International Journal of Applied Pharmaceutics

ISSN- 0975-7058 Vol 15, Issue 5, 2023

Original Article

PREDICTION OF HIGH-RISK NSSNPS ASSOCIATED WITH WISP3 GENE EXPRESSION: AN *IN SILICO* STUDY

SAUNDARYA M. S.[1], SUSHA DINESH[2*], SAMEER SHARMA[2]

[1]Manipal School of Life Sciences, Manipal Academy of Higher Education (MAHE), Manipal-576104, Karnataka, India. [2]Department of Bioinformatics, BioNome, Bengaluru-560043, Karnataka, India
*Corresponding author: Susha Dinesh; *Email: susha@bionome.in

Received: 08 May 2023, Revised and Accepted: 13 Jun 2023

ABSTRACT

Objective: The primary aim of this investigation is to comprehensively examine the detrimental effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on the WISP3 gene. This objective will be accomplished through intricate evaluations encompassing protein stability prediction, amino acid conservation analysis, investigation of protein-protein interactions (PPI), scrutiny of post-translational modifications (PTM), and the utilization of bioinformatics tools to forecast the potential association between nsSNPs and various diseases. By implementing these sophisticated methodologies, we aim to unveil the intricate mechanisms by which harmful nsSNPs influence the functionality and pathological implications of the WISP3 gene.

Methods: Retrieved rsIDs of SNPs from the dbSNP database and filtered using 5 *in silico* programs. Selected nsSNPs were subjected to further analysis i.e., protein stability and conservation analysis, solvent accessibility analysis, PPI and PTM analysis, prediction and evaluation of both native and mutant protein, and identification of cancer association and gene expression analysis.

Results: The study found that seven (C122Y, C145Y, C52Y, C78R, C75G, N233K, and R245I) of the nsSNPs are potentially vulnerable due to their higher conservancy and ability to reduce protein stability. Two (D271N and Q56H) of the nsSNPs from the initial screening were found to be associated with colon adenocarcinoma.

Conclusion: The study's findings could help researchers design experiments to validate the predictions and develop potential treatments for diseases associated with the WISP3 gene.

Keywords: SNPs, WISP3, CCN6, *In silico* analysis, Gene expression, Cancer

DOI: https://dx.doi.org/10.22159/ijap.2023v15i5.48269. Journal homepage: https://innovareacademics.in/journals/index.php/ijap

INTRODUCTION

WISP3, also known as WNT1-inducible signaling pathway protein 3, is a gene that encodes proteins belonging to the CCN family of proteins. CCN family is named after its first three members: cysteine-rich protein 61 (CYR61/CCN1), connective tissue growth factor (CTGF/CCN2), and nephroblastoma overexpressed (NOV/CCN3), with WISP3 also called CCN6 being the sixth member of this family [1]. This gene is located on chromosome 6q22, which encodes a 354 amino-acid protein [2]. This gene is essential for normal postnatal skeletal growth and cartilage homeostasis. It is characterized by four conserved cysteine-rich domains: insulin-like growth factor-binding domain (IGFBP), von Willebrand factor type C module (VWC), thrombospondin domain (TSP), and C-terminal cystine knot-like domain (CTCK) [3] (fig. 1).

Fig. 1: Schematic representation of the WISP3 protein

Amino acid residue numbers are indicated above each domain. WISP3 contains a signaling peptide (SP) and four conserved cysteine-rich domains: IGFBP, VWC, TSP, and CTCK.

Mutations in this gene are associated with a rare autosomal recessive skeletal disorder called progressive pseudo-rheumatoid dysplasia (PPD). PPD typically manifests in children between the ages of 3 and 8. The first symptoms are an unusual gait pattern, weakness, exhaustion, and rigidity in the knuckles and knees. Over time, the patients are often accompanied by symptoms including swollen finger and knee joints and a significant constriction of the region between the hip bones and knee joints [4]. It is overexpressed in a subset of colorectal cancers (CRCs) [5]. Also, loss of WISP3 expression is associated with inflammatory breast cancer, suggesting that this gene functions as a tumor suppressor in inflammatory breast cancer [6].

The most prevalent type of genetic variation in people is single nucleotide polymorphisms (SNPs). There are around 11.5 million SNPs in the human genome [7]. Non-synonymous SNPs (nsSNPs) are one of the types of SNPs present in the coding region, which accounts for changes in encoded amino acids. They are the key factors contributing to many Mendelian disorders. Identifying whether a single amino acid substitution will lead to a pathological effect or not is of great importance for the development of personalized medicines [8].

Various *in silico* approaches have been developed to speculate the deleterious influence of SNPs, including sequence-based and structure-based approaches. Sequence-based methods rely on analyzing features such as the conservation of amino acids across different species or the physicochemical properties of the mutated residue. In contrast, structure-based methods consider the three-dimensional structure of the protein and analyze the potential impact of an SNP on protein stability, folding, or interaction with other molecules. *In silico* methods utilize computational algorithms and models to assess the potential impact of genetic variations on protein function and structure [9].

In this study, we have attempted to identify the nsSNPs of the CCN6 gene and their influence on the structure and function of the protein using various bioinformatic tools and public datasets and to identify various cancers associated with certain nsSNPs. This study is the first systematic and extensive *in silico* analysis of nsSNPs of the CCN6 gene, which will be helpful in future extensive studies in this regard.

MATERIALS AND METHODS

Retrieval of SNPs

The SNPs of the gene CCN6 were retrieved from the dbSNP database (https://www.ncbi.nlm.nih.gov/snp/) [10], and the protein

sequence (UniProt ID: O95389) was obtained from UniProtKB (https://www.uniprot.org/) [11] in FASTA format.

A total of 329 nsSNPs of CCN6 (Gene ID: 8838) were retrieved from National Center for Biotechnology Information (NCBI) dbSNP. Several bioinformatic tools were then used to carry out SNP analysis on the data.

Identification and prediction of effects of deleterious SNPs

In order to analyze the structural and functional effects of harmful nsSNPs in the CCN6 gene, SIFT, SNAP2, Align GVGD, PANTHER, and PolyPhen-2 were employed sequentially.

SIFT (Sorting Intolerant from Tolerant) (https://sift.bii.a-star.edu.sg/) is a sequence homology-based tool that determines deleterious (probability score<0.05) and tolerated (probability score>0.05) missense SNPs. It is a process with multiple phases that starts with a search for related sequences, continues with the selection of closely related sequences that might have properties similar to the query sequence, aligns these chosen sequences, and then determines normalized probabilities for each potential substitution from the alignment. The input query consisted of rsIDs collected from the dbSNP database [12].

SNAP2 (Screening for Non-acceptable Polymorphisms) (https://rostlab.org/services/snap/) examines a variety of sequence and variant features in order to distinguish between effective and neutral variants. The input query was protein sequence in FASTA format. The outcomes were calculated as a score that represents the likelihood that a specific mutation will alter the native protein's functionality with the level of precision that is anticipated. The score ranges from -100 (strong neutral prediction) to +100 (strong effect prediction) and displays a heatmap of every potential replacement at each position [13].

Align GVGD (http://agvgd.hci.utah.edu/) is a web server that predicts whether an amino acid substitution is deleterious or neutral. This prediction sets a strong emphasis on the biophysical properties of amino acids and multiple sequence alignments of proteins. The input query was protein sequence in FASTA format and amino acid substitutions. It has several different classified variants (C0, C15, C25, C35, C45, C55, and C65), with C65 being the most likely to affect the function and C15 being the least probable [14].

PANTHER (Protein analysis through evolutionary relationship) (http://www.pantherdb.org/tools/csnpScoreForm.jsp) is a classification program based on molecular functions, interactions with other proteins, and evolutionary relationships. This tool computes position-specific evolutionary conservation (PSEC) scores by estimating the alignment of various proteins that are evolutionary-related. The input query given was plain protein sequence, amino acid substitutions, and human organism [15].

PolyPhen-2 (Polymorphism phenotyping v2) (http://www.pantherdb.org/tools/csnpScoreForm.jsp) was employed to investigate the potential impact of an amino acid substitution on the overall structure and function of the protein. Protein sequence, rsIDs, and information about amino acid substitution were provided as the input query to the server. The score can be between 0 and 1, with scores close to 0 being tolerated and scores close to 1 being detrimental. It carries out functional annotation of SNPs, maps coding SNPs to gene transcripts, extracts annotations and structural properties of protein sequences, and creates conservation profiles. Using a combination of all these properties, it then calculates the likelihood that the missense mutation will be harmful [16].

Determination of protein stability by MUpro

MUpro (https://mupro.proteomics.ics.uci.edu/) is a set of machine learning programs to predict how single-site amino acid change affects protein stability. This tool determines the sign of protein stability changes and associated Delta G values upon mutation. The input query was plain protein sequence and amino acid substitution [17].

Estimation of conservation profile by ConSurf

An amino acid's level of evolutionary conservation in a protein reflects an equilibrium between the need to maintain the structural integrity and function of the macromolecule and the amino acid's inherent propensity to mutate. ConSurf (https://consurf.tau.ac.il/) is a web server for identifying functional regions in macromolecules by examining the evolutionary dynamics of amino acid substitutions among homologous sequences. The evolutionary rate of an amino acid position is determined by either the empirical Bayesian method or the maximum likelihood method. The input query used was a protein sequence in FASTA format. The conservation score is represented using a color-coded scheme that ranges from 1 to 9, with 9 indicating highly conserved residue [18].

Prediction of solvent accessibility by NetsurfP-2.0

NetSurfP-2.0 (https://services.healthtech.dtu.dk/services/NetSurfP-2.0) employs a convolutional and long short-term memory neural network architecture that was developed using protein structure solutions as training data**.** It predicts solvent accessibility, secondary structure, disorder, and backbone dihedral angles for each residue of the input sequences. The input query used was protein sequence in FASTA format [19].

Post-translational modification (PTM) analysis by MusiteDeep

MusiteDeep (https://musite.net/) is an online tool that offers a deep-learning framework for predicting and visualizing protein post-translational modification (PTM) sites. A protein sequence in FASTA format was used as input. It provides the prediction for multiple PTMs simultaneously [20].

Analysis of protein-protein interactions

Proteins and their functional interactions form the backbone of the cellular machinery. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) (https://string-db.org/) is a database of curated biological pathway knowledge and databases of physical interactions. In addition to implementing well-known classification schemes like Gene Ontology and KEGG, it also provides fresh, new schemes based on both hierarchical clustering of the association network itself and high-throughput text mining. The input query used was protein sequence in FASTA format [21].

Prediction and evaluation of the 3D structure of CCN6 protein and mutated protein

I-TASSER (Iterative Threading ASSEmbly Refinement) (https://zhanglab.dcmb.med.umich.edu/I-TASSER/) is a hierarchical protocol for structure-based function annotation and automated protein structure prediction. With the aid of multiple threading alignments, iterative structural assembly simulations, and atomic-level structure refinement, it first creates full-length atomic structural models. Based on comparisons between the protein's sequence and structure profile, the biological functions of the protein, including its ligand-binding sites, enzyme commission number, and gene ontology terms, are then derived from databases of known protein functions. The FASTA sequence of CCN6 was the input file for this server [22].

SWISS-MODEL (https://swissmodel.expasy.org/) for homology modeling of protein structures, a fully automated server that uses the updated UniProtKB proteome to align targets with templates. In this instance, the FASTA sequence was used as the input query. The preferred Ramachandran plot region, QMEAN, and Molprobity score provided by this server can be used to validate the predicted structures [23].

PROCHECK and ERRAT were employed to evaluate the stereochemical quality of protein structure. PROCHECK (https://servicesn.mbi.ucla.edu/PROCHECK/) examines both the overall structural geometry and the geometry of individual residues, it assesses the stereochemical quality of a protein structure. ERRAT (https://servicesn.mbi.ucla.edu/ERRAT/) validates the statistical relationship of non-bonded interactions between various types of atoms based on typical atomic interactions and serves to validate the overall model quality. For both servers input query was predicted models in pdb format [24].

ProSA-web (https://prosa.services.came.sbg.ac.at/prosa.php) is another widely used tool for enhancing and validating experimental protein structures. The given input query was predicted models in pdb format [25].

Point mutation was made in the native protein sequence at specific locations and I-TASSER was used for the structural analysis of the mutated protein. The mutated model was evaluated using the TM-align tool (https://zhanglab.dcmb.med.umich.edu/TM-align/) compares protein structures based on superimposing the structures to assess structural similarity and computes the root mean square deviation (RMSD) and template modeling-score (TM score). TM-score ranges from 0 to 1, with 1 being a perfect match between two structures, 0.0<TM-score<0.30 means random structural similarity, and 0.5<TM-score<1.00 means both structures are in the same fold [26].

Identification of cancer associated with nsSNPs

cBioPortal (https://www.cbioportal.org/) is a database of cancer genomics that facilitates data exploration and analysis through the use of a variety of visualization and analytical tools. To determine the relationship between the specific nsSNPs and cancer, the distribution of CCN6 gene mutations in the server was searched [27].

CanSAR Black (https://cansarblack.icr.ac.uk/) is a comprehensive knowledgebase that combines data from various disciplines and employs artificial intelligence and machine learning techniques to produce predictions helpful in the drug discovery process [28].

Analysis of gene expression and overall survival rate

GEPIA (Gene Expression Profiling Interactive Analysis) (http://gepia.cancer-pku.cn/), an interactive database, analyzes the RNA sequencing expression data. It provides dot plots or box plots of a gene's expression profiles and survival analysis using the log-rank test. The gene name (CCN6) was used as input for both analyses, and a specific cancer name was selected [29].

RESULTS

Retrieval of nsSNPs

The dbSNP database was used to retrieve the SNPs for the human CCN6 gene. It contained 7156 SNPs in total, out of which 2 inframe deletions, 1 inframe insertion, 2 inframe indels, 6 initiator codon variants, 590 noncoding transcript variants, 5595 introns, 329 nsSNPs (missense), and 125 synonymous. For this study, only nsSNPs of CCN6 were retrieved, which contributed to only 4.5% of all SNPs known in the human CCN6 gene (fig. 2).

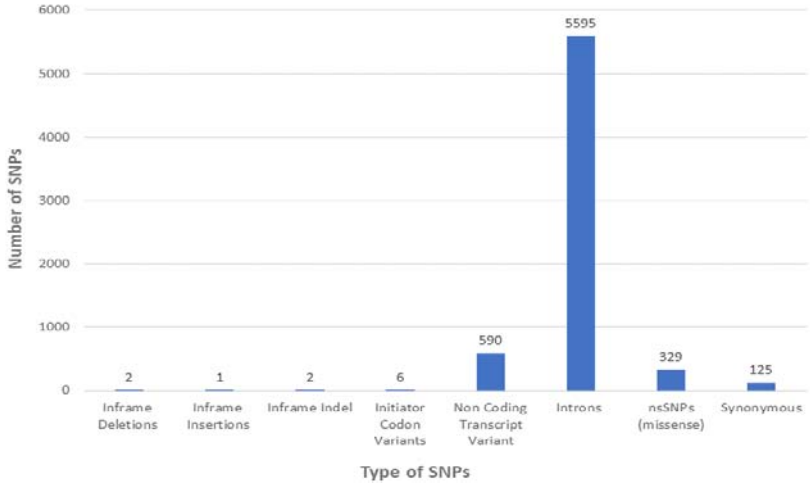


Fig. 2: SNP distribution in the CCN6 gene's various functional classes as found in the dbSNP database

Identification and prediction of effects of deleterious SNPs

Through the use of various tools like SIFT, SNAP2, Align GVGD, Polyphen-2, and PANTHER, *in silico* analysis of CCN6 SNPs obtained through dbSNP was carried out. Initial screening was done using SIFT, out of 329 nsSNPs it predicted 102 to be deleterious or tolerated, with the remaining nsSNPs not found. Among 102 nsSNPs, SIFT categorized the 52 nsSNPs as deleterious and 50 nsSNPs as tolerated. SNAP2, Align GVGD, PolyPhen-2, and PANTHER were used to filter the SIFT result. A total of 22 variants were significant according to SNAP2, while the other 15 had no effect. Align GVGD identified that out of 102 nsSNPs, 20 SNPs as being most likely to be affected, and 16 nsSNPs as being less likely to be affected. PolyPhen-2 predicted 15 as potentially harmful and 18 as benign and PANTHER revealed that 18 nsSNPs were probably damaging and 19 residues were probably benign (fig. 3).

16 significant nsSNPs were chosen based on the pathogenicity demonstrated in at least 4 out of 5 tools (table 1).

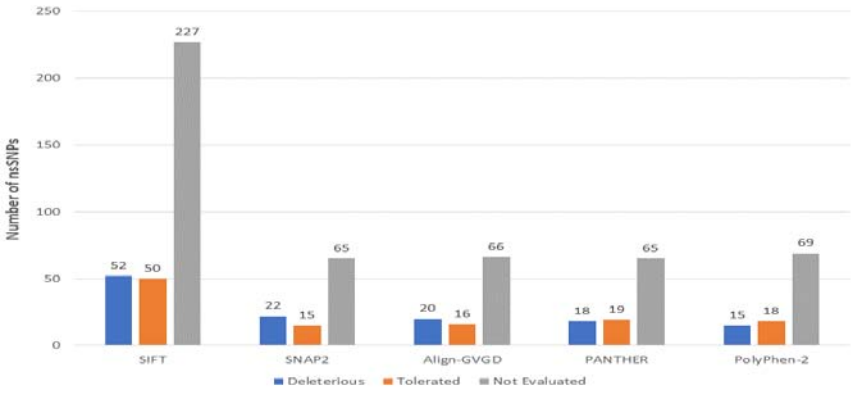


Fig. 3: Deleterious nsSNPs predicted by SIFT, SNAP2, Align-GVGD, PANTHER, and PolyPhen-2 software

Table 1: Selected 16 significant nsSNPs evaluated by 5 *in silico* programs

RsIDs	Amino acid change	SIFT	SNAP 2	Align-GVGD	Panther	PolyPhen-2
rs35914692	W50R	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs121908899	C145Y	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs121908899	C122Y	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs121908902	C78R	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs121908903	S334P	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs147337485	G83E	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs143511761	F348L	Deleterious	Effect	Class C15	Probably damaging	Probably damaging
rs144622585	D105V	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs146519527	R245I	Deleterious	Effect	Class C65	Probably damaging	Possibly damaging
rs149494426	T267I	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs199997447	C75G	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs201041023	N125K	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs371614814	N233K	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs372585876	L14P	Deleterious	Effect	Class C65	Probably damaging	
rs372770731	G97R	Deleterious	Effect	Class C65	Probably damaging	Probably damaging
rs377647286	C52Y	Deleterious	Effect	Class C65	Probably damaging	Probably damaging

Determination of protein stability

Based on the free energy change value (DDG value), the 16 nsSNPs that were chosen for analysis were examined by the MUpro server to determine the impact of point mutation on protein stability. The DDG value counts the energy shifts that occur between a protein's folded and unfolded states. Mutation is said to increase stability if the energy change is positive and vice versa. There were 16 variants, of which 14 (C122Y, C145Y, C52Y, C78R, N233K, R245I, C75G, D105V, F348L, G97R, N125K, S334P, L14P and W50R) were predicted to decrease protein stability and 2 (G83E, and T267I) to increase it (table 2).

Estimation of conservation profile

All the 16nsSNPs were examined using the ConSurf web server to assess evolutionary conservation and identify potential structural and functional residues (fig. 4). Out of 16 nsSNPs, 8 (C122Y, C145Y, C52Y, C78R, G83E, N233K, R245I, and C75G) were found to be highly conserved with the conservation score 9, 6 (D105V, F348L, G97R, N125K, S334P, and T267I) were found to be moderately conserved with the score 8, 7 and 6, and 2 (L14P and W50R) were predicted as a variable with the score 4 (table 3). Among these 8 highly conserved residues, 4 (C122Y, C145Y, C52Y, and C75G) were found to be structural and buried, and the remaining 4 (C78R, G83E, N233K, and R245I) were functional and exposed.

The output from MUpro and ConSurf was compared and examined to identify the most harmful nsSNPs. Based on this comparison, the 7nsSNPs (C122Y, C145Y, C52Y, C78R, C75G, N233K, and R245I) that were chosen as potentially harmful were analyzed further.

Fig. 4: Conservation profile of CCN6 by ConSurf server

Table 2: Analysis of protein stability of high-risk nsSNPs by MUpro

Mutation	DDG (KJ/mol)	Stability
C122Y	-2.9313	Decrease
C145Y	-5.0108	Decrease
C52Y	-3.2857	Decrease
C78R	-4.1539	Decrease
D105V	-2.4221	Decrease
F348L	-1.0958	Decrease
G83E	0.9351	Increase
G97R	-0.5008	Decrease
L14P	-7.4454	Decrease
N125K	-6.4003	Decrease
N233K	-5.5881	Decrease
R245I	-0.4975	Decrease
S334P	-2.4744	Decrease
T267I	1.0029	Increase
W50R	-2.8828	Decrease
C75G	-6.7074	Decrease

Table 3: Analysis of protein evolutionary conservation profile of high-risk nsSNPs by ConSurf

Mutation	Score	Buried/Exposed	Grade
C122Y	-1.279	b	9
C145Y	-1.279	b	9
C52Y	-1.278	b	9
C78R	-1.278	e	9
D105V	-0.444	e	7
F348L	-0.962	b	8
G83E	-1.026	e	9
G97R	-0.554	e	7
L14P	0.746	b	4
N125K	-0.433	e	6
N233K	-1.288	e	9
R245I	-1.158	e	9
S334P	-0.721	b	7
T267I	-0.197	e	6
W50R	0.706	e	4
C75G	-1.278	b	9

Prediction of solvent accessibility

NetSurfP evaluated the solvent accessibility and stability for the 7 variants (C122Y, C145Y, C52Y, C78R, C75G, N233K, and R245I) (fig. 5). All seven variants and their respective wild variants were buried (table 4). Along with the buried or exposed information of the residue, Relative Surface Accessibility (RSA) and Absolute Surface Accessibility (ASA) value was given in the output. RSA is a measurement that compares a residue's actual solvent accessibility to its maximum accessibility. RSA values are relative and normalized. Whereas ASA values represent the actual surface area of residue that is accessible to solvent molecules without normalization [30].

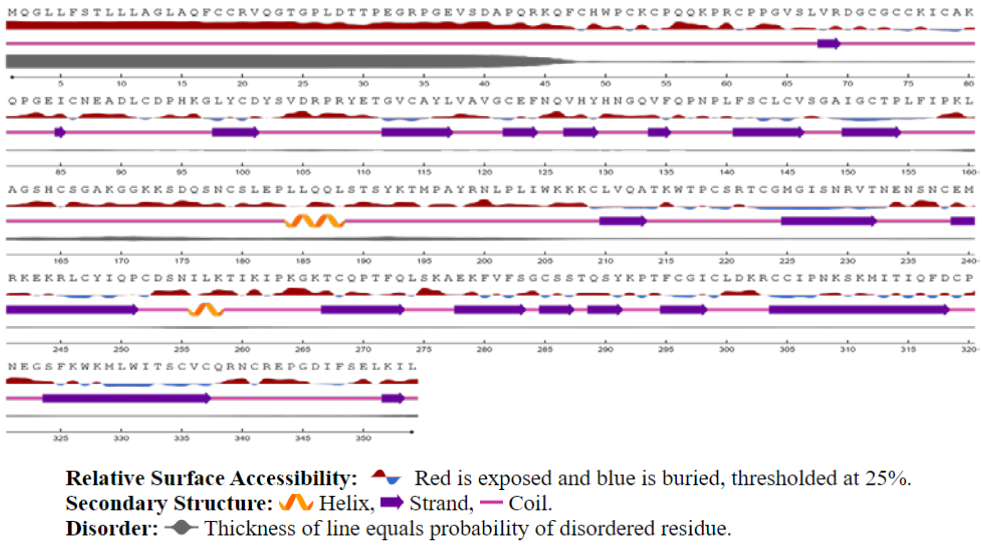


Fig. 5: Solvent accessibility and stability prediction by NetSurfP-2.0

Table 4: Prediction of solvent accessibility by NetSurfP-2.0

Mutation	Class assignment	RSA (%)	ASA (Å)
C122Y	Buried	0.073	10.276
C145Y	Buried	0.024	3.31
C52Y	Buried	0.111	15.651
C78R	Buried	0.125	17.496
N233K	Buried	0.132	19.257
R245I	Buried	0.174	39.797
C75G	Buried	0.105	14.78

Table 5: PTM sites predicted by MusiteDeep

Residue	Position	Predicted PTM
Glutamine (Q)	16	Pyrrolidone carboxylic acid
Cysteine (C)	18	Palmitoylation
Cysteine (C)	19	Palmitoylation
Glutamine (Q)	22	Pyrrolidone carboxylic acid
Threonine (T)	30	Phosphorylation
Proline (P)	59	Hydroxylation
Proline (P)	62	Hydroxylation
Proline (P)	63	Hydroxylation
Proline (P)	82	Hydroxylation
Lysine (K)	169	Acetylation
Asparagine (N)	178	Glycosylation
Serine (S)	180	Phosphorylation
Threonine (T)	222	Glycosylation
Proline (P)	251	Hydroxylation
Proline (P)	293	Hydroxylation
Proline (P)	307	Hydroxylation
Asparagine (N)	308	Glycosylation
Cysteine (C)	335	Palmitoylation
Cysteine (C)	337	Palmitoylation
Cysteine (C)	341	Palmitoylation
Seine(S)	349	Phosphorylation

Post-translational modification (PTM) analysis

PTMs are important in the folding and breakdown of proteins, in the regulation of gene expression, and in various biological pathways. The PTM predicted by MusiteDeep included pyrrolidone carboxylic acid, palmitoylation, phosphorylation, hydroxylation, acetylation, and glycosylation (table 5).

Analysis of protein-protein interactions

The interaction network of WISP3 was constructed using STRING (fig. 6), consisting of 11 nodes and 22 edges. The analysis predicted that WISP3 is associated with the following proteins: WNT1 (Wnt family member 1), COL10A1 (collagen type X alpha 1 chain), BMP4 (bone morphogenetic protein 4), RHOC (ras homolog family member C), LRP6 (LDL receptor-related protein 6), VWF (von Willebrand factor), ITGB (integrin beta), EBLN2 (endogenous Bornavirus like nucleoprotein 2), MRAP2 (melanocortin 2 receptor accessory protein 2), and SPARC (secreted protein acidic and cysteine-rich).

Prediction and evaluation of the 3D structure of CCN6 protein and mutated protein

The three-dimensional structure of human CCN6 was first modeled using SWISS-MODEL. It provided 5 structures and all of them were a partial structure of our targeted protein based on the best-aligned template from the UniProtKB database. Due to the poor GMQE and QMEAN Z-Scores, none of the structures were validated.

Hence, we used I-TASSER for the three-dimensional structural analysis of the CCN6 protein. The top 10 structural analogs in PDB were used as templates for modeling, of which the topmost template (PDB ID: 1W0R) covered 95% of the human CCN6 query sequence. The server provided the top 5 models for the targeted protein. The models' quality was assessed through additional analyses using the PROCHECK, ERRAT, and ProSA programs (table 6), and model number 1 was chosen as the best model (fig. 7).

In our study, the best model (model number 1) included all seven of the investigated mutations (C52Y, C122Y, C145Y, C75G, C78R, N233K, and R245I). Using the I-TASSER web server, models for the native and mutated proteins were generated. Using Biovia Discovery Studio, the native and mutated models were both visualized. TM-align, a program for aligning and comparing protein structures, was used to compare the structural analog. We received values for the TM-score and RMSD of 0.16734 and 5.02, respectively.

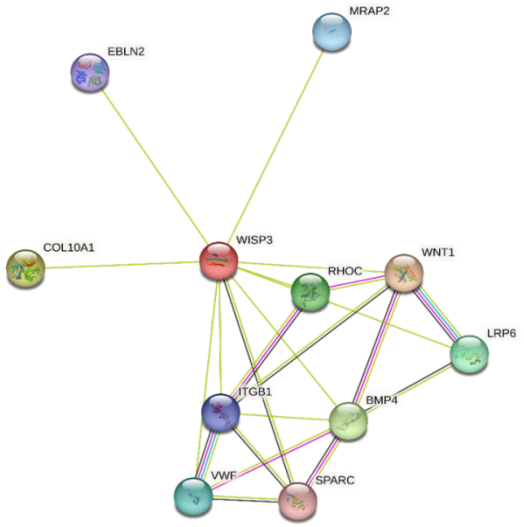


Fig. 6: Protein-protein interaction network of WISP3 by STRING server

Association of cancer with damaging nsSNPs

All the SIFT-predicted nsSNPs were checked for cancer association in cBioPortal and canSAR black. The mutation profile in both web

servers revealed that D271N and Q56H are associated with colon adenocarcinoma (COAD). canSAR black also revealed that the severity of both mutations is moderate.

Expression analysis of CCN6 gene

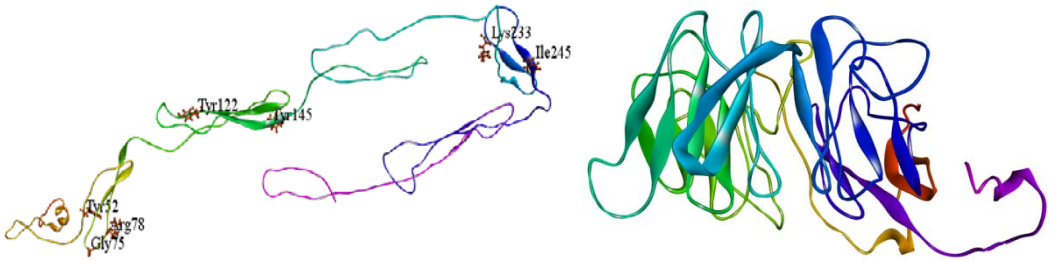
The results of the box plot analysis by GEPIA showed that colon adenocarcinoma (COAD) is caused by the overexpression of the CCN6 gene (fig. 8).

Survival analysis in COAD patients

For analyzing patient survival in cases of colon adenocarcinoma (COAD), we used the GEPIA database. Based on the median level of CCN6 expression, the patients were divided into high-expression and low-expression groups. Compared to patients with lower CCN6 expression, patients with higher expression in COAD reported longer survival times.

Table 6: Scores of different structural assessment tools for the predicted models from I-TASSER

Model number	I-TASSER C-score	PROCHECK ramachandran plot (%)	ProSA Z-score	ERRAT score
01	-2.66	51.7	-3.48	84.39
02	-3.17	44.3	-1.68	51.29
03	-3.22	51	-4.17	65.92
04	-4.32	43	-0.14	72.56
05	-3.91	39.3	-1.69	77.08
Mutated model-1	-2.04	53.2	-0.7	71.42



(a) **(b)**

Fig. 7: Homology models from I-TASSER server; (a) Mutated Model-1; (b) Model-1

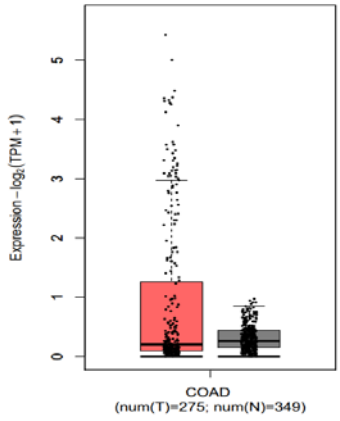


Fig. 8: Boxplot analysis of CCN6 gene expression in case of COAD for both tumor (red) and normal (grey) samples

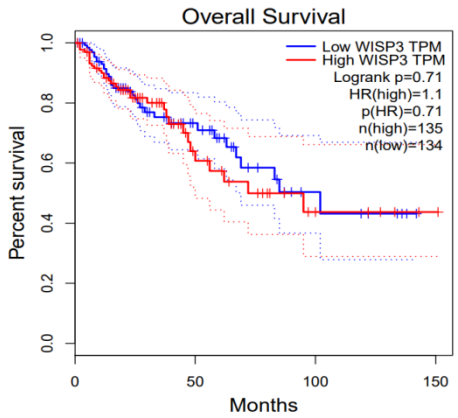


Fig. 9: Overall survival for COAD

DISCUSSION

WISP3 encodes Wnt1-inducible signaling protein 3, a secreted protein that is cysteine-rich and multidomain and whose paralogous CCN family members have been linked to a variety of biological processes, including the development of the skeleton, the vascular, and the nervous system. Progressive pseudo-rheumatoid dysplasia (PPD) is brought on by loss-of-function WISP3 mutations in humans. The identified variant (Chr6: 112382301; WISP3:c.156C>A p. Cys52*) will result in premature termination of the WISP3 protein [3]. Also, overexpression of this gene is associated with colorectal cancer.

SNPs represent the most frequent type of DNA variation in humans. nsSNPs, together with SNPs in regulatory regions, have been found to have the greatest influence on phenotype [31]. In our analysis, some nsSNPs are related to a disease condition, but others are not associated with any change in phenotype and are regarded as neutral. Unraveling their clinical significance will result in significant advances in the field of medical genetics.

In silico analysis of harmful SNPs from large datasets has recently become extremely significant due to the presence of harmful SNPs in several oncogenes [32]. *In silico* methods are beneficial for predicting the effects of SNPs because they can efficiently analyze large datasets of genetic variations and provide predictions for SNPs that have not yet been experimentally characterized. Moreover, these methods can provide insights into the molecular mechanisms underlying the effects of SNPs, which can help design experiments to further validate the predictions.

In this study, we used five prediction tools (SIFT, SNAP2, Align-GVGD, Polyphen-2, and PANTHER) to analyze the data to obtain a comprehensive picture of the pathogenic SNPs of the CCN6 gene. Due to the fact that each algorithm relies on a different set of parameters, we chose 16 nsSNPs (table 1) that were identified in this study as high-risk and predicted to be harmful by at least 4 out of 5 tools. The SNPs prediction accuracy can be improved by combining a variety of computationally based techniques.

A protein's structural stability has a significant impact on its function, activity, and regulation. Protein degradation, misfolding,

and aggregation are all caused by decreased protein stability, which ultimately results in dysfunction [33]. MUpro was used to assess the impact of the 16 harmful nsSNPs mentioned above on the stability of the WISP3 protein. Out of these 16 nsSNPs, 14 made the protein less stable and may affect protein dysfunction (table 2).

The severity of a harmful mutation can be estimated using a protein's evolutionary conservation profile. nsSNPs located in highly conserved regions are more likely to be functionally important, and mutations that affect these residues are more likely to have deleterious effects than nsSNPs located in variable regions [34]. We examined the 16 most harmful nsSNPs' potential effects using the ConSurf web server. ConSurf provides evolutionary conservation data with predictions of solvent accessibility for locating structural and functional sites. Additionally, highly conserved residues are classified as structural or functional depending on where they are present in the protein core or protein surface. 8 out of 16 nsSNPs had a high conservation score, according to ConSurf (table 3).

We concluded that 7 of these 16 nsSNPs (C122Y, C145Y, C52Y, C78R, C75G, N233K, and R245I) are potentially vulnerable due to their higher conservancy and capacity to reduce protein stability. Additionally, we used NetsurfP to examine how these seven high-risk nsSNPs affected the structure of the CCN6 protein. It makes predictions about the protein's secondary structure and solvent accessibility. It depicted all the variations as being buried (table 4).

Proteins may undergo reversible or irreversible chemical changes after translation, which is one of the last stages in protein biosynthesis, that is, post-translational modification (PTM). PTM increases the genome's coding capacity and produces highly varied and expansive proteomes. MusiteDeep was employed to predict PTM sites in the CCN6 gene. It predicted a total of 21 PTM sites which included pyrrolidone carboxylic acid, palmitoylation, phosphorylation, hydroxylation, acetylation, and glycosylation (table 5).

A key component of understanding cellular processes is the network of protein-protein interactions. In addition to providing an easy-to-use platform for interpreting the structural and functional characteristics of proteins, STRING plays a crucial role in filtering and evaluating functional genomics data. This database was used in the current study to show how the WISP3 protein interacts with other proteins that may be involved in various pathways and whose disruption may lead to disease.

Since the human CCN6 protein had no PDB ID, structural prediction techniques were used to estimate the protein's three-dimensional structure. Utilizing SWISS-MODEL, the three-dimensional structure of human CCN6 was first modeled. It offered 5 structures, all of which were fragments of the protein that was our target. This resulted in model 1 (with the highest GMQE score of 0.09) being identified as the best structure among the five. None of the structures were validated due to the inadequate GMQE and QMEAN Z-scores and low coverage.

Then we used the automated protein structure prediction tool I-TASSER, which generated the top 5 models using the FASTA sequence of the protein as an input file. The server's confidence score (C-score), which ranges from-5 to 2, gives a preliminary assessment of the caliber of the projected models. The models with the highest value are those that are most compatible. As a result, model 1 (with the highest C-score of-2.66), was chosen as the best structure. To create higher-quality targeted protein structures, experimental models must be validated. To be certain of it, several computational tools, including PROCHECK, ERRAT, and ProSA were used. Given that the Ramachandran plot displays the torsion angles of the predicted models' protein backbones, it is given the highest priority among all the verification matrices. PROCHECK divides the Ramachandran plot into four regions: the core, allowed, generously allowed, and disallowed region. These regions are used to determine the stereochemical quality of a specific protein structure [35]. Based on the average assessment of all validation software, Model-1 was chosen as the CCN6 protein's best structure (table 6).

Point mutations were introduced in the native protein sequence at specific locations and provided to I-TASSER. It generated five models, among which model 1 was chosen as the best, as it contained all 7 mutations. PROCHECK, ProSA, and ERRAT programs were used to verify its structure, it gave scores that were slightly different from model 1 of native protein (table 6). The structures of mutant and wild-type were compared using the TM-align program. The TM-score and RMSD values we received indicate random structural similarity between the two structures.

A protein mutation causes genomic instability, which can result in a variety of cancers. Different cancer prognostic tools were employed to investigate these correlations. The types of cancers connected to CCN6 are listed in the Cancer genomics database cBioPortal. According to this database, 30 different cancers are associated with the CCN6 gene due to various anomalies. D271N and Q56H were found to have mutation profiles in the web server that were related to colon adenocarcinoma (COAD). The same was also revealed by CanSAR Black, and both mutations are of moderate severity. Low levels of CCN6 expression were observed in metastasizing breast cancer cell lines, suggesting that this gene may function as a tumor suppressor [36].

The TCGA and GTEx data are used in the interactive web application GEPIA, which analyses gene expression in tumors and normal samples. We performed a box plot analysis of CCN6 gene expression in COAD, the result showed that colon adenocarcinoma is caused by the overexpression of the gene (fig. 8). Also, a survival curve was plotted, which plots the survival probability (percentage) against time, and provides an essential summary of the data needed to determine measures i.e., the median survival time. According to this analysis, COAD patients with low CCN6 levels had shorter survival times (fig. 9). Strong *in vivo* research is, however required in the future to confirm the association of a nsSNP with a particular cancer.

CONCLUSION

The functional SNPs in the CCN6 gene have never been thoroughly and systematically analyzed *in silico* before this study. Due to their presence in the highly conserved region and capacity to affect protein stability, we identified seven nsSNPs (C122Y, C145Y, C52Y, C78R, C75G, N233K, and R245I) as potentially harmful. The CCN6 gene products function as a tumor suppressor by limiting cell proliferation in a variety of cellular mechanisms. As a result, changes to this gene have been linked to a variety of diseases, including cancers. Two of the nsSNPs (D271N and Q56H) were found to be associated with colon adenocarcinoma. However, to characterize how these polymorphisms affect the protein's structure and function and to create effective, personalized treatment options, extensive experimental validation is required.

ABBREVIATION

nsSNPs: non-synonymous Single Nucleotide Polymorphisms; WISP3: WNT1-inducible signaling pathway protein 3; CCN6: Cellular Communication Network factor 6; SIFT: Sorting Intolerant from Tolerant; SNAP: Screening for Non-acceptable Polymorphisms; SNP: Single Nucleotide Polymorphisms; PANTHER: Protein analysis through evolutionary relationship; STRING: Search Tool for the Retrieval of Interacting Genes/Proteins; TCGA: The Cancer Genome Atlas; TM score: Template Modelling-score; ASA: Absolute Surface Accessibility; RSA: Relative Surface Accessibility; GEPIA: Gene Expression Profiling Interactive Analysis; GTEx: Genotype-Tissue Expression; PDB: Protein Data Bank; QMEAN: Qualitative Model Energy Analysis; GMQE: Global Model Quality Estimation; RMSD: Root Mean Square Deviation

ACKNOWLEDGEMENT

I thank BioNome (https://bionome.in/) and their scientific team for their computational services and assistance in this research.

FUNDING

Nil

AUTHORS CONTRIBUTIONS

All the authors have equally contributed to the study.

CONFLICTS OF INTERESTS

Declared none

REFERENCES

1. Kutz WE, Gong Y, Warman ML. WISP3, the gene responsible for the human skeletal disease progressive pseudo rheumatoid dysplasia, is not essential for skeletal function in mice. Mol Cell Biol. 2005;25(1):414-21. doi: 10.1128/MCB.25.1.414-421.2005, PMID 15601861.
2. Yu Y, Hu M, Xing X, Li F, Song Y, Luo Y. Identification of a mutation in the WISP3 gene in three unrelated families with progressive pseudo rheumatoid dysplasia. Mol Med Rep. 2015;12(1):419-25. doi: 10.3892/mmr.2015.3430, PMID 25738435.
3. Sailani MR, Chappell J, Jingga I, Narasimha A, Zia A, Lynch JL. WISP3 mutation associated with pseudo rheumatoid dysplasia. Cold Spring Harb Mol Case Stud. 2018;4(1):a001990. doi: 10.1101/mcs.a001990. PMID 29092958.
4. Mansuri MF, Sindhu MA, Hameed M, Javed MN, Laique K, Ullah SS. Progressive pseudo rheumatoid skeletal dysplasia presenting with proportionate short stature and positive WISP3 mutation: a case report. PJR. 2022;32(1).
5. Lu Y, Wang X, Sun X, Feng W, Guo H, Tang C. WISP3 is highly expressed in a subset of colorectal carcinomas with a better prognosis. Onco Targets Ther. 2016;9:287-93. doi: 10.2147/OTT.S97025. PMID 26834488.
6. Kleer CG, Zhang Y, Pan Q, van Golen KL, Wu ZF, Livant D. WISP3 is a novel tumor suppressor gene of inflammatory breast cancer. Oncogene. 2002;21(20):3172-80. doi: 10.1038/sj.onc.1205462, PMID 12082632.
7. Madsen BE, Villesen P, Wiuf C. A periodic pattern of SNPs in the human genome. Genome Res. 2007;17(10):1414-9. doi: 10.1101/gr.6223207, PMID 17673700.
8. Wang B, Tian W, Lei X, Perez Rathke A, Yuan Tseng Y, Liang J. Structure-based method for predicting deleterious missense SNPs. IEEE EMBS Int Conf Biomed Health Inform. 2019. doi: 10.1109/bhi.2019.8834504, PMID 34136829.
9. Wanarase SR, Chavan SV, Sharma S. Evaluation of SNPs from human IGFBP6 associated with gene expression: an in-silico study. J Biomol Struct Dyn. 2023:1-13. doi: 10.1080/07391102.2023.2192793.
10. Yazar M, Ozbek P. *In silico* tools and approaches for the prediction of functional and structural effects of single-nucleotide polymorphisms on proteins: an expert review. Omics. 2021;25(1):23-37. doi: 10.1089/omi.2020.0141, PMID 33058752.
11. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM. DbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-11. doi: 10.1093/nar/29.1.308, PMID 11125122.
12. UniProt Consortium. UniProt: the universal protein Knowledge Base in 2021. Nucleic Acids Res. 2021;49(D1):D480-9. doi: 10.1093/nar/gkaa1100, PMID 33237286.
13. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-4. doi: 10.1093/nar/gkg509, PMID 12824425.
14. Bromberg Y, Yachdav G, Rost B. SNAP predicts the effect of mutations on protein function. Bioinformatics. 2008;24(20):2397-8. doi: 10.1093/bioinformatics/btn435, PMID 18757876.
15. Tavtigian SV, Byrnes GB, Goldgar DE, Thomas A. Classification of rare missense substitutions, using risk surfaces, with genetic-and molecular-epidemiology applications. Hum Mutat. 2008;29(11):1342-54. doi: 10.1002/humu.20896, PMID 18951461.
16. Tang H, Thomas PD. Panther-Psep: predicting disease-causing genetic variants using position-specific evolutionary preservation. Bioinformatics. 2016;32(14):2230-2. doi: 10.1093/bioinformatics/btw222, PMID 27193693.
17. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter(7):Unit7.20. doi: 10.1002/0471142905.hg0720s76. PMID 23315928.
18. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins. 2006;62(4):1125-32. doi: 10.1002/prot.20810, PMID 16372356.
19. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 2016;44(W1):W344-50. doi: 10.1093/nar/gkw408, PMID 27166375.
20. Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sønderby CK. NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. Proteins. 2019;87(6):520-7. doi: 10.1002/prot.25674, PMID 30785653.
21. Wang D, Liu D, Yuchi J, He F, Jiang Y, Cai S. MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. Nucleic Acids Res. 2020;48(W1):W140-6. doi: 10.1093/nar/gkaa275, PMID 32324217.
22. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta Cepas J. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607-13. doi: 10.1093/nar/gky1131, PMID 30476243.
23. Yang J, Zhang Y. Protein structure and function prediction using I-TASSER. Curr Protoc Bioinformatics. 2015;52(1):5.8.1-5.8.15. doi: 10.1002/0471250953.bi0508s52. PMID 26678386.
24. Schwede T, Kopp J, Guex N, Peitsch MC. Swiss-model: an automated protein homology-modeling server. Nucleic Acids Res. 2003;31(13):3381-5. doi: 10.1093/nar/gkg520, PMID 12824332.
25. Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 1993;2(9):1511-9. doi: 10.1002/pro.5560020916, PMID 8401235.
26. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007;35:W407-10. doi: 10.1093/nar/gkm290, PMID 17517781.
27. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302-9. doi: 10.1093/nar/gki524, PMID 15849316.
28. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401-4. doi: 10.1158/2159-8290.CD-12-0095. PMID 22588877.
29. di Micco P, Antolin AA, Mitsopoulos C, Villasclaras Fernandez E, Sanfelice D, Dolciami D. canSAR: update to the cancer translational research and drug discovery knowledge base. Nucleic Acids Res. 2023;51(D1):D1212-9. doi: 10.1093/nar/gkac1004, PMID 36624665.
30. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45(W1):W98-W102. doi: 10.1093/nar/gkx247, PMID 28407145.
31. Shaji D. The relationship between relative solvent accessible surface area (rASA) and irregular structures in protean segments (ProSs). Bioinformation. 2016;12(9):381-7. doi: 10.6026/97320630012381, PMID 28250616.
32. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002;30(17):3894-900. doi: 10.1093/nar/gkf493, PMID 12202775.
33. Kamaraj B, Rajendran V, Sethumadhavan R, Purohit R. In silico screening of cancer-associated mutation on PLK1 protein and its structural consequences. J Mol Model. 2013;19(12):5587-99. doi: 10.1007/s00894-013-2044-0, PMID 24271645.
34. Deller MC, Kong L, Rupp B. Protein stability: a crystallographer's perspective. Acta Crystallogr F Struct Biol Commun. 2016;72(2):72-95. doi: 10.1107/S2053230X15024619, PMID 26841758.
35. Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet. 2001;10(21):2319-28. doi: 10.1093/hmg/10.21.2319, PMID 11689479.
36. Jha NK, Kumar P. Molecular docking studies for the comparative analysis of different biomolecules to target hypoxia-inducible factor-1α. Int J App Pharm. 2017;9(4):83, doi: 10.22159/ijap.2017v9i4.19505.

37. Tran MN, Kleer CG. Matricellular CCN6 (WISP3) protein: a tumor suppressor for mammary metaplastic carcinomas. J Cell Commun Signal. 2018;12(1):13-9. doi: 10.1007/s12079-018-0451-9, PMID 29357008.
38. Lim EC, Lim SW, Tan KJ, Sathiya M, Cheng WH, Lai KS. In silico analysis of deleterious SNPs of FGF4 gene and their impacts on protein structure, function and bladder cancer prognosis. Life (Basel). 2022;12(7):1018. doi: 10.3390/life12071018, PMID 35888106.
39. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H. The protein data bank. Nucleic Acids Res. 2000;28(1):235-42. doi: 10.1093/nar/28.1.235, PMID 10592235.