Selek, H. Celik, and M. Karaoglanoglu, "Increased oxidative stress and its relation with collagen metabolism in knee osteoarthritis," *Rheumatol. Int.*, 2007, doi: 10.1007/s00296-006-0247-8.

[19] V. Pejaver, W. L. Hsu, F. Xin, A. K. Dunker, V. N. Uversky, and P. Radivojac, "The structural and functional signatures of proteins that undergo multiple events of post-translational modification," *Protein Sci.*, 2014, doi: 10.1002/pro.2494.

[20] L. A. Kelley and M. J. E. Sternberg, "Protein structure prediction on the web: A case study using the phyre server," *Nat. Protoc.*, 2009, doi: 10.1038/nprot.2009.2.

[21] Y. Zhang and J. Skolnick, "TM-align: A protein structure alignment algorithm based on the TM-score," *Nucleic Acids Res.*, 2005, doi: 10.1093/nar/gki524.

[22] H. Venselaar, T. A. H. te Beek, R. K. P. Kuipers, M. L. Hekkelman, and G. Vriend, "Protein structure analysis of mutations causing inheritable diseases. An e-Science

approach with life scientist friendly interfaces," *BMC Bioinformatics*, 2010, doi: 10.1186/1471-2105-11-548.

[23] M. Bhagwat, "Searching NCBI's dbSNP database," *Curr. Protoc. Bioinforma.*, 2010, doi: 10.1002/0471250953.bi0119s32.

[24] S. Banka *et al.*, "Identification and characterization of an inborn error of metabolism caused by dihydrofolate reductase deficiency," *Am. J. Hum. Genet.*, 2011, doi: 10.1016/j.ajhg.2011.01.004.

[25] H. Cario *et al.*, "Dihydrofolate reductase deficiency due to a homozygous DHFR mutation causes megaloblastic anemia and cerebral folate deficiency leading to severe neurologic disease," *Am. J. Hum. Genet.*, 2011, doi: 10.1016/j.ajhg.2011.01.007.