Sequence and structure based analysis of K-RAS gene to study its regulatory role in Non-small cell lung cancer

Submitted in partial fulfilment of the requirement for the degree of

Masters of Science

In

Biotechnology

By

Nonita Sood

217815

Under the guidance of

Dr. Tiratha Raj Singh

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

HIMACHAL PRADESH-173234

Certificate of Originality

This is to certify that the thesis titled **Sequence and structure based analysis of K-RAS gene to study its regulatory role in Non-small cell lung cancer** is an original work of the student and is being submitted in partial fulfillment for the award of the Degree of **Master of Science (Biotechnology)**. This dissertation thesis has not been submitted earlier either to this University or to any other University/ Institution for the fulfillment of the requirement of any course of study.

(Signature of Supervisor) (**Signature of Student)**

Dr. Tiratha Raj Singh

Professor

Dept. of Biotechnology and Bioinformatics

JUIT, Waknaghat, H.P.

173234

Nonita Sood

DECLARATION

I, Nonita Sood, present the project entitled “Sequence and Structural based analysis of K-RAS gene to study its regulatory role in Non-Small Cell Lung Cancer”. Though care has been taken while writing this report, there may be still some errors (typographical or otherwise) which are inadvertent on my part.

Place: **Jaypee University of Information Technology, Waknaghat, Distt.- Solan, H.P.**

Signature of Candidate

Nonita Sood

CERTIFICATE OF UNIVERSITY FACULTY GUIDE

Enrollment No. - 217815

This is to certify that Ms. Nonita Sood of M.Sc (Biotechnology) has completed this Research/ Dissertation Project under my supervision in partial fulfillment for the award of Master of Science Degree in Biotechnology from Jaypee University of Information Technology, Waknaghat, Distt.- Solan, Himachal Pradesh.

Course Name: DISSERTATION

Course Code: 20MS9BT491

SIGN OF FACULTY GUIDE	**SIGN OF STUDENT**
Dr. Tiratha Raj Singh	**Nonita Sood**

PLACE: Jaypee University of Information Technology, Waknaghat, Solan, H.P.

DATE:

ACKNOWLEDGEMENT

I, Nonita Sood would like to express my profound gratitude to my guide **Dr. Tiratha Raj Singh** for his guidance, support and constant encouragement throughout the course of this project work. He has been more than just my project guide; at times a mentor to rescue me out of my doubts. He has always helped me to work hard and also taught me how to implement different ideas to deal with the problem. Moreover, he taught me to not give up and many other valuable lessons.

Furthermore, I would like to acknowledge Vice-Chancellor **Prof. (Dr.) Rajendra Kumar Sharma**, **Prof. (Dr.) Ashok Kumar Gupta**, Dean of academics & research for providing me with an opportunity to be a part of the institute and to complete my Master's Degree.

I also want to mention the HOD of Biotechnology and Bioinformatics **Prof. (Dr.) Sudhir Syal** has been a source of immense motivation and inspiration both for my academic and personal life. He was never, and I know will never be, more than just a phone call away. He has helped me in almost every aspect I have asked him for.

In addition, I would like to thank all the faculty members of the BT/BI Department of JUIT, who have helped me whenever I needed and also would like to thank all the lab engineers and specially Ms. Somlata Sharma ma'am for providing me with a workplace and for always motivating me.

I would also like to appreciate the part that my classmates (Ankita and Gargi) have played in shaping this project work. They have been my constant support and cheered me up at hard times. They helped me whenever I had any doubts. Thanks a lot!

A big thanks to all open-source tool providers because of which I was able to complete my project.

I would like to thank the almighty God for his grace throughout my life. Last but not the least I would like to thank my Mother and Father who have always supported me through thick and thin and have been a constant source of encouragement and support; also, my elder sister who has never given up on me and always motivated me.

[Thanks to JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY]

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

NSCLC- Non Small Cell Lung Cancer

KRAS- Kristen ras oncogene

MAPK- Mitogen-activated protein kinase

NCBI- National Centre of Biotechnology Information

SNP- Single nucleotide polymorphism

ADC- Adenocarcinoma

SCC- Squamous Cell Lung Cancer

LCLC- Large Cell Lung Carcinoma

EGFR- Epidermal Growth Factor Receptor

ALK- Anaplastic lymphoma kinase

GDP- Guanosine diphosphate

GTP- Guanosine triphosphate

GEF- Guanine nucleotide exchange factors

INDEX

Abstract

KRAS is the particularly prevalent oncogene in non-small cell lung cancer (NSCLC).Single nucleotide polymorphism (SNP) can add, take away, or alter protein coding sites and can lead to complicated illnesses. KRAS stands for Kirsten Ras that provides instructions for the production of K-Ras protein which plays an important part in RAS/MAPK signaling pathway. This study entailed an in-depth investigation on human KRAS nsSNPs suspected of illness. From NCBI's dbSNP, the nsSNPs were obtained, and numerous sequence and structure based analysis were carried out to screen out the most affecting SNPs. Tools like Poly-Phen-2, Meta-SNP, PMut,SNAP2, PhD-SNP, SNP&GO, I-Mutant2.0, Align GVGD were used for the analysis. Then phenotypic consequences were analyzed for these SNPs. Finally three extremely deleterious nsSNPs were predicted which are G60R, T58I and V152G. These mutations were found in highly conserved regions according to the functional and evolutionary conservation analysis performed using ConSurf. The structural effects caused by these mutations were also analyzed using HOPE server. The sorting of genes and SNPs will be expected to reap rewards from this work and it is believed that the information generated at genomic and proteomic level will help the experimental scientists to design efficient experiments for further validations.

Chapter-1

Introduction

Majority of the cancer related deaths occur from lung cancer at global level. The main variable which triggers lung cancer is smoking whereas other causes includes radon i.e., secondhand exposure to secondhand smoke or substances like nickel, chromium, soot, tar; genetics (higher risk of lung cancer has been attributed to unique chromosomal regions); DNA damage is the basic cause of cancer, as we know most of the damages can be repaired but there are some remaining damages caused due to smoking which cause NSCLC. Approximately 85% of lung cancer deaths were triggered by non-small cell lung cancer. (NSCLS)[1], which includes all epithelial lung tumors. There are other NSCLC subtypes (including pleomorphic, carcinoid tumor, salivary gland carcinoma) however the following subtypes are the most prevalent:

- Squamous-cell carcinoma
- Large-cell carcinoma
- Adenocarcinoma

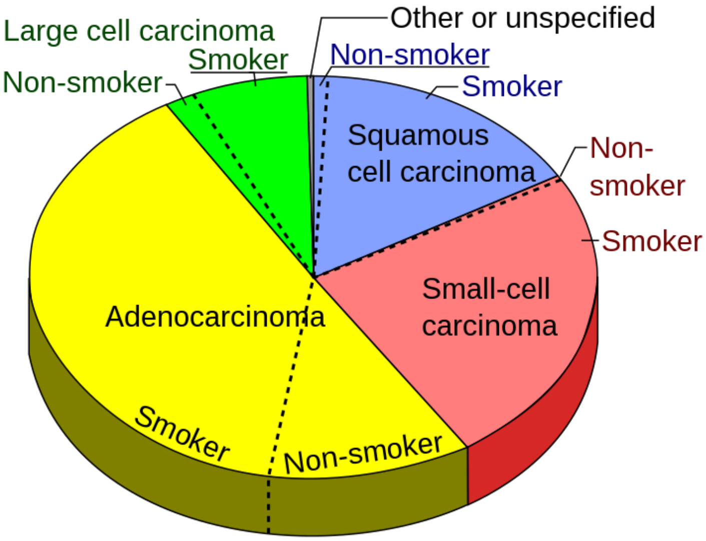


Figure -1. Types of NSCLC are depicted in a pie graph. (Stefanov T., 2019)

They can occur as discrete histomorphic phenotypes or as mixed cell type combinations.

Adenocarcinoma is the most prevalent kind of lung cancer currently found among never smokers or lifelong non-smokers is adenocarcinoma of the lung which accounts for about 40% of lung malignancies [2]. ADCs generally develop in more distant airways than SCCs. Men are more likely than women to acquire **squamous cell lung cancer**. It has strong correlation with history of tobacco use. Acc. to the Nurses' Health Study, when compared to a non-smoker, an individual who had smoked for 20-30 years in the past had a relative risk of SCC of about 5.5 whereas of 30-40years smoking history is about 16.SCCs closely correlating with smoking and ongoing inflammation and are manifested in more proximal airways. The expression "large cell lung carcinoma" indicates an array of undifferentiated malignant tumors that originate from altered epithelial cells. Historically, ten percent of NSCLCs served as LCLCs.[3]. If the tumor cells do not have a glandular or squamous structure, the large cell carcinoma can be excluded. EGFR, KRAS, MET, PIK30A, ALK, BRAF are some genes involved in NSCLC.

In the above mentioned genes the top two genes are:

- EGFR i.e., with 10-35% occurrence
- KRAS i.e., with 25% occurrence

EGFR

Critical growth factor signaling circulates across the extracellular environment to the cell by the transmembrane protein identified as the epidermal growth factor receptor (EGFR) that additionally displays cytoplasmic kinase activity. Since, over sixty percent of NSCLCs exhibit EGFR recognition, making it a crucial therapeutic target for the curative use of these cancerous tumors.[11]

KRAS

The protein conveys impulses sent to the cell's nucleus from outside its walls. These signals convey instructions to the cell on how to develop and replicate (proliferate) or age and take on specified obligations (differentiate).

The identification of KRAS and BRAF mutations, Patients with lung ADC were all learned to have EGFR mutations, and these alterations were related to how smoothly the patients replied to EGFR inhibitors. K-Ras, a protein which assists in cell growth and division, is synthesized in a significant way by the KRAS gene. KRAS gene is often appearing alongside with STK11 mutation. The H-ras, K-ras, and N-ras genes; the three variables constitute the three ras genes; soon after one of their respective codons (12, 13, or 61) undergoes modification, they emerge to functional oncogenes .Quick tests for identifying the presence of the point alterations are being used to explore the contribution of mutant Ras genes in the beginning of human tumors. Although the incidence varies greatly, it showed that an assortment of tumor types might display ras gene adjustments. The most frequent adenocarcinomas consist of those of the pancreas (90%), colon (50%), and lung (30%), thyroid tumors (50%), and myeloid leukemia (30%).[12]A K-ras mutation in codon 12 of the gene is found in about one-third of lung adenocarcinoma.[13]

Objectives

Major objectives of this research work were designed and categorized as follows:

- To study the interactions of KRAS gene and protein.
- To identify regulatory targets for KRAS gene.
- To identify SNPs for KRAS gene.
- Sequence and structure level analysis of non-synonymous SNPs and their functional and evolutionary analysis.

CHAPTER-2

Literature Review

KRAS Gene

Ras proteins have resisted all attempts at therapeutic intervention, and they have actually been labeled as "undruggable" for a long time. [4] The RAS gene basically encodes G protein that has lower molecular weight of about 21 Kd and size of 189 amino acids, containing activity for guanosine triphosphatase. [5] This acts as a molecular switch for signal transduction, helping to control cell development and differentiation. The activation and deactivation of RAS involves GDP\GTP exchange and hydrolysis of GTP, involving other regulatory proteins like GAP and GEFs. Two proteins control the activation of RAS: GTPase-activating proteins (GAP) speed up GTP hydrolysis; GEF elicit the exchange of GDP for GTP and RAS, correspondingly. [14]. Hence RAS proteins control different cellular functions by activating different effectors. The small G protein binds to GTP and thus prevents GAP from increasing and slowing the reformation of GTP to GDP, resulting in boosting the active state of KRAS connecting to GTP. As a result unlimited cell growth occurs, thus inducing tumourigenesis. Codon 12 is where maybe 80% of Genetic alterations occur; three mutations which are common- KRAS G12C, KRAS G12D and KRAS G12V. KRAS G12C is a glycine to cysteine conversion; KRAS G12D is a glycine to aspartic acid change whereas KRAS G12V is a mutation of glycine to valine. Patients suffering from the median OS and two-year survivorship are both worse during NSCLC with KRAS mutations. PI3K and MEK transmission can be induced via KRAS G12D mutations(6). A distinct downstream signaling may be impacted by different types of point mutations. Patients with G12C mutation suffer from bone metastases spread and patients with G12V mutation experience pleura-pericardial metastases (10). Chemotherapy has caused a slightest boost in the patients' overall survival with advanced

NSCLC. Except for the GDP/GTP-binding site, the RAS proteins did not appear to contain adequate pockets for drug binding. [4]A number of biological functions are triggered and inhibited by the KRAS gene. In addition to managing signal transduction pathways, it also controls tumor growth, differentiation, and death. The management of transcription factors and the management of gene expression all involve the KRAS gene. There seem to be four exons in the KRAS gene. The KRAS protein, a small GTPase is transcribed by the first three exons. The Ras-related protein which is similarly a small GTPase is encoded by the fourth exon. The KRAS protein is split into two domains: the C-terminal domain which is in charge of the protein's GTPase activity, and the N-terminal domain that controls charge of the protein's GTP binding. Basic functions of KRAS gene include regulation of cell proliferation. Ras proteins possess intrinsic GTPase activity and basically bind to GTP\GDP. Mutations in KRAS have a negative prognostic factor when compared to KRAS wild type tumors. [5,7,8] Due to the variety of the studies, its significance is still debatable. KRAS has a significant role in both pancreatic and lung adenocarcinoma. Because KRAS appears to be the tumors' initiating event, this seems like the most likely explanation. For instance, the frequent G12C mutation is a telltale sign of exposure to tobacco smoke in lung cancer[4]. A meta-analysis of twenty eight studies that make up three thousand six hundred and twenty patients revealed the negative fate of KRAS in lung adenocarcinoma, in contrast to in squamous-cell carcinoma. Lung cancer may occur owing to mutations in the KRAS gene that engage oncogenic pathways. Researchers from the University of Michigan confirmed in a similar research that colorectal cancer can also arise in response to alterations in the KRAS gene. The most frequent KRAS co-mutational partners encountered in NSCLC are TP53 (40%), CDKN2A (19.8%) and STK11\LKB1 (32%).[9] KRAS mutations are typically discovered in regions where GAPs bind to RAS.[4]

Lung cancer may occur via mutations in the KRAS gene that engage oncogenic pathways. Researchers from the University of Michigan confirmed in a similar research that colorectal cancer can also arise through mutations in the KRAS gene.It is anticipated that drugs that inhibit all four Ras isoform- HRAS, NRAS, KRAS4A, and KRAS4B—will be excessively hazardous [4].

Table: 1. Information on KRAS gene.

Total SNPs	20081
Missense SNPs	432
Chromosome No.	12
Molecular Weight	21656 Da
Length	189 aa protein
Gene ID	3845
UniProt ID	P01116
NCBI	dbSNP

Activation of KRAS

Human malignancies typically have mutations in the RAS oncogenes, which trigger the RAS-RAF-MEK-ERK pathway and facilitate cell survival as well as proliferation.[4]Multiple downstream signaling pathways essential in cell growth, differentiation, and survival can be triggered by mutant RAS. [4]

Upon activation of cell-membrane receptors, the adaptable proteins Grb2 employs GEFs to the cell membrane whereby RAS proteins can be found via prenylation, augmenting the variety of active GTP-RAS proteins. The serine/threonine kinase RAF (A-RAF, B-RAF, and C-RAF/RAF-1) is then activated through dimerization as a result of the interaction between the activated RAS and RAF (A-RAF, B-RAF, and C-RAF/RAF-1). Afterwards the direct stimulation and activation of MEK1 and MEK2, tyrosine and serine/threonine dual-specificity kinases, and these activates nuclear and cytoplasmic targets that govern transcription, cell proliferation, differentiation, along with metabolism, ERK1 and ERK2 molecules are then activated.

Future prospects of NSCLC

Notwithstanding the simple fact that RAS mutations have been documented for more than 30 years, effective treatments centered on the mutant RAS protein haven't yet been identified. [4]The concurrent restriction of numerous signaling circuits and induced lethality could represent the secret to breaking out of this therapeutic impasse, but they must first go through an exhaustive examination. This cautious faith ought to be seen not as the destination of the journey, instead fostering optimism for the future. The logical establishment of highly targeted medications and their combinations utilized on effectively selected individuals is a promising strategy that could lead to substantial enhancements in clinical outcomes. [17]

In comparison to individuals with EGFR-mutant NSCLC, patients suffering from KRAS-mutant NSCLC have a shorter median survival time. [15,16]

CHAPTER: 3

METHODOLOGIES

3.1 DATA COLLECTION

The information about K-RAS gene with **Gene ID: 3845** was obtained from different databases like NCBI, UniProt, PDB (Protein Data Bank). The missense nsSNPs were 432; obtained from NCBI database (**https://www.ncbi.nlm.nih.gov**). The UniProt database (**https://www.uniprot.org**) was used to obtain human KRAS sequence **[UniProt ID: P01116]**. The KRAS protein structure has been acquired from the Protein Data Bank **(PDB ID: 4EPV).** The Accession No. was **AAB41942**.

The length of protein sequence was **188 a.a.** and the nucleotide sequence for the gene is **45684 bp**.

3.2 INTERACTION ANALYSIS OF KRAS PROTEIN

Cellular life depends on a complex web of interdependent bimolecular connections. These linkages are particularly important for protein-protein interactions because of their flexibility, specificity, and variety.[18]Using the STRING (Search Tool for the Retrieval of Interacting Genes/Protein)database(**http://string-db.org/)**the KRAS protein interaction investigation was carried out. STRING is helpful to understand protein-protein interactions which are helpful to study structural, functional and evolutionary properties of protein. The input used was the name of the protein i.e. **KRAS** protein. Further organism was selected which in this case is ***Homo sapiens*** and the results were obtained. STRING also displays a network statistics table comprising of average node degree, average local clustering coefficient, PPI enrichment score etc.; for the search made. STRING provides a comprehensive coverage as the upcoming version 11.5 of the library consists of more than 14000 creatures.[18]

3.3 IDENTIFICATION OF REGULATORY TARGETS FOR KRAS GENE

(http://iregulon.aertslab.org)

Identification of regulatory targets was done using I-Regulon which is a plugin of Cytoscape. A transcription factor (TF) and its specific transcriptional targets make up a Regulon. To unravel the transcriptional regulatory network underpinning a group of associated genes, it leverages cis-regulatory sequence analysis. I-Regulon employs a genome-wide ranking-and-recovery methodology to locate abundant transcription factor motifs and their best combinations of direct targets.[42] With the help of i Regulon plugin, we were able to locate the regulons using motifs while tracking their discovery in a network that previously exists or in a collection of co-regulated genes. I-Regulon explores the regulatory areas enclosing each gene in the gene set with the objective of recognizing enriched TF motifs or ChIP-seq peaks via datasets encompassing virtually 10.000 TF motifs as well as 1000 ChIP-seq datasets or "tracks".[42]The most efficient combinations of direct targets and motif-specific transcription factors have been determined via GLI1.

3.4 DETECTION OF NETWORK MOTIFS

RAND-ESU, a method that Wernicke established[44], offers an important boost over mfinder.[45] This algorithm, that relies on the extremely accurate enumeration algorithm ESU, has been resorted within an application called FANMOD.[44]FANMOD tool was used for the identification of network motifs. Network motifs involving up to eight vertices can be recognized by FANMOD[19]. It is a useful concept to understand and study the structural design principles of complex networks.[43]A graphical user interface, the ability to evaluate colored networks, as well as the capacity for transfer to an assortment of accessible to humans and machine-accessible file formats, notably values between commas and HTML , is further advantages of FANMOD[19].It uses Z-score to detect motifs that are substantially more common in the network than those that were initially identified in 1000 random networks. The input used was obtained from KEGG pathway. Input consists of gene IDs and the relation type (represented by 1 for association,-1 for dissociation/ inhibition or 0 for missing interactions).Network motifs involving till the size of eight vertices can be detected by FANMOD.[19]

3.5 DELETERIOUS nsSNPs IDENTIFICATION USING SEQUENCE BASED PREDICTION TOOLS

Five servers and tools that use sequences were employed to forecast whether the nsSNPs on the protein might impact the protein's functionality. The approaches used were Meta-SNP, Predict SNP. The disease pathogenicity related to the nsSNPs was predicted using SNAP 2, SIFT, SNPs&GO, PhD-SNP and PolyPhen-2.

3.5.1 Meta-SNP

(**http://snps.biofold.org/meta-snp**)

The most widespread category of genetic variation detected in human DNA has the designation as SNP. The collection of SNPs seen in each of a diploid organism's two copies of a certain chromosome is known as a haplotype.[20] Haplotype DNA has several applications, including drug development and genetic disease detection.[22]Meta-SNP helps to predict whether the given single protein variation is classified as disease associated or not. Meta-SNP consists of other approaches that are PANTHER, PhD-SNP, SIFT, SNAP. If the value is reported to >0.5, the nsSNP is diseased whereas if the value is<0.5, the given nsSNP is neutral. Hence a missense mutation's severity can be anticipated by Meta-SNP.

3.5.2 Predict SNP

(**https://loschmidt.chemi.muni.cz/predictsnp1**)

Predict SNP consists of eight computational prediction tools which are PANTHER, MAPP, nsSNP, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT and SNAP. Predict SNP helps in predicting the effects of altering an amino acid. The value of the mutations between 0, 1 are considered to be deleterious. More the gap between the score and zero, more harmful are the mutations. [21]

3.5.3 SNAP-2

(**https://github.com/Rostlab/SNAP2**)

SNAP-2 stands for “Screening for Non-Acceptable Polymorphisms”. It is a neural network-derived tool that accurately predicts the functional effects of nsSNPs. Only sequence information needs to be supplied for SNAP's input [22].A thoroughly standardized measure for the preciseness of each prediction was introduced by SNAP[22] .SNAP-2 is based on Hidden Markov Models (HMM).Score calculations are made around -100 and +100 where -100 being strong neutrality and +100 being diseased condition [23]. It also demonstrates the prediction results in the form of heatmap. [24]

3.5.4 SIFT

(**http://sift-dna.org**)

SIFT stands for "Sorting Intolerant From Tolerant". It helps to identify whether to swap an amino acid has an influence on how proteins function so that users can determine which variations to focus their research on [25].Changes at sites that are preserved well are typically projected to be damaging by SIFT because it anticipates that crucial amino acids would be conserved across the protein family [25].SIFT is capable of reliably recognizing items even when they have been partially obscured by clutter and other objects. [26,27,28]In order to pick local scale invariant reference frames, local scale selection is used in the SIFT descriptor, which is based on receptive field measurements of pictures.[28]Sequence homology is employed by the SIFT server to analyze the impact of amino acid replenishment and to recognize both advantageous and detrimental SNPs. A substitution of an amino acid at a certain position with a likelihood of less than 0.05 is deemed harmful and intolerable, while a chance of more than 0.05 is anticipated to be tolerable.[47] The protein composition and the intended replacement constitute the data inputs for SIFT.

3.5.5 SNPs&GO

(**http://snps-and-go.biocomp.unibo.it/snps-and-go/index.html**)

SNPs&GO helps by utilizing the protein functional annotation that helps to determine whether or not a mutation is linked to a disease. SNPs&GO has a score of efficacy up to 82% and over a large pool of proteins with specified non-synonymous mutations. [25]SNPs&GO beats other known predictive techniques by accumulating in particular framework information gathered through protein sequence, evolutionary information, and function as recorded in the Gene Ontology terms[29].SNP&GO predicts the insurgence of disease in humans by identifying the mutations with the help of SVM classifier. The server is based on

Support Vector Machines (SVM).[30] The server's output lists the likelihood that each protein variant will be linked to human diseases.[31]

3.5.6 PhD-SNP

(http://snps.biofold.org/phd-snp/)

In order to forecast the implications of human SNVs throughout both coding along with the non-coding areas, PhD-SNP leverages sequence-based characteristics.[29]PhD-SNP is a support vector machine (SVM). The protein sequence and the residue alterations were entered into the PhD-SNP for review. An intuitive interface has been offered by the PhD-SNP web service to forecast the effects of SNVs, which in coded and without coding domains. [32] Protein alignments in FASTA format, as well as wild-type and mutant residues at stipulated amino acid positions, were used as input.

3.5.7 PolyPhen-2

http://genetics.bwh.harvard.edu/pph2/

PolyPhen-2 aids in anticipating how an amino acid transformation could influence a protein's biological function. Sequence number, phylogeny, and structural traits that define the replacement are used to identify potentially harmful SNPs.[33]

3.6 DELETERIOUS nsSNPs IDENTIFICATION USING STRUCTURAL BASED PREDICTION TOOLS

The variation in protein stability generated by the detrimental nsSNPs has been evaluated using seven sequence-based forecasting approaches. I-Mutant2.0, Align GVGD, CUPSAT, and MUpro were the tools employed for evaluating whether nsSNPs altered protein stability and dynamics.

3.6.1 I-Mutant 2.0

http://folding.biofold.org/i-mutant/i-mutant2.0.html

Protein structure and functional activity depends on protein stability.[46] I-Mutant 2.0 is used to anticipate protein stability in response to single nucleotide change. Torsion angle and atom potentials that are particular to a given structural environment are utilized by ΔΔG to foresee the variation in the spontaneous release of energy amongst mutant and wild type proteins. The Pro-Therm derived dataset, the largest repository of experimental information on protein mutations, is used in this programme.[48] The input consists of protein sequence or structure along with its position and new residue. Depending on whether structural or sequence information is utilized, I-Mutant2.0 adequately predicts eighty percent or seventy seven percent of the data set. [34]

3.6.2 Align GVGD

http://agvgd.hci.utah.edu/

It uses protein multiple sequence alignments and the physical and chemical attributes of amino acids to determine where missense mutations in important genes would lie on a spectrum from enriched deleterious to enhance neutral. The physiological properties of the recognized amino acids at each site are assessed by Align-GVGD via an MSA (composition, polarity, and volume). Input consists of FASTA format of the protein sequence and the substitution or mutation list.

3.6.3 CUPSAT

http://cupsat.tu-bs.de

It is a method to anticipate how point mutations might influence protein stability. The prediction model examines the mutation site's conditions in terms of amino acids utilizing torsion angle distribution and amino acid-atom potentials. It consists of two modules for the prediction of mutant stability: 1) From existing PDB structures, 2) From custom protein structures. Torsion angle perturbation in protein mutants has been accommodated using the Gaussian apodization function. In addition to information with respect to the alteration site and it’s structural features (solvent accessibility, secondary structure, and torsion angles), the outcome provides broad details regarding the modifications to protein stability for nineteen distinct hypothetical alterations of a single amino acid. As well as it examines how well mutant amino acids can adjust to the torsion angles that are being recorded.[35]

3.6.4 MUpro

http://mupro.proteomics.ics.uci.edu/

A group of machine-learning programmes called MUpro can forecast how a single-site amino acid change would affecting the stability of a protein. Input consists of alteration position, original amino acid, substituting amino acid and sequence or structure file. The confidence score for the prediction is between -1 and1. A score of less than 0 demonstrates that the mutation renders the protein less stable. The outlook is more certain when the score is lower. A score greater than 0, implies that the mutation promotes stability of protein. The prediction is more certain the higher the score. [36]

3.7 CONSERVED REGIONS IDENTIFICATION

The evolutionarily conserved components of the KRAS gene have been identified employing the ConSurf server. Using phylogenetic connections across comparable sequences, ConSurf assesses evolutionary conservation of amino or nucleus locations across an amino acid, or DNA or RNA molecule. One benefit of ConSurf over alternative approaches is the precise calculation of the evolutionary rate using either an empirical Bayesian method. Input consists of PDB ID and identified chain. The lowest score indicates the most conserved position in a protein.

3.8 ANALYSIS OF PHENOTYPIC CONSEQUENCES OF THE SCREENED MUTATIONS

FATHMM tool was used for the analysis of phenotypic consequences of 5 mutations. The forecasting method employed was the weighted algorithm under the “Inherited Disease” section, whereas the chosen morphological correlations were “Human Phenotype Ontology”. Input consists of PDB ID and substitution. To make interpretation easier and to concentrate analysis on a selection of high-confidence predictions, FATHMM assigns each prediction a confidence value (a p-score). [37]

3.9 STRUCTURAL EFFECTS OF DELETERIOUS nsSNPs

HOPE server was used to visualize and analyze the mutations. Have Your Protein Explained server embodies what HOPE stands for HOPE elucidates the molecular basis of a trait associated with a disease produced by changes in human proteins. [38].HOPE receives data drawn from a variety of sources of information, such as projections from DAS services, classifications retrieved from the UniProt repository of the protein's significantly sequence, and estimations on the protein's 3D dimensions from WHAT IF Web services.[49]

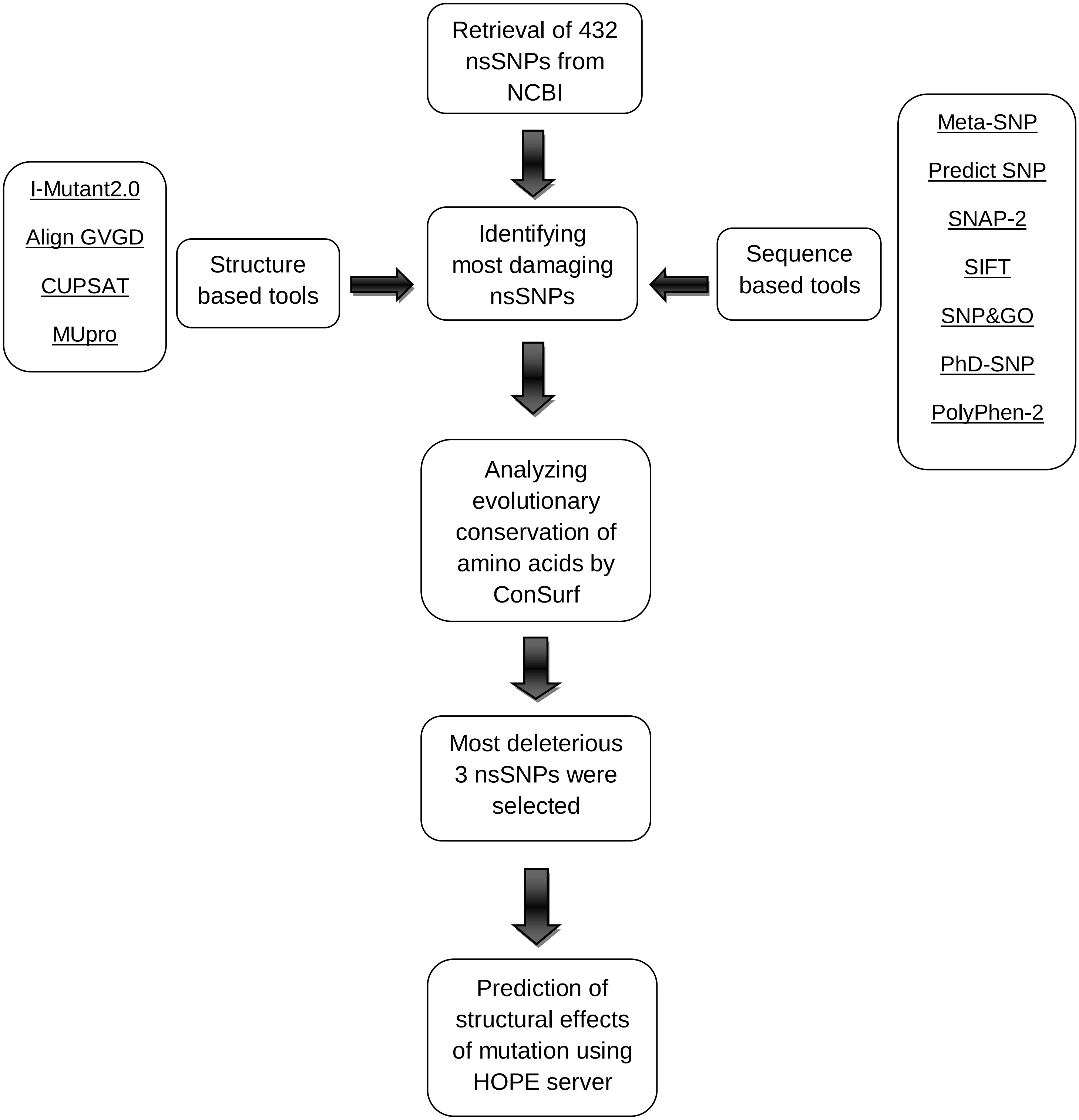

Retrieval of 432 nsSNPs from NCBI
I-Mutant2.0
Align GVGD
CUPSAT
MUpro
Structure based tools
Identifying most damaging nsSNPs
Sequence based tools
Meta-SNP
Predict SNP
SNAP-2
SIFT
SNP&GO
PhD-SNP
PolyPhen-2
Analyzing evolutionary conservation of amino acids by ConSurf
Most deleterious 3 nsSNPs were selected
Prediction of structural effects of mutation using HOPE server

Figure- 2. Schematic diagram summarizing the protocol for the study.

CHAPTER-4

Results and Discussion

4.1 Availability of KRAS SNPs

A total of 20,081 SNPs were obtained from NCBI database, out of these 432 SNPs were missense SNPs, which were selected for screening to identify the deleterious SNPs associated with NSCLC. Results obtained from NCBI can be seen in Table-2.

Table:2. Representation of number of SNPs obtained from NCBI

Total SNPs	20081
Inframe Deletions	10
Missense SNPs	432
Introns	16541
Synonymous	161

4.2 Interaction Analysis of KRAS

Results are displayed in the form of nodes and edges, where nodes represent the proteins and edges represents the association between the proteins. STRING also represents known as well as predicted interactions. The interaction partner for KRAS can be seen in the figure-3.The results obtained shows that 10 genes are associated with KRAS gene.

STRING also provides information in a tabular format regarding the resultant network. This information consists of number of nodes, edges, avg. node degree, and avg. local clustering coefficient and PPI enrichment value (Table-3).

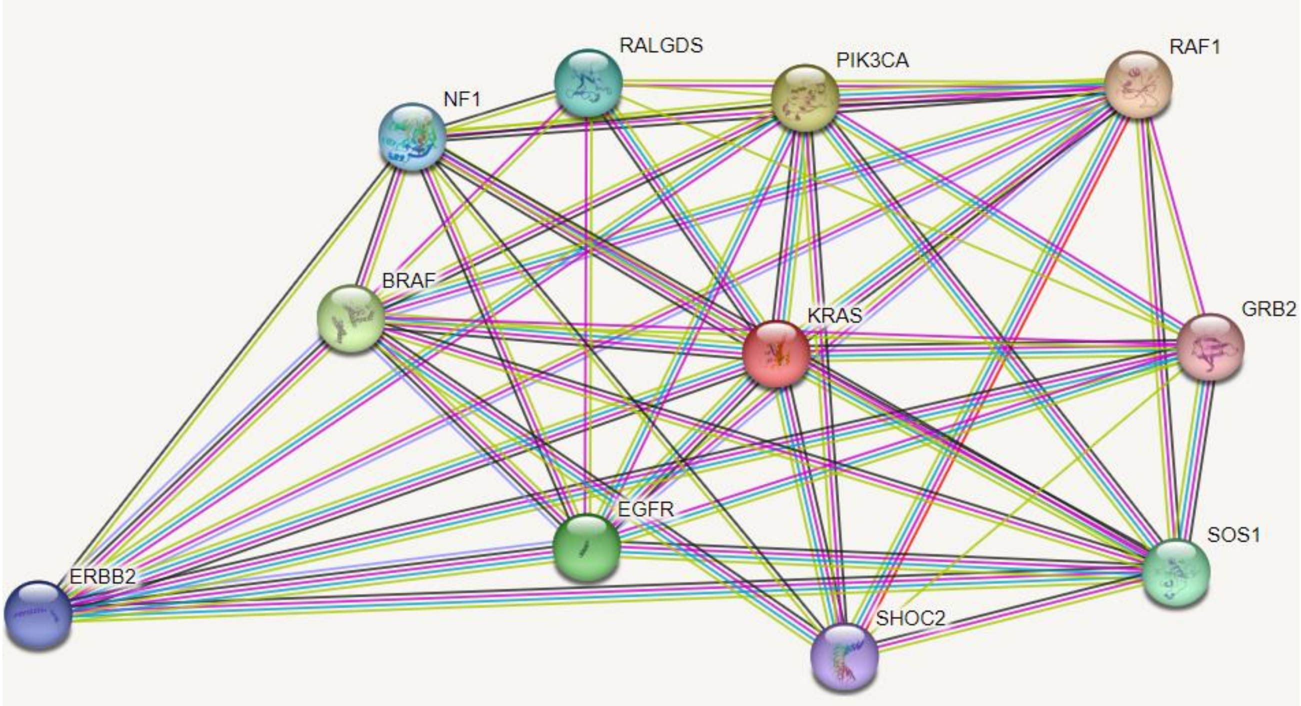


Figure-3. An interacting network of KRAS protein obtained through STRING database.

Table:3. Information obtained about the network in STRING database.

Number of nodes	11
Number of edges	49
Avg. node degree	8.91
Avg. local clustering coefficient	0.899
PPI enrichment p-value	8.95e-10

When the results were obtained, it was seen that 10 genes were in association with KRAS gene as shown in table-4.

Table:4. Information about the interacting partners of KRAS gene.

Functional Partner	**Score**	**Description**
RAF1	0.999	RAF proto-oncogene serine/threonine-protein kinase; This vital regulatory link operates as a switch that controls how cells behave, involving development, differentiation, a process called survival, and cancer-promoting transformation. Serine/threonine-protein kinase governs Ras GTPases that happen to be membrane-associated linked to the MAPK/ERK chain. A mitogen-activated protein kinase, or MAPK cascade, notably consists of recurrent degradation of the two distinctive MAPK kinases (MAP2K1/MEK1 and MAP2K2/MEK2) in addition to the extra-cellular signal controlled kinase commences, whenever RAF1 is activated.
PIK3CA	0.999	Alpha isoform of Phosphatidylinositol 4, 5-bisphosphate 3-kinase as well as phosphoinositide-3-kinase (PI3K), that has been phosphorylated in order to generate PIP3
BRAF	0.999	Serine/threonine-protein kinase B-raf is a proto-oncogene which takes part with conveyance of mitogenic signals via the cell membrane towards mucus.
EGFR	0.998	The epidermal growth factor receptor latches to the EGF family of ligands as well as it stimulates a variety of signal cascades which convert extracellular signals into the appropriate cellular responses.
SOS1	0.998	Son of sevenless homolog 1; fosters Ras-bound GDP exchange by GTP. Governs the phosphorylation of MAP kinase MAPK3 in anticipation of EGF, presumably via boosting Ras activation.
RALGDS	0.997	Facilitates GTP binding and GTPase activation through stimulating GDP dissociation between the Ras-related RalA and RalB GTPases. Engage in interconnections and operates like an impacting component for Rap, K-Ras, H-Ras, and R-Ras.
NF1	0.994	Elevates Ras's GTPase activity.
ERBB2	0.992	Essential component of neuregulin-receptor complex. Regulates outgrowth and stabilization of peripheral microtubules.
SHOC2	0.990	It is regulatory subunit of protein phosphatase 1 (PP1c), which precisely dephosphorylate the 'Ser-259' regulatory site of RAF1 activity at specialized signaling complexes.

4.3 IDENTIFICATION OF REGULATORY TARGETS FOR KRAS GENE

The idea of i-Regulon is to enable direct mapping of gene regulatory networks utilizing motif enrichment in a set of co-expressed genes. The transcription factor that is master regulon for KRAS gene taken is GLI1. A network of twenty three regulatory targets for KRAS gene is obtained as seen in figure-4.

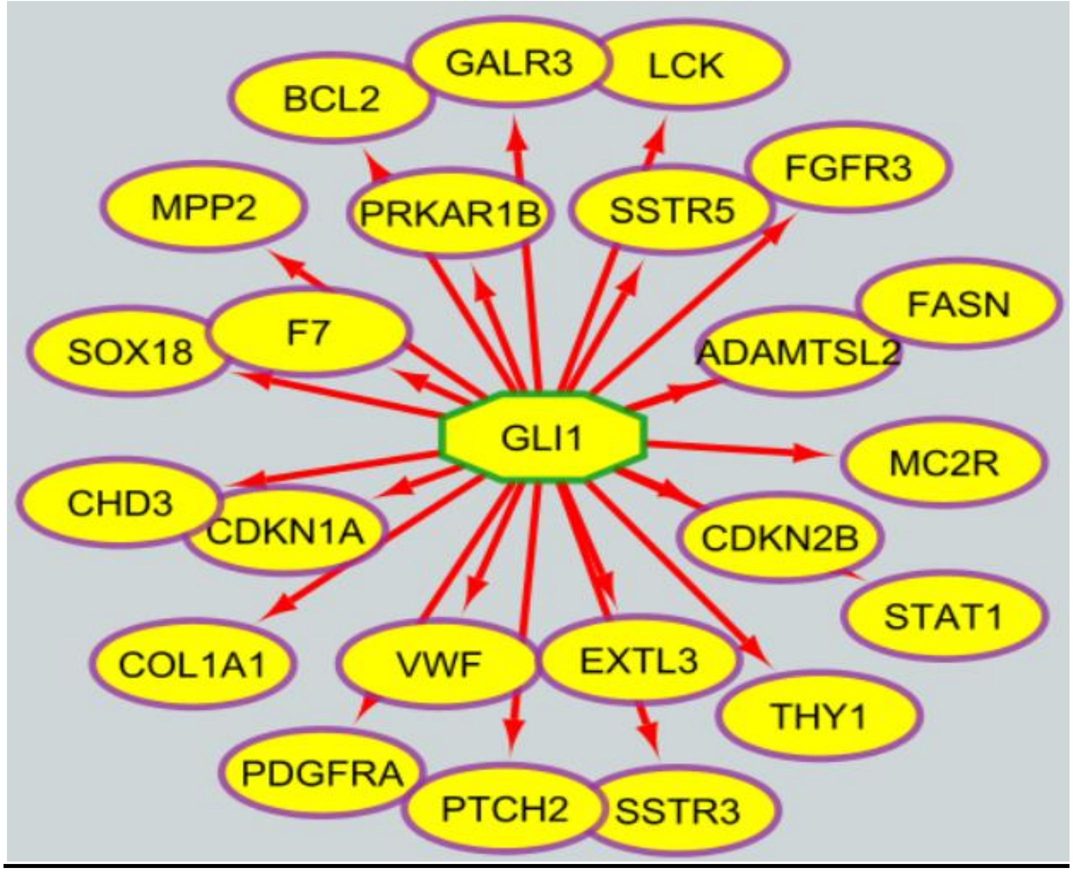


Figure- 4. Interaction network obtained consisting of different regulatory targets.

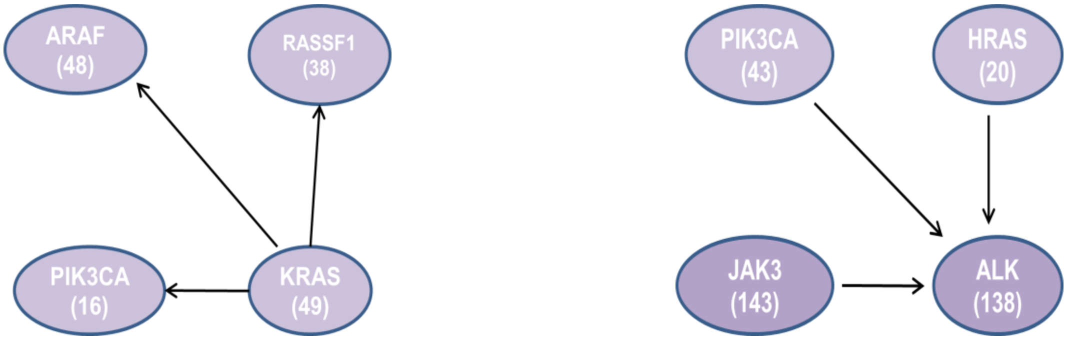
The nodes were selected and prediction was run to obtain a table consisting of Motif ID, AUC, NES, cluster code and transcription factors (Table-5). The maximum NES score for each enriched TF is shown together with the cluster code.

Table:5. Representation of Motif IDs and its transcription factors.

RANK	MOTIF ID	AUC	NES	CLUSTER CODE	TRANSCRIPTION FACTOR
1	transfac_pro-M01037	0.472845	11.0699	M1	GLIS1,GLI1,GLI3,GLI2,ZIC3,ZIC1,ZIC2
2	swissregulon-GLI1…3.p2	0.389823	10.1446	M1	GLI2,GLI1,GLI3,GLIS1,ZIC2,ZIC1,ZIC3
3	homer-M00079	0.349776	8092492	M1	GLI3,GLI1,GLI2,GLIS1,ZIC2,ZIC1,ZIC3
4	transfac_pro-M01703	0.349776	7.89027	M1	GLI2,GLI3,GLI1,GLIS1,ZIC2,ZIC1,ZIC3
5	transfac_pro-M01706	0.333832	7.47834	M1	GLI2,GLI3,GLI1,GLIS1,ZIC2,ZIC3,ZIC1
6	transfac_pro-M01706	0.328538	7.34157	M1	GLI3,GLI2,GLI1,GLIS1,ZIC3,ZIC1,ZIC2

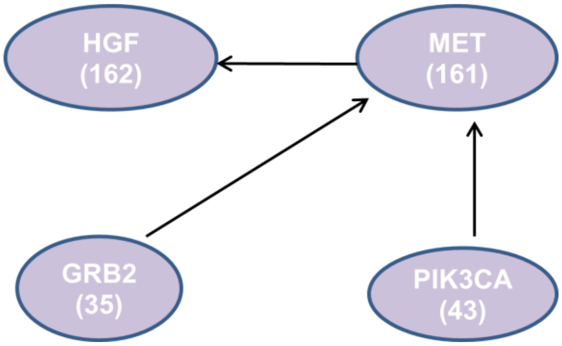
4.4 DETECTION OF NETWORK MOTIFS

Only a simple text file comprising each line identifying a network edge is utilised by FANMOD for statistical analysis. Network motifs for different sizes from 3-8 were obtained and filtered. Results were downloaded in html format. Here, figure-5 shows the nine detected motifs.

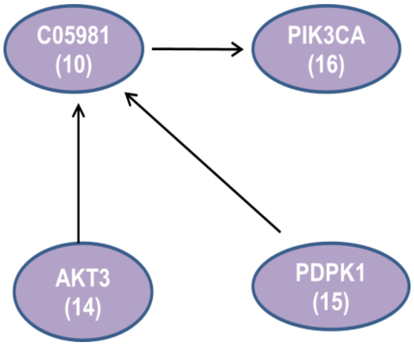


1)

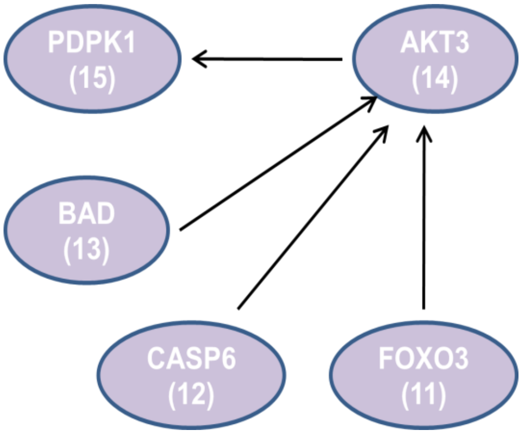
2)


HGF (162)
MET (161)
GRB2 (35)
PIK3CA (43)

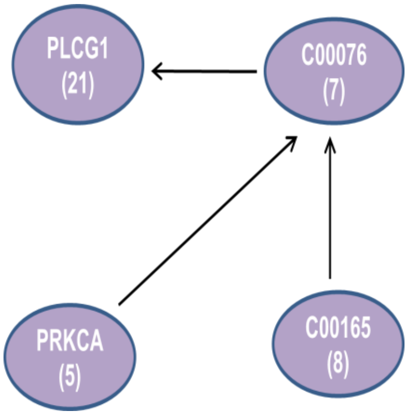
3)


C05981 (10)
PIK3CA (16)
AKT3 (14)
PDPK1 (15)

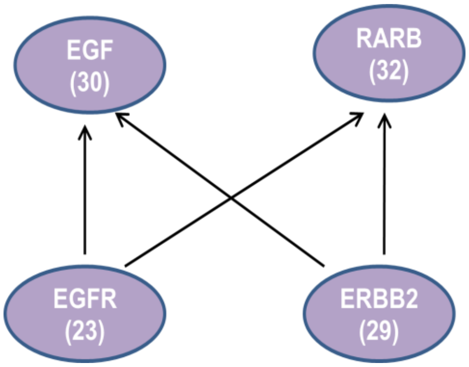
4)


PDPK1 (15)
AKT3 (14)
BAD (13)
CASP6 (12)
FOXO3 (11)

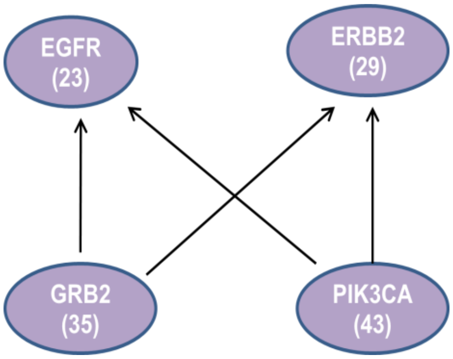
5)


PLCG1 (21)
C00076 (7)
PRKCA (5)
C00165 (8)

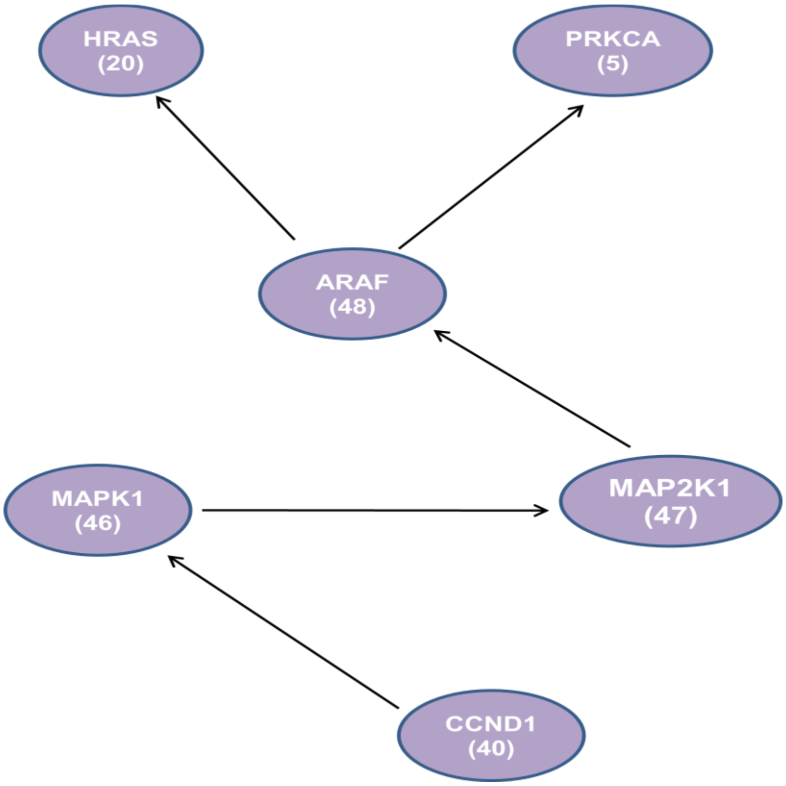
6)


EGF (30)
RARB (32)
EGFR (23)
ERBB2 (29)

7)


EGFR (23)
ERBB2 (29)
GRB2 (35)
PIK3CA (43)

8)



9)

Figure- 5. Representation of4,5 and 6 node of network Motifs detected using FANMOD.

The resultant table consists of motif ID, nodes, Z-score, p-value and significant value for the nine detected motifs. If a motif's z-score is more than 2.0, it is considered statistically overrepresented. [39]

Table: 6. Information of the genes (obtained from FANMOD) along with chromosomal location, protein formed and its function.

Gene	**Chromosomal location**	**Protein formed**	**Functions**
ARAF	Xp11.3	Serine/threonine-protein kinase A-Raf	It is a factor in the mechanism by which mitogenic signals get transmitted from the nuclear envelope to the cell membrane. [50]
RASSF1	3p21.31	Ras association domain-containing protein 1	Regulates the abnormal cell proliferation in psoriasis.
PIK3CA	3q26.3	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform	PIP3 generates proteins with PH domains to the membrane, particularly AKT1 and PDPK1, the fact that it attracts them to the membrane is what renders PIP3 crucial for growth of cells, the continuation of life, division, movement, and morphology. A cell's ability to signal in response to different growth inputs.
BAD	N/A	Bcl2-associated agonist of cell death	Triggers cell death.
CASP6	17p13	Caspase-6	Cysteine protease that is crucial for innate immunity, development, axonal degeneration, and programmed cell death[51]
HGF	7q21	Hepatocyte growth factor	It appears to be a hepatotrophic factor, an efficient mitogen for parenchymal hepatocellular that have already completed maturity in addition to a stimulant of growth for numerous additional tissues and varieties of cells. Binding to and encouraging dimerization of the MET receptor tyrosine kinase activating ligand.
MET	7q21- q31	Hepatocyte growth factor receptor	HGF ligand-binding receptor tyrosine kinase that transmits correspondence into the interior of the cell from the matrix that surrounds the cells. It modulates an assortment of physiological processes, comprising of survival, proliferative growth, and dispersal.
GRB2	17q25.1	Growth factor receptor-bound	Ras transmission as well as the surface of cells growth inhibitor receptors are

		protein 2	connected by an adapter protein.
FOXO3	6q21	Forkhead box protein O3	A transcriptional activator that controls several processes, including apoptosis and autophagy, by recognizing and binding to the DNA sequence 5'-[AG]TAAA[TC]A-3'.[52]
ALK	2p23	ALK tyrosine kinase receptor	An crucial part of the development and differentiation of the nervous system is performed by a neuronal tyrosine kinase receptor complex that is specifically and intermittently expressed throughout certain central and peripheral neurological areas.[53]
MAPK1	22q11.22	Mitogen-activated protein kinase 1	The serine/threonine kinase is a vital component within the MAP kinase transmission pathway. The 2 MAPKs that are important to the MAPK/ERK pathways were MAPK1/ERK2 and MAPK3/ERK1. Additionally, they take part in a signaling cascade that is started by KIT along with KITLG/SCF activation.
MAP2K1	15q22.31	Dual specificity mitogen-activated protein kinase 1	An important component of the MAP kinase signal transduction system is the dual affinity protein kinase. Growth factors, cytokines, and hormones which join with cell-surface receptors induce RAS to start being triggered, which in turn promotes RAF1 to become active.
HRAS	11p15.5	GTPase H-Ras	Associated with the signal transduction activation of the Ras protein. [54]Intrinsic GTPase exists in Ras proteins to conduct a pass and bind GDP/GTP .[55]
PDPK1	16p13.3	3-phosphoinositide-dependent protein kinase 1	Serves as the signal's activating phosphorylation for PKB/AKT1, which stretches the signal's reach to downstream targets influencing cell survival and proliferation in addition to facilitating the absorption and storage of glucose. Plays a crucial part in the transduction of signals from insulin.
AKT3	1q43-q44	RAC-gamma serine/threonine-protein kinase	Enhances adipocyte differentiation and PPARG transcriptional activity. Phosphorylates IKKB, which subsequently in turn activates the NF-kappa-B pathway.

PLCG1	20q12	1-phosphatidylinositol 4,5-bisphosphate	Facilitates production of the 2nd messenger's inositol 1, 4, 5-trisphosphate (IP3) and di-acylglycerol (DAG). Modulates intracellular signaling cascades in a vital way.
PRKCA	17q24.2	Protein kinase C alpha type	A protein kinase that is instantly phosphorylated by targets and activated by calcium, phospholipids, and di-acylglycerol (DAG). It plays a part in controlling the progression of swelling, tumourigenesis, cardiac remodeling, blood vessel development, distinctiveness, rather than migration, and adherence, in addition to the beneficial and adverse effects of these mechanisms.
EGF	4q25-q27	Pro-epidermal growth factor	EGF enhances the growth of certain fibroblasts in cell culture as well as epidermal and epithelial tissues of various types, both in situ and in laboratory.
RARB	3p24	Retinoic acid receptor beta	Due to its poor binding to co repressors, it primarily functions as a gene-expression stimulator in the presence or absence of hormone ligand. [56] Receptor for retinoic acid
JAK3	19p13.11	Tyrosine-protein kinase JAK3	Tyrosine kinase without a receptor participates in a number of procedures, notably cell formation, differentiation, and growth. Performs a key function in hematopoiesis during T-cell maturation and mediates crucial signaling events in innate and adaptive immunity.
CCND1	11q13	G1/S-specific cyclin-D1	The transcription of E2F target genes, which are essential for proceeding through the G(1) phase, is made possible by the phosphorylation of RB1, because it allows the transcription factor E2F to break free from the RB/E2F complex.[57]
ERBB2	17q12	Receptor tyrosine-protein kinase erbB-2	Despite the fact that neuregulin do not interact with it alone, it is a crucial part of a neuregulin-receptor complex. A possible ligand for this receptor is GP30.
		Epidermal growth	EGF group ligands engage with the target tyrosine kinase, the following then triggers

EGFR	7p11	factor receptor	a number of transmission cascades to get engaged on in order translate outside signals into suitable cellular reactions. [58]
PDPK1	16p13.3	3-phosphoinositide-dependent protein kinase-1	Regulates the activity of a group of insulin and growth factor-stimulated protein kinases

If a network motif's frequency has a P-value of less than 0.01, this will be deemed statistically efficient. [40]

Table: 7. Table showing Z-Score, P-value and significant value for detected motifs.

Sr. No.	**Motif_ID**	**Nodes**	**Z-Score**	**P-Value**	**SignificanceProfile**
1.	204	4	17.757	0	0.672
2.	1056880	6	6.5064	0	0.153
3.	2184	4	3.8253	0	0.144
4.	532744	5	4.4562	0	0.224
5.	2116	4	3.5399	0	0.134
6.	2184	4	3.8253	0	0.144
7.	2116	4	3.5399	0	0.134
8.	204	4	17.757	0	0.672
9.	2116	4	3.5399	0	0.134

A vector of z-scores represents the significance profile of a set of network motifs.[41]

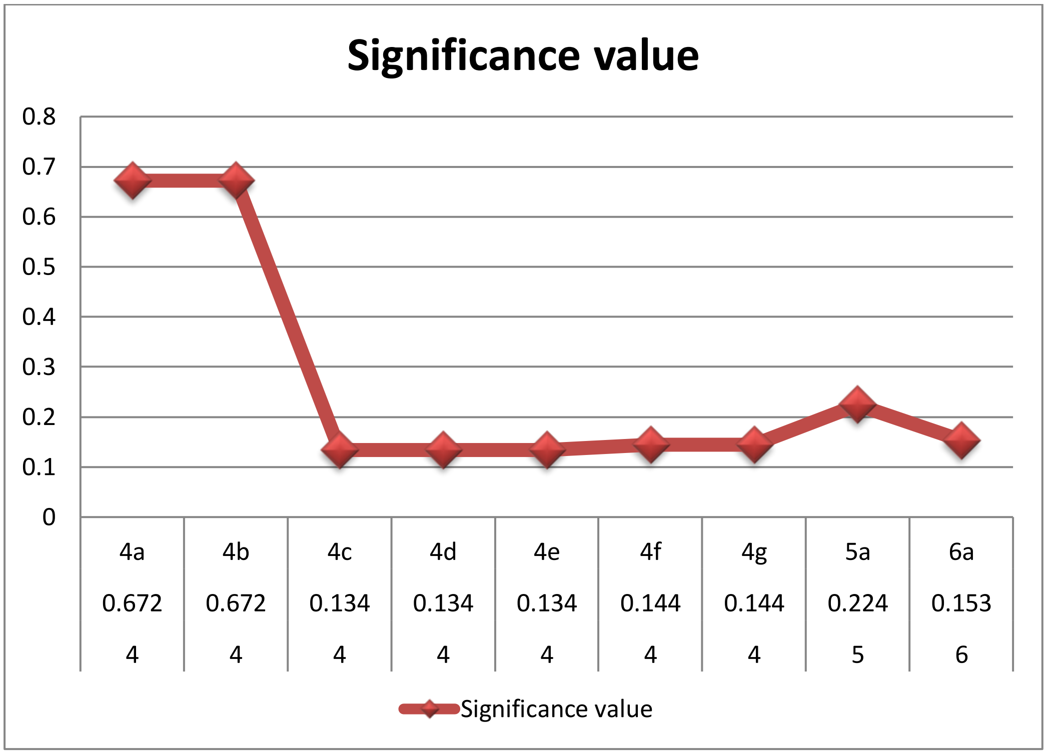


Figure- 6. Graph showing significance value of different nodes

4.5 Screening of nsSNPs

The structural and functional influence of the nsSNPs was predicted for the KRAS protein while this work centered on the nsSNPs of KRAS and associated destructive role. Total nsSNPs retrieved from dbSNP database were 432, which were further evaluated to identify their deleterious nature using sequence based tools. Out of 432 nsSNPs, 14 nsSNPs resulted as dangerous or disease associated or probably damaging after utilizing the seven sequence-based tools which were SNP&GO, Meta SNP, Predict SNP, PolyPhen 2, PhD SNP, SNAP-2 and SIFT. The results can be seen in Table 8.

Table: 8. List of missense nsSNPs indicated by six out of seven methods predicted to be harmful based on sequence based analysis.

Variant ID	**Mutation**	**Meta-SNP**	**Predict SNP**	**SNAP-2**	**SIFT**	**SNP&GO**	**PhD-SNP**	**PolyPhen**
rs104886029	A59V	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs104886030	G60S	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs104886031	G60R	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs104886032	F156L	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs104886033	T58I	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs104886034	V14I	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs104886035	P34R	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs104894367	V152G	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs121913238	Q61E	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs121913527	A146T	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs121913538	L19F	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs202247812	N116S	Disease	Deleterious	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs373500216	A134G	Disease	Neutral	Deleterious	Deleterious	Disease	Deleterious	Probably Damaging
rs387907205	Y71H	Disease	Deleterious	Deleterious	Tolerated	Disease	Deleterious	Probably Damaging

The resultant 14 nsSNPs were thought to be important mutations so were further used in structure based analysis. The structure-based tools used were Align GVGD, CUPSAT, MUpro and I-Mutant 2.0.

The structure powered tools analyzed all the 14 nsSNPs selected from Table 8. After the analysis was completed 5 nsSNPs were predicted as seen in Table 9. For further analysis of conserved nsSNPs, ConSurf was used.

Table: 9. List of nsSNPs along with the predictions obtained from four structure-based tools showing deleterious, destabilizing or damaging effect.

Variant ID	**Mutation**	**I-Mutant 2.0**	**CUPSAT**	**Align GVGD**	**MUpro**		
					SVM[a]	**SVM**[b]	**Neural Networks**
rs104894366	P34R	Decrease	Most likely	Destabilizing	Decrease	Decrease	Decrease
rs104894367	V152G	Decrease	Most likely	Destabilizing	Decrease	Decrease	Decrease
rs387907205	Y71H	Decrease	Most likely	Destabilizing	Decrease	Decrease	Decrease
rs104894364	T58I	Increase	Most likely	Destabilizing	Decrease	Decrease	Decrease
rs104894359	G60R	Decrease	Most likely	Destabilizing	Decrease	Increase	Decrease

Total 12-13 computational methods or biotools were used to identify or filter the important SNPs. After combining all the results of the analysis of the above mentioned tools, three nsSNPs were found to be most deleterious which were P34R, V152G, and Y71H. These five nsSNPs: rs104894366 (P34R), rs104894367 (V152G), rs387907205 (Y71H), rs104894364 (T58I), rs104894359 (G60R) were selected for MDS analysis.

4.7 Functional and Evolutionary Conservation Analysis of KRAS

Conservation analysis gives a basic idea regarding the damage caused by the deleterious mutations on the function as well as the structure of the proteins. Basically it helps to determine the nature of deleterious mutations in a highly conserved residue. For the ConSurf database, the protein structure 4EPV was employed as input.

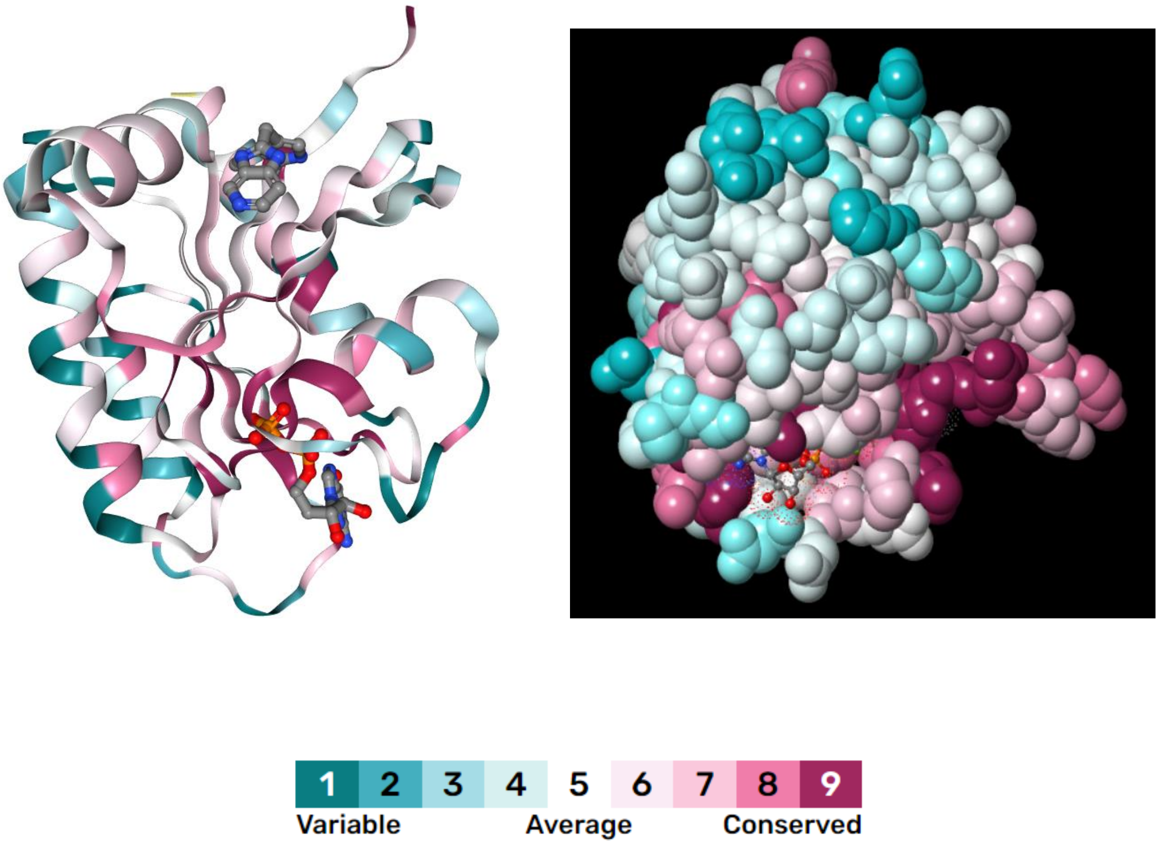


Figure-7. Five substitutions in the chain A of the 4EPV assumed with ConSurf support can be observed in the conservation analysis results.

With the least preserved places (grade 1) coloured turquoise, the half-preserved places (grade 5) coloured white, and the majority of the preserved places (grade 9) coloured maroon., the constant conservation scores have been separated into a unique scale of nine categories.

Table: 10. ConSurf results showing the color score for the mutations.

Sr.No	**Mutation**	**Residue**	**Colour**
1	P34R	P	8
2	V152G	V	9
3	Y71H	Y	8
4	T58I	T	9
5	G60R	G	9

4.8 Analysis of Phenotypic Consequences of deleterious nsSNPs

FATHMM gives prediction based on the score obtained for the mutation. Every minimal change in the underlying amino acid is conveyed by a score with a value close to zero.

Table :11. Using the FATHMM tool, phenotypic repercussions of mutations are anticipated.

Variant ID	**Mutation**	**Prediction**	**Score**	**Human Phenotype Ontology Information**
rs104894359	G60R	Damaging	-1.81	Phenotypic abnormality
rs104894364	T58I	Damaging	-2.31	Phenotypic abnormality
rs104894367	V152G	Damaging	-1.36	Phenotypic abnormality

4.9 Structural Effects of deleterious nsSNPs

Input for HOPE server consists of mutation and the sequence. HOPE starts to gather data from a variety of data sources. The outcome exhibits the effects of the hypothesized mutation on the protein's three-dimensional structure with the proper description of amino acid properties and domains as shown in table-12.

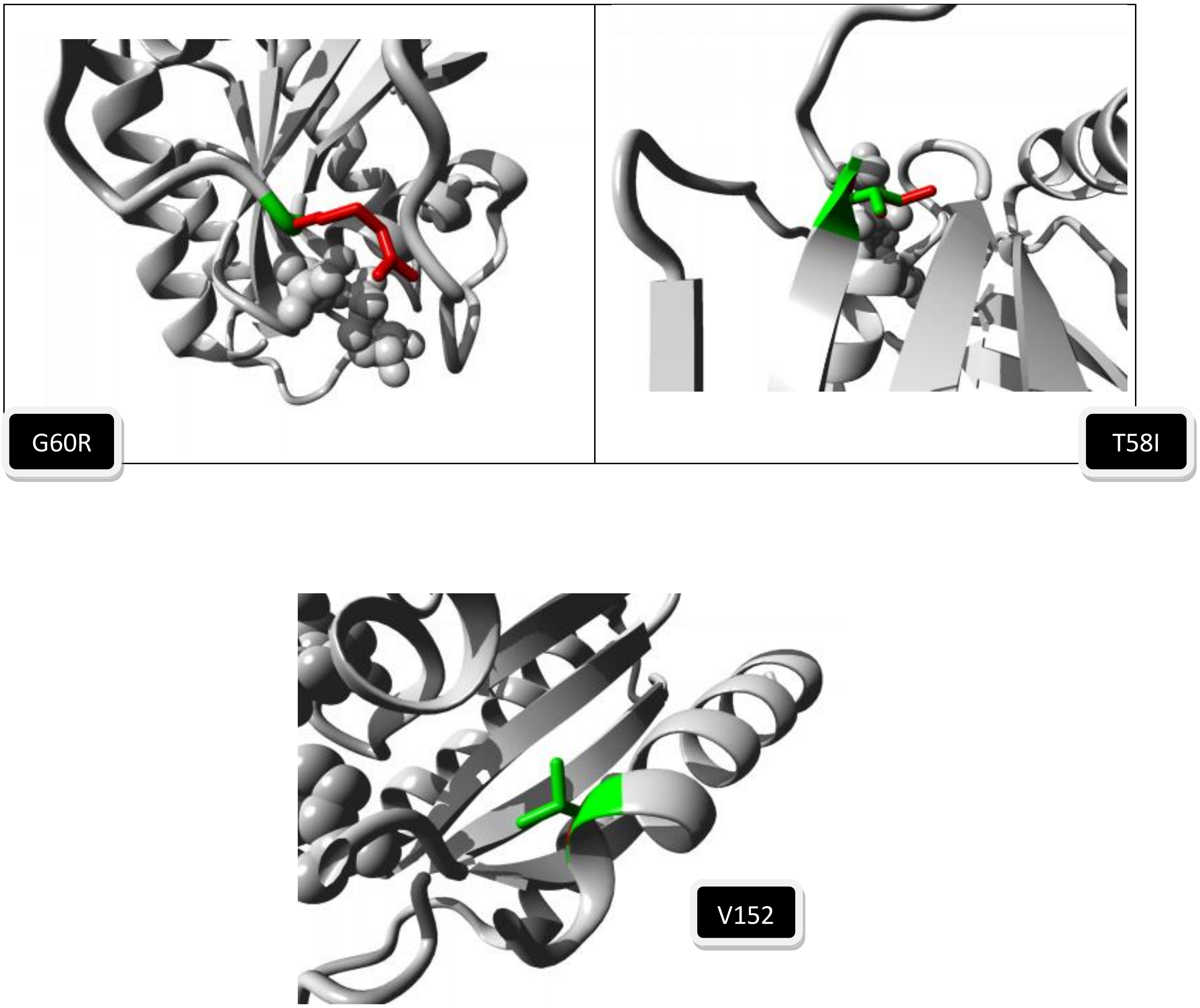
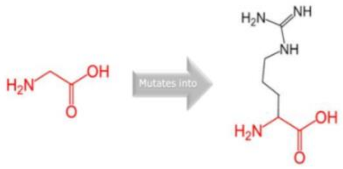
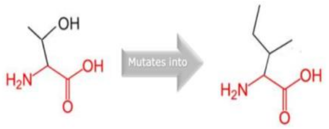
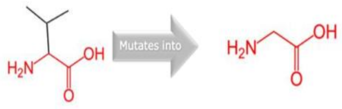


Figure -8. Visualization using HOPE server of both mutated residues of protein (red) and wild type a.a residue (green)

Table: 12. Structural effects on KRAS obtained from HOPE server

Mutation	**Modification of structure**	**Protein traits**	**Domain**
G60R	 H₂N OH O Mutates into H₂N NH NH H₂N OH O	Wild type residue is greater in comparison to mutant residue. Despite the charge within the mutant residue is positive, it was neutral when compared to the WT residue. The local stability might be altered by the mutation, which could ultimately have an influence on the ligand conversations generated through an adjacent residue. Additional hydrophobic than mutant residue is wild-type residue.	The mutant residue is in touch between residues in a domain deemed crucial for interacting to other molecules and is situated in that domain. The protein's functionality may be impacted by the mutation if it disrupts the interaction between each of these domains. The modified residue resides in a region crucial for the association of extra molecules and is in touch between residues in a region crucial for how the protein works. This interaction may be altered by the mutation, which might interfere with the signal's advancement from the binding site to the activity site.

T58I		Greater in size compared to wild type residue is mutant residue. Increased hydrophobic in comparison to wild-type residue is mutant residue. The protein's centre obscured the wild-type residue. Because it’s larger, the mutant residue probably isn’t going fit. Although the transformed residue was not in immediate proximity with a nucleotide, one of its neighbors turned out to do so. The mutation might have an influence on this interaction. Position 8 valine and the wild-type residue unite to establish a hydrogen bond. Because of the disparity in size across the new residue and the prior wild-type residue, the new residue is unable to establish an identical hydrogen bond as the former wild-type residue. The building of hydrogen bonds will be altered by the hydrophobicity difference.	The modified residue is in connection with receptors in a domain that is crucial for engaging to other molecules and is positioned in that domain. The protein's functioning may be impaired by the mutation if it alters the communication amongst these two domains. The impacted residue resides in a region that is essential to the association of other molecules which is in correspondence with residues in domain crucial for the protein's function. This interaction may be compromised by the mutation that might interfere with the signal's transition from the binding site to the activation domain.
V152G		The ratios of the wild-type and mutant amino acids vary. When the altered molecule is the wild-type residue, it grows more hydrophobic. The altered residue gets compressed more than the wild-type residue. The alteration will trigger the protein's core to break down and become vacant. The hydrophobicities of the residues in the wild-type strain and variant	A domain which is crucial in facilitating the attachment of other molecules comprises altered residue. Residues form another domain is in touch with the mutant residue. The aforementioned relationships may be disturbed due to the alterations. The altered residue is in correspondence with residues in a domain that is crucial for interacting to other molecules and has a

		proteins vary. The protein's core is going to eliminate its hydrophobic interactions as an outcome of the mutation.	location in that domain. The protein's performance may be adversely affected by the mutation if it alters the interaction amongst these two domains. The modified residue is situated in a region that's crucial for the association of other molecules and is in touch with residues in a region crucial for the function of the protein. This interaction may be disrupted by the mutation, which could interfere with the signal's advancement from the binding site to the functional domain.

Conclusion

NSCLC constitutes roughly about 85% of all lung cancers. A number of biological functions are triggered and inhibited by the KRAS gene and mutations in KRAS alter its functions. Therefore, we studied the 10 interaction partners of KRAS protein. Further, we analyzed the KRAS putative mutations using structure and sequence based tools.7 sequence-based methods were used to analyze a total of 432 nsSNPs that had been extracted from the dbSNP database. We uncovered 14 mutations that turned out to be detrimental in nature as a result of this analysis. Structure-based analysis was subsequently utilized to these adjustments. Five mutations that might alter KRAS conformations have been discovered by the structure level inquiries. These 5 mutations were then analyzed for phenotypic consequences where 3 of them were predicted to be damaging. Thus from these analysis we can propose that G60R, T58I and V152Gare the mutations which can be studied more thoroughly using in vitro methods for their experimental verifications.

REFRENCE

1. Subramanian J, Govindan R (Feburary 2007) “ Lung Cancer in never smokers: a review”. Journal of Clinical Oncology. 25(5):56170. Doi: 10.1200/JCO.2006.06.8015.PMID 17290066
2. Kenfield SA, Wei EK, Stampfer MJ, Rosner BA, Colditz GA (June 2008). “ Comparison of aspects of smoking among the four histological types of lung cancer”. Tobacco Control. 17(3) : 198- 204. Doi : 10.1136/tc.2007.022582.PMC 3044470. PMID 18390646.
3. Popper HH(2011).” Large cell carcinoma of the lung- a vanishing entity?” – Magazine of European Medical Oncology. 4: 4-9. Doi: 10.1007/s12254-011-0245-8. S2CID 71238993.
4. Okumura S, Jänne PA. Molecular pathways: the basis for rational combination using MEK inhibitors in KRAS-mutant cancers. Clin Cancer Res. 2014 Aug 15;20(16):4193-9. doi: 10.1158/1078-0432.CCR-13-2365. Epub 2014 Jun 6. PMID: 24907112.
5. Reck M, Carbone DP, Garassino M, Barlesi F. Targeting KRAS in non-small-cell lung cancer: recent progress and new approaches. Ann Oncol. 2021 Sep;32(9):1101-1110. doi: 10.1016/j.annonc.2021.06.001. Epub 2021 Jun 2. PMID: 34089836.
6. Skoulidis F, et al. Co-occurring genomic alterations define major subsets of KRAS-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities. Cancer Discov. 2015;5(8):860–77.
7. Rodenhuis, S., et al., Mutational activation of the K-ras oncogene. New England Journal of Medicine, 1987. 317(15): p. 929-935
8. Goulding, R.E., et al., KRAS mutation as a prognostic factor and predictive factor in advanced/metastatic non-small cell lung cancer: A systematic literature review and metaanalysis. Cancer Treatment and Research Communications, 2020: p. 100200.
9. Adderley H, Blackhall FH, Lindsay CR. KRAS-mutant non-small cell lung cancer: Converging small molecules and immune checkpoint inhibition. EBioMedicine. 2019 Mar;41:711-716. doi: 10.1016/j.ebiom.2019.02.049. Epub 2019 Mar 7. PMID: 30852159; PMCID: PMC6444074.
10. Román, M., Baraibar, I., López, I. *et al.* KRAS oncogene in non-small cell lung cancer: clinical perspectives on the treatment of an old target. *Mol Cancer* **17**, 33 (2018). https://doi.org/10.1186/s12943-018-0789-x

11. da Cunha Santos G, Shepherd FA, Tsao MS. EGFR mutations and lung cancer. Annu Rev Pathol. 2011;6:49-69. doi: 10.1146/annurev-pathol-011110-130206. PMID: 20887192.
12. Bos JL. ras oncogenes in human cancer: a review. Cancer Res. 1989 Sep 1;49(17):4682-9. Erratum in: Cancer Res 1990 Feb 15;50(4):1352. PMID: 2547513
13. Slebos RJ, Rodenhuis S. The ras gene family in human non-small-cell lung cancer. J Natl Cancer Inst Monogr. 1992;(13):23-9. PMID: 1327034.
14. Pylayeva-Gupta Y, Grabocka E, Bar-Sagi D. RAS oncogenes: weaving a tumorigenic web. Nat Rev Cancer 2011;11:761–74.
15. Sun JM, Hwang DW, Ahn JS, Ahn MJ, Park K. Prognostic and predictive value of KRAS mutations in advanced non-small cell lung cancer. PLoS ONE 2013;8:e64816.
16. Brugger W, Triller N, Blasinska-Morawiec M, Curescu S, Sakalauskas R, Manikhas GM, et al. Prospective molecular marker analyses of EGFR and KRAS from a randomized, placebo-controlled study of erlotinib maintenance therapy in advanced non-small-cell lung cancer. J Clin Oncol 2011;29:4113–20.
17. Matikas A, Mistriotis D, Georgoulias V, Kotsakis A. Targeting KRAS mutated non-small cell lung cancer: A history of failures and a future of hope for a diverse entity. Crit Rev Oncol Hematol. 2017 Feb;110:1-12. doi: 10.1016/j.critrevonc.2016.12.005. Epub 2016 Dec 9. PMID: 28109399.
18. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research. 2021 Jan 8;49(D1):D605-12.
19. Sebastian Wernicke, Florian Rasche, FANMOD: a tool for fast network motif detection, *Bioinformatics*, Volume 22, Issue 9, May 2006, Pages 1152–1153, https://doi.org/10.1093/bioinformatics/btl038
20. Capriotti E, Altman RB, Bromberg Y. Collective judgment predicts disease-associated single nucleotide variants. BMC genomics. 2013 May;14:1-9.
21. Leong IU, Stuckey A, Lai D, Skinner JR, Love DR. Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination,

and two metaservers to classify long QT syndrome gene mutations. BMC medical genetics. 2015 Dec;16:1-3.

22. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35(11):3823-35. doi: 10.1093/nar/gkm238. Epub 2007 May 25. PMID: 17526529; PMCID: PMC1920242.

23. "Large scale identification, Mapping, and Genotyping of Single-Nucleotide Polymorphism in the Human Genome." https://www.science.org/doi/10.1126/science.280.5366.1077.

24. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. BMC genomics. 2015 Dec;16(8):1-2.

25. Pauline C. Ng, Steven Henikoff, SIFT: predicting amino acid changes that affect protein function, *Nucleic Acids Research*, Volume 31, Issue 13, 1 July 2003, Pages 3812–3814, https://doi.org/10.1093/nar/gkg509

26. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic acids research. 2012 Jul 1;40(W1):W452-7.

27. Yurcik W. Visualizing NetFlows for Security at Line Speed: The SIFT Tool Suite. InLISA 2005 Dec 4 (pp. 169-176).

28. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009 Jul;4(7):1073-81.

29. Frousios K, Iliopoulos CS, Schlitt T, Simpson MA. Predicting the functional consequences of non-synonymous DNA sequence variants—evaluation of bioinformatics tools and development of a consensus strategy. Genomics. 2013 Oct 1;102(4):223-8.

30. Manfredi M, Savojardo C, Martelli PL, Casadio R. E-SNPs&GO: embedding of protein sequence and function improves the annotation of human pathogenic variants. Bioinformatics. 2022 Dec 1;38(23):5168-74.

31. Capriotti E, Fariselli P. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. Nucleic acids research. 2017 Jul 3;45(W1):W247-52.

32. Capriotti E, Fariselli P. PhD-SNPg: a webserver and lightweight tool for scoring singlenucleotide variants. Nucleic acids research. 2017 Jul 3;45(W1):W247-52.

33. **I.**Adzhubei,D. M. Jordan, and S.R. Sunyaev,"Predicting functional effect of Human Missense Mutations Using PolyPhen-2", Curr. Protoc. Hum Genet.

Editor. Board Jonathan Haines Al, vol.0 7,p. Uni7.20, Jan.2013,doi:10.1002/0471142905.hg0720s76.

34. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W306-10. doi: 10.1093/nar/gki375. PMID: 15980478; PMCID: PMC1160136.

35. Gromiha, M., V., P.&D., S.(2006). CUPSAT: Prediction of protein stability upon point mutations. NucleicAcidsResearch,34(WEB. SERV. ISS.):239-242.doi: 10.1093/nar/gkl190.

36. J. Cheng, A. Z. Randall, M. Sweredoski, and P. Baldi. SCRATCH: a Protein Structure and Structural Feature Prediction Server. Nucleic Acids Research, vol. 33, w72-76, 2005.

37. Mark F Rogers, Hashem A Shihab, Matthew Mort, David N Cooper, Tom R Gaunt, Colin Campbell, FATHMM-XF: accurate prediction of pathogenic point mutations via extended features, *Bioinformatics*, Volume 34, Issue 3, February 2018, Pages 511–513, https://doi.org/10.1093/bioinformatics/btx536

38. Venselaar, H., te Beek, T.A., Kuipers, R.K. *et al.* Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* **11**, 548 (2010). https://doi.org/10.1186/1471-2105-11-548

39. Kashani Z., Ahrabian H., Elahi E. et al.: “ Kavosh: a new algorithm for finding network motifs”, BMC Bioinf., 2009 10,(1), p. 318.

40. Timothy LaRock, Ingo Scholtes, Tina Eliassi-Rad, Sequential motifs in observed walks, *Journal of Complex Networks*, Volume 10, Issue 5, October 2022, cnac036, https://doi.org/10.1093/comnet/cnac036

41. Milo R., Itzkovitz S., Kashtan N. et al.: “ Superfamilies of evolved and designed networks”.

42. Janky R, Verfaillie A, Imrichová H, Van de Sande B, Standaert L, Christiaens V, Hulselmans G, Herten K, Naval Sanchez M, Potier D, Svetlichnyy D, Kalender Atak Z, Fiers M, Marine JC, Aerts S. iRegulon: from a gene list to a gene regulatory network using large motif and track collections. PLoSComput Biol. 2014 Jul 24;10(7):e1003731. doi: 10.1371/journal.pcbi.1003731. PMID: 25058159; PMCID: PMC4109854.

43. Masoudi-Nejad A, Schreiber F, Razaghi MK Z (2012). "Building Blocks of Biological Networks: A Review on Major Network Motif Discovery Algorithms". *IET Systems Biology*. **6** (5): 164–74. doi:10.1049/iet-syb.2011.0011. PMID 23101871

44. Wernicke S (2006). "Efficient detection of network motifs". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. **3** (4): 347–

359. CiteSeerX 10.1.1.304.2576. doi:10.1109/tcbb.2006.51. PMID 17085844. S2CID 6188339.

45. Kashtan N, Itzkovitz S, Milo R, Alon U (2004). "Efficient sampling algorithm for estimating sub-graph concentrations and detecting network motifs". *Bioinformatics*. **20** (11): 1746–1758. doi:10.1093/bioinformatics/bth163. PMID 15001476.

46. Deller, M. C., Kong, L. & Rupp, B. Protein stability: a crystallographer's perspective. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **72**, 72–95 (2016).

47. Khokhlatchev, A. *et al.* Identification of a novel Ras-regulated proapoptotic pathway. *Curr. Biol.* **12**, 253–265 (2002).

48. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* **32**, D120–D121. https://doi.org/10.1093/nar/gkh082 (2004).

49. Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics. 2010 Nov 8;11:548. doi: 10.1186/1471-2105-11-548. PMID: 21059217; PMCID: PMC2992548.

50. Su W, Mukherjee R, Yaeger R, Son J, Xu J, Na N, Merna Timaul N, Hechtman J, Paroder V, Lin M, Mattar M, Qiu J, Chang Q, Zhao H, Zhang J, Little M, Adachi Y, Han SW, Taylor BS, Ebi H, Abdel-Wahab O, de Stanchina E, Rudin CM, Jänne PA, McCormick F, Yao Z, Rosen N. ARAF protein kinase activates RAS by antagonizing its binding to RASGAP NF1. Mol Cell. 2022 Jul 7;82(13):2443-2457.e7. doi: 10.1016/j.molcel.2022.04.034. Epub 2022 May 24. PMID: 35613620; PMCID: PMC9271631.

51. Orth K, Chinnaiyan AM, Garg M, Froelich CJ, Dixit VM. The CED-3/ICE-like protease Mch2 is activated during apoptosis and cleaves the death substrate lamin A. J Biol Chem. 1996 Jul 12;271(28):16443-6. PMID: 8663580.

52. Brunet A, Bonni A, Zigmond MJ, Lin MZ, Juo P, Hu LS, Anderson MJ, Arden KC, Blenis J, Greenberg ME. Akt promotes cell survival by phosphorylating and inhibiting a Forkhead transcription factor. Cell. 1999 Mar 19;96(6):857-68. doi: 10.1016/s0092-8674(00)80595-4. PMID: 10102273.

53. Souttou B, Carvalho NB, Raulais D, Vigny M. Activation of anaplastic lymphoma kinase receptor tyrosine kinase induces neuronal differentiation through the mitogen-activated protein kinase pathway. J Biol Chem. 2001 Mar 23;276(12):9526-31. doi: 10.1074/jbc.M007333200. Epub 2000 Dec 19. PMID: 11121404.

54. Gripp KW, Bifeld E, Stabley DL, Hopkins E, Meien S, Vinette K, Sol-Church K, Rosenberger G. A novel HRAS substitution (c.266C>G; p.S89C) resulting in

decreased downstream signaling suggests a new dimension of RAS pathway dysregulation in human development. Am J Med Genet A. 2012 Sep;158A(9):2106-18. doi: 10.1002/ajmg.a.35449. Epub 2012 Jul 20. PMID: 22821884; PMCID: PMC4166655.

55. Williams JG, Pappu K, Campbell SL. Structural and biochemical studies of p21Ras S-nitrosylation and nitric oxide-mediated guanine nucleotide exchange. Proc Natl Acad Sci U S A. 2003 May 27;100(11):6376-81. doi: 10.1073/pnas.1037299100. Epub 2003 May 9. PMID: 12740440; PMCID: PMC164454.
56. Hauksdottir H, Farboud B, Privalsky ML. Retinoic acid receptors beta and gamma do not repress, but instead activate target gene transcription in both the absence and presence of hormone ligand. Mol Endocrinol. 2003 Mar;17(3):373-85. doi: 10.1210/me.2002-0340. Epub 2002 Dec 23. PMID: 12554770.
57. Lew DJ, Dulić V, Reed SI. Isolation of three novel human cyclins by rescue of G1 cyclin (Cln) function in yeast. Cell. 1991 Sep 20;66(6):1197-206. doi: 10.1016/0092-8674(91)90042-w. PMID: 1833066.
58. Chen WS, Lazar CS, Lund KA, Welsh JB, Chang CP, Walton GM, Der CJ, Wiley HS, Gill GN, Rosenfeld MG. Functional independence of the epidermal growth factor receptor from a domain required for ligand-induced internalization and calcium regulation. Cell. 1989 Oct 6;59(1):33-43. doi: 10.1016/0092-8674(89)90867-2. PMID: 2790960.

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: ...19/May/23

Type of Document (Tick): PhD Thesis | M.Sc. M.Tech Dissertation/ Report | B.Tech Project Report | Paper

Name: Nonita Sood Department: (M.Sc) Biotechnology Enrolment No 217815

Contact No. 7807307237 E-mail. nia.sood.84680@gmail.com

Name of the Supervisor: Dr. Jitendra Raj Singh

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): SEQUENCE AND STRUCTURE BASED ANALYSIS OF KRAS GENE TO STUDY ITS REGULATORY ROLE IN NON-SMALL CELL LUNG CANCER

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages = 54
- Total No. of Preliminary pages = 10
- Total No. of pages accommodate bibliography/references = 5

Nonita Sood

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at ....4.........(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

19/05/23

(Signature of Guide/Supervisor)

19/05/2023

Signature of HOD

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
19/05/2023	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String	05%	Word Counts	7205
Report Generated on			Character Counts	42127
		Submission ID	Total Pages Scanned	40
20/05/2023		2097026650	File Size	1.75 M

Checked by
Name & Signature

Librarian

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com