A Comprehensive in Silico Analysis of the Functional and Structural Impact of Non-Synonymous Single Nucleotide Polymorphisms in the Human KRAS Gene

Nahla E. Abdelraheem[1], Ghada M. El-Tayeb[2], Lamia O. Osman[3], Samar A. Abedlrhman[3],
Aisha S. Ali[4], Ahmed H. Elsadig[5], Sofia B. Mohamed[6,*]

[1]Department of Histopathology, National University, Khartoum, Sudan
[2]Department of Pharmaceutical Chemistry, National University, Khartoum, Sudan
[3]Department of Clinical Chemistry, National University, Khartoum, Sudan
[4]Department of Microbiology, National University, Khartoum, Sudan
[5]Faculty of Medicine, University of Khartoum, Khartoum, Sudan
[6]Department of Vector and biomedical studies, Tropical Medicine Research Institute, Khartoum, Sudan

Abstract   The protein product of the normal KRAS gene performs an essential function in normal tissue signaling, and the mutation of a KRAS gene is an essential step in the development of many cancers. Despite the reported association of KRAS gene mutations with human disease susceptibility, the comprehensive computational analysis of coding, noncoding and regulatory SNPs, and their functional impacts on protein level, still remains unknown. In this study, we performed a comprehensive analysis of the functional and structural impact of non-synonymous (nsSNP) that are deleterious to KRAS structure and function. A sequence homology-based approach was adopted for screening nsSNPs in KRAS, including nine different in silico prediction algorithms; SIFT, PROVEAN, MutPred, SNP&GO, PhD-SNP, PANTHER, I-Mutant 2.0, MUpo and DUET. Moreover, to understand how mutations can affect the strength of the interactions that bind proteins together we submitted data to ELASPIC servers and evolutionary conservation data was obtained from the ConSurf web server. The PDB structure of KRAS protein was obtained from RCSB Protein Data Bank (PDB ID: 4QL3) and Structural theoretical models of KRAS were created using mutation3D and chimera servers. Our results demonstrate that 14 nsSNPs in the antiviral KRAS gene may be deleterious to KRAS structure and function. 14 deleterious and high-risk nsSNPs were identified in KRAS gene. 11 of these are located at highly conserved amino acid sites in the protein domain (RAS). Additionally, we detected 4 nsSNPs in the core residues of the protein and 10 SNPS at interface. Finally, we classified six SNPs as top high-risk (Q61E, A146T, N116S, K147E, G60R and A59V) which may alter the putative structure of KRAS's domain, particularly in G1, G2, G3, G4 and G5 motif regions. These same regions are critical to the protein's structure and function. This study adopted an extensive in silico analysis of the highly polymorphic KRAS gene and will be a valuable resource for future targeted mechanistic and population-based studies.

Keywords   KRAS gene, Insilico analysis, MutPred, PhD-SNP, DETU, ELASPIC, Mutation3D and UCSF chimera program

1. Introduction

Single Nucleotide Polymorphism, also known as SNP, accounts for the most common form of genetic mutation in human. It has been reported that ~93% of all human genes represent at least one SNP [1]. Therefore, they are liable for generating the majority of biological variations among individuals. An understanding of the relationship between these genetic variations and their phenotypic effects could therefore be a step toward exploring the causes of various disorders or diseases. SNPs can fall within the coding regions (coding SNPs) or noncoding regions of genes (noncoding SNPs), or in the intergenic region between two genes [2, 3]. While the two others are quite natural in the human genome and phenotypically neutral [4], nonsynonymous coding SNPs (nsSNPs) are thought to have the principal impact on phenotype by changing the protein sequence. As they cause amino acid alteration in the corresponding protein product, it may exert deleterious effects on the structure, function, solubility, or stability of proteins [5]. Beside these the nsSNPs perturb gene regulation by modifying DNA and transcriptional binding factors [6] and the maintenance of the formational integrity

of cells and tissues [7]. Thus, it is likely that nsSNPs play a major role in the functional diversity coded proteins in human populations and often associated with human diseases. Indeed, earlier studies have revealed that more than 50% of the mutations associated with inherited genetic disorders are resulted by nsSNPs [8]. Disease-causing mutations are frequently observed at either core or interface residues mediating protein interactions. Mutations at core residues frequently destabilize protein structure while mutations at interface residues can specifically affect the binding energies of protein-protein interactions. Missense mutations at protein–protein interaction sites, called interfaces, are important contributors to human disease. Interfaces are non-uniform surface areas characterized by two main regions, "core" and "rim", which differ in terms of evolutionary conservation and physicochemical properties. Moreover, within interfaces, only a small subset of residues ("hot spots") is crucial for the binding free energy of the protein–protein complex [9]. Protein-protein interactions are critical for nearly every process in the cell and deleterious mutations hindering these interactions can have severe consequences for the associated cellular function. A variety of efforts from personalized medicine to understand viral evolution require knowing how specific mutations effect the protein-protein interactions. Conversely, designing proteins with improved binding or altered specificity requires that the impact of mutations on the native interface be understood. Currently this information is not available experimentally on the proteome-wide scale necessary for these tasks. Towards this end, considerable effort has been devoted towards developing methods to predict the impact of mutations on binding affinity. Most of these approaches rely on physics based methods that attempt to faithfully model on the atomic level the interactions determining protein-protein binding affinity. However, a major obstacle of such approaches is the need for the reconstruction of the full-atomic model for every mutant complex, which limits the accuracy of the approach (since the position of the side-chains is difficult to model) and reduces the computational speed and the range of applications(since rebuilding the full-atomic model is generally them most time-consuming step) [10]. The KRAS (Kirsten rat sarcoma viral oncogene homolog) (ID: 3845) gene belongs to a family of oncogenes called the Ras family, located on the short (p) arm of chromosome 12 at position 12.1 and consists of six exons. It provides instructions for making a protein called K-Ras protein (~21 kDa) which is involved primarily in regulating cell division, differentiation, and apoptosis. As a part of a signaling pathway known as the RAS/MAPK pathway, the protein relays signals from outside the cell to the cell's nucleus, these signals instruct the cell to grow and divide or to mature and take on specialized functions (differentiate). K-Ras protein is a GTPase; acting like a switch to transmit signals, it must be turned on by attaching to a molecule of GTP, and turned off (inactivated) when it converts the GTP to GDP [11]. A Point mutation in KRAS codons 12 and 13 are predictive of poor prognosis of lung cancer while mutation at codons 61 and 146 were associated with shorter progression-free survival in colorectal cancer patients receiving anti-epidermal growth factor receptor treatment [12]. Bioinformatics has become a very important part of different areas of biology. Bioinformatics tools and techniques to aid in the processing and extraction of useful results from large amounts of raw data; textual mining of biological literature, analysis of gene and protein expression, simulation and modeling of DNA, RNA, and protein structures, comparison of genetic and genomic data and helps analyze and catalogue the biological pathways and networks that are an important part of systems biology [13]. Although there are presently several published articles describing the association of SNPs in the KRAS gene with different types of diseases, computational analysis has not yet been undertaken on the functional and structural consequences of nsSNPs in this gene. In the current study, we employed different publicly available bioinformatics tools and databases for a comprehensive analysis of nsSNPs in KRAS gene. As the majority of disease mutations affect protein stability, we also proposed modeled protein structures of the mutant proteins and compared them with the native protein in order to evaluate stability changes.

2. Material and Methods

2.1. Dataset Collection

The amino acid sequence [Uniprot: P01116] of KRAS protein was obtained from UniProt Protein Database (http://www.uniprot.org). The information about SNPs of KRAS gene of Homo sapiens was obtained from the db-SNP (http://www.ncbi.nlm.nih.gov/SNP/). Structure of KRAS protein [PDB IDs: 4QL3] was retrieved from RCSB Protein Data Bank.

2.2. Non-Synonymous SNPs Analysis

Functional effects of nsSNPs were predicted using the following in silico algorithms: SIFT (http://sift.jcvi.org/), PROVEAN (http://provean.jcvi.org), PhD-SNP (http://snps.biofold.org/phdsnp/phd-snp.html), SNPs&GO (http://snps-and-go.biocomp.unibo.it/snps-and-go/), and MutPred (http://mutpred.mutdb.org/) [14]. nsSNPs predicted to be deleterious by at least four in silico algorithms were categorized as high-risk nsSNPs and were selected for further analysis.

2.2.1. Assessment of the Functional Impacts of Deleterious nsSNPs by SIFT and PROVEAN

SIFT (Sorting intolerant from tolerant)
Predicts whether an amino acid substitution affects protein function based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences [15]. The SIFT scores range from 0 to 1, and scores ≤0.05 are predicted by the algorithm to be damaging amino acid substitutions, whereas scores >0.05 are considered to be tolerated.

PROVEAN (Protein **V**ariation **E**ffect **An**alyzer):

Is a software tool that predicts whether an amino acid substitution has an impact on the biological function of a protein grounded on the alignment-based score [16]. The score measures the change in sequence similarity of a query sequence to a protein sequence homolog between without and with an amino acid variation of the query sequence. If the PROVEAN score ≤-2.5, the protein variant is predicted to have a "deleterious" effect, while if the PROVEAN score is >-2.5, the variant is predicted to have a "neutral" effect.

2.2.2. Prediction of Disease Related Mutations SNPs by SNPs&GO

SNPs&GO [17] is also a support vector machine (SVM) based on the method to accurately predict the mutation related to disease from protein sequence. The input is the FASTA sequence of the whole protein, the output is based on the difference among the neutral and disease related variations of the protein sequence. The RI (reliability index) with value of greater than 5 depicts the disease related effect caused by mutation on the function of parent protein. The PHD SNP and PANTHER, algorithms were also used in the display of output.

2.2.3. Prediction of Harmful Mutations by MutPred

The MutPred server (http://mutpred.mutdb.org/) [18] was employed to classify an amino acid substitution (AAS) as disease-associated or neutral. In addition, it predicts molecular cause of disease/deleterious AAS. MutPred is based upon SIFT and a gain/loss of 14 different structural and functional properties. The output of MutPred contains a general score (g), i.e., the probability that the amino acid substitution is deleterious/disease-associated, and top 5 property scores (p), where p is the P-value that certain structural and functional properties are impacted.

2.3. Protein Stability Analysis

The stability of KRAS protein was predicted using the following in silico algorithms: I-mutant 2.0 (http://folding.biofold.org/i-mutant/i-mutant2.0.html), Upro (http://www.ics.uci.edu/~baldig/mutation. html) and DUET server http://structure.bioc.cam.ac.uk/duet

2.3.1. Analysis of the Effects of nsSNPs on the Protein Stability by I-Mutant 2.0 and MUpro

I-Mutant 2.0 is a SVM based tools i.e., support vector machine based tool that leads to automatic protein stability change prediction which is caused by single point mutation [19]. The initiations were done either by using protein structure or more precisely from the protein sequence. The output is a free energy change value (ΔΔG). Positive ΔΔG value infers that the protein being mutated is of higher stability and vice versa is also true.

MUpro: is also a support vector machine-based tool for the prediction of protein stability changes upon nonsynonymous SNPs. The value of the energy change is predicted, and a confidence score between -1 and 1 for

measuring the confidence of the prediction is calculated. A score <0 means the variant decreases the protein stability; conversely, a score >0 means the variant increases the protein stability.

2.3.2. Predicted Change in Folding and Unfolding Free Energy up on Mutation Using DUET Server

DUET: A web server for an integrated computational approach to study missense mutations in proteins. DUET consolidates two complementary approaches (mCSM and SDM) in a consensus prediction; obtained by combining the results of these parate methods in an optimized predictor using Support Vector Machines (SVM).

SDM: The method SDM, introduced in relies on amino acid propensities derived from environment-specific substitution tables for homologous protein families that feed a statistical potential energy function and encompass an evolutionary view of the constraints from the immediate residue environment. The approach compares amino acid propensities for the wild-type and mutant proteins in the folded and unfolded states in order to estimate the free energy differences between wild type and mutant. The website is at: http://www-cryst.bioc.cam.ac.uk/sdm/sdm.php.

mCSM: mCSM is a machine learning method to predict the effects of missense mutations based on structural signatures .The mCSM signatures were derived from the graph-based concept of Cutoff Scanning Matrix (CSM), originally proposed to represent network topology by distance patterns in the study of biological systems. mCSM uses a graph representation of the wild-type residue environment to extract geometric and physicochemical patterns (the last represent edinterms of pharmacophores) that are then used to represent the 3D chemical environment during supervised learning. These signatures have been successfully applied in a range of tasks including protein structural classification and function prediction, as well as large-scale receptor based protein ligand prediction [20]. The mCSM website is available at: http://structure.bioc.cam.ac.uk/mcsm.

2.4. Prediction of the Stability Effects upon Mutation in Both, Domain Cores and Domain-Domain Interfaces

ELASPIC is a novel ensemble machine learning approach that predicts the effects of mutations on protein folding and protein-protein interactions. Here we present the ELASPIC web server, which makes the ELASPIC pipeline available through a fast and intuitive interface. The web server can be used to evaluate the effect of mutations on any protein in the Uniprot database, and allows all predicted results, including modeled wild-type and mutated structures, to be managed and viewed online and downloaded if needed. It is backed by a database which contains improved structural domain definitions, and a list of curated domain-domain interactions for all known proteins, as well as homology models of domains and domain-domain interactions for the human

proteome. Homology models for proteins of other organisms are calculated on the fly, and mutations are evaluated within minutes once the homology model is available [10].

2.5. Identification of Functional SNPs in Conserved Regions by Using ConSurf Web Server

Evolutionary conservation of amino acid residues in KRAS was determined using the ConSurf web server (consurf.tau.ac.il/). ConSurf-DB provides evolutionary conservation profiles for proteins of known structure in the PDB. Amino acid sequences similar to each sequence in the PDB were collected and multiply aligned using CSI-BLAST and MAFFT, respectively. The evolutionary conservation of each amino acid position in the alignment was calculated using the Rate 4Site algorithm, implemented in the ConSurf web-server. The algorithm takes explicitly into account the phylogenetic relations between the aligned proteins and the stochastic nature of the evolutionary process. Rate 4 Site assigns a conservation level for each residue using an empirical Bayesian inference. Visual inspection of the conservation patterns on the 3-dimensional structure often enables the identification of key residues that comprise the functionally-important regions of the protein [21].

2.6. Structural Analysis

2.6.1. Detection of nsSNPs Location in Protein Structure Using Mutation3D

Mutation3D (http://mutation3d.org) is a functional prediction and visualization tool for studying the spatial arrangement of amino acid substitutions on protein models and structures. It is intended to be used to identify clusters of amino acid substitutions arising from somatic cancer mutations across many patients in order to identify functional hotspots and fuel downstream hypotheses. It is

also useful for clustering other kinds of mutational data, or simply as a tool to quickly assess relative locations of amino acids in proteins.

2.6.2. 3D Modeling

RMSD Protein Data Base was used to generate 3D structural models for wild type KRAS (UniProtKB 4QL3) and each of the 13 high-risk nsSNPs in KRAS's RAS domain.

2.6.3. Modeling Amino Acid Substitution; H-Bonding and Clash

UCSF Chimera is a highly extensible program for interactive visualization and analysis of molecular structures and related data. Chimera (version 1.8) software was used to scan the 3D (three-dimensional) structure of specific protein, and hence modifies the original amino acid with the mutated one to see the impact that can be produced. The outcome is then a graphic model depicting the mutation [22]. Chimera (version 1.8) currently available within the Chimera package and available from the Chimera web site http://www.cgl.ucsf.edu/chimera/.

3. Results

3.1. Retrieval of SNPs

The dbSNP was utilized for retrieving the SNPs in the human KRAS gene. A total of 137 SNPs were founded in the coding region, among them 45 were synonymous, 83 non-synonymous and missense, 6 non-synonymous and nonsense, and 3 frame-shift mutations (Fig. 1). Only missense non-synonymous coding SNPs were chosen for further analysis.
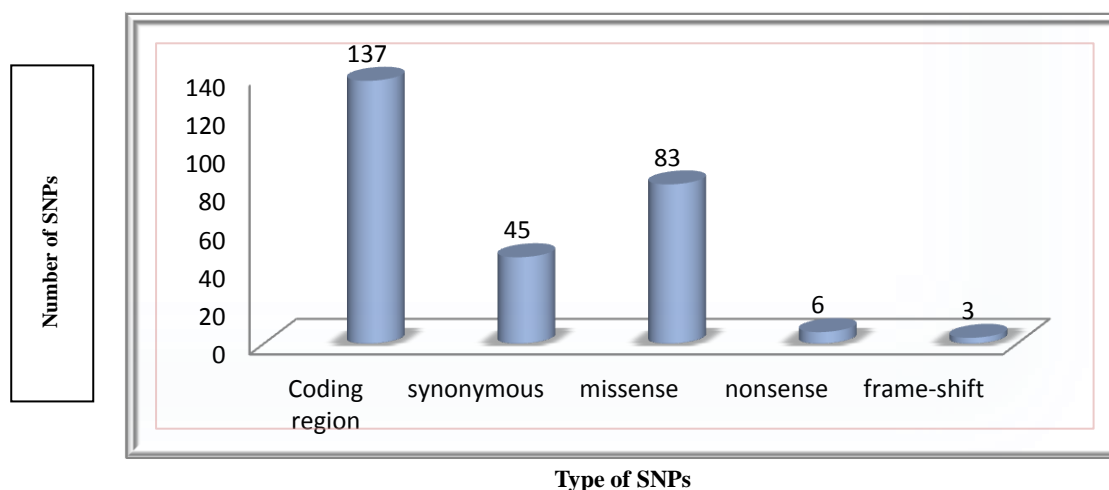

Figure 1. A graphical representation of the distribution of missense, synonymous, nonsense and frame shift SNPs for KRAS gene (based on the dbSNP database)

3.2. Prediction of Tolerated and Deleterious nsSNPs

nsSNPs were submitted to the SIFT program to predict their effect on protein function, out of the 83 SNPs screened, 19 variants were found to be deleterious and 8 tolerated. Among the SNPs analyzed, SIFT couldn't find 57 SNPs.

3.3. Deleterious nsSNPs by PROVEAN Server

Missense nsSNPs were also submitted to the PROVEAN server. 17 SNPs were considered to be Deleterious and Six of them were neutra. A significant correlation between the results obtained from the evolutionary-based approach SIFTS and the structural based approach PROVEAN was detected. Out of the total 15SNPs were predicted as Deleterious by PROVEAN and SIFT suggesting that these nsSNPs may disrupt both the protein function and structure. The detailed result has been depicted in Table 1.

3.4. Damaging nsSNPs Found by SNPs&GO, PHD-SNP and PANTHER

SNPs&GO is a server for the prediction of single point protein mutations likely to be involved in the insurgence of diseases in humans. The results of SNPs & GO for the inputs of all 26 nsSNPs are given in Table 1. The results are displayed in terms of neutral or disease causing mutation. 18out of 26 mutations were predicted to have disease causing abilities while the rest were neutral by PHD-SNP. Also, 12 nsSNPs were disease while the rests were neutral by SNP& GO. PANTHER server was also used. Out of 26 SNPs, the server could not predict the disease causing ability of 11 SNPs. Among the 15 SNPs predicted, 9 were found to be disease causing and the rest were neutral (Table 2).

3.5. Disease Related Amino Acid Substitution Prediction by MutPred

MutPred analysis was done to determine the degree of tolerance for each amino acid substitution on the basis of physio-chemical properties. Table 3 summarizes the result obtained from MutPred server. This implied that some nsSNPs may account for potential structural and functional alterations of KRAS.

Table 1. List of non-synonymous SNPs of the human KRAS gene analyzed by SIFT and PROVEAN

rsSNP NO	Nucleotide change	Amino acid change	SIFT prediction	SIFT score	PROVEAN prediction	PROVEAN score (cutoff = −2.5)
rs104886028	C/T	M72I	TOLERATED	0.358	DELETERIUOS	-3.376
rs104886029	G/A	A59V	DELETERIOUS	0	DELETERIUOS	-3.839
rs104894359	C/G	G60R	DELETERIOUS	0	DELETERIUOS	-7.758
rs373500216	G/C	A134G	DELETERIOUS	0.016	DELETERIOUS	-3.378
rs104894361	T/A	K5N	DELETERIOUS	0	DELETERIUOS	-3.836
rs104894364	G/A	T58I	DELETERIOUS	0	DELETERIUOS	-5.823
rs104894365	C/T	V14I	DELETERIOUS	0.001	NEUTRAL	-0.819
rs104894366	G/C	P34R	DELETERIOUS	0	DELETERIOUS	-7.598
rs121913238	G/C	Q61E	DELETERIOUS	0.003	DELETERIUOS	-2.772
rs121913527	C/T	A146T	DELETERIOUS	0	DELETERIUOS	-3.628
rs202247812	T/C	N116S	DELETERIOUS	0.001	DELETERIUOS	-4.564
rs387907205	A/G	Y71H	TOLERATED	0.061	DELETERIUOS	-4.454
rs387907206	T/C	K147E	DELETERIOUS	0.001	DELETERIUOS	-3.624
rs121913538	C/A	L19F	DELETERIOUS	0.001	DELETERIOUS	-3.799
rs138669124	A/T	F141L	DELETERIOUS	0.035	DELETERIUOS	-4.534
rs372793780	C/T	R164Q	DELETERIOUS	0.016	DELETERIOUS	-2.610
rs374681135	G/A	P178S	TOLERATED	0.507	NEUTRAL	-0.199
rs368557003	T/C	Q165R	TOLERATED	0.949	NEUTRAL	-0.013
rs369501492	T/A	M170L	TOLERATED	0.33	-	-
rs201170656	T/G	M189L	TOLERATED	1	NEUTRAL	-0.162
rs200970347	C/T	G179S	TOLERATED	0.11	NEUTRAL	-1.374
rs193929331	T/C	K5E	DELETERIOUS	0.001	DELETERIOUS	-3.727
rs121913528	C/T	A59T	DELETERIOUS	0.024	DELETERIOUS	-3.727
rs104894367	A/C	V152G	DELETERIOUS	0.004	-	-
rs104894362	G/C	F156L	DELETERIOUS	0.001	-	-
rs104894360	T/A	D40V	DELETERIOUS	0.001	NEUTRAL	-0.52

Table 2. List of non-synonymous SNPs analyzed for disease association by SNP&GO, PHD-SNP and PANTHER

| AA change	PHD-SNP			PANTHER			SNPs&GO		
	Prediction	RI	Probability	Prediction	RI	Probability	Prediction	RI	Probability
M72I	Neutral	2	0.390	Neutral	9	0.058	Neutral	8	0.117
A59V	Disease	6	0.801	Neutral	0	0.497	Disease	1	0.555
G60R	Disease	7	0.846	Disease	8	0.910	Disease	5	0.752
A134G	Disease	6	0.780	Unclassified	NA	NA	Disease	0	0.513
K5N	Disease	0	0.523	Neutral	1	0.448	Neutral	2	0.376
T58I	Disease	6	0.793	Disease	5	0.765	Disease	3	0.664
V14I	Disease	3	0.633	Neutral	5	0.253	Neutral	3	0.343
P34R	Disease	5	0.734	Disease	8	0.920	Disease	5	0.755
Q61E	Disease	6	0.716	Neutral	7	0.133	Neutral	5	0.252
A146T	Disease	7	0.830	Unclassified	NA	NA	Disease	3	0.646
N116S	Disease	6	0.803	Unclassified	NA	NA	Disease	2	0.584
Y71H	Disease	5	0.744	Neutral	4	0.317	Neutral	2	0.411
K147E	Disease	6	0.825	Unclassified	NA	NA	Disease	3	0.629
L19F	Disease	7	0.854	Neutral	2	0.397	Disease	1	0.573
F141L	Disease	5	0.737	Unclassified	NA	NA	Neutral	3	0.359
R164Q	Disease	2	0.587	Unclassified	NA	NA	Neutral	2	0.393
P178S	Neutral	2	0.197	Unclassified	NA	NA	Neutral	9	0.071
Q165R	Neutral	5	0.225	Unclassified	NA	NA	Neutral	7	0.147
M170L	-	-	-	-	-	-	-	-	-
M189L	Neutral	9	0.031	Unclassified	NA	NA	Neutral	10	0.024
G179S	Neutral	4	0.301	Unclassified	NA	NA	Neutral	7	0.147
K5E	Disease	4	0.689	Neutral	2	0.387	Disease	3	0.658
A59T	Disease	6	0.748	Neutral	1	0.447	Disease	1	0.525
V152G	Disease	7	0.830	Unclassified	NA	NA	Disease	0	0.507
F156L	Disease	6	0.785	Unclassified	NA	NA	Neutral	0	0.481
D40V	-	-	-	-	-	-	-	-	-
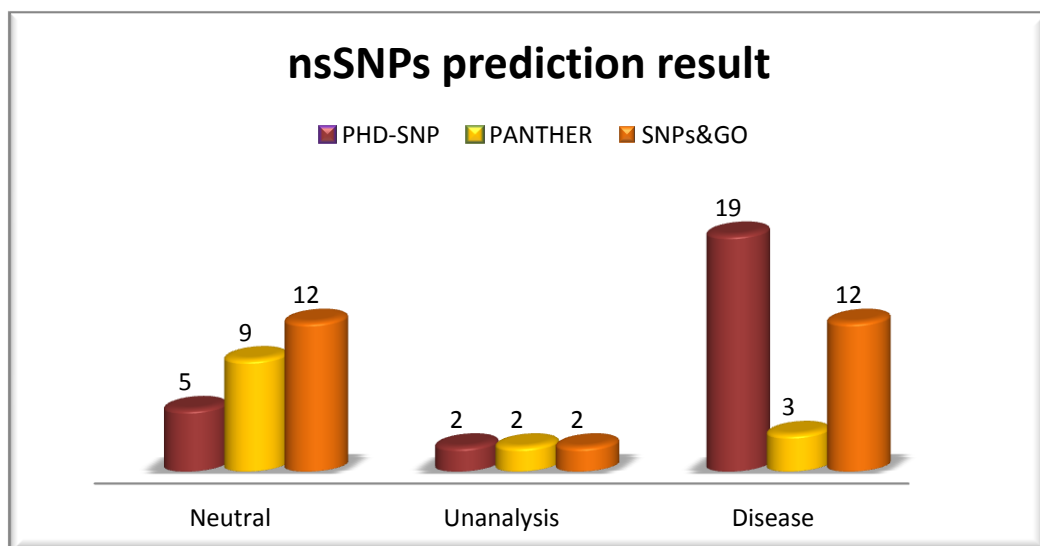
Figure 2. Graphical representation of Disease variations by PHD-SNP, PANTHER and SNP&GO

Table 3. Prediction of functional effects of nsSNPs using MutPred

Mutation	Probability Score	Top 5 features	Confident Hypotheses
M72I	-	-	
A59V	0.793 highly harmful	Loss of catalytic residue at A59 (P = 0.0406) Loss of disorder (P = 0.0741) Loss of helix (P = 0.1299) Loss of sheet (P = 0.1398) Gain of loop (P = 0.2045)	Loss of catalytic residue at A59 (P = 0.0406)
G60R	0.919 highly harmful	Gain of phosphorylation at T58 (P = 0.1068) Loss of glycosylation at T58 (P = 0.123) Gain of solvent accessibility (P = 0.1319) Loss of catalytic residue at A59 (P = 0.1348) Gain of loop (P = 0.2045	
A134G	0.864 (highly harmful)	Loss of stability (P = 0.0657) Gain of disorder (P = 0.1296) Gain of phosphorylation at Y137 (P = 0.2028) Loss of MoRF binding (P = 0.3951) Gain of catalytic residue at L133 (P = 0.4402)	
K5N	0.888(highly harmful)	Loss of methylation at Y5 (P = 0.0037) Loss of ubiquitination at G5 (P = 0.021) Gain of glycosylation at T2 (P = 0.1048) Loss of disorder (P = 0.1189) Gain of sheet (P = 0.1208)	Loss of ubiquitination at G5 (P = 0.021)
T58I	0.940 (highly harmful)	Loss of glycosylation at T58 (P = 0.0493) Loss of disorder (P = 0.0697) Loss of sheet (P = 0.1398) Loss of phosphorylation at T58 (P = 0.1979) Gain of loop (P = 0.204	Loss of glycosylation at T58 (P = 0.0493)
V14I	0.897 (highly harmful)	Gain of catalytic residue at L19 (P = 0.0448) Loss of ubiquitination at K16 (P = 0.0727) Gain of methylation at K16 (P = 0.0983) Loss of sheet (P = 0.302) Loss of disorder (P = 0.3882)	Gain of catalytic residue at L19 (P = 0.0448)
P34R	0.880(highly harmful)	Loss of catalytic residue at D30 (P = 0.0714) Loss of sheet (P = 0.0817) Loss of glycosylation at T35 (P = 0.0874) Gain of solvent accessibility (P = 0.1708) Loss of disorder (P = 0.1763)	
Q61E	0.715 (harmful)	Gain of disorder (P = 0.0755) Gain of helix (P = 0.132) Loss of sheet (P = 0.1398) Gain of phosphorylation at T58 (P = 0.1442) Loss of glycosylation at T58 (P = 0.1859)	
A146T	0.919(highly harmful)	Gain of phosphorylation at A146 (P = 0.0405) Gain of glycosylation at S145 (P = 0.0545) Gain of methylation at K147 (P = 0.1163) Loss of MoRF binding (P = 0.1357) Gain of disorder (P = 0.1387)	Gain of phosphorylation at A146 (P = 0.0405)
N116S	0.889 (highly harmful)	Loss of catalytic residue at N116 (P = 0.0079) Gain of disorder (P = 0.0374) Loss of stability (P = 0.046) Gain of phosphorylation at N116 (P = 0.1088) Gain of ubiquitination at K117 (P = 0.1111)	Gain of disorder (P = 0.0374) Loss of stability (P = 0.046)
Y71H	0.771 (highly harmful)	Gain of disorder (P = 0.029) Loss of phosphorylation at Y71 (P = 0.0387) Gain of catalytic residue at Y71 (P = 0.0732) Loss of stability (P = 0.1315) Loss of helix (P = 0.1706	
K147E	0.854 (highly harmful)	Loss of methylation at K147 (P = 0.0184) Loss of ubiquitination at K147 (P = 0.0218) Loss of MoRF binding (P = 0.0249)	Loss of methylation at K147 (P = 0.0184) Loss of ubiquitination at K147 (P = 0.0218) Loss of MoRF binding (P = 0.0249)

Mutation	Probability Score	Top 5 features	Confident Hypotheses
		Gain of sheet (P = 0.1451) Gain of disorder (P = 0.1965)	
L19F	0.746 (harmful)	Gain of catalytic residue at L19 (P = 0.0722) Gain of methylation at K16 (P = 0.0942) Loss of ubiquitination at K16 (P = 0.1191) Loss of helix (P = 0.2271) Loss of stability (P = 0.2868)	
F141L	0.807 (highly harmful)	Gain of catalytic residue at F141 (P = 0.0801) Gain of sheet (P = 0.1208) Loss of phosphorylation at T144 (P = 0.1409) Gain of helix (P = 0.1736) Gain of disorder (P = 0.2017)	
R164Q	0.588 (harmful)	Loss of MoRF binding (P = 0.0435) Gain of methylation at K169 (P = 0.1503) Loss of helix (P = 0.1706) Gain of loop (P = 0.2045) Gain of disorder (P = 0.2484)	
P178S	0.303 (Normal)	Gain of MoRF binding (P = 0.0449) Gain of phosphorylation at P178 (P = 0.0478) Loss of sheet (P = 0.0817) Gain of glycosylation at P178 (P = 0.0857) Gain of methylation at K173 (P = 0.0983)	
Q165R	0.289 (Normal)	Gain of MoRF binding (P = 0.1191) Gain of disorder (P = 0.1918) Loss of stability (P = 0.3831) Loss of phosphorylation at T158 (P = 0.4074) Gain of methylation at R161 (P = 0.6389)	
M170L	-	-	
M189L	0.603 (harmful)	Loss of MoRF binding (P = 0.0735) Loss of disorder (P = 0.078) Loss of sheet (P = 0.1907) Gain of loop (P = 0.2045) Loss of methylation at K184 (P = 0.2419)	
G179S	0.342 (Normal)	Gain of phosphorylation at G179 (P = 0.0276) Gain of disorder (P = 0.03) Loss of ubiquitination at K176 (P = 0.0675) Gain of sheet (P = 0.0827) Gain of MoRF binding (P = 0.122)	
K5E	0.884 (highly harmful)	Loss of methylation at Y5 (P = 0.0037) Loss of ubiquitination at G5 (P = 0.021) Gain of catalytic residue at K5 (P = 0.1128) Gain of sheet (P = 0.1208) Gain of disorder (P = 0.1334)	Loss of ubiquitination at G5 (P = 0.021
A59T	0.782(highly harmful)	Gain of glycosylation at T58 (P = 0.0659) Gain of phosphorylation at A59 (P = 0.153) Loss of catalytic residue at A59 (P = 0.1563) Gain of disorder (P = 0.1652) Loss of stability (P = 0.1753)	

3.6. Identification of Functional nsSNP

Changes in the protein stability of missense variants were examined using I-Mutant 2.0 and MUpro software. The results for the outputs of Deleterious missense SNPs by SIFT and PROVEN analysis (G60R, A134G, K5N, T58I, P34R, Q61E, A146T, N116S, K147E, L19F, F141, R164Q, K5E and A59V) and one SNP was found deleterious by PROVEN and disease by PHD-SNP (Y71H). The predicted results were either increase or decrease of the free energy change upon mutation. 13 out of 14 SNPs were found decreasing for

the free energy by I-Mutant and 10 SNPs by Mupro. MUpro couldn't calculate 2 nsSNPs. And Y71H SNP was found decreasing by both softwares. The results are shown in (Table 4).

3.7. Impact of nsSNPS on KRAS Protein Stability

DUET is the effects of nsSNPs on protein stability. The results display the predicted change in folding free energy upon mutation (Ginkcal/mol). A positive value (and red writing) corresponds to a mutation predicted as destabilizing;

while a negative sign (and blue writing) corresponds to a mutation predicted as stabilizing. The information displayed includes them CSM(i) and SDM(ii) individually predicted protein stability changes, the combined DUET prediction (iii), a structural summary of the mutation highlighting the wild-type residue and position number, the mutation and its 3D environment (iv). The protein and mutation can also be visualized (v). The results are shown in figures (3-17), each figure shows regression analysis for the stability predictions generated by DUET in comparison with the experimentally measured variation in stability for the considered datasets

3.8. Prediction of Stability Effects upon Mutation in Both, Domain Cores and Domain-Domain Interfaces by ELASPIC

We classified nsSNPs into two groups depending on their structural location: core or interface. To designate a residue as participating in a protein interaction interface, we used ELASPIC server. We obtained core residue positions for 4 snSNPs and interfaces for 9 snSNPs. The result is explained in (Table 5).

3.9. Distributions of nsSNPs in RAS Domain by Mutation3D Server

We Detected 11nsSNPs located in a protein domain structure (G60R, A134G, P34R, Q61E, A146T, N116S, K147E, L19F, F141, Y71H and A59V), 2 nsSNPs were uncovering K5E and Q164R and one SNP hasn't founded T58I. This 11snSNP were considered higher risk for KRAS protein. The result is shown in (fig 18).

Table 4. Prediction of mutant protein stability by I-Mutant 2.0 and MUpro

Amino acid change	Amino acid position	I.Mutant Stability prediction	DDG value (kcal/mol)	MUpro Stability prediction	Score
A/V	59	DECREASE	-1.36	INCREASE	1
R/Q	164	DECREASE	-0.59	DECREASE	-0.4372822
G/R	60	DECREASE	-1.36	DECREASE	-0.5152381
K/N	5	INCREASE	-0.23	-	-
P/R	34	DECREASE	-0.60	DECREASE	-0.48127684
Q/E	61	INCREASE	0.22	INCREASE	0.78593395
A/T	146	DECREASE	-0.97	DECREASE	-0.97001511
N/S	116	DECREASE	-0.81	DECREASE	-1
L/F	19	DECREASE	-0.55	DECREASE	-1
F/L	141	DECREASE	-1.05	-	-
K/E	147	DECREASE	-0.73	DECREASE	-0.20800439
A/G	134	DECREASE	-1.27	DECREASE	-0.82581505
K/E	5	DECREASE	-0.23	DECREASE	-0.35902125
A/T	59	DECREASE	-1.36	INCREASE	0.42365096
Y/H	71	DECREASE	-0.77	DECREASE	-1

Figure 3.   GLY60ARG SNP analysis of DUET server

mCSM Predicted Stability Change (ΔΔG):

-0.564 Kcal/mol (*Destabilizing*)

SDM Predicted Stability Change (ΔΔG):

0.12 Kcal/mol (*Stabilizing*)

DUET Predicted Stability Change (ΔΔG):

-0.506 Kcal/mol (*Destabilizing*)

Mutation:

Wild-type: **ALA**
Position: **59**
Mutant-type: **VAL**
Chain: **A**
Relative solvent accessibility: **8.70 %**
Secondary structure: **Loop or irregular**
Side-chain hydrogen bond: **No Hbonds**

Figure 4. ALA59VAL SNP analysis of DUET server

mCSM Predicted Stability Change (ΔΔG):

-0.453 Kcal/mol (*Destabilizing*)

SDM Predicted Stability Change (ΔΔG):

-0.77 Kcal/mol (*Destabilizing*)

DUET Predicted Stability Change (ΔΔG):

-0.264 Kcal/mol (*Destabilizing*)

Mutation:

Wild-type: **LYS**
Position: **5**
Mutant-type: **ASN**
Chain: **A**
Relative solvent accessibility: **48.80 %**
Secondary structure: **Extended beta-strand**
Side-chain hydrogen bond: **No Hbonds**
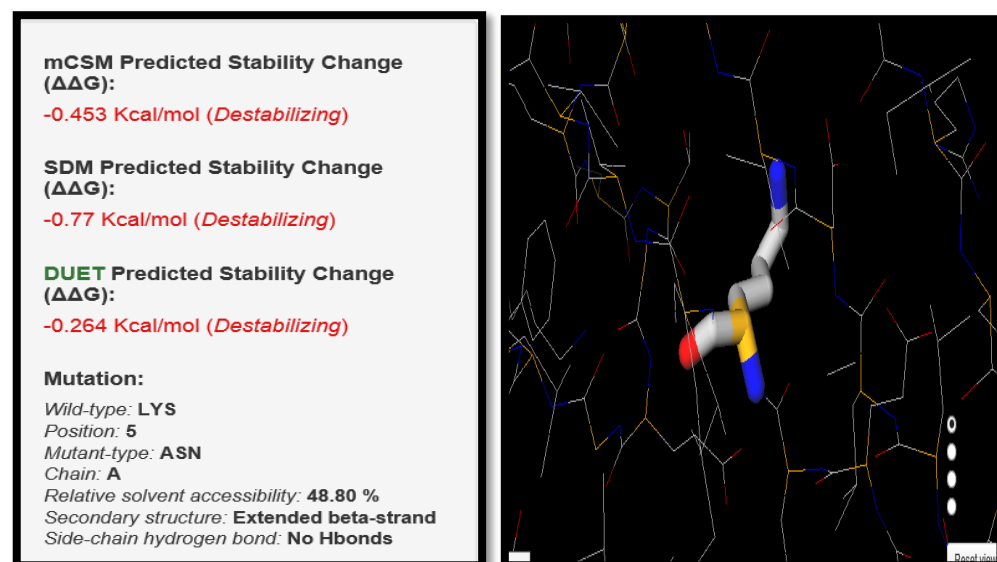
Figure 5. LYS5ASN SNP analysis of DUET server

mCSM Predicted Stability Change (ΔΔG):

-0.038 Kcal/mol (*Destabilizing*)

SDM Predicted Stability Change (ΔΔG):

1.07 Kcal/mol (*Stabilizing*)

DUET Predicted Stability Change (ΔΔG):

-0.036 Kcal/mol (*Destabilizing*)

Mutation:
Wild-type: **PRO**
Position: **34**
Mutant-type: **ARG**
Chain: **A**
Relative solvent accessibility: **33.00 %**
Secondary structure: **H-bonded turn**
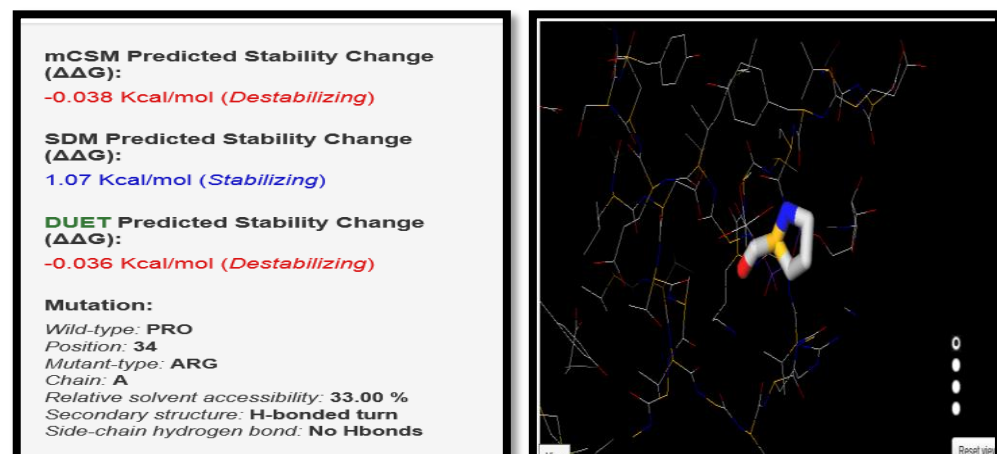*Side-chain hydrogen bond:* **No Hbonds**

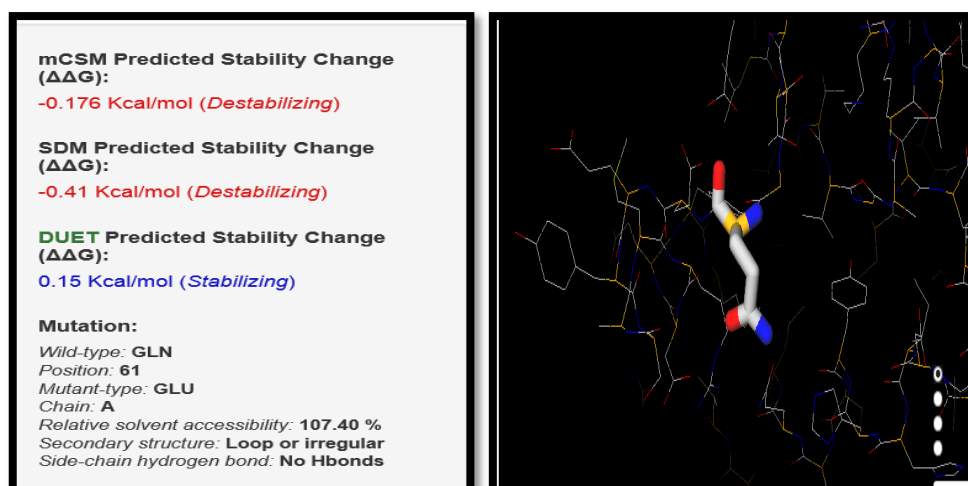**Figure 6.** PRO34ARG SNP analysis of DUET server

Figure 7.   GLN61GLU SNP analysis of DUET server
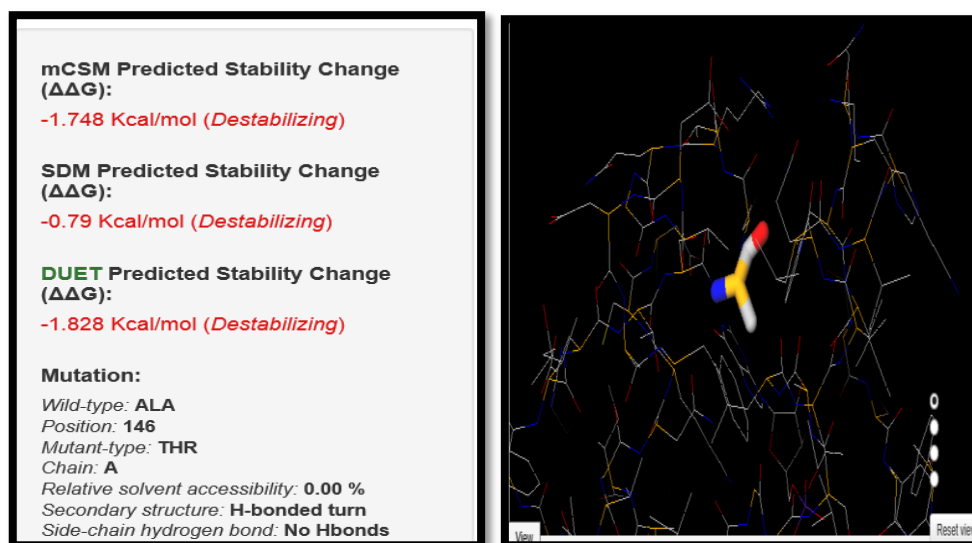
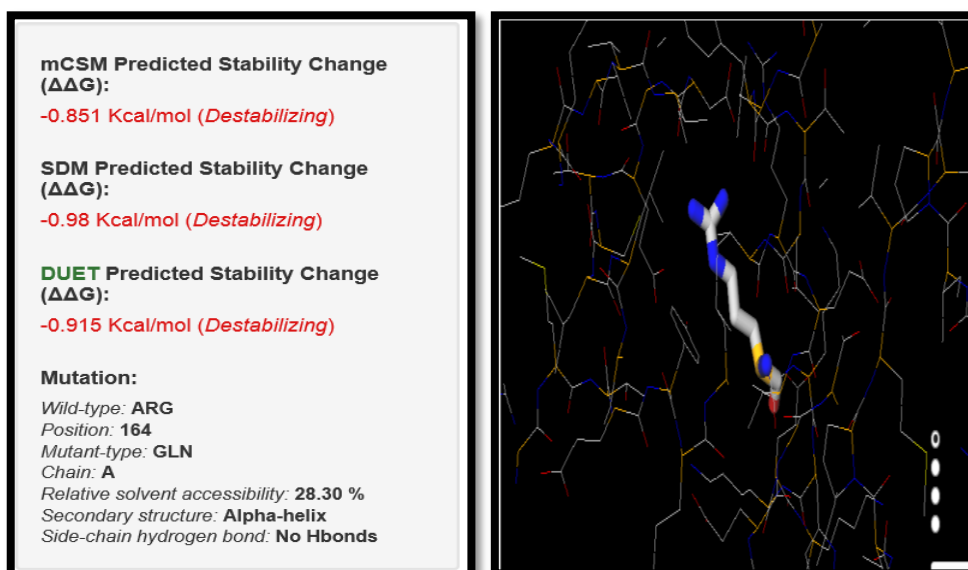**Figure 8.**   ALA146THR SNP analysis of DUET server
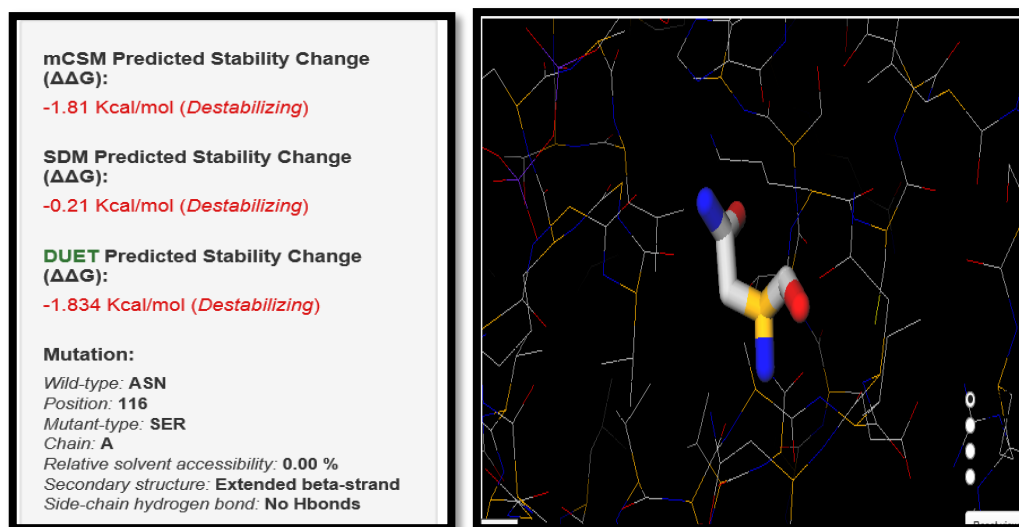

Figure 9.   ARG164GLN SNP analysis of DUET server

mCSM Predicted Stability Change (ΔΔG):

-1.81 Kcal/mol (*Destabilizing*)

SDM Predicted Stability Change (ΔΔG):

-0.21 Kcal/mol (*Destabilizing*)

DUET Predicted Stability Change (ΔΔG):

-1.834 Kcal/mol (*Destabilizing*)

Mutation:

Wild-type: **ASN**
Position: **116**
Mutant-type: **SER**
Chain: **A**
Relative solvent accessibility: **0.00 %**
Secondary structure: **Extended beta-strand**
Side-chain hydrogen bond: **No Hbonds**

Figure 10.   ASN116SER SNP analysis of DUET server

mCSM Predicted Stability Change (ΔΔG):

-1.736 Kcal/mol (*Destabilizing*)

SDM Predicted Stability Change (ΔΔG):

-0.84 Kcal/mol (*Destabilizing*)

DUET Predicted Stability Change (ΔΔG):

-1.885 Kcal/mol (*Destabilizing*)

Mutation:

Wild-type: **LEU**
Position: **19**
Mutant-type: **PHE**
Chain: **A**
Relative solvent accessibility: **0.20 %**
Secondary structure: **Alpha-helix**
Side-chain hydrogen bond: **No Hbonds**
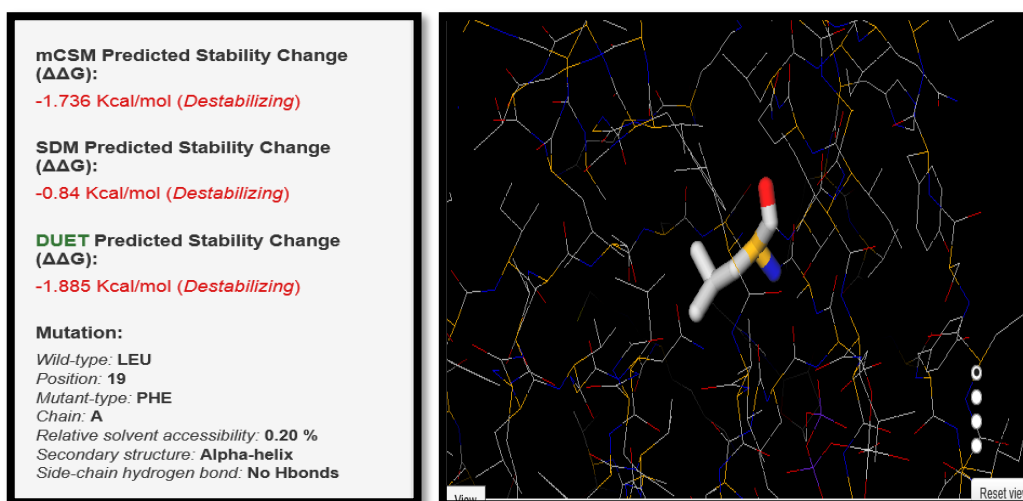
Figure 11.   LEU19PHE SNP analysis of DUET server

mCSM Predicted Stability Change (ΔΔG):

-1.296 Kcal/mol (*Destabilizing*)

SDM Predicted Stability Change (ΔΔG):

-0.04 Kcal/mol (*Destabilizing*)

DUET Predicted Stability Change (ΔΔG):

-1.398 Kcal/mol (*Destabilizing*)

Mutation:

Wild-type: **PHE**
Position: **141**
Mutant-type: **LEU**
Chain: **A**
Relative solvent accessibility: **7.70 %**
Secondary structure: **Extended beta-strand**
Side-chain hydrogen bond: **No Hbonds**
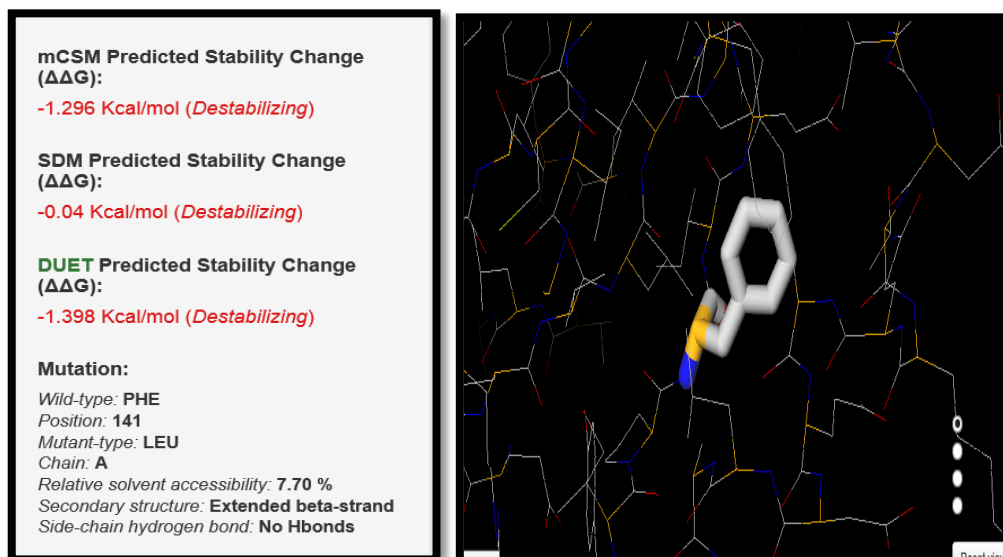
Figure 12.   PHE141LEU SNP analysis of DUET server
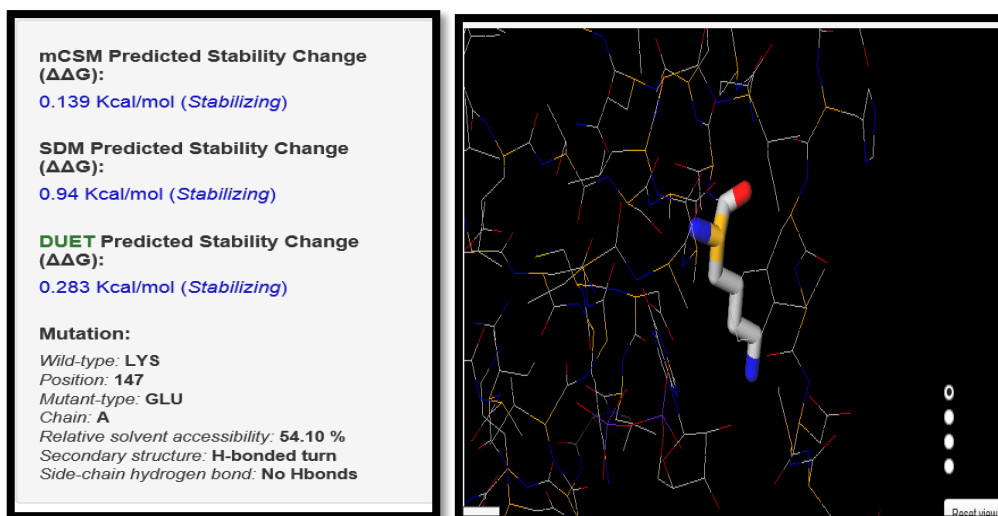
Figure 13.   LYS147GLU SNP analysis of DUET server
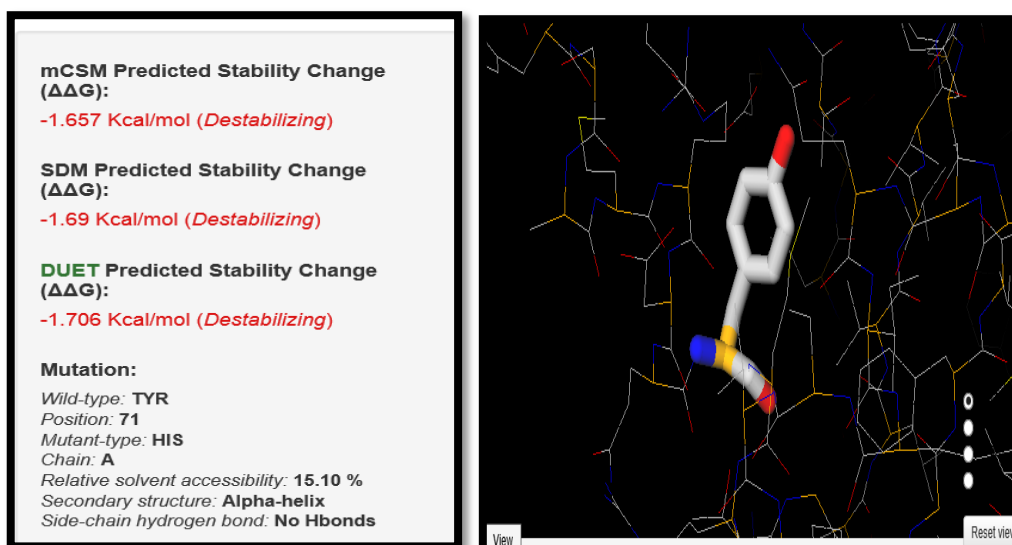
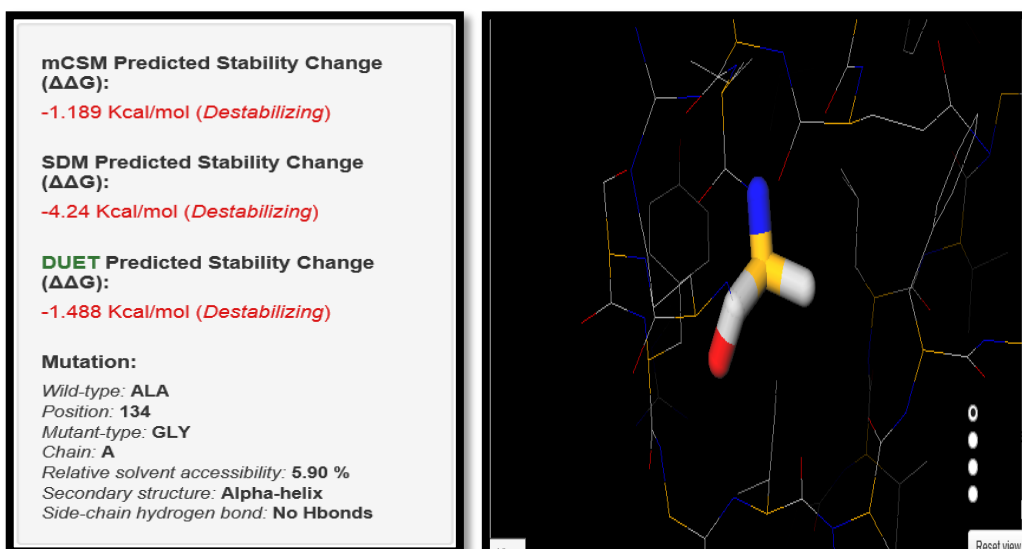**Figure 14.**   TYR71HIS SNP analysis by DUET server

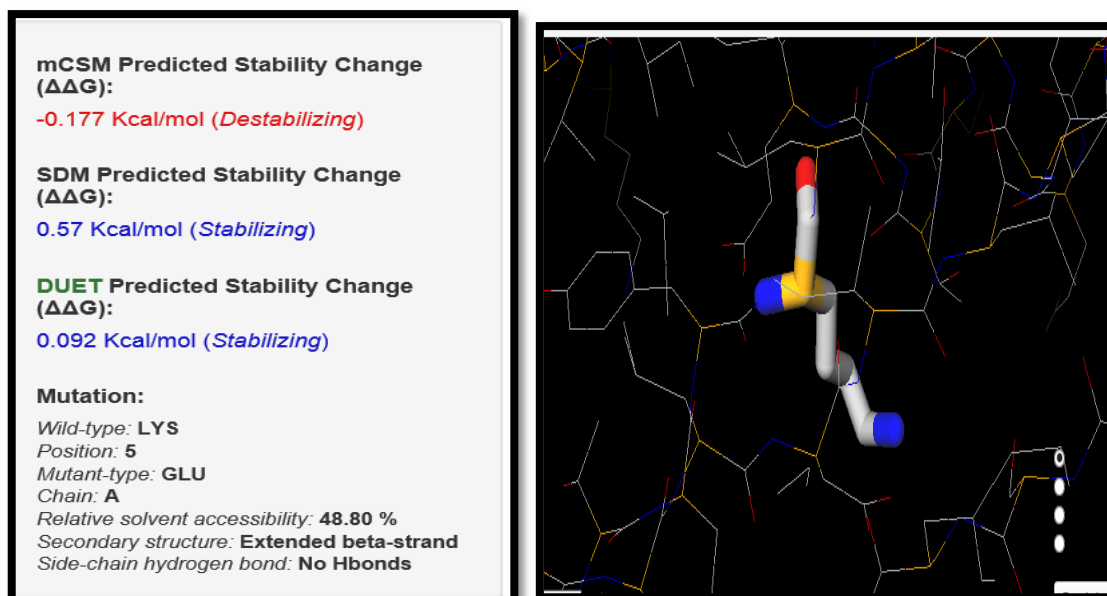**Figure 15.**   ALA134GLY SNP analysis of DUET server

mCSM Predicted Stability Change (ΔΔG):
-0.177 Kcal/mol (*Destabilizing*)

SDM Predicted Stability Change (ΔΔG):
0.57 Kcal/mol (*Stabilizing*)

DUET Predicted Stability Change (ΔΔG):
0.092 Kcal/mol (*Stabilizing*)

Mutation:
Wild-type: **LYS**
Position: **5**
Mutant-type: **GLU**
Chain: **A**
Relative solvent accessibility: **48.80 %**
Secondary structure: **Extended beta-strand**
Side-chain hydrogen bond: **No Hbonds**

Figure 16. LYS5GLU SNP analysis of DUET server

mCSM Predicted Stability Change (ΔΔG):
-0.709 Kcal/mol (*Destabilizing*)

SDM Predicted Stability Change (ΔΔG):
-0.1 Kcal/mol (*Destabilizing*)

DUET Predicted Stability Change (ΔΔG):
-0.567 Kcal/mol (*Destabilizing*)

Mutation:
Wild-type: **ALA**
Position: **59**
Mutant-type: **THR**
Chain: **A**
Relative solvent accessibility: **8.70 %**
Secondary structure: **Loop or irregular**
Side-chain hydrogen bond: **No Hbonds**
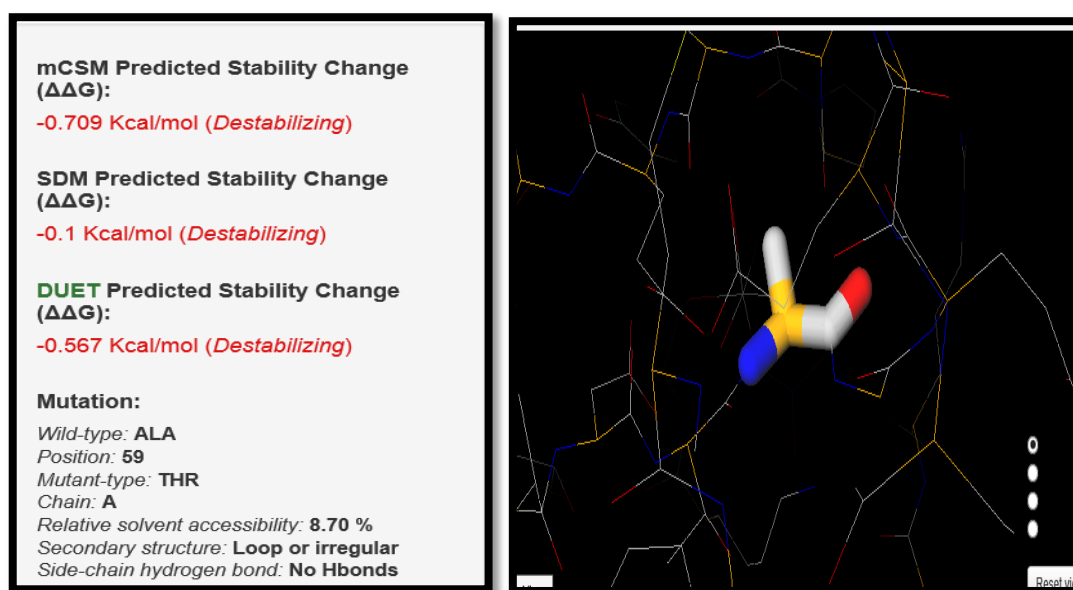
Figure 17. ALA59THR SNP analysis of DUET server

Table 5. Stability effects nsSNPs in domain cores and domain-domain interfaces by ELASPIC

Mutation	Type	ΔG_{wt}	ΔG_{mut}	$\Delta\Delta G$
A134G	Core	25.9877	27.7076	1.505
A146T	Core	25.8833	30.4648	1.363
L19F	Core	25.043	32.9852	1.862
R164Q	Core	25.792	27.1035	1.430
A59V	Interface w/ GNDS isoform 2	-8.14228	-8.89411	0.098
A59V	Interface w/ RASGRP2	-12.7098	-12.9341	1.455
A59V	Interface w/ GRP2 isoform 2	-19.5184	-20.4798	1.666
A59V	Interface w/ RGL2	22.4762	21.7994	0.782
A59V	Interface w/ RASGRP3	5.60172	8.49014	-0.708
A59V	Interface w/ GRP3 isoform 2	7.04479	11.572	-1.225
A59V	Interface w/ RASGRF2	-11.8625	-12.2989	0.598
A59V	Interface w/ Q6PCE1	-12.3988	-14.446	0.228

Mutation	Type	ΔG_{wt}	ΔG_{mut}	$\Delta\Delta G$
F141L	Interface w/ EPHB2 isoform 2	2.91738	3.01623	0.914
F141L	Interface w/ EPHB2 isoform 3	4.72306	4.79971	0.928
F141L	Interface w/ FGFR1 isoform 6	0.998398	1.30955	1.786
F141L	Interface w/ FGFR1 isoform 7	0.998398	1.30955	1.786
F141L	Interface w/ FGFR3	-2.44454	-2.42097	0.316
F141L	Interface w/ FGFR3 isoform 2	-2.44454	-2.42097	0.316
F141L	Interface w/ FGFR3 isoform 3	0.260501	1.38867	0.522
F141L	Interface w/ CAMK2B	3.70003	4.30686	0.812
F141L	Interface w/ KCC2B isoform 2	5.13709	5.44805	1.847
F141L	Interface w/ KCC2B isoform 3	5.13709	5.44805	1.847
F141L	Interface w/ KCC2B isoform 4	5.13709	5.44805	1.847
F141L	Interface w/ KCC2B isoform 5	5.13709	5.44805	1.847
F141L	Interface w/ KCC2B isoform 6	5.13709	5.44805	1.847
F141L	Interface w/ KCC2B isoform 8	5.13709	5.44805	1.847
F141L	Interface w/ RAF1	10.4884	10.5214	0.506
F141L	Interface w/ RAF1 isoform 2	18.6163	18.8813	2.103
F141L	Interface w/ RPS6KA2	5.94346	6.17709	1.455
F141L	Interface w/ MAP3K3	0.734483	0.97636	0.629
F141L	Interface w/ M3K3 isoform 2	-4.95877	-4.99415	1.426
F141L	Interface w/ JAK1	-0.893071	-0.904665	0.865
F141L	Interface w/ self	0.0958425	0.169964	1.462
F141L	Interface w/ MAPK1	-0.243383	-0.0880824	0.809
F141L	Interface w/ self	-1.02989	-1.01777	2.566
F141L	Interface w/ RPS6KA2	0.400825	1.20373	1.526
F141L	Interface w/ KS6A2 isoform 2	4.17322	4.3234	1.401
F141L	Interface w/ KS6A2 isoform 3	2.33006	2.60007	1.504
F141L	Interface w/ KS6A2 isoform 3	7.4631	7.67391	3.182
F141L	Interface w/ RPS6KA6	5.02975	5.10512	1.992
F141L	Interface w/ RPS6KA6	-3.36752	-3.29276	0.381
F141L	Interface w/ NTRK1	2.23428	2.33305	0.910
F141L	Interface w/ NTRK1 isoform 2	0.754929	0.870462	1.710
F141L	Interface w/ NTRK1 isoform 3	2.23428	2.33305	0.910
F141L	Interface w/ NTRK1 isoform 4	1.46169	1.6638	1.495
F141L	Interface w/ RET	4.28189	3.95185	0.645
F141L	Interface w/ RET isoform 2	5.65239	6.21458	2.626
F141L	Interface w/ STK11	4.30991	4.35824	0.891
F141L	Interface w/ STK11 isoform 2	3.45535	3.7581	0.902
F141L	Interface w/ ARAF	4.59639	5.79579	0.297
F141L	Interface w/ Q96II5	5.688	5.59322	2.489
F141L	Interface w/ B4E0X2	10.0544	9.81124	1.609
F141L	Interface w/ L7RRS6	10.3486	11.0333	0.954
G60R	Interface w/ GNDS isoform 2	-10.1092	-7.02888	0.574
G60R	Interface w/ RASGRP2	-12.5889	-9.55914	1.603
G60R	Interface w/ GRP2 isoform 2	-20.3335	-16.6697	1.506
G60R	Interface w/ RGL2	22.4251	43.241	1.299
G60R	Interface w/ RASGRP3	5.61069	23.2786	-0.065
G60R	Interface w/ RASGRF2	-12.0759	-5.5571	1.200
G60R	Interface w/ Q6PCE1	-11.6684	-6.03049	1.044
K147E	Interface w/ RASGRP3	6.17638	6.8121	0.385
K5E	Interface w/ GRP2 isoform 2	-19.6072	-19.7599	0.187
K5E	Interface w/ RASGRP3	6.24692	6.65376	-0.476
K5N	Interface w/ GRP2 isoform 2	-19.6072	-19.7246	0.175
K5N	Interface w/ RASGRP3	6.24692	6.51511	-0.847

Mutation	Type	ΔG_{wt}	ΔG_{mut}	$\Delta\Delta G$
N116S	Interface w/ RASGRP3	6.20943	6.39417	-0.116
P34R	Interface w/ GNDS isoform 2	-7.99343	-6.86063	1.301
P34R	Interface w/ RASGRP2	-13.0652	-13.1025	2.437
P34R	Interface w/ GRP2 isoform 2	-20.4933	-18.1574	2.105
P34R	Interface w/ RGL2	21.2632	29.4011	2.604
P34R	Interface w/ RASGRP3	5.85446	4.38821	-0.661
P34R	Interface w/ RASGRF2	-11.7697	-12.1523	0.699
P34R	Interface w/ Q6PCE1	-13.3996	-13.0016	2.457
Q61E	Interface w/ GNDS isoform 2	-6.28822	-7.50674	-0.872
Q61E	Interface w/ RASGRP2	-12.0988	-10.2192	0.094
Q61E	Interface w/ GRP2 isoform 2	-19.7656	-20.506	-0.129
Q61E	Interface w/ RGL2	17.6733	17.3831	-0.426
Q61E	Interface w/ RASGRP3	5.77864	5.61189	0.855
Y71H	Interface w/ GNDS isoform 2	-6.40214	-6.56891	0.776
Y71H	Interface w/ RASGRP2	-12.5031	-12.0581	0.135
Y71H	Interface w/ GRP2 isoform 2	-19.2923	-19.2433	0.874
Y71H	Interface w/ RGL2	22.7957	21.1824	0.618
Y71H	Interface w/ RASGRP3	5.77191	6.31831	-0.679
Y71H	Interface w/ RASGRF2	-11.5777	-11.6553	1.345
Y71H	Interface w/ Q6PCE1	-13.6448	-14.4746	-0.181
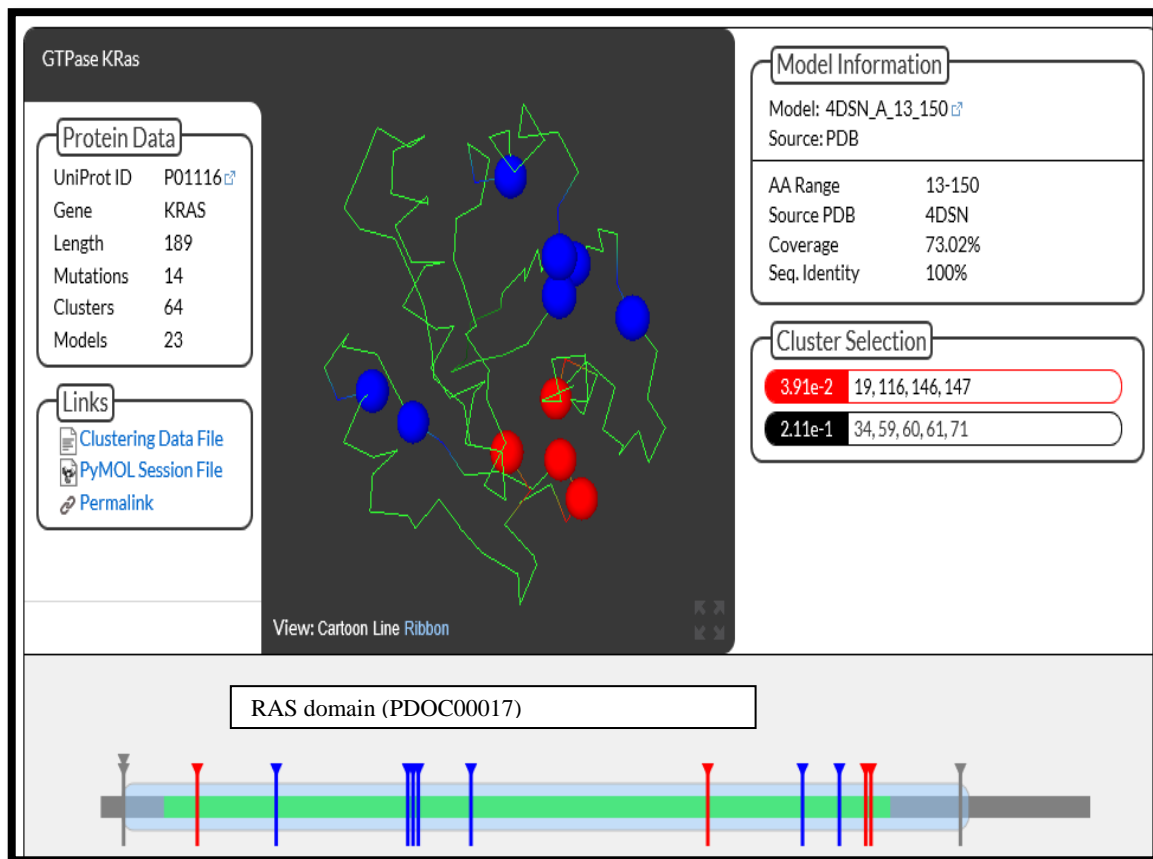
Figure 18. Structural models for wild type KRAS and 11 high-risk nsSNPs in the RAS domain

3.10. Conservation Profile of High-Risk Non-Synonymous SNPs

The nsSNPs that are located at highly conserved amino acid positions tend to be more deleterious than nsSNPs that are located at non-conserved sites. To further investigate the potential effects of the 11 high-risk nsSNPs plus 5K and Q164 SNPs, we calculated the degree of evolutionary conservation of all amino acid sites in the KRAS protein using the ConSurf web server. ConSurf analysis revealed that all residues are highly conserved (Conservation Score of 7–9) except Q164 SNP is conserved (Conservation Score of 6) (Table 6).

3.11. Homology Modeling of New and Wild Amino Acids of Deleterious nsSNPs

Protein 3D structural information is an important feature for predicting the effects of deleterious nsSNPs. Analysis of the protein structure provides information about the environment of the mutation.

The H-bonding interactions and clashes of wild type and mutant of 14 nsSNPS were calculated by Chimer program. The native and mutant protein structure for the 14 nsSNPs of human KRAS (PDB ID: 4QL3) are shown in Figures (19-32).

Table 6. Conservation profile of amino acids in KRAS that coincide in location with high-risk nsSNPs

Residues position	Residues	CS Score normalized	Color	RESIDUE VARIETY	Function
5	K	-0.706	9	R,N,K	F
19	L	-0.659	8	L,M,F	S
34	P	-0.760	9	P,S,T	E
59	A	-0.834	9	T,A	B
60	G	-0.852	9	R	B
61	Q	-0.827	9	Q,,T	B
116	N	-0.869	9	S	B
134	A	-0.655	8	S,A	F
141	F	-0.294	7	H,Y,W,F	S
146	A	-0.834	9	A,S	S
147	K	-0.763	8	A,K,R	S
164	R	-0.267	6	S,N,R,K,D,I	B
71	Y	-0.692	9	H,Y	S

CS: conservation score (1–4= variable, 5= average, 6–9= conserved); (f): predicted functional site, (s): predicted structural site; (E):exposed ;(B): Buried.
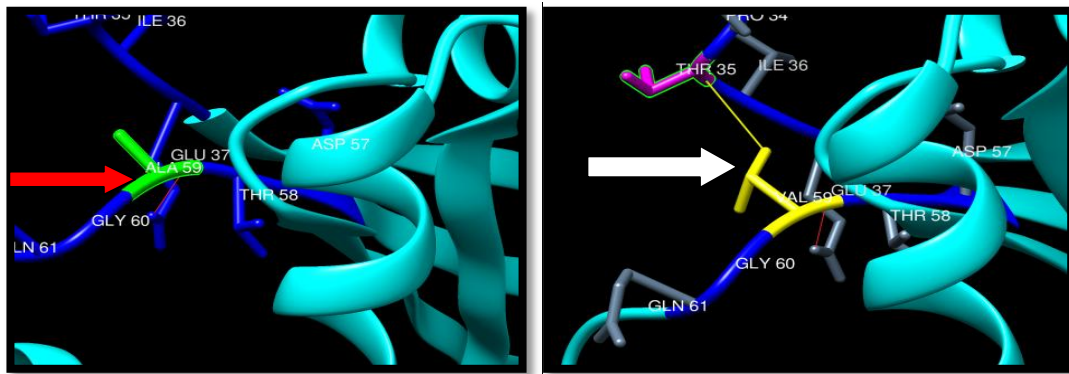
Figure 19. Rs121913527 (A59V) SNP structure: Single H-bonding interaction in wild type and mutant, one clash between mutant and THR35. Wide type green color, mutant Yellow color, H-bonding (Red line) and Clash (Yellow line)
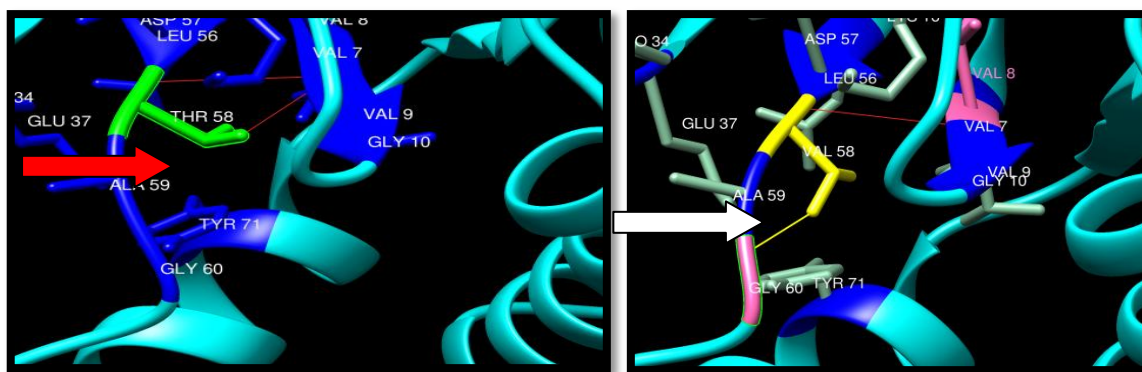
Figure 20. Rs121913527 (T58V) SNP structure: 2 H-bonding interaction in wild type and one in mutant, one clash between mutant and GLY60. Wide type green color, mutant Yellow color, H-bonding (Red line) and Clash (Yellow line)
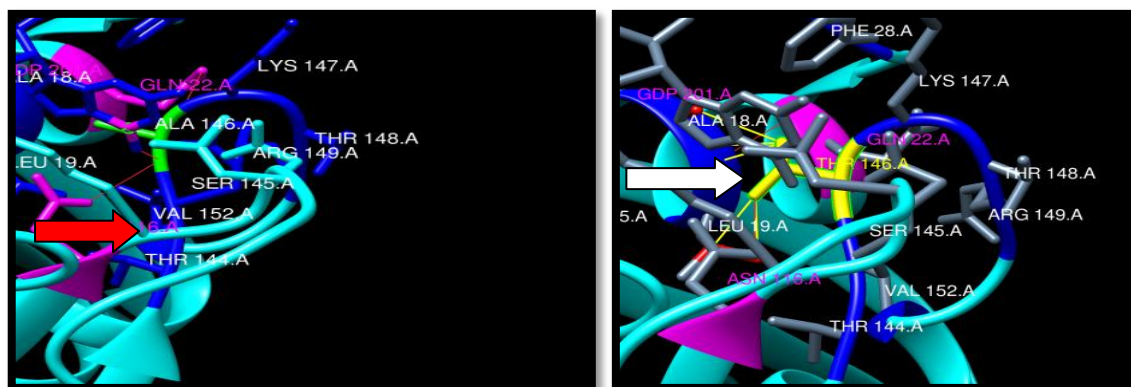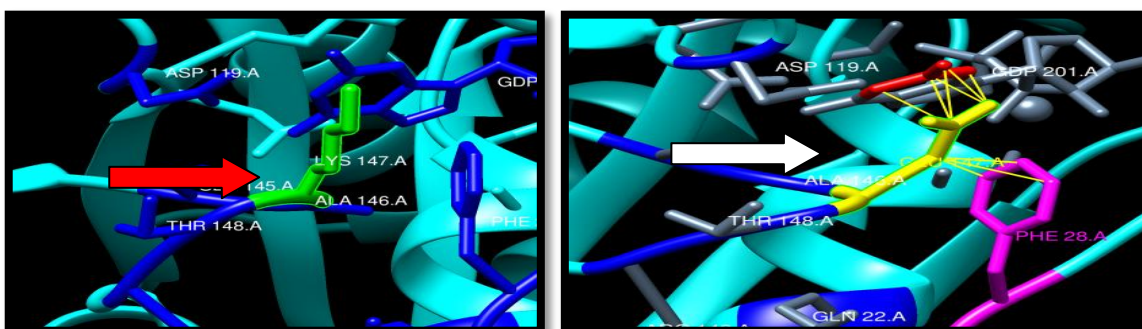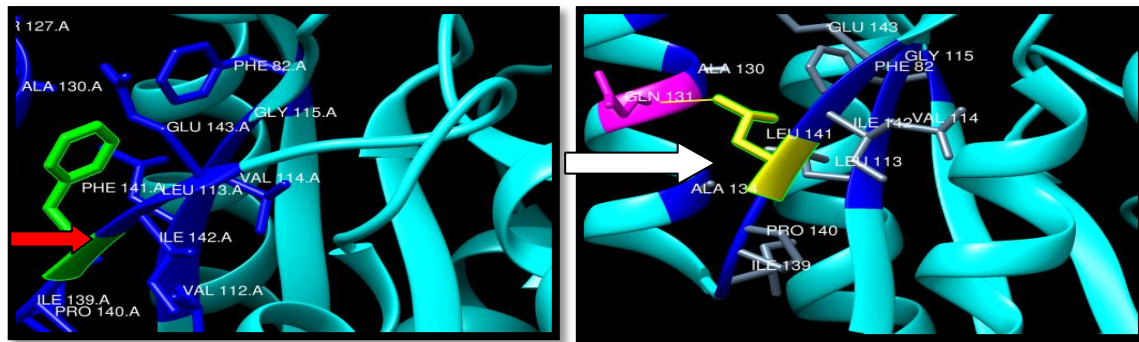

Figure 21. Rs121913527 (A146T) SNP structure: 3 H-bonding interaction in wild type and 2 in mutant, 5 clash between mutant and CDP 201 Logan, LEU19 and ALA18. wild type green color, mutant Yellow color, H-bonding (Red line) and Clash (Yellow line)

Figure 22. Rs387907206 (K147E) SNP structure:9 clash between mutant and CDP 201 Logan and PHE28 . wild type green color, mutant Yellow color and Clash (Yellow line)
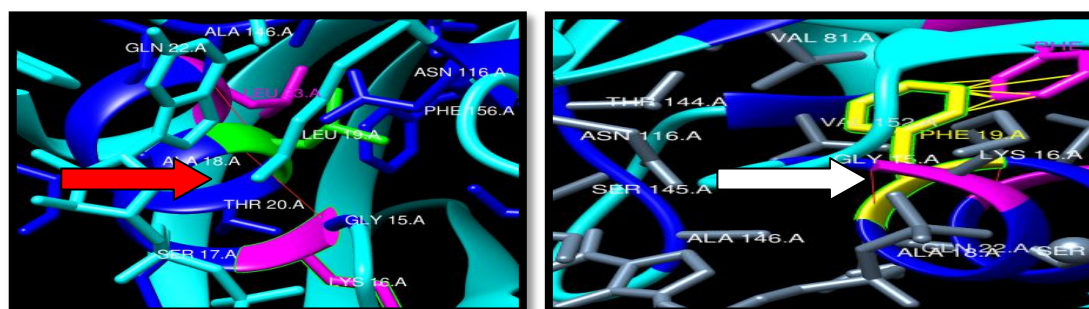
Figure 23. Rs121913538 (L19F) SNP structure: 2 H-bonding interaction in wild type and mutant. 8 clash between mutant and PHE 56. wild type green color, mutant Yellow color, H-bonding (red line) and Clash (Yellow line)

Figure 24.   Rs138669124 (F141L) SNP structure: wild type green color, mutant Yellow color and single clash between mutant and GLN 131 (Yellow line)
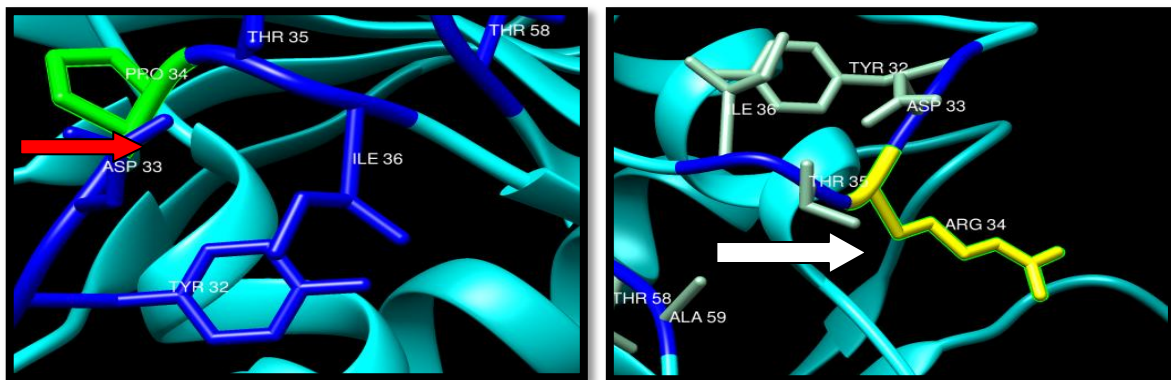
Figure 25.   Rs104894366 (P34R) SNP structure: wild type green color and mutant Yellow color
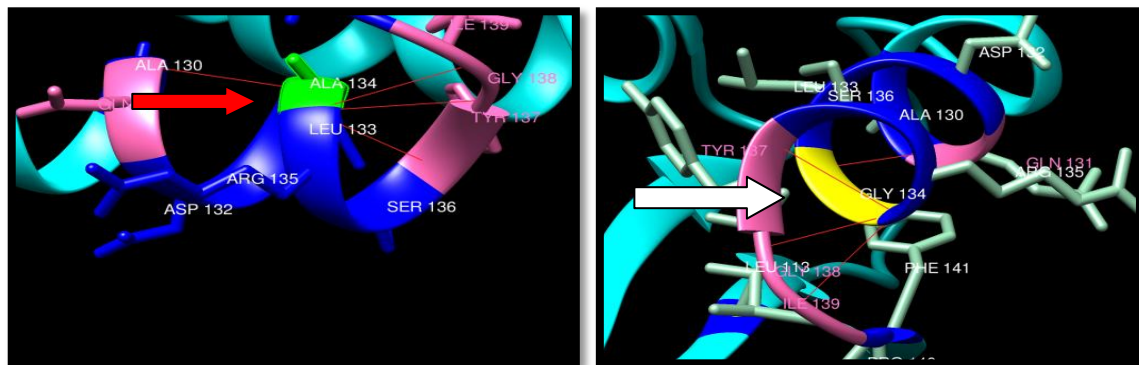
Figure 26.   Rs373500216 (A134G) SNP structure: 4H-bonding interaction in wild type and 4 -bonding interaction in mutant. wild type green color, mutant Yellow color, H-bonding (red line)
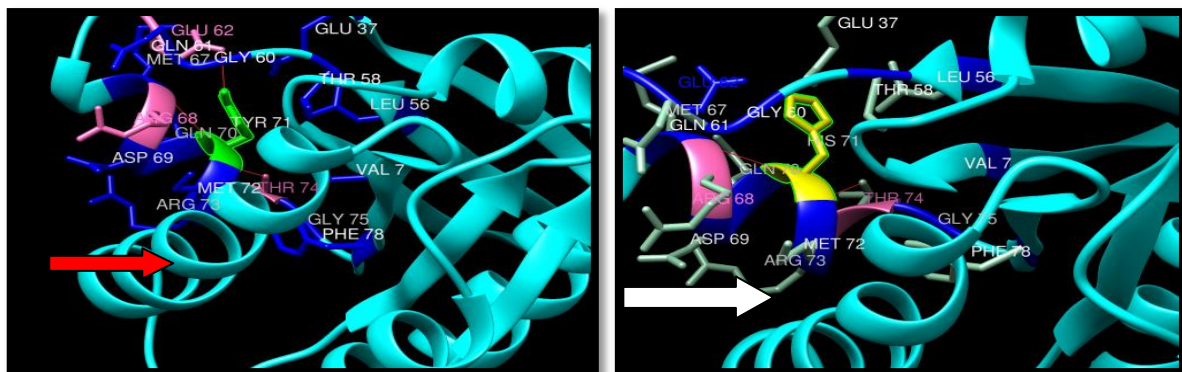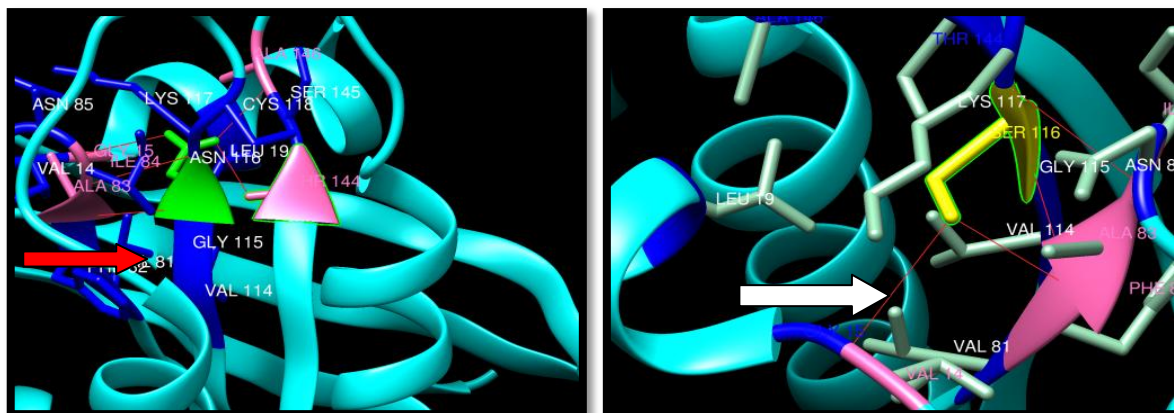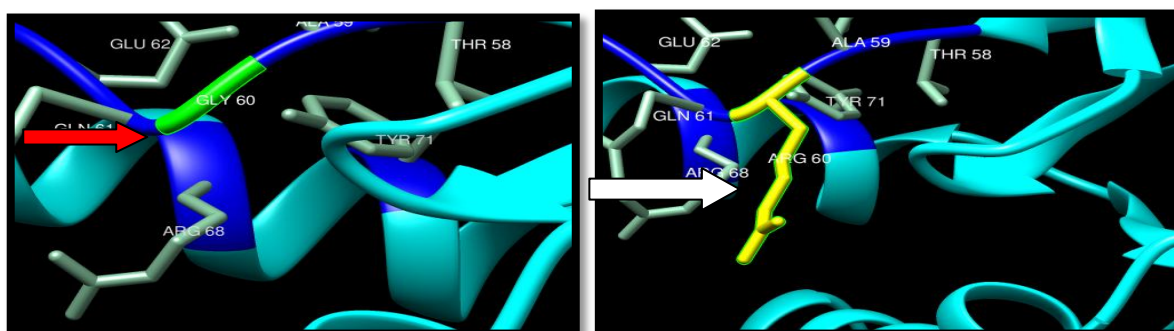

Figure 27.   Rs387907205 (Y71H) SNP structure: 3 H-bonding interaction in wild type and 2 -bonding interaction in mutant. wild type green color, mutant Yellow color, H-bonding (red line)

Figure 28. Rs202247812 (N116S) SNP structure: 5 H-bonding interaction in wild type and 3 -bonding interaction in mutant , wild type green color, mutant Yellow color, H-bonding (red line)

Figure 29. Rs104894359 (G60R) SNP structure: wild type green color and mutant Yellow color
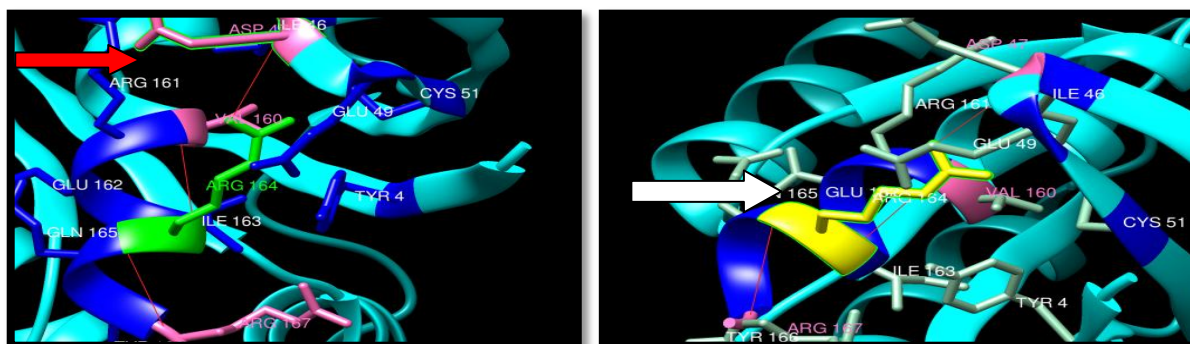
Figure 30. Rs372793780 (R164Q) SNP structure: 3 H-bonding interaction in wild type and mutant. wild type green color, mutant Yellow color, H-bonding (red line)
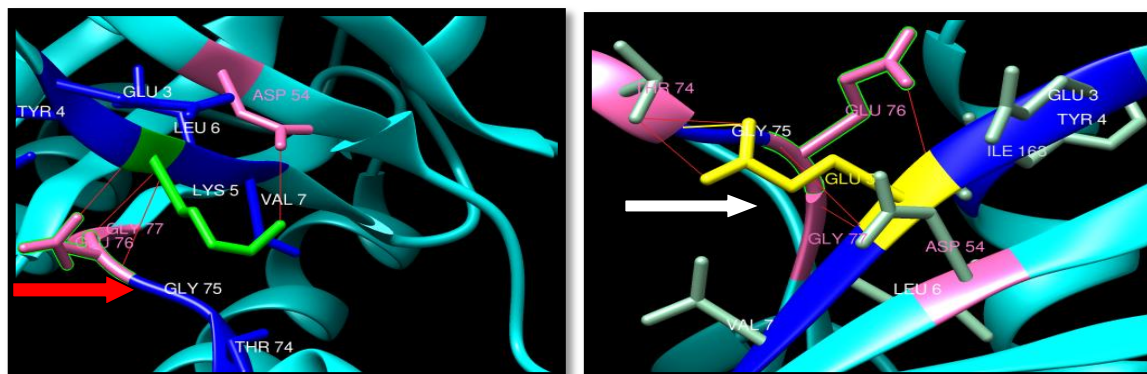

Figure 31. Rs193929331 (K5E) SNP structure: 4 H-bonding interaction in wild type and 5 H-bonding interaction in mutant and one clash THE74. wild type green color, mutant Yellow color, H-bonding (red line) and clash by yellow color
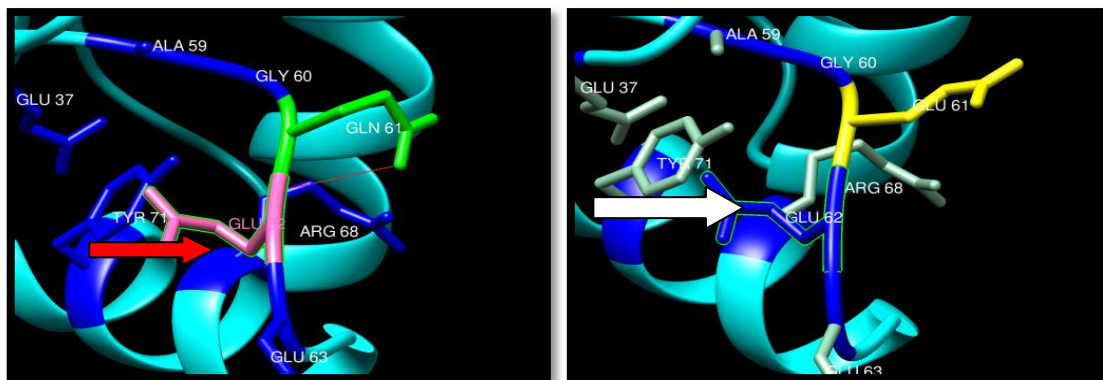
Figure 32. Rs121913238 (Q61E) SNP structure: Single H-bonding interaction in wild type. wild type green color, mutant Yellow color and H-bonding (red line)

4. Discussion

Most of the nsSNPs of KRAS are still uncharacterized in terms of their disease causing potential. So, this study has been conducted to explore and extend the knowledge related to the effect of nsSNPs on the stability and function of the KRAS gene using bioinformatics approach. The methods used are based on different aspects and parameters describing the pathogenicity and provide clues on the molecular level about the effect of mutations. It is not easy or even possible to predict the pathogenic effect of SNPs using single method. Therefore, we used multiple methods to compare and rely on the results predicted. In the present study we used nine different in silico prediction algorithms; SIFT, PROVEAN, MutPred, SNP&GO, PhD-SNP, PANTHER, I-Mutant 2.0, MUpo and DUET, To sort tolerant and intolerant SNPs, our result showed 14 out of 26 SNPs were deleterious by SIFT and PROVEN (G60R, A134G, K5N, T58I, P34R, Q61E, A146T, N116S, K147E, L19F, F141, R164Q, K5E and A59V). While 3 nsSNPs were found disease causing by SNP&GO, PhD-SNP and PANTHER (G60R, T58I and P34R) and 13 nsSNPs were found disease causing by SNP&GO, PhD-SNP (G60R, A134G, T58I, P34R, V152G, A146T, N116S, K147E, L19F, F141, R164Q , K5E and A59V). By comparing the output of the 5different insilico tools SIFT, PROVEAN, SNP&GO, PhD-SNP and PANTHER (G60R, A134G, T58I, P34R, A146T, N116S, K147E, L19F, F141, R164Q, K5E and A59V) mutations were found to be functionally significant. The differences in prediction capabilities can be attributed to the fact that every method uses different sets of sequences and alignments. For further analysis we used MutPred tool to determine the degree of tolerance for each amino acid substitution on the basis of physio-chemical properties, the result showed (G60R, A134G, K5N, T58I, P34R, Q61E, A146T, N116S, K147E, L19F, F141, R164Q , K5E and A59V) were harmful and highly harmful. Prediction of Loss of ubiquitination at; G5($P = 0.021$ for K5E, Loss of catalytic residue at A59 ($P = 0.0406$) for A59V, Loss of ubiquitination at G5($P = 0.021$) for K5N, Loss of glycosylation at T58 ($P = 0.0493$) for T58I, Gain of phosphorylation at A146 ($P = 0.0405$) for A146T,

Gain of disorder ($P = 0.0374$), Loss of stability ($P = 0.046$) for N116S, while loss of methylation at K147 ($P = 0.0184$), ubiquitination at K147 ($P = 0.0218$) and MoRF binding ($P = 0.0249$) were predicted for K147E and Loss of ubiquitination at G5 ($P = 0.021$ was predicted at K5E.This implied that some nsSNPs may account for potential structural and functional alterations of KRAS. For further confirmation, we analyzed the effect of these nsSNPs using I-Mutant 2.0 and MUpo for 14 nsSNPs which were detected deleterious by SIFT and PROVEN to determinate the stability of protein upon mutation. Above results demonstrated that protein stability was changed due to SNPs alteration. Additionally, we performed DUET analysis to determinate the effects of nsSNPs on protein stability, and the data that expressed as the change in Gibbs Free Energy (DDG) showed that 11 out of 14 nsSNPS affected the protein stability. The combination of the prediction tools results in 14nsSNPs (G60R, A134G, K5N, T58I, P34R, Q61E, A146T, N116S, K147E, L19F, F141, R164Q, K5E and A59V) indicated as the most damaging mutations in KRAS gene. Furthermore, This 14 SNPs were analyzed, structurally and functionally by 5bioinformatic tools; ELASPIC servers, ConSurf web server, mutation3D, PDB and Chimera. Interface residues play different roles in protein-protein interactions and display both different conservation patterns and different types of interactions depending on their relative position within the interface [23]. The "core" residues are defined as residues which are exposed in the monomeric protein but buried in the protein complex. Core residues are typically hydrophobic with a composition strongly divergent from the composition of the remainder of the protein surface [24]. Core residues supply the bulk of the energy driving association by hydrophobic interactions [25]. The hydrophobic interactions within the complex cause the core region to become tightly packed upon complex association with little room for conformational variability. For these reasons, the core residues are strongly conserved during evolution [26] and mutations in this region are usually more strongly unfavorable when compared to mutations at the periphery of the interface. In cancer, 3D location of mutations at an interface has served as evidence that protein interactions may

be important for metastasis site determination. So ELASPIC server applied to classify the nsSNPs according to interface and core residues in KRAS protein, the result showed 4 SNPs were in core (A134G, A146T, L19F and Q164R) and the rest SNPs were in interface. The SNP that occurs in the protein core affects all the protein interactions, and the SNP occurs at an interface only affect the interactions mediated by the interface. Our findings suggest that perturbation to cancer pathways may in fact be mutation-specific and point to the need for analysis methods for tumor-specific network topologies. Mutation3D server was used for nsSNPs location in RAS domain, the result showed 11 SNPs out of 14 located in RAS (G60R, A134G, Y71H, P34R, Q61E, A146T, N116S, K147E, L19F, F141 and A59V). Our ConSurf and Clustal Omega results indicated that 11 nsSNPs were found in the highly conserved region and predicted to have a potential impact on KRAS protein also it predicted that 116N, G60, 59A and 61Q were important structural residue (highly conserved and buried) and that 34P was an important functional residue (highly conserved and exposed). For the 14SNPs 3D structure of protein, H bond and clash was shown between wild type and mutant using the visualized chimera program. 12 nsSNPs out of 14 (Y71H, T58I, P34R, Q61E, A146T, N116S, K147E, L19F, F141, R164Q, K5E and A59V) have differing number of H-bonding between wild and mutant residues indicating that mutation will affect the protein stability, also clash was detected in mutant structure and that indicates the change in the environment of the protein which may affect structure and function. KRAS is G protein and have the Ras domain, it contains a six-stranded beta sheet, 5 alpha helices: and G domain (166 amino acids) which binds guanosine nucleotides, about 20kDa. The G domain contains five G motifs that bind GDP/GTP directly, G1- P-loop binds the beta phosphate of GDP and GTP, G2 - Also called Switch I, consists of Threonine-35 that binds the terminal phosphate (γ-phosphate) of GTP and the divalent Magnesium ion present in the active site, but makes no contacts with GDP, G3- Called the Switch II, has the DXXGQ motif, where the Aspartate-57 is specific for guanine binding rather than adenine and Glutamine-61 is the crucial residue that activates a catalytic water molecule for hydrolysis of GTP to GDP,G4 - LVGNKxDL motif, provides the specific interaction to guanine and G5 - SAK consensus sequence, the alanine-146 is specific for guanine rather than adenine. The two switch motifs, G2 and G3 are the main parts of the protein that move upon activation by GTP. This change in the conformation of the two switch motifs is what mediates the basic functionality as a molecular switch protein. This GTP bound state of Ras is On-state and GDP bound state is the Off-state. Ras also binds a magnesium ion which helps to coordinate nucleotide binding [27]. Out of 14 deleterious SNPs we found 6 nsSNPs are located in these motifs and it has big roles in altering protein functions.

Rs104886029 A59V: Mutant residue is smaller than the wild-type residue. This will cause a possible loss of external interactions. In the 3D-structure the mutation is located in a loop in G1 motif. Single H-bonding in wild type and mutant. One clash with THE in 35 in G2 motif, so this SNP can affect the protein functions. The SNP is located in highly conserved residue which made it highly damaging for the protein. The mutant-type residue was buried in the interface with GNDS isoform 2, RASGRP2, GRP2 isoform 2, RGL2, RASGRP3, GRP3 isoform 2, RASGRF2 and Q6PCE1 proteins indicating that the mutation will cause loss of catalytic residues P= 0.0406.

Rs121913238 Q61E: the mutant residue is smaller than the wild-type, in addition to that the wild type residue is neutral, and in comparison with the negatively charged mutantant residue this can cause repulsion between the mutated residue and neighboring residues. In the 3D-structure the SNP is located in a loop in G3 motif. Single H-bond in wild type and no H-bonding in mutant that may affect the stability of the protein. The SNP is highly conserved and harmful. This indicates that the mutation is damaging to the protein. Also, the variant is annotated with disease, damaging and increase the stability of protein. The mutant-type residue was buried in the interface with GNDS isoform 2, RASGRP2, GRP2 isoform 2, RGL2 and RASGRP3 proteins. The mutation will cause Gain to disorder (P = 0.0755), Gain of helix (P = 0.132), Loss of sheet (P = 0.1398) Gain of phosphorylation at T58 (P = 0.1442). Loss of glycosylation at T58 (P = 0.1859).

Rs104894359 G60R: the mutated residue is smaller than the wild-type, in addition, the wild type residue was neutral, in comparison with the positively charged of mutant this can cause repulsion between the mutated residue and neighboring residues. In the 3D structure the SNP is located in a loop in G3 motif. The wild-type residue is a glycine, the most flexible of all residues. This flexibility might be necessary for the protein's function. Mutation of this glycine can abolish this function. No H-bonding and clash between the wide and mutant residues. The SNP is highly conserved and high harmful indicating that the mutation is damaging to the protein. The variant is annotated with disease, damaging and decrease the stability of protein. The mutant-type residue was buried in the interface with GNDS isoform2,RASGRP2, GRP2 isoform2, RGL2, RASGRP3, RASGRF2 and Q6PCE1. Also mutation will cause Gain of phosphorylation at T58 (P = 0.1068), Loss of glycosylation at T58 (P = 0.123), Gain of solvent accessibility (P = 0.1319), Loss of catalytic residue at A59 (P = 0.1348) and Gain of loop (P = 0.2045).

Rs202247812 N116S: The mutant residue is smaller than the wild-type residue. In the 3D-structure the SNP is located in extended beta-strand in G4 motif. 5 H-bonding interaction in wild type and 3 -bonding interaction in mutant. The SNP is highly conserved and high harmful. This indicates that the mutation is damaging to the protein. The variant is annotated with disease, damaging and decrease the stability of protein. The mutant-type residue was a structural site in the interface with RASGRP3 proteins. The mutation will cause Gain of disorder (P = 0.0374), Loss of stability (P = 0.046).

Rs121913527 A146T: The mutant residue is bigger than the wild-type residue. The wild-type residue was more

hydrophobic than the mutated residue. The hydrophobicity and size difference between wild-type and mutant residue can disrupt the H-bonding interactions with the neighboring residues, and hence the protein framework. In the 3D-structure the SNP is located in a loop in G5 motif.3 H-bonding interaction in wild type and 2 in mutant, 5 clashes between mutant and CDP 201logan, LEU19 and ALA18 The SNP is highly conserved and high harmful. This indicates that the mutation is damaging to the protein. The variant is annotated with disease, damaging and decrease the stability of protein. The mutant-type residue was structural in the core. The mutation will cause Gain of phosphorylation at A146 (P = 0.0405).

Rs387907206 K147E: The mutant residue is positively charged and wild-type is negative. In the 3D-structure the SNP is located in a loop in G5 motif. 9 clash between mutant and CDP 201 Logan and PHE28. The SNP is highly conserved and high harmful, which indicates that the mutation is damaging to the protein. The variant is annotated with disease, damaging and decrease the stability of protein. The mutant-type residue was structured in the interface with RASGRP3 protein. The mutation will cause Loss of methylation at K147 (P = 0.0184), Loss of ubiquitination at K147 (P = 0.0218) and Loss of MoRF binding (P = 0.0249).

5. Conclusions

The available data from the NCBI dbSNP database for KRAS gene has been analyzed through several SNP analyzing tools and the predicted deleterious SNPs were evaluated for their deleterious effect on the protein function and stability. 14 SNPs were predicted deleterious; those are (G60R, A134G, K5N, T58I, P34R, Q61E, A146T, N116S, K147E, L19F, F141, R164Q, K5E and A59V). Also, we found these mutations cause a change in protein affinity that can result in malfunction of the protein interaction network. Among them, 6 SNPs were predicted to be located in the motif regions G1, G2, G3, G4 and G5 (G60R, Q61E, A146T, N116S, K147E and A59V) and have the most probability to increase cancer susceptibility.

REFERENCES

[1]   A. Chakravarti, "To a future of genetic medicine," Nature, vol. 409, pp.822–823, 2001.

[2]   Carninci. P, Kasukawa. T, Katayama. S et al., 2005. "The transcriptionall and scape of the mammalian genome," Science. 309:1559–1563

[3]   Liu. J, Gough. J, and Rost. B. 2006. "Distinguishing protein-coding from non-coding RNAs through support vector machines," PLoS Genetics.2:4-29.

[4]   Ng .P. C. and Henikoff. S. 2006. "Predicting the effects of amino acid substitutions on protein function," Annual Review of Genomics and Human Genetics. 7:61–80.

[5]   Smith .E.P, Boyd. J, Frank. G.R et al., 1994. "Estrogen resistance caused by a mutation in the estrogen-receptor gene in a man," The New England Journal of medicine. 331. pp. 1056– 1061.

[6]   Lander E. S. 1996. "The new genomics: global views of biology," Science.274: 536–539.

[7]   Thomas. R, R. McConnell, J. Whittacker, P. Kirkpatrick, Bradley. J, and. Sand ford. R. 1999. "Identification of mutations in the repeated part of the autosomal dominant polycystic kidney disease type 1 gene, PKD1, by long-range PCR," The American Journal of Human Genetics 65:39–49.

[8]   Doniger S.W, Kim H. S., Swaine D. etal. 2008. "Acatalog of neutral and deleterious polymorphism in yeast. PLoS Genetics.vol. 4, no.8, ArticleIDe1000183.

[9]   BILLUR H. E, HOFREE. M, CARTER.H .2015. Identifying Mutation Specific Cancer Pathways Using A Structurally Resolved Protein Interaction Network. Pacific Symposium on Biocomputing. 20: 84-95

[10]   Jeffrey R. B and Yang Z. 2015. Predicting the Effect of Mutations on Protein Protein Binding Interactions through Structure-Based Interface Profiles. PLOS Computational Biology. DOI:10.1371. 1-25

[11]   https://ghr.nlm.nih.gov/gene/KRAS.accssed in february2016.

[12]   Gregory J. Riely, Mark G. Kris, Daniel Rosenbaum, Jenifer Marks, Allan Li, Dhananjay A. Chitale, Khedoudja Nafa, Elyn R. Riedel, Meier Hsu, William Pao, Vincent A. Miller, and Marc Ladanyi. 2008. Frequency and Distinctive Spectrum of KRAS Mutations in Never Smokers with Lung Adenocarcinoma. Clin Cancer Res.14: 5731–5734.

[13]   http://www.biotecharticles.com/Bioinformatics-Article/Importance-and-Applications-of-Bioinformatics-in-Molecular-Medicine-214.html.

[14]   Adzhubei. IA, Schmidt .S, Peshkin. L, Ramensky. VE, Gerasimova. A, et al. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249.

[15]   Ng P. C., Henikoff. S. 2006. Predicting the Effects of Amino Acid Substitutions on Protein Function. Annu Rev Genomics Hum Genet. 7: 61-80.

[16]   Narayana S. A, Harika. V and Sreenivasulu. K. 2015. In silico Evaluation of Nonsynonymous Single Nucleotide Polymorphisms in the ADIPOQ Gene Associated with Diabetes, Obesity, and Inflammation. Avicenna J Med Biotech 7: 121-127.

[17]   Calabrese R., Capriotti. E, Fariselli. P, Martelli. P. L, Casadio. R. 2009. Functional Annotations Improve the Predictive Score of Human Disease-related Mutations in Proteins. Human Mutation. 30: 1237-1244.

[18]   Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 25: 2744–2750.

[19]   Capriotti E., P. Fariselli, R. Casadio. 2005. I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure. Nucleic Acids Res.33: 306-310.

[20]   Douglas E.V. Pires, David B. Ascher and Tom L. Blundell.

2014. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Research. 42: 314–319.

[21] Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. 38: 529–33.

[22] Sofia B. M, Mohamed M. H. 2016. Insilico Validation of Babesia Bovis Merozoite Surface Antigen-1, Merozoite Surface Antigen-2b and Merozoite Surface Antigen-2c Proteins for Vaccine and Drug Development. International Journal of Bioinformatics and Biomedical Engineering. 2: 30-39.

[23] Levy ED. 2010. Asimple definition of structural regions in proteins and it s use in analyzing interface evolution. Journal of molecular biology. 403: 660–70. doi:10.1016/j.jmb.09.02 8PMID:20868694.

[24] Talavera D, Robertson DL, Lovell SC. 2011. Characterization of protein-protein interaction interfaces from a single species. PloSone. 6: e21053. doi:10.1371 /journal.pone 0021053 PMID :21738603.

[25] Andreani J, Faure G, Guerois R. Versatility and in variance in the evolution of homologous heteromeric interfaces.2012. PLoS computational biology. 8: e1002677. doi:10.1371/journal.pcbi.1002677 PMID: 22952442.

[26] Guharoy. M, Chakrabarti. P. 2005. Conservation and relative importance of residues a cros protein-protein interfaces. Proceedings of the National Academy of Sciences of the United States of America. 102: 15447–52.PMID:16221766.

[27] Omim- Neuroblastoma Ras Viral Oncogene Homolog; Nras.