A comprehensive *in silico* analysis of deleterious SNPs of Paraplegin protein associated with heredity spastic paraplegia through Mitochondrial Dysfunction

Ammara Akhtar[a], Sobia Nazir Ch[b], Rana Muhammad Mateen[c], Mureed Hussain[d]

[a] Department of Life Sciences, University of Management and Technology, Lahore, Pakistan, 54770

Email Address: ammaraakhtar3@gmail.com, sobianazirch63@gmail.com , Muhammad.mateen@umt.edu.pk , Mureed.hussain@umt.edu.pk.

Corresponding Author: Mureed.hussain@umt.edu.pk , ammaraakhtar3@gmail.com

Abstract

Heredity spastic paraplegia is a heterogenous neurological disorder primarily associated with progressive spasticity in the lower limb area. The mutations in the some of the mitochondrial proteins are directly associated with heredity spastic paraplegia. Paraplegin is a mitochondrial protein, functional SNPs associated with this protein can lead to heredity spastic paraplegia. Over the past many years, massive efforts and high-throughput methods have been formulated to find the complicated heterogeneity, mutations, and functional variants associated with different diseases. The need for well-defined and organized strategies with tremendous genetic information is imperative in this research field. In this study, the *in silico* analysis was carried out to identify the pathogenic variants of SPG7 (Paraplegin protein) that lead to mitochondrial dysfunction and are associated with the heredity spastic paraplegia. To find novel mutation, the variants were collected from gnomAD database, which were then analyzed with CADD after applying the frequency filter (≤ 0.002). The results of CADD were then compared with several freely available bioinformatic tools to validate our findings. The mutations that affect the splicing region were also studied with *in sillico* tools. The impact of mutations on the structure of proteins were visualized with the help of Chimera. To find novel mutations in the recent work, the list of mutations was further examined and compared with the help of *in sillico* tool ClinVar.

Keyword: SPG7, Paraplegin, SNPs, Heredity spastic paraplegia, *in sillico* analysis

1. Introduction

HSP (hereditary spastic paraplegias) is a genetically neurodegenerative syndrome which is also known as Strumpell-Lorrain disease. The disease is described on the basis of one of the major

hallmarks, axonal degeneration, which later lead to abnormalities in corticospinal region. Because of corticospinal dysfunction, it will ultimately cause weakness and spasticity in the lower limb area (Depienne *et al.*, 2007;Giudice *et al.*, 2014).

The heterogeneity associated with the clinical manifestations of heredity spastic paraplegia reflects that, some proteins of HSP contribute in one of many processes or pathways at cellular and molecular level, so the abnormalities or mutations may spur defects in multiple processes (Morgan *et al.*, 2006;Paisan-Ruiz *et al.*, 2008). The heredity spastic paraplegia genes are also associated with mitochondrial functions, so mitochondrial abnormalities have been implicated in the development of neurological disorder.

Paraplegin (SPG7) shows autosomal recessive form of heredity spastic paraplegia, is a metalloprotease protein with 17 exons which have a AAA domain was first identified in 1998 (Casari *et al.*, 1998). The protein consists of 52 kilobases (De Michele *et al.*, 1998) and present in the inner membrane of mitochondria and get mutated in heredity spastic paraplegia. The protein has a prevalence rate of 2–6/100,000. It is primarily present with homologous AFG3L2 protein, which is a protease in nature, in the mitochondria. The main functions perform by the paraplegin is controlling the quality of protein, enzyme processing, maturation of mitochondrial protein and have role in the ribosomal assembly. The other important function played is the degradation of misfolded proteins. The mutation in paraplegin can lead to defects in the oxidative phosphorylation process of mitochondria and it has been seen in the muscle biopsies of HSP patient (Koppen *et al.*, 2007).The gene is mapped on chromosomal position 16q24.3, the gene code for a 795 amino acid protein named paraplegin and the size of the protein is 3.2 kb in both the fetus and adult tissues (Casari *et al.*, 1998).

The mutations in the SPG7 gene primarily contribute about 1.5-6 % of autosomal recessive form of heredity spastic paraplegia. This is associated with the both complicated and uncomplicated phenotypes with the onset of HSP from the early childhood age or even in adulthood (Elleuch *et al.*, 2006). The mutation in the protein can either reduce or diminish the activity of respiratory complex I, which is result cause increase in the oxidative stress by directly producing ROS. This will lead to less ATP production and ultimately cause loss of energy (Atorino *et al.*, 2003). The mutated mice with paraplegin can cause motor defects and neurodegeneration, lead to

mitochondrial dysfunction and swelling of axons (Ferreirinha *et al.,* 2004). The complication associated with the oxidative stress can be rescued with the exogenous production of paraplegin protein (Atorino *et al.*, 2003). Patients with mutation in the SPG7 gene have shown signs of cerebellar atrophy as well as optic neuropathy, are the key factors that lead to diagnosis of spastic paraplegia 7 gene (Klebe *et al.*, 2012).

The identification of the novel pathogenic variants is a serious challenge with a wide variety of variants being available. It leads to a major responsibility for the scientist to identify potentially damaging variants with the context of all available evidences. For this purpose, a variety of computer-aided bioinformatic tools are present that have the potential to predict the likely damaging or pathogenic ability of variants (Flanagan *et al.*, 2010).

To find out and predict that either an SNP will lead to disease or not, computational based bioinformatics tools can be used. Since the start, sequencing of the whole human genome, there are many projects being carried out to analyze and predict the effect of various genetic variation using computational based methods to understand the genetic basis of the disease. Until now, according to an estimate there are more than four million proven single nucleotide polymorphism (SNPs) in the human genome. This vast volume of available genetic variation aids in analyzing the basis of disease through bioinformatic tools. The bioinformatic tools are a great source for functional analysis on the basis of human genome (Mooney, 2005).

This study has been carried out to find novel mutations of gene *SPG7,* related to heredity spastic paraplegia through mitochondrial dysfunction with the help of *in silico* tools.

2. Material and methods

2.1. Data Retrieval

To study and identify mutations that lead to heredity spastic paraplegia, variants of SPG7 gene associated with heredity spastic paraplegia through mitochondrial dysfunction were retrieved through *in silico* analysis by using different bioinformatics tools like gnomAD (Genome Aggregation Database), dbSNP and Variation View. The protein sequence of paraplegin protein were retrieved from NCBI.

2.2. Variants Selection

For selecting variants, certain filters were applied for the selection of a set of variants for further analysis. In this study Loss of function, non-synonymous, missense and heterozygous mutations were selected through applying filters in gnomAD.

The frequency of 100 pathogenic mutation related to autosomal recessive form of heredity spastic paraplegia were collected from Variation View, and a cut off value of <0.002 allelic frequency was selected. The variants below 0.002 allelic frequency were further selected for CADD analysis.

2.3. CADD Analysis

The CADD tool was used for scoring the variants, different filters were applied to get desired variants. The cut-off value was selected for C-score (Phred score >15), and variants were further selected and analyzed with a variety of bioinformatics tools (Table 1). The analysis was done for the mutation in the missense and splicing sites.

2.4. Missense variant analysis

After C scoring, the missense variants were analyzed with PhD-SNP^{g}, PredictSNP2, UMD-predictor, SNP&GO and PROVEAN for further validation of results.

2.5. Splicing Variant analysis

The analysis of splicing SNPs of Paraplegin protein was done by using several computer- aided tools (Human Splice Finder, Spice and Spliceman). To find novel mutations in this study, the mutations obtained from this study were checked with already reported mutations in the ClinVar.

2.6. 3D Modelling of Paraplegin

To check the effect of certain mutations on the protein structure, function and its stability, modelling of the protein were performed through I-TASSER. The structural models of proteins were obtained through I-TASSER. Based on C-score and RMSD, model with high confidence level was selected. C-score basically shows a confidence level for calculating the quality of all predicted structural models of the protein. The score is ranges from -5 to 2, and higher the score value, the higher the confidence level of the predicted protein model.

TM-score as well as RMSD values provide a set of standards for estimating the level of similarity between two structures and to calculate the accuracy of the predicted model. If the structure of the protein is already not known, it is mandatory to evaluate the quality of the model. Based on C-

score, TM-score and RMSD is used to define the quality of the predicted model. In this study, the best model (first one) was selected among five models and further used for the analysis of mutation on the protein structure (Figure 1). For SPG7, the estimated C-score=-2.06, estimated TM-score = 0.47±0.15 and estimated RMSD = 13.5±4.0Å (Yang and Zhang, 2015).

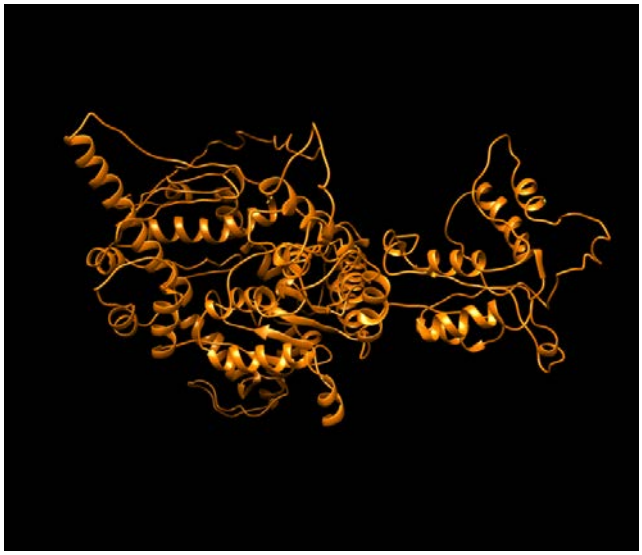


Figure 1. 3D Model of Paraplegin protein generated through I-TASSER

2.7. Effect of mutations on the stability of protein

The visualization of all protein structure was done with the help of UCSF-Chimera. The mutations were introduced in the structure by using target sequence in Chimera. Although mutation can be visualized in the structure using Chimera, but it does not give any sophisticated prediction whether the change in amino acid structure will affect the stability of function properties of the protein. To find the influence of certain mutations on the stability of protein, I-Stable was used.

2.8. Clashes and contacts

Clashes/contacts were also analyzed based on VDW radii to check if introduced mutation showed unfavorable interaction with nearby residues. Contacts tells that if there is any direct interaction either polar or non-polar between residues which may be necessarily not favorable, is present (Pettersen *et al.*, 2004).

2.9. Hydrophobicity analysis

Hydrophobicity analysis was done according to kdHydrophobicity scale by using Chimera.

Table 1. List of bioinformatic tools used to carry out variant analysis

	Program	Input	Output	URL	Reference
MUTATION ANALYSIS	**CADD**	List of variants in text tab delimited (VCF6) format	Integration of 60 different annotation Exhibit deleteriousness of Single nucleotide variants including other deletion or insertion variants	https://cadd.gs.washington.edu/	(Rentzsch *et al.*, 2018)
	PHD-SNP^{g}	List of variants in VCF, CSV or Mut file	Probability is greater than 0.5 means that variation is pathogenic or below this, its benign	http://snps.biofold.org/phd-snpg/index.html	(Capriotti and Fariselli, 2017)
	SNPs&GO	Protein sequence and amino acid substitution	Disease probability (if >0.5 mutation is predicted Disease)	http://snps.biofold.org/snps-and-go/snps-and-go.html	(Capriotti *et al.*, 2013)
	PROVEAN	Protein structure and amino acid substitution	the scoring threshold is -2.5, in which greater than -2.5 is neutral, While score smaller than -2.5 is deleterious	http://provean.jcvi.org/index.php	(Choi and Chan, 2015)
	UMD-PREDICTOR	Gene name, variant list of nucleotide substitution	<50 polymorphism; (ii) 50–64 probable polymorphism; (iii) 65–74 probably pathogenic mutation; and (iv) >74 pathogenic mutation.	http://umd-predictor.eu/	(Salgado *et al.*, 2016)
	PREDICT-SNP2	Variant list with nucleotide substitution	Score in the form of percentage. Red color exhibit deleteriousness. Score is based on the combination of consensus score based on the	https://loschmidt.chemi.muni.cz/predictsnp2/	(Bendl *et al.*, 2016)

			result achieved through 5 tools that exhibit best performance.		
SPLICING	**SPICE**	Variant list with nucleotide substitution	Score ranging from 0-1 (low-high) Exhibit 2 types of thresholds: optimal sensitivity threshold and optimal specificity threshold which are 0.115 and 0.749.	https://sourceforge.net/p/spicev2-1/wiki/SPICE%20wiki/	(Yeo and Burge, 2004)
	SPLICEMAN	>seq 5 flanking nucleotides (wildtype allele/mutant allele)5 flanking nucleotides	Score in the form of percentage. As the percentage increases, the chance of the mutation to disrupt the splicing process increases.	http://fairbrother.biomed.brown.edu/spliceman/	(Lim and Fairbrother, 2012)
	HUMAN SPLICE FINDER	Protein sequence and amino acid substitution	Impact of amino acid substitution on the splicing process, either introducing new splicing site or break a splicing site.	http://www.umd.be/HSF/HSF.shtml	(Desmet *et al.*, 2009)
Stability	**i-Stable**	Protein sequence and amino acid substitution	Score is either in negative or positive numbers, whereas negative number predict the destabilizing effect on the protein while positive number shows stabilizing effect	http://predictor.nchu.edu.tw/istable/	(Chen *et al.*, 2013)
Modelling	**I-TASSER**	Protein sequence	If the C-score value is higher, it means the high confidence level of the model.	https://zhanglab.ccmb.med.umich.edu/I-TASSER/	(Yang and Zhang, 2015)

3. Results

3.1. Missense variant analysis of SPG7

Variant for SPG7 gene were retrieved through gnomAD. Total 2021 variants were obtained. An allelic frequency filter (< 0.002) was applied, and a total of 1977 variants were further uploaded in the CADD. After CADD analysis, missense and PHRED score filter (≥15) were applied, 463 variants were achieved.

3.1.1. Deleterious missense SNPs identified through different bioinformatic tools

The obtained variants were further analyzed with different bioinformatic tools SNPs&GO, PHD-SNPg, PREDICTSNP2, PROVEAN and UMD-Predictor. The specific set of filters "deleterious, pathogenic, probably pathogenic" were applied to obtain highly pathogenic variants among variety of variants, 32 variants were obtained. The purpose of using these filters were to obtain those variants that were predicted to have an effect on the activity of gene by all the *in silico* tools. A cut off value was set for each software score value (PHD.SNP^{g}: 0.9, PROVEAN: -4.00, SNP&GO: 0.6, Predict SNP: 0.8. UMD-predictor: 80), 19 variants were obtained (Table 2).

3.1.2 Stability Predictions

The variants were further analyzed through I-Stable, to check the effect of these mutation on the stability of the protein. A filter of decrease was applied, and 9 variants were finally obtained (Table 3, Figure 2). The 3D model of protein was further used to analyze the clashes and contacts and visualized the mutation in the protein structure with the help of Chimera.

Table 2. List of high-pathogenic variants of paraplegin (*SPG7*) after applying cut off value (PHD-SNP^g ≥0.9: PROVEAN ≤-4.00: Predict-SNP2 ≥0.8: SNP&GOs ≥0.6: UMD-Predictor ≥80)

Chr	Pos	Ref	Alt	Substitution	PHD-SNP^g	Score	PROVEAN	Score	Predict-SNP2	Score	SNP&GO	Score	UMD	Prediction	Phred
16	89620349	T	C	p.Leu695Pro*	Pathogenic	0.988	Deleterious	-6.25	deleterious	1	disease	0.853	84	Pathogenic	17.79
16	89620279	G	A	p.Gly672Arg*	Pathogenic	0.996	Deleterious	-7.49	deleterious	1	disease	0.849	93	Pathogenic	18.39
16	89620261	G	C	p.Gly666Arg*	Pathogenic	0.982	Deleterious	-7.49	deleterious	1	disease	0.863	99	Pathogenic	18.51
16	89620241	A	G	p.Tyr659Cys	Pathogenic	0.97	Deleterious	-7.82	deleterious	1	disease	0.843	90	Pathogenic	18.56
16	89614447	T	C	p.Leu530Pro*	Pathogenic	0.996	Deleterious	-6.44	deleterious	1	disease	0.812	84	Pathogenic	22.7
16	89613160	G	A	p.Gly515Glu	Pathogenic	0.977	Deleterious	-7.78	deleterious	1	disease	0.838	100	Pathogenic	22.8
16	89613142	T	G	p.Leu509Arg	Pathogenic	0.997	Deleterious	-5.74	deleterious	1	disease	0.8	93	Pathogenic	22.9
16	89611095	C	T	p.Thr455Met	Pathogenic	0.988	Deleterious	-5.58	deleterious	1	disease	0.806	93	Pathogenic	23.5
16	89599042	A	G	p.Asp441Gly*	Pathogenic	0.993	Deleterious	-6.58	deleterious	1	disease	0.825	99	Pathogenic	23.7
16	89599030	T	G	p.Leu437Arg*	Pathogenic	0.994	Deleterious	-6	deleterious	1	disease	0.853	93	Pathogenic	23.8
16	89598969	C	T	p.Arg417Cys	Pathogenic	0.974	Deleterious	-7.79	deleterious	1	disease	0.824	99	Pathogenic	24.1
16	89598937	A	G	p.Tyr406Cys	Pathogenic	0.994	Deleterious	-8.34	deleterious	1	disease	0.867	93	Pathogenic	24.2
16	89598943	A	G	p.Asp408Gly	Pathogenic	0.995	Deleterious	-6.6	deleterious	1	disease	0.859	100	Pathogenic	24.2
16	89598918	C	T	p.Arg400Trp	Pathogenic	0.954	Deleterious	-7.23	deleterious	1	disease	0.813	96	Pathogenic	24.3
16	89598891	C	T	p.Arg391Trp*	Pathogenic	0.993	Deleterious	-7.77	deleterious	1	disease	0.841	93	Pathogenic	24.5
16	89598885	C	T	p.Arg389Cys*	Pathogenic	0.991	Deleterious	-7.7	deleterious	1	disease	0.842	100	Pathogenic	24.5
16	89598382	G	A	p.Cys353Tyr	Pathogenic	0.996	Deleterious	-10.09	deleterious	1	disease	0.856	100	Pathogenic	25
16	89598369	G	A	p.Gly349Arg	Pathogenic	0.999	Deleterious	-7.62	deleterious	1	disease	0.833	100	Pathogenic	25.2
16	89598370	G	T	p.Gly349Val	Pathogenic	0.991	Deleterious	-8.54	deleterious	1	disease	0.848	100	Pathogenic	25.2
16	89598355	G	A	p.Gly344Asp	Pathogenic	0.989	Deleterious	-6.64	deleterious	1	disease	0.845	90	Pathogenic	25.3
16	89598336	G	T	p.Gly338Cys	Pathogenic	0.992	Deleterious	-8.47	deleterious	1	disease	0.834	93	Pathogenic	25.4
16	89592798	G	C	p.Arg227Pro	Pathogenic	0.987	Deleterious	-6.21	deleterious	1	disease	0.847	100	Pathogenic	27.1
16	89590561	T	C	p.Leu175Pro*	Pathogenic	0.995	Deleterious	-6.61	deleterious	1	disease	0.829	84	Pathogenic	28.4

Asterisk (*) sign with amino acid substitutions indicates that the mutations are already reported in the ClinVar, all other mutations are novel.

Table 3. High risk pathogenic mutation identified through *in-silico* tools causing decrease in the stability of paraplegin (*SPG7*)

Chr	Pos	Ref	Alt	Substitution	i-Mutant 2.0	DDG	MUpro	Conf. Score	iStable
16	89598336	G	T	p.Gly338Cys	Decrease	-1.36	Decrease	-0.0666	Decrease
16	89599030	T	G	p.Leu437Arg	Decrease	-1.05	Decrease	-0.0123	Decrease
16	89598918	C	T	p.Arg400Trp*	Decrease	-0.39	Decrease	-0.77	Decrease
16	89599042	A	G	p.Asp441Gly	Decrease	-0.59	Decrease	-0.105	Decrease
16	89598885	C	T	p.Arg389Cys	Decrease	-1.01	Decrease	-0.7189	Decrease
16	89592798	G	C	p.Arg227Pro	Decrease	-0.65	Decrease	-0.4989	Decrease
16	89620349	T	C	p.Leu695Pro*	Decrease	-1.64	Decrease	-0.9667	Decrease
16	89614447	T	C	p.Leu530Pro	Decrease	-1.4	Decrease	-1	Decrease
16	89590561	T	C	p.Leu175Pro*	Decrease	-1.81	Decrease	-1	Decrease

Asterisk (*) indicates that the mutations are already reported in the ClinVar, all other mutations are novel.

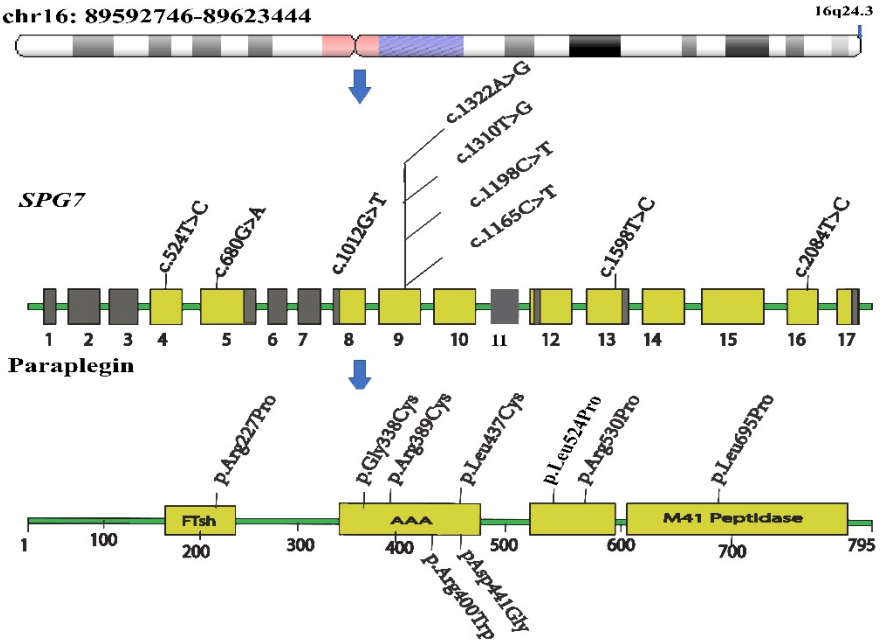


Figure 2. List of mutations of paraplegin from cDNA (SNV) to protein level (amino acid substitution). The gene position on the chromosome is 16q24.3. The *SPG7* gene constitute three main domains FTsh domain (144-237), AAA domain (341-481) and M41 domain (562-745).

3.2. Clashes and contacts

Clashes and contacts were analyzed to study the interaction of mutant residue with the neighboring residues by using Chimera. Four missense mutations exhibited interaction with neighboring residues. These interactions have the potential to destabilize the structure of protein, thus leading to disruption of the normal protein structure (Figure 3).

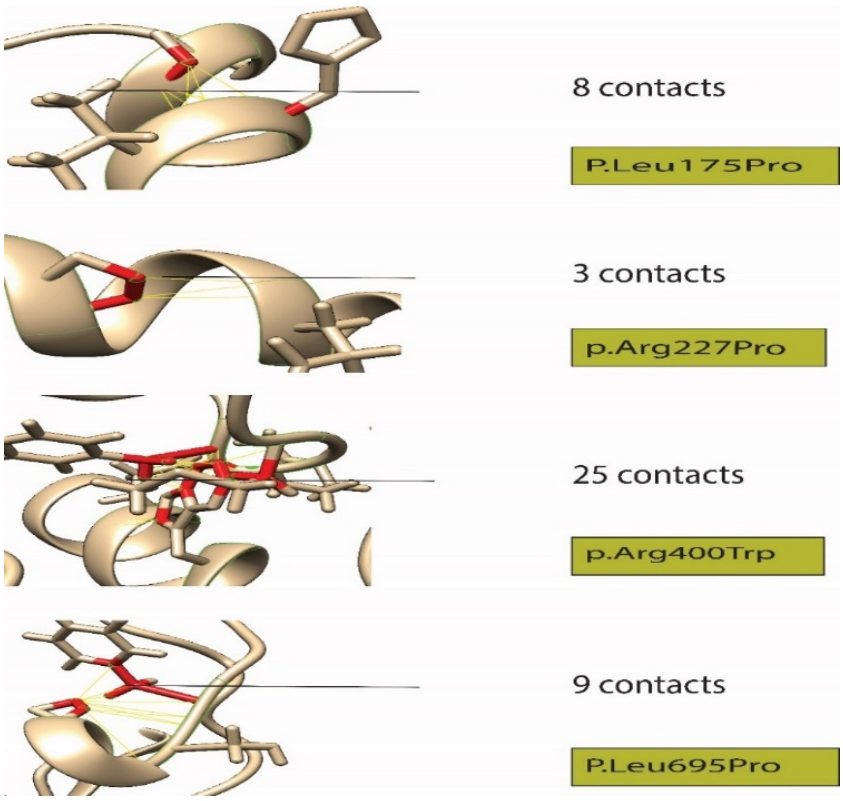
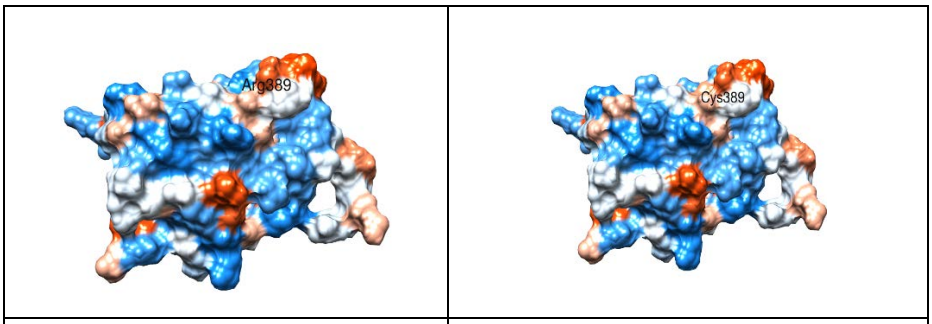


Figure 3. Analysis of clashes/contacts found in result of amino acid substitiution of paraplegin (*SPG7*)

3.3. Hydrophobicity analysis

Chimera was used to carry out the hydrophobicity analysis. In this process, amino acid residues were given characteristically features and properties according to kdHydrophobicity, values were given to each amino acid residue based on the hydrophobicity scale generated by Kyte and Doolittle. A color was allocated to each amino acid according to its hydrophilic or hydrophobic nature (Blue color: most hydrophilic, White: Neutral, orange red: most hydrophobic) along with a value according to hydropathy scale (Hydrophobicity value: More positive, hydrophobic: More negative, hydrophilic) (Table 4, Figure 4).



Variation: p.Arg389Cys Hydrophobicity value: -4.5	Variation: p.Arg389Cys Hydrophobicity value: 2.5

Figure 4. An example of Hydrophobicity analysis of AAS with color depiction ranging from blue to orange to red exhibiting the effect of mutation on the hydropathy properties of protein

Table 4. Hydrophobicity values of AAS of Paraplegin (*SPG7*) according to Kyte-Doolittle hydrophobicity scale

Chr	Pos	Ref	Alt	AAS	Hydrophobicity value	
					Wild	Mutant
16	89590561	T	C	p.Leu175Pro	3.8	-1.6
16	89592798	G	C	p.Arg227Pro	-4.5	-1.6
16	89598336	G	T	p.Gly338Cys	-0.4	2.5
16	89598885	C	T	p.Arg389Cys	-4.5	2.5
16	89598918	C	T	p.Arg400Trp	-4.5	-0.9
16	89599030	T	G	p.Leu437Arg	3.8	-4.5
16	89599042	A	G	p.Asp441Gly	-3.5	-0.4
16	89614447	T	C	p.Leu530Pro	-4.5	-1.6
16	89620349	T	C	p.Leu695Pro	-4.5	-1.6

3.4. Splicing SNPs identified through HSF, SPICE & SPLICEMAN

The variants were retrieved through gnomAD, a total of 2403 variants were obtained. By applying allelic frequency and PASS filter (< 0.002), 2351 variants were left, which were analyzed through CADD score. The CADD score filter (≥ 15) were applied, and a total of 148 variants were obtained which were further analyzed through applying filters (Canonical-Splice, Splice donor and Splice acceptor), 23 variants were obtained in the end. The variants were further analyzed through different bioinformatics tools named Human Splice Finder, Spicev2.15 and Spliceman to check the effect of the mutation on the splicing regions of the paraplegin (Table 5).

Table 5. Analysis of the effect of mutations on the splicing sites of the paraplegin (*SPG7*) through various bioinformatic tools

Chr	Position	Ref	Alt	Consequence	Spliceman		SPICE						HSF	
					Spliceman L1 distance	Ranking (L1)	SSF_ wt	SSF_ mut	MES _wt	MES _mut	SPiCEprobability	SPiCEinter_2thr	WT score	CV Variation (%)
16	89595988	G	C	c.861+1G>C	37791	81%	74.43	0	7.4	-0.87	1	high	82.4	-32.56
16	89597089	A	G	c.862-2A>G	35137	68%	90.22	0	9.4	1.44	1	high	89.5	-32.35
16	89590413	G	C	c.377-1G>C	33813	61%	78.2	0	5.04	-3.03	1	high	83.09	-2.21
16	89597217	G	A	c.987+1G>A	32812	56%	78.98	0	4.51	-3.67	1	high	83.39	-32.19
16	89613064	A	G	c.1450-2A>G*	35709	71%	85.92	0	6.82	-1.14	1	high	86.98	-33.28
16	89592877	G	T	c.758+1G>T	34295	64%	76.63	0	6.93	-1.57	1	high	66.43	-18.53
16	89598311	G	A	c.988-1G>A*	34233	64%	80.84	0	4.97	-3.78	1	high	84.68	-34.19
16	89614410	G	A	c.1553-1G>A	33968	62%	87.91	0	5.74	-3.01	1	high	54.55	+22.31
16	89619386	G	A	c.1780-1G>A	34955	67%	80.8	0	6.36	-2.39	1	high	46.24	+62.59
16	89598310	A	G	c.988-2A>G	35306	69%	80.84	0	4.97	-2.99	1	high	54.55	+22.31
16	89614409	A	G	c.1553-2A>G*	35112	68%	87.91	0	5.74	-2.22	1	high	88.53	-32.7
16	89613169	G	A	c.1552+2dupT	30624	45%	77.38	0	8.61	0.42	1	high	54.55	+14.25
16	89577001	G	A	c.286+1G>A	31794	51%	85	0	7.76	-0.42	1	high	89.21	-30.08
16	89613169	G	T	c.1552+2dupT	30624	45%	77.38	0	8.61	0.1	1	high	84.69	-31.69
16	89619545	T	A	c.1936+6_1936+8delGAG	32723	56%	84.36	0	8.3	0.11	1	high	49.15	+58.88
16	89620200	A	G	c.1937-2A>G*	37197	79%	84.31	0	7.25	-0.7	1	high	92.49	-31.29
16	89576897	G	T	c.184-1G>T	37529	80%	95.65	0	11.88	3.28	1	high	93.93	-30.82
16	89576896	A	T	c.184-2A>T	39870	92%	95.65	0	11.88	3.51	1	high	93.93	-30.82
16	89595883	A	G	c.759-2A>G	35041	68%	87.01	0	9.51	1.55	1	high	89.54	-32.33
16	89579446	G	A	c.376+1G>A	34153	63%	76.56	0	7.84	-0.34	1	high	78.08	-34.38
16	89579446	G	T	c.376+1G>T*	34795	66%	76.56	0	7.84	-0.66	1	high	41.15	+65.22
16	89579446	G	C	c.376+1G>C	30178	43%	76.56	0	7.84	-0.43	1	high	78.08	-34.38
16	89592735	A	G	c.619-2A>G	37927	82%	86.12	0	9.68	1.73	1	high	89.4	-3.45

The threshold value is set 65 in HSF. For a mutation, if the wild type score is above the set value and the variation score is below -10%, then it will break the splice site. And if the wild type score is below it and variation score is above +10%, the mutation will produce a new splice site. Asterisk (*) sign with amino acid substitutions indicates that the given mutations are already reported in the ClinVar, all other mutations are novel.

4. Discussion

Heredity spastic paraplegia is preliminary a heterogenous disorder associated with lower limb weakness followed by progressive spasticity and with passage of time, condition may get worsen. Presently, there has been not a particular effective therapy found for heredity spastic paraplegia, although there are drugs (anti-spastic effect) for example diazepam and baclofen that play an important role in lessening the complications including pain and fractures associated with the disease as well as improving the living of patients (Finsterer *et al.*, 2012;Klebe *et al.*, 2015). Today, there is a dire need to understand the pathophysiology associated with the disease as well as a detailed analysis of genetic mutation leading to heredity spastic paraplegia.

The importance of identifying novel variants in the field of genetic testing related to a particular disease is increasing day by day, ultimately producing prominent changes by enhancing the sensitivity of the genetic testing process (Judkins *et al.*, 2005). Although commonly used approaches for this purpose is gene sequencing, but due to the vast variety of newly found variants, it is really difficult to analyze the pathogenicity of each variant and also not every variant produce alteration in the functional properties of the protein (Campuzano *et al.*, 2015;Young and Fields, 2015). So, a detailed analysis in the field of genetic diagnosis is required, where variant should be tested *in-vitro* to study its effect (either deleterious or benign). However experimental methods are rather laborious, time taking, difficult to perform and expensive as well as it is not practicable to perform the validation of a very large number of variants through experimentation and find their impact on protein structure, function and stability.

To deal with the above mentioned limitations, and to find novel efficient methods and robust strategies to study the effect of variants on the function and structure of the protein, many computational methods have been developed in the past few years (Choi *et al.*, 2012;Gromiha, 2007;Tokuriki and Tawfik, 2009). With the passage of time, the importance of bioinformatic tools are increasing rapidly, as they provide an efficient and easy methods to get a detailed analysis of the structural and functional properties of the proteins.

In this study, variants of the genes associated with heredity spastic paraplegia were assembled through gnomAD (the Genome Aggregation Database). An allelic frequency filter (<0.002) was applied and below this, variants were selected for further analysis. CADD analysis was further done and CADD filter of ≥15 was applied. Then after selection, missense and splicing variants were analyzed through a variety of computer-aided tools.

Several bioinformatics tools (SNP&GO, PHD-SNPg, PredictSNP2, UMD-Predictor and PROVEAN) were utilized to investigate the pathogenicity associated with missense variants of Paraplegin protein. After a detailed analysis of missense variants for pathogenicity and their effect on stability of protein, highly pathogenic mutation, in case of paraplegin protein, a total number of 9 mutations were obtained. The mutation of SPG7 were analyzed using NM_003119.4. The mutations were visualized through Chimera.

Bioinformatics tools gain information about the protein through amino acid conservation methods by aligning and comparing the related sequences of the specific protein among different species. The functional properties of the protein are predicted in this manner through amino acid conservation, databases, structural alteration at amino acid level, annotations and through physicochemical analysis (Geourjon *et al.*, 2001;Laskowski *et al.*, 2016;Wallqvist *et al.*, 2000). In this work, to analyze the effect of missense mutation on the stability of protein related to heredity spastic paraplegia through mitochondrial dysfunction, I-Stable tool was used. For visualizing the mutations, Chimera was used. In Chimera, each mutation was visualized by introducing the amino acid substitution in the given sequence of the protein. Clashes\Contacts were also analyzed to see the type of interaction with the neighboring residues (Figure 3).

To analyze the impact of mutation on the splicing region of these mitochondrial genes, the bioinformatics tools named Spliceman, Human splice finder and SPICE were used. After thorough analysis, 25 mutation were identified in case of SPG7, through CADD analysis. All these mutations have the potential to produce deleterious effect in case of autosomal recessive form of heredity spastic paraplegia through mitochondrial dysfunction.

One of the continuous problems faced by researchers in the protein chemistry to understand and identify different forces that are directly involved in the protein folding process of polypeptide chains. The detailed examination of the protein structures can be used to understand the various kind of forces. It is primarily accepted that two opposite forces in the final structure of the protein. The hydrophilic chains have direct access to water while at the same time avoiding the contact of hydrophobic chains with aqueous medium. Through amino acid sequence of the protein, it can be understood whether a specific amino acid residue will reside interior or outside of the membrane. There are various empirical techniques that have been used to analyze the distribution of all the twenty amino acids in different conformation while formation of three- dimensional structure of a specific protein. By devising a scheme, the probability of secondary structures to adopt a general

shape can be evaluated. The amino acids can be either hydrophilic and hydrophobic based on their particular chemical characteristics, for example alanine (A), valine(V), leucine (L), proline (P), phenylalanine (F) and cysteine (Cys) are considered to be hydrophobic, while lysine and arginine are hydrophilic in nature, both consisting of positive charge at the neutral pH (Scheiner *et al.*, 2002). The SOAP is an online computer program which is written in C-language, it allocates a proper hydropathy value to every residue in a specific amino acid sequence. According to the scale, as the value increases, the more the hydrophobic the residue (Kyte and Doolittle, 1982).

To find novel mutation in this work, ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) was used to compare the variation with already reported data found in this database. It is basically publicly available archive which provide a platform for quick interpretation and analysis of genetic variants. It collects data of a wide range of genetic variants of clinically importance, which are mainly submitted by the testing labs, researchers, experts and other individual groups (Landrum *et al.*, 2017). In this study, after elaborated series of analysis, each variant was further analyzed in ClinVar to check the novelty of the work. In case of novel mutation found in SPG7 through this work, a total of 14 missense mutations and 18 splicing mutations were found which has been not previously reported in any kind of scientific work, thus, can be considered novel.

5. Conclusion

Heredity spastic paraplegia, a neurodegenerative disease, can be characterized with a number of disorders in which lower limb complexity associated with progressive spasticity are prefatory symptoms. The protein associated with mitochondrial dysfunction are directly involved in causing heredity spastic paraplegia. In this study, the mutation analysis of paraplegin protein (SPG7), was carried out through bioinformatic tools. The mutations related to paraplegin protein found in this study are directly involved in causing heredity spastic paraplegia through mitochondrial abnormalities. Different bioinformatic tools were utilized to analyze the effect of missense and splicing mutation on the structure and stability of the protein. The results showed that these mutations have the potential to cause autosomal recessive form of heredity spastic paraplegia and its associated complications. The mutations were further compared with the ClinVar data to find if they have been reported previously or not. The mutations reported in this work can be further analyzed for the *in vitro* analysis.

6. References

1. Atorino, L., Silvestri, L., Koppen, M., Cassina, L., Ballabio, A., Marconi, R., Langer, T. and Casari, G. 2003. Loss of m-AAA protease in mitochondria causes complex I deficiency and increased sensitivity to oxidative stress in hereditary spastic paraplegia. *The Journal of cell biology*, **163** (4): 777-787
2. Bendl, J., Musil, M., Štourač, J., Zendulka, J., Damborský, J. and Brezovský, J. 2016. PredictSNP2: a unified platform for accurately evaluating SNP effects by exploiting the different characteristics of variants in distinct genomic regions. *PLoS computational biology*, **12** (5): e1004962
3. Campuzano, O., Allegue, C., Fernandez, A., Iglesias, A. and Brugada, R. 2015. Determining the pathogenicity of genetic variants associated with cardiac channelopathies. *Scientific reports*, **5**: 7953
4. Capriotti, E., Calabrese, R., Fariselli, P., Martelli, P. L., Altman, R. B. and Casadio, R. 2013. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC genomics*, **14** (3): S6
5. Capriotti, E. and Fariselli, P. 2017. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic acids research*, **45** (W1): W247-W252
6. Casari, G., De Fusco, M., Ciarmatori, S., Zeviani, M., Mora, M., Fernandez, P., De Michele, G., Filla, A., Cocozza, S. and Marconi, R. 1998. Spastic paraplegia and OXPHOS impairment caused by mutations in paraplegin, a nuclear-encoded mitochondrial metalloprotease. *Cell*, **93** (6): 973-983
7. Choi, Y. and Chan, A. P. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, **31** (16): 2745-2747
8. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. and Chan, A. P. 2012. Predicting the functional effect of amino acid substitutions and indels. *PloS one*, **7** (10): e46688
9. De Michele, G., De Fusco, M., Cavalcanti, F., Filla, A., Marconi, R., Volpe, G., Monticelli, A., Ballabio, A., Casari, G. and Cocozza, S. 1998. A new locus for autosomal recessive hereditary spastic paraplegia maps to chromosome 16q24. 3. *The American Journal of Human Genetics*, **63** (1): 135-139
10. Depienne, C., Stevanin, G., Brice, A. and Durr, A. 2007. Hereditary spastic paraplegias: an update. *Current opinion in neurology*, **20** (6): 674-680

11. Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M. and Béroud, C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research*, **37** (9): e67-e67
12. Elleuch, N., Depienne, C., Benomar, A., Hernandez, A. O., Ferrer, X., Fontaine, B., Tallaksen, C., Zemmouri, R., Stevanin, G. and Durr, A. 2006. Mutation analysis of the paraplegin gene (SPG7) in patients with hereditary spastic paraplegia. *Neurology*, **66** (5): 654-659
13. Ferreirinha, F., Quattrini, A., Pirozzi, M., Valsecchi, V., Dina, G., Broccoli, V., Auricchio, A., Piemonte, F., Tozzi, G. and Gaeta, L. 2004. Axonal degeneration in paraplegin-deficient mice is associated with abnormal mitochondria and impairment of axonal transport. *The Journal of clinical investigation*, **113** (2): 231-242
14. Finsterer, J., Löscher, W., Quasthoff, S., Wanschitz, J., Auer-Grumbach, M. and Stevanin, G. 2012. Hereditary spastic paraplegias with autosomal dominant, recessive, X-linked, or maternal trait of inheritance. *Journal of the neurological sciences*, **318** (1-2): 1-18
15. Flanagan, S. E., Patch, A.-M. and Ellard, S. 2010. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genetic testing and molecular biomarkers*, **14** (4): 533-537
16. Geourjon, C., Combet, C., Blanchet, C. and Deléage, G. 2001. Identification of related proteins with weak sequence identity using secondary structure information. *Protein Science*, **10** (4): 788-797
17. Giudice, T. L., Lombardi, F., Santorelli, F. M., Kawarai, T. and Orlacchio, A. 2014. Hereditary spastic paraplegia: clinical-genetic characteristics and evolving molecular mechanisms. *Experimental neurology*, **261**: 518-539
18. Gromiha, M. 2007. Prediction of protein stability upon point mutations. *Biochemical Society Transactions*, **35** (6): 1569-1573
19. Judkins, T., Hendrickson, B. C., Deffenbaugh, A. M. and Scholl, T. 2005. Single nucleotide polymorphisms in clinical genetic testing: the characterization of the clinical significance of genetic variants and their application in clinical research for BRCA1. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **573** (1-2): 168-179

20. Klebe, S., Depienne, C., Gerber, S., Challe, G., Anheim, M., Charles, P., Fedirko, E., Lejeune, E., Cottineau, J. and Brusco, A. 2012. Spastic paraplegia gene 7 in patients with spasticity and/or optic neuropathy. *Brain*, **135** (10): 2980-2993
21. Klebe, S., Stevanin, G. and Depienne, C. 2015. Clinical and genetic heterogeneity in hereditary spastic paraplegias: from SPG1 to SPG72 and still counting. *Revue neurologique*, **171** (6-7): 505-530
22. Koppen, M., Metodiev, M. D., Casari, G., Rugarli, E. I. and Langer, T. 2007. Variable and tissue-specific subunit composition of mitochondrial m-AAA protease complexes linked to hereditary spastic paraplegia. *Molecular and cellular biology*, **27** (2): 758-767
23. Kyte, J. and Doolittle, R. F. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, **157** (1): 105-132
24. Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D. and Jang, W. 2017. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, **46** (D1): D1062-D1067
25. Laskowski, R. A., Tyagi, N., Johnson, D., Joss, S., Kinning, E., McWilliam, C., Splitt, M., Thornton, J. M., Firth, H. V. and Study, D. 2016. Integrating population variation and protein structural analysis to improve clinical interpretation of missense variation: application to the WD40 domain. *Human molecular genetics*, **25** (5): 927-935
26. Lim, K. H. and Fairbrother, W. G. 2012. Spliceman—a computational web server that predicts sequence variations in pre-mRNA splicing. *Bioinformatics*, **28** (7): 1031-1032
27. Mooney, S. 2005. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in bioinformatics*, **6** (1): 44-56
28. Morgan, N. V., Westaway, S. K., Morton, J. E., Gregory, A., Gissen, P., Sonek, S., Cangul, H., Coryell, J., Canham, N. and Nardocci, N. 2006. Erratum: PLA2G6, encoding a phospholipase A2, is mutated in neurodegenerative disorders with high brain iron (Nature Genetics (2006) 38 (752-754)). *Nature Genetics*, **38** (8): 957
29. Paisan-Ruiz, C., Dogu, O., Yilmaz, A., Houlden, H. and Singleton, A. 2008. SPG11 mutations are common in familial cases of complicated hereditary spastic paraplegia. *Neurology*, **70** (16 Part 2): 1384-1389

30. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. and Ferrin, T. E. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, **25** (13): 1605-1612
31. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. and Kircher, M. 2018. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, **47** (D1): D886-D894
32. Salgado, D., Desvignes, J. P., Rai, G., Blanchard, A., Miltgen, M., Pinard, A., Lévy, N., Collod-Béroud, G. and Béroud, C. 2016. UMD-predictor: a high-throughput sequencing compliant system for pathogenicity prediction of any human cDNA substitution. *Human mutation*, **37** (5): 439-446
33. Scheiner, S., Kar, T. and Pattanayak, J. 2002. Comparison of various types of hydrogen bonds involving aromatic amino acids. *Journal of the American Chemical Society*, **124** (44): 13257-13264
34. Tokuriki, N. and Tawfik, D. S. 2009. Stability effects of mutations and protein evolvability. *Current opinion in structural biology*, **19** (5): 596-604
35. Wallqvist, A., Fukunishi, Y., Murphy, L. R., Fadel, A. and Levy, R. M. 2000. Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics*, **16** (11): 988-1002
36. Yang, J. and Zhang, Y. 2015. Protein structure and function prediction using I-TASSER. *Current protocols in bioinformatics*, **52** (1): 5.8. 1-5.8. 15
37. Yeo, G. and Burge, C. B. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology*, **11** (2-3): 377-394
38. Young, D. L. and Fields, S. 2015. The role of functional data in interpreting the effects of genetic variation. *Molecular biology of the cell*, **26** (22): 3904-3908

Figures and Tables

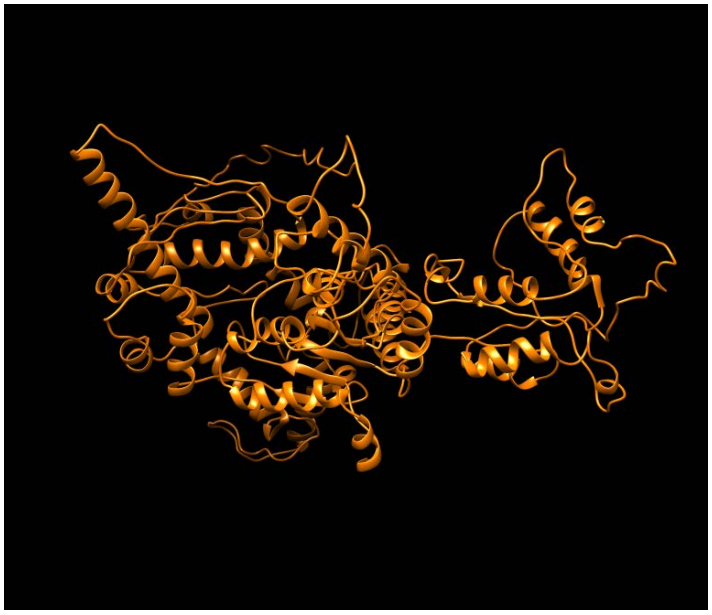


Figure 1. 3D Model of Paraplegin protein generated through I-TASSER

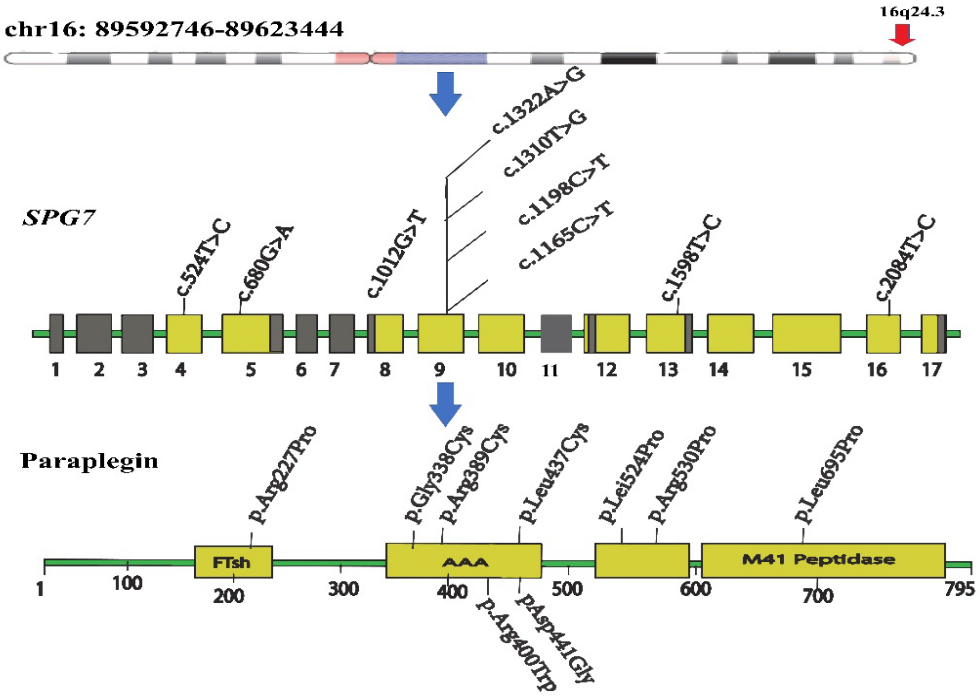


Figure 2. List of mutations of paraplegin from cDNA (SNV) to protein level (amino acid substitution). The gene position on the chromosome is 16q24.3. The *SPG7* gene constitute three main domains FTsh domain (144-237), AAA domain (341-481) and M41 domain (562-745)

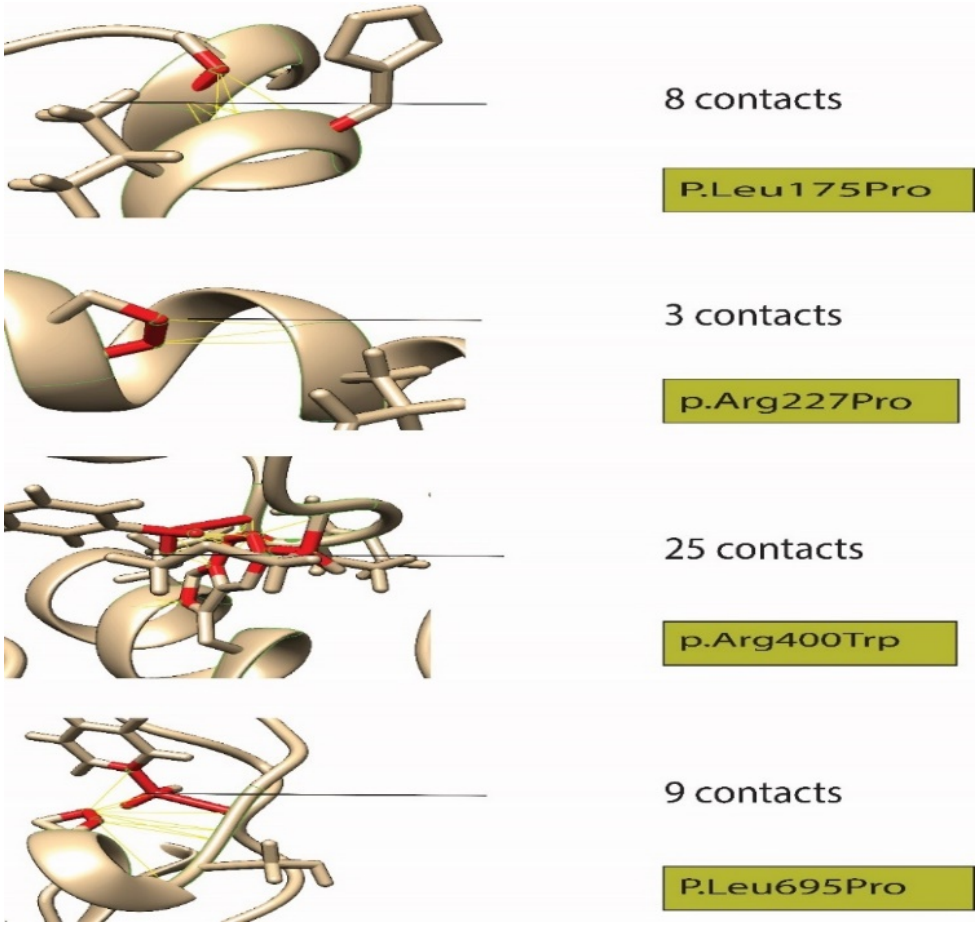


Figure 3. Analysis of clashes/contacts found in result of amino acid substitiution of paraplegin (*SPG7*)

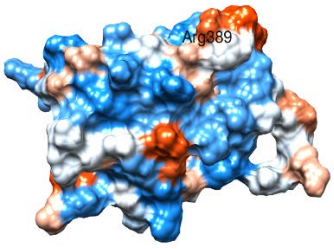
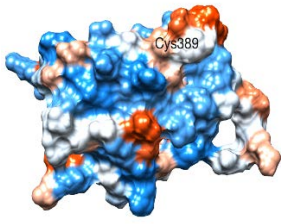
	
Variation: p.Arg389Cys Hydrophobicity value: -4.5	Variation: p.Arg389Cys Hydrophobicity value: 2.5

Figure 4. An example of Hydrophobicity analysis of AAS with color depiction ranging from blue to orange to red exhibiting the effect of mutation on the hydropathy properties of protein

Table 1. List of bioinformatic tools used to carry out variant analysis

	Program	Input	Output	URL	Reference
MUTATION ANALYSIS	**CADD**	List of variants in text tab delimited (VCF6) format	Integration of 60 different annotation Exhibit deleteriousness of Single nucleotide variants including other deletion or insertion variants	https://cadd.gs.washington.edu/	(Rentzsch *et al.*, 2018)
	PHD-SNP^{g}	List of variants in VCF, CSV or Mut file	Probability is greater than 0.5 means that variation is pathogenic or below this, its benign	http://snps.biofold.org/phd-snpg/index.html	(Capriotti and Fariselli, 2017)
	SNPs&GO	Protein sequence and amino acid substitution	Disease probability (if >0.5 mutation is predicted Disease)	http://snps.biofold.org/snps-and-go/snps-and-go.html	(Capriotti *et al.*, 2013)
	PROVEAN	Protein structure and amino acid substitution	the scoring threshold is -2.5, in which greater than -2.5 is neutral, While score smaller than -2.5 is deleterious	http://provean.jcvi.org/index.php	(Choi and Chan, 2015)
	UMD-PREDICTOR	Gene name, variant list of nucleotide substitution	<50 polymorphism; (ii) 50–64 probable polymorphism; (iii) 65–74 probably pathogenic mutation; and (iv) >74 pathogenic mutation.	http://umd-predictor.eu/	(Salgado *et al.*, 2016)
	PREDICT-SNP2	Variant list with nucleotide substitution	Score in the form of percentage. Red color exhibit deleteriousness. Score is based on the combination of consensus score based on the result achieved through 5 tools that exhibit best performance.	https://loschmidt.chemi.muni.cz/predictsnp2/	(Bendl *et al.*, 2016)
	SPICE	Variant list with nucleotide substitution	Score ranging from 0-1 (low-high)	https://sourceforge.net/p/spicev2-	(Yeo and Burge, 2004)

SPLICING			Exhibit 2 types of thresholds: optimal sensitivity threshold and optimal specificity threshold which are 0.115 and 0.749.	1/wiki/SPICE%20wiki/	
	SPLICEMAN	>seq 5 flanking nucleotides (wildtype allele/mutant allele)5 flanking nucleotides	Score in the form of percentage. As the percentage increases, the chance of the mutation to disrupt the splicing process increases.	http://fairbrother.biomed.brown.edu/spliceman/	(Lim and Fairbrother, 2012)
	HUMAN SPLICE FINDER	Protein sequence and amino acid substitution	Impact of amino acid substitution on the splicing process, either introducing new splicing site or break a splicing site.	http://www.umd.be/HSF/HSF.shtml	(Desmet *et al.*, 2009)
Stability	**i-Stable**	Protein sequence and amino acid substitution	Score is either in negative or positive numbers, whereas negative number predict the destabilizing effect on the protein while positive number shows stabilizing effect	http://predictor.nchu.edu.tw/istable/	(Chen *et al.*, 2013)
Modelling	**I-TASSER**	Protein sequence	If the C-score value is higher, it means the high confidence level of the model.	https://zhanglab.ccmb.med.umich.edu/I-TASSER/	(Yang and Zhang, 2015)

Table 2. List of high-pathogenic variants of paraplegin (*SPG7*) after applying cut off value (PHD-SNP^g ≥0.9: PROVEAN ≤-4.00: Predict-SNP2 ≥0.8: SNP&GOs ≥0.6: UMD-Predictor ≥80)

Chr	Pos	Ref	Alt	Substitution	PHD-SNP^g	Score	PROVEAN	Score	Predict-SNP2	Score	SNP&GO	Score	UMD	Prediction	Phred
16	89620349	T	C	p.Leu695Pro*	Pathogenic	0.988	Deleterious	-6.25	deleterious	1	disease	0.853	84	Pathogenic	17.79
16	89620279	G	A	p.Gly672Arg*	Pathogenic	0.996	Deleterious	-7.49	deleterious	1	disease	0.849	93	Pathogenic	18.39
16	89620261	G	C	p.Gly666Arg*	Pathogenic	0.982	Deleterious	-7.49	deleterious	1	disease	0.863	99	Pathogenic	18.51
16	89620241	A	G	p.Tyr659Cys	Pathogenic	0.97	Deleterious	-7.82	deleterious	1	disease	0.843	90	Pathogenic	18.56
16	89614447	T	C	p.Leu530Pro*	Pathogenic	0.996	Deleterious	-6.44	deleterious	1	disease	0.812	84	Pathogenic	22.7
16	89613160	G	A	p.Gly515Glu	Pathogenic	0.977	Deleterious	-7.78	deleterious	1	disease	0.838	100	Pathogenic	22.8
16	89613142	T	G	p.Leu509Arg	Pathogenic	0.997	Deleterious	-5.74	deleterious	1	disease	0.8	93	Pathogenic	22.9
16	89611095	C	T	p.Thr455Met	Pathogenic	0.988	Deleterious	-5.58	deleterious	1	disease	0.806	93	Pathogenic	23.5
16	89599042	A	G	p.Asp441Gly*	Pathogenic	0.993	Deleterious	-6.58	deleterious	1	disease	0.825	99	Pathogenic	23.7
16	89599030	T	G	p.Leu437Arg*	Pathogenic	0.994	Deleterious	-6	deleterious	1	disease	0.853	93	Pathogenic	23.8
16	89598969	C	T	p.Arg417Cys	Pathogenic	0.974	Deleterious	-7.79	deleterious	1	disease	0.824	99	Pathogenic	24.1
16	89598937	A	G	p.Tyr406Cys	Pathogenic	0.994	Deleterious	-8.34	deleterious	1	disease	0.867	93	Pathogenic	24.2
16	89598943	A	G	p.Asp408Gly	Pathogenic	0.995	Deleterious	-6.6	deleterious	1	disease	0.859	100	Pathogenic	24.2
16	89598918	C	T	p.Arg400Trp	Pathogenic	0.954	Deleterious	-7.23	deleterious	1	disease	0.813	96	Pathogenic	24.3
16	89598891	C	T	p.Arg391Trp*	Pathogenic	0.993	Deleterious	-7.77	deleterious	1	disease	0.841	93	Pathogenic	24.5
16	89598885	C	T	p.Arg389Cys*	Pathogenic	0.991	Deleterious	-7.7	deleterious	1	disease	0.842	100	Pathogenic	24.5
16	89598382	G	A	p.Cys353Tyr	Pathogenic	0.996	Deleterious	-10.09	deleterious	1	disease	0.856	100	Pathogenic	25

16	89598369	G	A	p.Gly349Arg	Pathogenic	0.999	Deleterious	-7.62	deleterious	1	disease	0.833	100	Pathogenic	25.2
16	89598370	G	T	p.Gly349Val	Pathogenic	0.991	Deleterious	-8.54	deleterious	1	disease	0.848	100	Pathogenic	25.2
16	89598355	G	A	p.Gly344Asp	Pathogenic	0.989	Deleterious	-6.64	deleterious	1	disease	0.845	90	Pathogenic	25.3
16	89598336	G	T	p.Gly338Cys	Pathogenic	0.992	Deleterious	-8.47	deleterious	1	disease	0.834	93	Pathogenic	25.4
16	89592798	G	C	p.Arg227Pro	Pathogenic	0.987	Deleterious	-6.21	deleterious	1	disease	0.847	100	Pathogenic	27.1
16	89590561	T	C	p.Leu175Pro*	Pathogenic	0.995	Deleterious	-6.61	deleterious	1	disease	0.829	84	Pathogenic	28.4

Asterisk (*) sign with amino acid substitutions indicates that the mutations are already reported in the ClinVar, all other mutations are novel.

Table 3. High risk pathogenic mutation identified through *in-silico* tools causing decrease in the stability of paraplegin (*SPG7*)

Chr	Pos	Ref	Alt	Substitution	i-Mutant 2.0	DDG	MUpro	Conf. Score	iStable
16	89598336	G	T	p.Gly338Cys	Decrease	-1.36	Decrease	-0.0666	Decrease
16	89599030	T	G	p.Leu437Arg	Decrease	-1.05	Decrease	-0.0123	Decrease
16	89598918	C	T	p.Arg400Trp*	Decrease	-0.39	Decrease	-0.77	Decrease
16	89599042	A	G	p.Asp441Gly	Decrease	-0.59	Decrease	-0.105	Decrease
16	89598885	C	T	p.Arg389Cys	Decrease	-1.01	Decrease	-0.7189	Decrease
16	89592798	G	C	p.Arg227Pro	Decrease	-0.65	Decrease	-0.4989	Decrease
16	89620349	T	C	p.Leu695Pro*	Decrease	-1.64	Decrease	-0.9667	Decrease
16	89614447	T	C	p.Leu530Pro	Decrease	-1.4	Decrease	-1	Decrease
16	89590561	T	C	p.Leu175Pro*	Decrease	-1.81	Decrease	-1	Decrease

Asterisk (*) indicates that the mutations are already reported in the ClinVar, all other mutations are novel.

Table 4. Hydrophobicity values of AAS of Paraplegin (*SPG7*) according to Kyte-Doolittle hydrophobicity scale

Chr	Pos	Ref	Alt	AAS	Hydrophobicity value	
					Wild	Mutant
16	89590561	T	C	p.Leu175Pro	3.8	-1.6
16	89592798	G	C	p.Arg227Pro	-4.5	-1.6
16	89598336	G	T	p.Gly338Cys	-0.4	2.5
16	89598885	C	T	p.Arg389Cys	-4.5	2.5
16	89598918	C	T	p.Arg400Trp	-4.5	-0.9
16	89599030	T	G	p.Leu437Arg	3.8	-4.5
16	89599042	A	G	p.Asp441Gly	-3.5	-0.4
16	89614447	T	C	p.Leu530Pro	-4.5	-1.6
16	89620349	T	C	p.Leu695Pro	-4.5	-1.6

Table 5. Analysis of the effect of mutations on the splicing sites of the paraplegin (*SPG7*) through various bioinformatic tools

Chr	Position	Ref	Alt	Consequence	Spliceman		SPICE						HSF	
					Spliceman L1 distance	Ranking (L1)	SSF_ wt	SSF_ mut	MES _wt	MES _mut	SPiCEprobability	SPiCEinter_2thr	WT score	CV Variation (%)
16	89595988	G	C	c.861+1G>C	37791	81%	74.43	0	7.4	-0.87	1	high	82.4	-32.56
16	89597089	A	G	c.862-2A>G	35137	68%	90.22	0	9.4	1.44	1	high	89.5	-32.35
16	89590413	G	C	c.377-1G>C	33813	61%	78.2	0	5.04	-3.03	1	high	83.09	-2.21
16	89597217	G	A	c.987+1G>A	32812	56%	78.98	0	4.51	-3.67	1	high	83.39	-32.19
16	89613064	A	G	c.1450-2A>G*	35709	71%	85.92	0	6.82	-1.14	1	high	86.98	-33.28
16	89592877	G	T	c.758+1G>T	34295	64%	76.63	0	6.93	-1.57	1	high	66.43	-18.53
16	89598311	G	A	c.988-1G>A*	34233	64%	80.84	0	4.97	-3.78	1	high	84.68	-34.19
16	89614410	G	A	c.1553-1G>A	33968	62%	87.91	0	5.74	-3.01	1	high	54.55	+22.31
16	89619386	G	A	c.1780-1G>A	34955	67%	80.8	0	6.36	-2.39	1	high	46.24	+62.59
16	89598310	A	G	c.988-2A>G	35306	69%	80.84	0	4.97	-2.99	1	high	54.55	+22.31
16	89614409	A	G	c.1553-2A>G*	35112	68%	87.91	0	5.74	-2.22	1	high	88.53	-32.7
16	89613169	G	A	c.1552+2dupT	30624	45%	77.38	0	8.61	0.42	1	high	54.55	+14.25

16	89577001	G	A	c.286+1G>A	31794	51%	85	0	7.76	-0.42	1	high	89.21	-30.08
16	89613169	G	T	c.1552+2dupT	30624	45%	77.38	0	8.61	0.1	1	high	84.69	-31.69
16	89619545	T	A	c.1936+6_1936+8delGAG	32723	56%	84.36	0	8.3	0.11	1	high	49.15	+58.88
16	89620200	A	G	c.1937-2A>G*	37197	79%	84.31	0	7.25	-0.7	1	high	92.49	-31.29
16	89576897	G	T	c.184-1G>T	37529	80%	95.65	0	11.88	3.28	1	high	93.93	-30.82
16	89576896	A	T	c.184-2A>T	39870	92%	95.65	0	11.88	3.51	1	high	93.93	-30.82
16	89595883	A	G	c.759-2A>G	35041	68%	87.01	0	9.51	1.55	1	high	89.54	-32.33
16	89579446	G	A	c.376+1G>A	34153	63%	76.56	0	7.84	-0.34	1	high	78.08	-34.38
16	89579446	G	T	c.376+1G>T*	34795	66%	76.56	0	7.84	-0.66	1	high	41.15	+65.22
16	89579446	G	C	c.376+1G>C	30178	43%	76.56	0	7.84	-0.43	1	high	78.08	-34.38
16	89592735	A	G	c.619-2A>G	37927	82%	86.12	0	9.68	1.73	1	high	89.4	-3.45

The threshold value is set 65 in HSF. For a mutation, if the wild type score is above the set value and the variation score is below -10%, then it will break the splice site. And if the wild type score is below it and variation score is above +10%, the mutation will produce a new splice site. Asterisk (*) sign with amino acid substitutions indicates that the given mutations are already reported in the ClinVar, all other mutations are novel.