LGU Journal of
LIFE SCIENCES

LGU Society of Life Sciences

DOI: https://doi.org/10.54692/lgujls.2023.0701240

Research Article
Vol 7 Issue 1 January - March 2023

LGU J. Life. Sci
ISSN 2519-9404
eISSN 2521-0130

In Silico Analysis of MARS1 Gene to Elucidate Low-Frequency Variants Associated with Interstitial Lung and Liver Diseases

Aamna Syed[1], Rana Muhammad Mateen[1], Ayman Naeem[1], Zainab Asif Mirza[1], Muhammad Usman Ghani[2], Mureed Hussain[1]*

1. Department of Life Sciences, University of Management and Technology, Lahore, Pakistan
2. Center of Applied Molecular Biology, University of the Punjab, Lahore

***Corresponding Author's Email: mureed.hussain@umt.edu.pk**

ABSTRACT: *Mutation in MARS1 gene is linked to the development of Interstitial lung and liver disease. The current study aimed in silico analysis to predict the most harmful missense and spliced variants of MARS1 that damage the functionality of Methionyl-tRNA synthetase 1 (MARS 1), catalyses the ligation of methionine to tRNA and is essential for protein biosynthesis. A total of 492 variants were retrieved from the gnomAD database and analysed by CADD, 308 missense variants with PHRED score ≥ 20 were further analysed by CAPICE, META-SNP and CONDEL.85 SNPs detected with deleterious impact on protein structure by screening nsSNPs. Moreover, in-silico stability analysis was done by different tools like DynaMut, DUET, i-Stable2.0 and YASARA. MARS1 protein structure obtained from RCSB PDB (PDB ID: 5GL7) and UCSF Chimera was used for its visualisation. NetSurf-2.0 obtained the analysis of protein functioning by position of residue in the structure. Our results showed that the structure of proteins was significantly deleterious and protein motif and function were changed, we proceeded to use the PROSITE database to forecast the post-translation modification sites and four significant nsSNPs with protein structure change effects. Splice analysis was conducted by SPiCE, Human Splice Finder. It concludes in silico analysis, genes can determine likely pathogenic variation for further in vitro experimental study.*

Keyword: *Interstitial, Methionyl-tRNA synthetase, loss of mutation, gnomAD, Mutation prediction*

INTRODUCTION

Interstitial lung and liver disease (ILLD) is an autosomal recessive disorder characterised by its lipoprotein growth within alveoli, leading to constricted

lung and respiratory failure caused by alterations in the methionyl-tRNA synthetase 1(*MARS1)* gene. *MARS1* is a considerable candidate gene for association with Interstitial Lung and Liver Disease (Rips et al., 2018).

MARS codes methionyl-tRNA synthetase, which belongs to the class 1 family of the aminoacyl-tRNA synthetase (ARSs); such enzyme plays a significant role in protein synthesis by charging tRNAs with their cognate amino acids (Lenz et al., 2020).

Analysis of pathogenic variants is crucial to detect deleterious mutations that are found in the human genome. The human genome consists of the intronic and exonic regions, but the pathogenicity ratio is higher in the coding region (Blackstone 2018). Mutations are mainly related to single nucleotide polymorphism (SNPs) at their coding region, which includes the alteration of an amino acid that emanates the deformity of the protein's function (Bao et al., 2020).

GnomAD is an alliance of investigators seeking to organise exome and genome data from a broader scale into a summary that can provide comprehensive information to the scientific community (Karczewski et al., 2020). These mutations are filtered in CADD, widely used to detect deleterious missense mutations, and can score SNVs. In addition, it works on machine learning between *de novo* variants and the variants that are arisen and become anchored in the human population (Kircher et al., 2014).

Several missense tools like Meta-SNP, CAPICE, and CONDEL are used. Meta-SNP is based on the value of the Reliability Index (RI). The RI value ranges from 0 to 1; the mutations with RI scores less than 0.5 were expected to be harmful, whereas those with RI scores more than 0.5 were projected to be tolerated (Kumar et al., 2018). CAPICE is a new machine-learning-based technique for prioritising pathogenic variants such as SNVs and short InDels (Li et al., 2020). CONDEL is a copy number variation software program (CNV). Its output is composed of five other predictive tools such as Log R Pfam E-value (Clifford et al., 2004), MAPP (Stone and Sidow 2005,

Binkley et al., 2010), Mutation Assessor (Massessor) (Reva et al., 2007), Polyphen2 (PPH2),7 (Adzhubei et al., 2010) and SIFT,13–15 (Ng and Henikoff 2001, Kumar et al., 2009). SIFT was 0.85, Logre was 0.51, MAPP SIFT was 0.85, Logre was 0.51, MAPP (Clifford et al., 2004) was 0.06, Polyphen2 was 0.28, and Massessor was 0.26. Although the intrinsic scores of the five prediction tools differ in form, they all indicate the likelihood that an amino acid change would be approved at a specific place in a protein sequence.

On pathogenic variants, stability tests are applied. The objective is to calculate the difference in free energy upon protein folding. Furthermore, change in Gibbs free energy evaluates the impact of missense mutations on protein's stability. For intendment, various bioinformatics tools such as DYNAMUT (Rodrigues et al., 2018), DUET (Pires et al., 2014) and i-Stable2.0 (Chen et al., 2013) are practised.

The Fold X algorithm, one of the strong determinants of protein stability used for stabilising and destabilising estimation, calculated free energy change upon mutation (Li et al., 2009). UCSF Chimera is used to visualising the retrieved 3D structures and the unwanted interactions due to mutated residues (Pettersen et al., 2004). Post-translational modification analysis increases the complexity of proteomes. PTM sites are involved in mutations, specifically at phosphorylation sites. PTM was confirmed by ScanPROSITE (a protein database) in the *MARS1* Proteins to discover motifs, domains and interactions with other proteins (Hulo et al., 2006).

Different conservation-based tools, such as Netsurf 2.0**,** predicted solvent accessibility (Klausen et al., 2019), secondary structure, structural disorder, and backbone dihedral angles (Petersen et al., 2009). Consurf outputs a score, with 9 being the most conserved amino acid and 1 representing the most varied amino acid. (Ashkenazy et al., 2016) were performed. PROTEIN PLUS was used for ligand binding analysis by observing the change in ligand interaction with protein. (Fährrolfes et al., 2017).

To validate variation in 5'/3' splice sites, two bioinformatic tools SPiCE (Leman et al., 2020) and Human Splice Finder (HSF) (Tang et al., 2016), were performed. Several articles were published related to the role of SNPs in the MARS1 gene in different diseases, but there is still a need for computational analysis.

The recent study was aimed to determine the functional and structural consequences of nsSNPs in the coding region of the *MARS1* gene that is crucial in disease susceptibility through the bioinformatics tool. The majority of the mutations produce an effect on protein stability. The bioinformatics tool also predicted post-translational modification sites on protein structure.

MATERIAL AND METHODS

Retrieval and selection of variants of the *MARS1* gene

The gnomAD v2.1, Variation Viewer and DbSNP database were utilised to assess human *MARS1* variants. The UniProt database has been used to derive protein sequence (PDB ID: P56192) and SNPs Information from the *MARS1* gene. For the retrieving of variants, we implemented gnomAD in our analysis. The schematic diagram is in the following Fig.1.

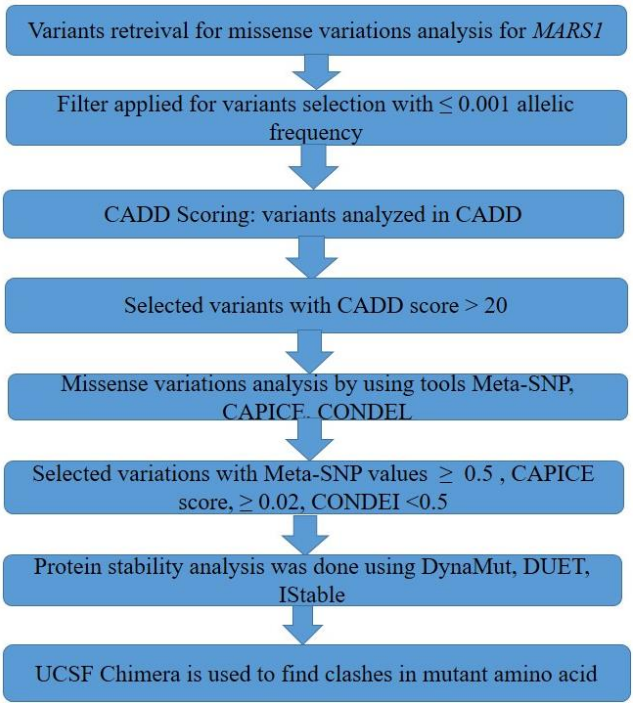

Fig.1 Streamline followed for missense variations analysis in *MARS1* gene

NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS

The following in silico tools are used to predict the effect of SNPs on protein function: Meta-SNP, CAPICE, and CONDEL.

Predicting the deleterious nsSNPs by Meta-SNP

Meta-SNP trained as a random forest-based binary classifier on the output capacity of four tools: SNAP, SIFT, PANTHER and PhD-SNP. As a meta classifier of the tool, its prediction is more accurate than the single tool. The predicted pathogenicity of a mutation depends upon the value of the Reliability Index (RI). The range of RI value is from 0 to 1. The pathogenicity score with an RI value < 0.5 was predicted as pathogenic, while the others with an RI value > 0.5 were predicted as tolerated.

Predicting the functional impact of deleterious nsSNPs by CAPICE

CAPICE is a supervised machine-learning-based model for prioritising pathogenic variants, including SNVs

and short InDels. It is trained on balanced data to annotate SNVs on 11 categorical features. Then, the retrieved variants are annotated accordingly to these features. Outputs are in the form of CAPICE scores predicting the pathogenicity of variations with a cutoff>0.02.

Predicting the functional impact of pathogenic nsSNPs by CONDEL

CONDEL is a software tool that predicts copy number variation (CNV)**.** The classifier is based on a combined predictive score of (Log R Pfam E-value (Logre), MAPP, Mutation Assessor (Massessor), Polyphen2 (PPH2), and SIFT. All retrieved variants were analysed with a combined classifier with output from Score 0 to 1; the higher score was characterised as deleterious.

PREDICTION OF PROTEIN STABILITY CHANGE

In silico analysis on protein stability aims to calculate the difference in free energy upon protein folding. The impact of missense mutations on protein's stability was evaluated by change in Gibbs free energy ($\Delta\Delta G$).

Predicting the protein stability changed by DynaMut, DUET and iStable 2.0

Missense mutation introduces a new amino acid which sometimes is incompatible with neighbouring residue, destabilising the protein and affecting its function. The retrieved variants were analysed with different stability prediction tools to predict the change in protein stability due to missense mutation. The DynaMut server predicts changes in free energy upon the folding and unfolding of the protein. Changes in free energy ($\Delta\Delta G$) of protein arise destabilised effect possessing threshold value >0.

PDB structure, wild type, mutant protein code and chain identification for visualising PDB structures and computing free-energy change ($\Delta\Delta G$) are needed to calculate the protein stability by DUET servers. The higher negative value will indicate a more impact of destabilisation.

iStable2.0 predicted protein stability depending on the vector-supported machine algorithm. Two input types,

structural or sequential protein information, can be supplied.

Prediction of energy minimisation and free energy change due to mutation by YASARA Fold X Program

YASARA software was downloaded along with plugin programs Fold X and Python by default installation to determine the effect of mutations on stability. The $\Delta\Delta G > 0$ value showed decreased stability while $\Delta\Delta G < 0$ value showed increased SNP stability.

PREDICTION OF PROTEIN STRUCTURAL PERSPECTIVE

Molecular Modeling by UCSF Chimera

For the confirmation of SNPs and 3D visualisation, we performed the analysis of the UCSF Chimera program. PDB ID fetched 3D structure (5GL7). The change in the amino acid at a specific position located in *MARS 1* chain A was generated by selecting an option of structure editing and then choosing rotamers. Native and mutant structures are visualised and downloaded as save PDB.

Clashes, Minimisation and labelling of molecule prediction

UCSF Chimera program found clashes that are unwanted interactions between native and mutant residue in Methionyl tRNA synthetase 1 protein. Energy minimisation is utilised to construct or refinish H- bond networks, eliminate unwanted contacts and lower total system energy in protein molecular modelling. A 3-letter amino acid code labelled residues with their perspective position.

Prediction of Post Translational Modification Sites on protein structure by UniProt and Prosite Scan

These are all the alterations following mRNA protein translation. These changes are important for the functioning of proteins or any other enzymes to be identified. The PTM enzymes detected a particular consensus sequence or motifs for these changes. In these specific locations, random mutation can also occur for change and cannot be noticed through activating enzymes, resulting in no modification of enzymes that cannot enable the protein

to work properly. We used ScanProsite to determine the PTM sites in the MARS1 Proteins to discover motifs, domains and interactions with other proteins in the FASTA or UniProt Sequence accession No. (P_56191) sequence. The results detailed changes in the location indicated.

ANALYSIS OF PROTEIN CONSERVATION

Prediction of Conserved residues by ConSurf

It is a method used in bioinformatics to compute the evolution of amino acid conservation by using an empirical Bayesian inference in the protein sequence. Assessment of the homologous sequences is dependent on evolutionary history. The importance of a residue can be observed through the conservational score: the more conserved residue, the more severe effect of its mutation on protein function. The degree of amino acid residue conservation with 50 homologous sequences was estimated. The method was chosen those significantly conserved residues for additional study at the high-risk nsSNPs sites. The conservation score is given together with a colour scheme. For example, the most conserved is score 9, while score 1 is the most varied amino acid.

Prediction of solvent accessibility prediction by NetSurf -2.0

The tool predicted secondary structure, structural disorder, backbone dihedral angles and solvent accessibility of amino acids to find the active site in the completely folded protein. Furthermore, characterising the position of residue in the protein, especially in the catalytic site, can be predicted by NetSurf -2.0. The mechanism for this prediction is based on Z, which can estimate the surfaces of proteins, but not their secondary structures. Its output consists of three subclasses, i.e., buried, partially buried and exposed protein regions in the protein.

Ligand Binding Analysis by Protein Plus

In some areas, the protein binds with a ligand known as a ligand binding site. These ligands are essential for allosteric conformation changes in the protein function. The mutation will impair

ligand binding to protein at these specific sites. To analyse the change in the interaction of the ligand with protein, we apply UCSF and PROTEIN PLUS. Proteins Plus focuses on interactions between proteins and ligands at the binding site. It may also detect protein pockets, generate ensembles or predict metal coordination.

EFFECT OF PREDICTED MUTATIONS ON SPLICING

Splicing proteins identify specific locations, but splicing doesn't happen to owe to undiscovered sites if a mutation replaces it. An estimation of spliceogenic variations that affect pre-mRNA splicing have been identified as the critical index of splicing in the silicon analyses, interrupting 5 and 3' splice sites or changing regulatory components 5 and 3.' Recent research has analysed the locations of regions that have predicted a prediction by using SPiCE, Human Splice Finder (HSF), for each variant to predict (5') donors and (3") accepters.

Prediction of Splice Sites by SPiCE

We used mRNA (NM_004990) transcript, chromosome position and reference position as input in the SPiCE server and modified amino acid variants. The resulting classifier calculates labels in protein sequences after uploading the FASTA file. Then, SPICE calculates the accepter and donor *MARS 1* gene site, and SPICE interpretation evaluates the probability score. The output is in the form of MES and SSF-Like scores that are calculated along with graphical representation. The SPiCE probability score range from 0-1; the higher value more will be the probability of disrupting the splice region.

Prediction of the impact of SNP located in splice site by HSF tool.

Human Splice Finder (HSF) identifies as well as forecasts mutations to increase or repress splicing patterns, identifying splicing locations for the acceptor and donor, and also the branch point and auxiliary sequences known for Exonic Splicing Enhancers (ESE) and Exonic Splicing Silencers (ESS).

NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS

The following in silico tools are used to predict the effect of SNPs on protein

function: Meta-SNP, CAPICE, and CONDEL.

Predicting the deleterious nsSNPs by Meta-SNP

Meta-SNP trained as a random forest-based binary classifier on the output capacity of four tools: SNAP, SIFT, PANTHER and PhD-SNP. As a meta classifier of the tool, its prediction is more accurate than the single tool. The predicted pathogenicity of a mutation depends upon the value of the Reliability Index (RI). The range of RI value is from 0 to 1. The pathogenicity score with an RI value < 0.5 was predicted as pathogenic, while the others with an RI value > 0.5 were predicted as tolerated.

Predicting the functional impact of deleterious nsSNPs by CAPICE

CAPICE is a supervised machine-learning-based model for prioritising pathogenic variants, including SNVs and short InDels. It is trained on balanced data to annotate SNVs on 11 categorical features. Then, the retrieved variants are annotated accordingly to these features. Outputs are in the form of CAPICE scores predicting the pathogenicity of variations with a cutoff>0.02. **Predicting the functional impact of pathogenic nsSNPs by CONDEL**

CONDEL is a software tool that predicts copy number variation (CNV)**.** The classifier is based on a combined predictive score of (Log R Pfam E-value (Logre), MAPP, Mutation Assessor (Massessor), Polyphen2 (PPH2), and SIFT. All retrieved variants were analysed with a combined classifier with output from Score 0 to 1; the higher score was characterised as deleterious.

PREDICTION OF PROTEIN STABILITY CHANGE

In silico analysis on protein stability aims to calculate the difference in free energy upon protein folding. The impact of missense mutations on protein's stability was evaluated by change in Gibbs free energy ($\Delta\Delta G$).

Predicting the protein stability changed by DynaMut, DUET and iStable 2.0

Missense mutation introduces a new amino acid which sometimes is incompatible with neighbouring residue, destabilising the protein and affecting its

function. The retrieved variants were analysed with different stability prediction tools to predict the change in protein stability due to missense mutation. The DynaMut server predicts changes in free energy upon the folding and unfolding of the protein. Changes in free energy ($\Delta\Delta G$) of protein arise destabilised effect possessing threshold value >0.

PDB structure, wild type, mutant protein code and chain identification for visualising PDB structures and computing free-energy change $(\Delta\Delta G)$ are needed to calculate the protein stability by DUET servers. The higher negative value will indicate a more impact of destabilisation.

iStable2.0 predicted protein stability depending on the vector-supported machine algorithm. Two input types, structural or sequential protein information, can be supplied.

Prediction of energy minimisation and free energy change due to mutation by YASARA Fold X Program

YASARA software was downloaded along with plugin programs Fold X and Python by default installation to determine the effect of mutations on stability. The $\Delta\Delta G > 0$ value showed decreased stability while $\Delta\Delta G < 0$ value showed increased SNP stability.

PREDICTION OF PROTEIN STRUCTURAL PERSPECTIVE

Molecular Modeling by UCSF Chimera

For the confirmation of SNPs and 3D visualisation, we performed the analysis of the UCSF Chimera program. PDB ID fetched 3D structure (5GL7). The change in the amino acid at a specific position located in *MARS 1* chain A was generated by selecting an option of structure editing and then choosing rotamers. Native and mutant structures are visualised and downloaded as save PDB.

Clashes, Minimisation and labelling of molecule prediction

UCSF Chimera program found clashes that are unwanted interactions between native and mutant residue in Methionyl tRNA synthetase 1 protein. Energy minimisation is utilised to construct or refinish H- bond networks, eliminate unwanted contacts and lower total

system energy in protein molecular modelling. A 3-letter amino acid code labelled residues with their perspective position.

Prediction of Post Translational Modification Sites on protein structure by UniProt and Prosite Scan

These are all the alterations following mRNA protein translation. These changes are important for the functioning of proteins or any other enzymes to be identified. The PTM enzymes detected a particular consensus sequence or motifs for these changes. In these specific locations, random mutation can also occur for change and cannot be noticed through activating enzymes, resulting in no modification of enzymes that cannot enable the protein to work properly. We used ScanProsite to determine the PTM sites in the MARS1 Proteins to discover motifs, domains and interactions with other proteins in the FASTA or UniProt Sequence accession No. (P_56191) sequence. The results detailed changes in the location indicated.

ANALYSIS OF PROTEIN CONSERVATION

Prediction of Conserved residues by ConSurf

It is a method used in bioinformatics to compute the evolution of amino acid conservation by using an empirical Bayesian inference in the protein sequence. Assessment of the homologous sequences is dependent on evolutionary history. The importance of a residue can be observed through the conservational score: the more conserved residue, the more severe effect of its mutation on protein function. The degree of amino acid residue conservation with 50 homologous sequences was estimated. We have chosen those significantly conserved residues for additional study at the high-risk nsSNPs sites. The conservation score is given together with a colour scheme. For example, the most conserved is score 9, while score 1 is the most varied amino acid.

Prediction of solvent accessibility prediction by NetSurf -2.0

The tool predicted secondary structure, structural disorder, backbone dihedral

angles and solvent accessibility of amino acids to find the active site in the completely folded protein. Furthermore, characterising the position of residue in the protein, especially in the catalytic site, can be predicted by NetSurf -2.0. The mechanism for this prediction is based on Z, which can estimate the surfaces of proteins, but not their secondary structures. Its output consists of three subclasses, i.e., buried, partially buried and exposed protein regions in the protein.

Ligand Binding Analysis by Protein Plus

In some areas, the protein binds with a ligand known as a ligand binding site. These ligands are essential for allosteric conformation changes in the protein function. The mutation will impair ligand binding to protein at these specific sites. To analyse the change in the interaction of the ligand with protein, we apply UCSF and PROTEIN PLUS. Proteins Plus focuses on interactions between proteins and ligands at the binding site. It may also detect protein pockets, generate ensembles or predict metal coordination.

EFFECT OF PREDICTED MUTATIONS ON SPLICING

Splicing proteins identify specific locations, but splicing doesn't happen to owe to undiscovered sites if a mutation replaces it. An estimation of spliceogenic variations that affect pre-mRNA splicing have been identified as the critical index of splicing in the silicon analyses, interrupting 5 and 3' splice sites or changing regulatory components 5 and 3.' Recent research has analysed the locations of regions that have predicted a prediction by using SPiCE, Human Splice Finder (HSF), for each variant to predict (5') donors and (3") accepters.

Prediction of Splice Sites by SPiCE

We used mRNA (NM_004990) transcript, chromosome position and reference position as input in the SPiCE server and modified amino acid variants. The resulting classifier calculates labels in protein sequences after uploading the FASTA file. Then, SPICE calculates the accepter and donor *MARS 1* gene site, and SPICE interpretation evaluates the probability score. The output is in the form of MES and SSF-Like scores that

are calculated along with graphical representation. The SPiCE probability score range from 0-1; the higher value more will be the probability of disrupting the splice region.

Prediction of the impact of SNP located in splice site by HSF tool.

Human Splice Finder (HSF) identifies as well as forecasts mutations to increase or repress splicing patterns, identifying splicing locations for the acceptor and donor, and also the branch point and auxiliary sequences known for Exonic Splicing Enhancers (ESE) and Exonic Splicing Silencers (ESS).

RESULTS AND DISCUSSION

In silico analysis of deleterious nsSNPs were analysed by missense tools to predict the most pathogenic nsSNPs involved in disease. The resulting missense nsSNPs were performed with stability analysis to determine the stability changes upon mutation.

Analysis of Missense Variants

Total numbers of 492 variants were retrieved from these databases combined. 492 variants after applying VEP annotation filter and selecting missense variants with allelic frequency ≤ 0.001 and allelic count < 50 were then analysed using CADD. 308 variants were left after applying the filter on variants with CADD scoring ≥ 20. CAPICE, CONDEL and Meta-SNP were applied to measure the effect of single nucleotide polymorphism on pathogenicity and figure out the SNPs associated with the disease. Meta-SNP predicted 85 nsSNPs to be pathogenic with a threshold score of> 0.5. Our CAPICE analysis showed that 85 nsSNPs were identified with deleterious effects with a threshold score > 0.02.

Whereas CONDEL detected 85 nsSNPs as a pathogenic effect with a cutoff score > 0.5 on *MARS1* gene.

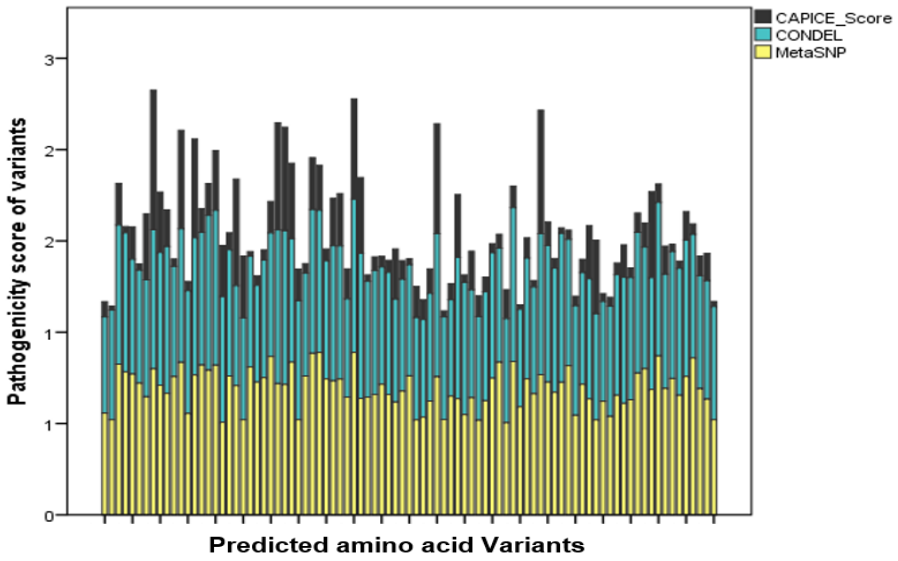

Fig.2 Graphical Representation of pathogenic missense variants in *MARS1*

After applying a combined cutoff filter of all missense analyses, we get 85 pathogenic variants. A study in 2019 was conducted in which Superoxide dismutase 3 (SOD3) were analysed through *in silico* analysis. The two mutations p.A91T and p.R231G were discovered to be deleterious for ligand binding analysis as a result of molecular dynamic simulation (Pereira et al., 2019).

Stability Analysis of Protein

Filter 85 missense variants were then analysed for stability change using protein stability predicting tools; DynaMut, DUET and iStable. The combined outcome of these tools demonstrates that 38 variants destabilised the protein structure, as shown in Fig. 3.

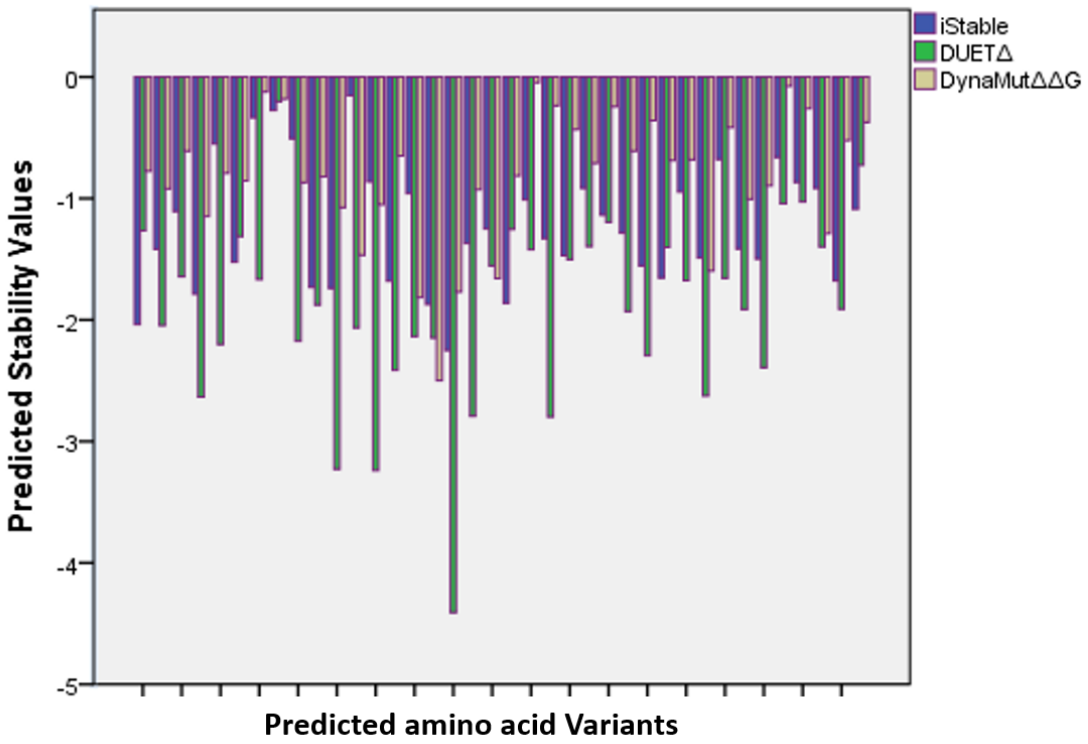

Fig. 3. Distribution of stabilising and destabilising nsSNPs upon stability change

Clashes Findings by UCSF Chimera

To visualise amino-acid mutations that disrupt the wild-type interaction with other residues, UCSF Chimera was used. Clashes which are unwanted interactions were identified when native amino acid was modified to a mutant amino acid for visualisation of this change in protein structure. Any change produced due to wild and mutant amino-acid residue may disturb the domain and have a loss of interactions that cause damage to protein structure. We analysed 85 filtered pathogenic variants, of which only 14 mutations showed clashes. Clashes between wild and mutant-type residues, as red-coloured lines between residues shown in Fig. 4. Results indicated that 14 mutations revealed clashes between wild and mutant type, i.e., *p.Ile285Phe*, *p.Gly310Trp*, *p.Pro329His*, *p.Cys408Trp*, *p.Ser484Pro*, *p.Phe551Leu*, *p.Tyr589His*, *p.Arg618His*,

p.Leu622Pro, *p.Arg625Trp*, *p.Leu681Pro*, *p.Ser712Phe*, *p.Asn716Lys*, *p.Gly796Val*.

Five mutations showed a higher no. of clashes (*p.Ile285Phe* (11 clashes), *p.Gly310Trp*(16 clashes), *p.Pro329His (*27 clashes), *p.Cys408Trp (*26 clashes), *p.Tyr589His* (26 clashes) have significantly affected the destabilisation of protein structure. Some mutations show a lesser number of clashes, such as *p.Ser484Pro* (7 clashes), *p.Phe551Leu* (4 clashes), *p.Arg618His* (8 clashes),), *p.Arg625Trp* (8 clashes), *p.Arg625Trp* (8 clashes) whereas the two remaining show only clash i.e.,.*p.Leu622Pro* (1 clash), and *p.Leu681Pro* (1 clash).

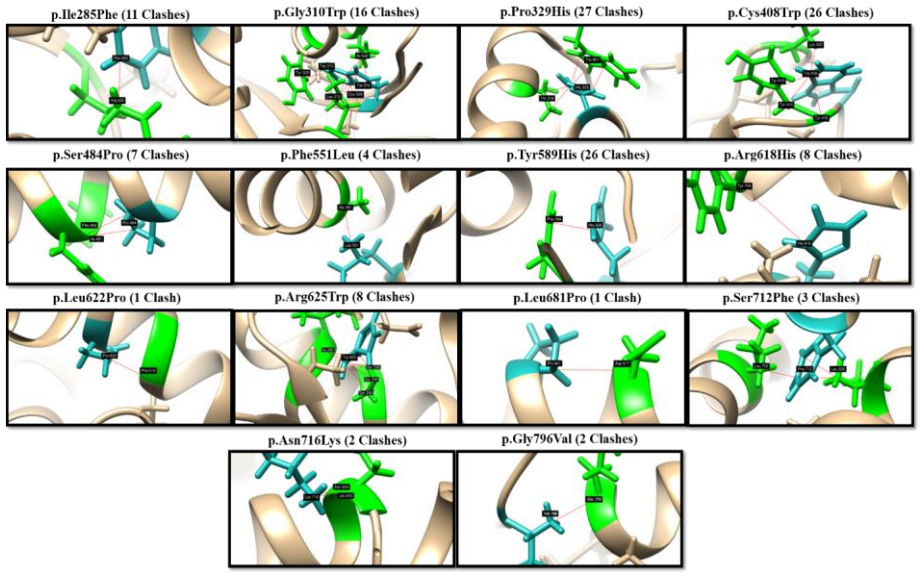

Fig. 4. Clashes finding due to mutations in *MARS 1* protein structure using chimera

Post Translational Modification (PTM) Analysis of *MARS1* Gene:

There have also been some mutations detected in protein Post-translational sites for modification. The 3 PTMs were found to be phosphorylation sites with the various types of kinases with their unique circumstances for our reported amino-acid substitution positions. The use of PROSITE (protein database) PTM (phosphorylation sites involving several kinases) has been confirmed, and these sites are monitored and visualised with UCSF Chimera protein structure. Fig.5a shows changes to the protein domain of methionyl-tRNA synthetase1 that involved UniProt predicted N-linked Glycosylation at p.Tyr532Cys. *Tyr532Cys* (ScanProsite site) showing (yellow) protein in fig.5a. As mutation stated, *p.Tyr532Cys* (a), Acidic Tyrosine, if mutated into Cysteine, is an uncharged amino acid and thus modifies the biochemical properties of nearby residues and may initiate the creation of a new N-linked Glycosylation site due to cysteine presence. In Fig. 5(b), ScanProsite predicted two phosphotheronin-implied sites at *p.Gln330Pro* and *p.Thr328Ile* highlighted in protein at (green). *p.Gln330Pro*, as indicated by mutation. Polar non charged glutamine mutating into proline which is uncharged amino acid, thus changing the residue's biochemical properties.

Another phosphorylation site (phosphothreonine) was observed at position 328 that may be disrupted by *p.Thr328Ile* reported mutations. (Orange and blue coloured site of the phosphothreonine affected. Also, in *p.Arg414Trp* Fig. 5(c) ScanProsite predicted one site participating in the RSD (cell attachment sequence).

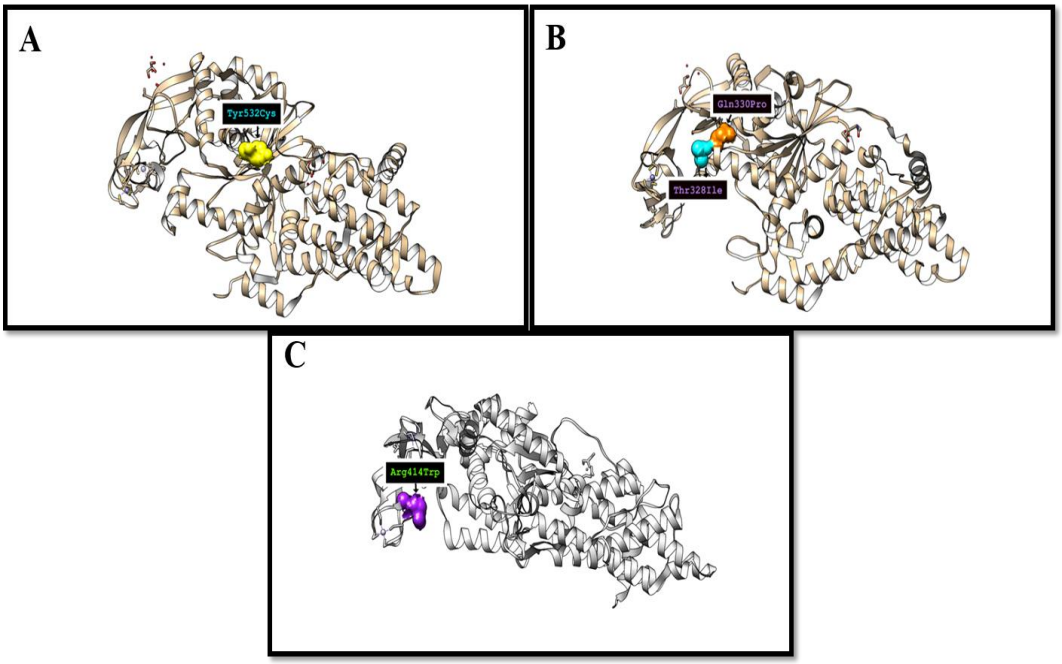

Fig. 5. Missense mutation highlighted in predicted PTMs sites

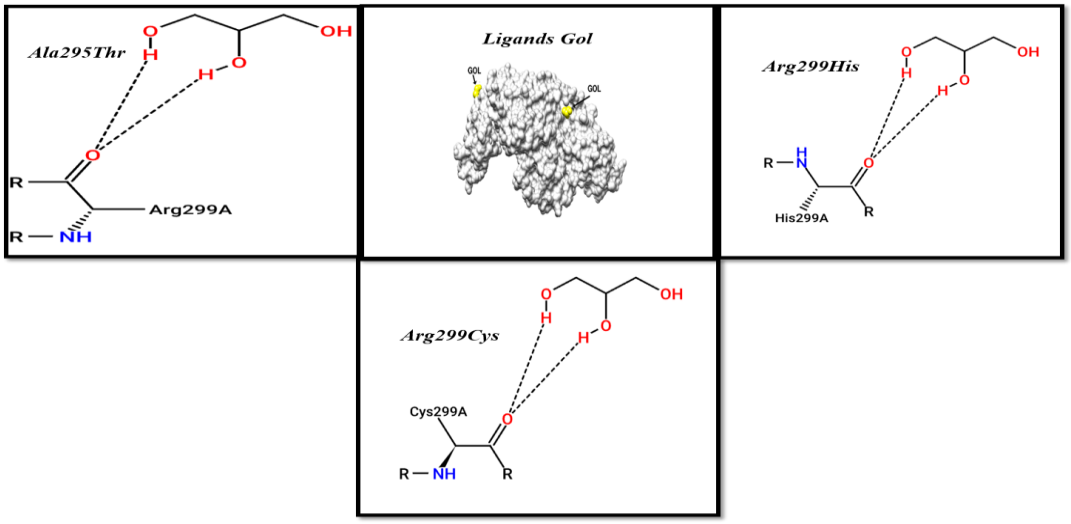

Fig. 6. The ligand highlighted with yellow (by using chimera), while the right end shows that residue **GOL_A_903** binding can interrupt the binding by mutation at *p.Arg299Cys, p.Arg299His, and p.Al295Thr* position

Protein Conservation and Secondary Structure Analysis of MARS1 Gene:

Predicted mutations at Ptm sites are then further validated by structural tools. The variant should be exposed and conserved for exposure to the modification enzyme. Consurf was used to anticipate our candidate protein structures for the conservation area or sequence. The more a protein is conserved, the more likely the protein is to mutate. In evolutionary research, conservation is vital. The slower the rate of conservation, the greater the protein structure will be conserved. The Consurf tool has been used to confirm whether the reported mutation is in the conservation area. NetSurf-2 anticipated secondary structures or to examine the buried or exposed amino acid on the surface.

Table.1 Protein conservation prediction using ConSurf and solvent accessibility, secondary structure prediction by NetSurfP-2.0

ConSurf			**NetSurfP-2.0**
Protein Change	**Score**	**Color**	**Sec. Structure/exposed/Buried**
p.Tyr532Cys	0.499	3	Exposed
p.Arg414Trp	-0.139	6	Exposed
p.Gln330Pro	-0.454	7	Exposed
p.Thr328Ile	-0.441	7	Exposed

Ligand Binding Analysis

The protein interacts with a ligand in certain places, known as a binding site. The mutation in these particular locations will disrupt ligand binding to protein. UCSF Chimera and PROTEIN PLUS are being used. The results showed that 3 variants at ligand binding sites could affect the binding of ligands to the protein. these variants p.Arg299His, p.Arg299Cys, and p.Ala295Thr as shown in the Fig.6.

Splice Site Variant analysis

10 variants were obtained from the gnomAD database to predict splice site defects.

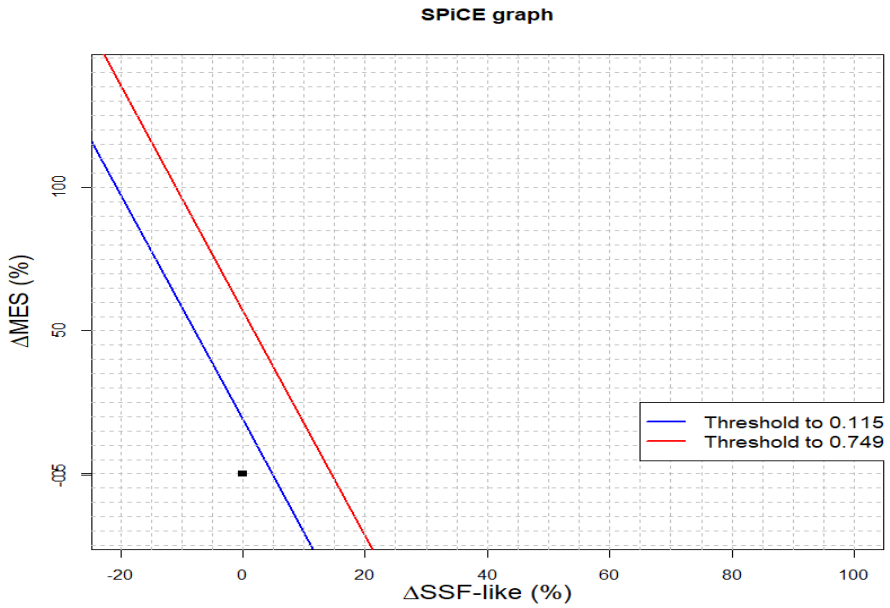

Fig.7 Graphical Representation of SPiCE Results

These variants have begun identifying the 5'/3' slice sites in the *MARS1* gene using several splicing tools. In 2014, a study was conducted to compare the missing rates of splice mutations which showed HSF with a minimum missing rate of 66 (Jian et al., 2014).

In this study utilising two bioinformatics methods to validate 5'/3' splice site (SPiCE and HSF). SPICE showed significant results in fig.7. However, HSF does not have a substantial *MARS1* gene interpretation suggesting that the splicing role of the *MARS1* gene is irrelevant.

The red colour schematic indicating the limit values and the points showing variations below the red scheme are the variants that are very unlikely or unlikely to alter the splicing mechanism. Our study can be a preliminary step for laboratory study. These identified variants can be the confident mutation of *MARS1* associated with Interstitial lung and liver disease. The structural analysis and functional analysis of these missense variants specify the effect of the variant on the protein.

CONCLUSION

Interstitial lung and liver disease is an autosomal recessive condition characterised by lipoprotein growth within the alveoli, leading to restrictive lung and respiratory failure. *MARS1* is a gene that codes for proteins. Interstitial Lung and Liver Disease associated with *MARS1*. An *in silico* study showed out of 492 retrieved variants 85 are significant highly pathogenic mutations with a valuable role in the function and structure of the *MARS1* gene. In addition, 38 missense pathogenic variants are responsible for disrupting the stability of protein structure. 14 variants have shown unwanted interactions with neighboring residues in the protein. Among highly pathogenic nsSNPs, only four nsSNPs predicted by PROSITE database to be in motif of Methionyl-tRNA Synthetase 1 protein and cause higher probability of interstitial lung and liver disease. *In silico* analysis in this study would help the researchers to understand the genetics of Interstitial lung and liver disorder and identify the mutations in *MARS1* gene that cause this disorder.

The reported SNPs analysis in *MARS1* gene can be further analysed by following wet lab experimental work or can be observed in animal models.

ACKNOWLEDGEMENTS

Authors acknowledge the Head of Department for the smooth conduct of the study.

CONFLICT OF INTEREST

The authors declared no conflict of interest.

REFERENCES

1. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010). A method and server for predicting damaging missense mutations. Nat. Methods. 7(4): 248-9.
2. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucl. Acids Res. 8: 44(W1):W344-50.
3. Bao S, Zhao H, Yuan J, Fan D, Zhang Z, Su J, Zhou M (2020). Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer. Brief. Bioinformat. 21(5): 1742-1755.
4. Binkley J, Karra K, Kirby A, Hosobuchi M, Stone EA, Sidow A (2010). ProPhylER: a curated online resource for protein function and structure based on evolutionary constraint analyses. Genome Res. 20(1): 142-154.
5. Blackstone C (2018). Hereditary spastic paraplegia. Handbook of clinical neurology 148: 633-652.
6. Chen CW, Lin J, Chu YW (2013). iStable: off-the-shelf predictor integration for predicting protein stability changes. Bio. Med. Cent. InBMC bioinformatics 14(2): 1-14.
7. Clifford RJ, Edmonson MN, Nguyen C, Buetow KH (2004). Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformat. 20(7): 1006-1014.

8. Fährrolfes R, Bietz S, Flachsenberg F, Meyder A, Nittinger E, Otto T, Volkamer A, Rarey M (2017). Proteins Plus: a web portal for structure analysis of macromolecules. Nucl. Acids Res. 45(W1): W337-W343.
9. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ (2006). The PROSITE database. Nucl. Acids Res. 34(1): D227-D230.
10. Jian X, Boerwinkle E, Liu X (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. Nucl. Acids Res. 42(22): 13534-13544.
11. Karczewski, K. J., et al. (2020). "The mutational constraint spectrum quantified from variation in 141,456 humans." Nature 581(7809): 434-443.
12. Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Gen. 46(3): 310-315.
13. Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Soenderby CK, Sommer MO, Winther O, Nielsen M, Petersen B, Marcatili P (2019). NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. Prot. 87(6): 520-527.
14. Kumar DT, Emerald LJ, Doss CG, Sneha P, Siva R, Jebaraj WC, Zayed H (2018). Computational approach to unravel the impact of missense mutations of proteins (D2HGDH and IDH2) causing D-2-hydroxyglutaric aciduria 2. Metabol. Brain Dis. 33(5): 1699-1710.
15. Kumar P, Henikoff S, Ng PC (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protocol. 4(7): 1073-1081.
16. Leman R, Gaildrat P, Le Gac G, Ka C, Fichou Y, Audrezet MP, Caux-Moncoutier V, Caputo SM, Boutry-Kryza N, Léone M, Mazoyer S (2020). Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined

in silico/in vitro studies: an international collaborative effort. Nucl. Acids Res. 48(3): 1600-1601.

17. Lenz D, Stahl M, Seidl E, Schöndorf D, Brennenstuhl H, Gesenhues F, Heinzmann T, Longerich T, Mendes MI, Prokisch H, Salomons GS (2020). Rescue of respiratory failure in pulmonary alveolar proteinosis due to pathogenic MARS1 variants. Pediat. Pulmonol. 55(11): 3057-3066.
18. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25(21): 2744-2750.
19. Li S, van der Velde KJ, De Ridder D, Van Dijk AD, Soudis D, Zwerwer LR, Deelen P, Hendriksen D, Charbon B, Van Gijn ME, Abbott K (2020). CAPICE: a computational method for Consequence-Agnostic Pathogenicity Interpretation of Clinical Exome variations. Genome Med. 12(1): 1-11.
20. Ng PC, Henikoff S (2001). "Predicting deleterious amino acid substitutions." Genome research 11(5): 863-874.
21. Pereira GRC, Da Silva AN R, Do Nascimento SS, De Mesquita JF (2019). In silico analysis and molecular dynamics simulation of human superoxide dismutase 3 (SOD3) genetic variants. J. Cell. Biochem. 120(3): 3583-3598.
22. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struc. Biol. 9(1): 1-10.
23. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004). UCSF Chimera—a visualisation system for exploratory research and analysis. J. Comput. Chem. 25(13): 1605-1612.
24. Pires DE, Ascher DB, Blundell TL (2014). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucl. Acids Res. 42(W1): W314-W319.

25. Reva B, Antipin Y, Sander C (2007). Determinants of protein function revealed by combinatorial entropy optimisation. Gen. Biol. 8(11): 1-15.
26. Rips J, Meyer-Schuman R, Breuer O, Tsabari R, Shaag A, Revel-Vilk S, Reif S, Elpeleg O, Antonellis A, Harel T (2018). MARS variant associated with both recessive interstitial lung and liver disease and dominant Charcot-Marie-Tooth disease. Europ. J. Med. Gen. 61(10): 616-620.
27. Rodrigues CH, Pires DE, Ascher DB (2018). DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucl. Acids Res. 46(W1): W350-W355.
28. Stone EA, Sidow A (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Gen. Res. 15(7): 978-986.
29. Tang R, Prosser DO, Love DR (2016). Evaluation of bioinformatic programmes for the analysis of variants within splice site consensus regions. Adv. Bioinform.