An Analysis of Gap Junction Beta-1 Protein

1. Introduction:

Gap Junction Beta-1(GJB1) protein, also known as Connexin-32, is a membrane spanning protein. Six connexins assemble into hexa-dimers to form gap junction channels which play a crucial role in the transfer of ions and small molecules between cells (McKusick, "Gap Junction Protein, Beta-1; GJB1). Since connexin proteins can be considered the building blocks of gap junctions, which are crucial for intracellular communication, changes in the shape and/or function of these proteins have the capacity to be pathogenic. An example of this is the Charcot Marie Tooth Disease X-Linked Dominant 1 (CMTX1). CMTX1 is caused by mutations to the GJB1 gene, which makes the GJB1 protein. Mutations to the gene cause a variety of phenotypes to the protein, some are smaller than usual, some are inhibited in their ability to form gap junctions and many more that are still being studied(McKusick). The GJB1 protein plays a role in forming gap junctions between Schwann cells in the peripheral nervous system and when it does not function correctly it leads to the demyelination of axons(McKusick, "Gap Junction Protein, Beta-1; GJB1). Demyelinated axons have a slow signaling rate therefore, patients with CMTX1 are unable to effectively control their arms, hands, legs and feet. Exome sequencing can be used to diagnose CMTX1; however, there are also common features such as onset between the ages 14 and 40, muscle weakness and atrophy in the upper and lower limbs(McKusick). Hereditary neuropathies are approximately 30 in 100,000 and CMTX1 accounts for 10-20% (Abrams). The aim of this project is to identify the most conserved domains of the GJB1 protein in order to determine the most conserved residues, on which pathogenic mutations can occur, and model possible mutations on the protein.

2. Results:

Literature review showed that there are five main ways in which mutations to the Gap Junction Beta 1 protein can present itself. Firstly, the protein may not be synthesized at all. The protein may be synthesized in limited amounts which could possibly explain the "late" onset of the disease. A mutant version of the GJB1 protein can be synthesized, which either does not get properly transported and accumulates in the Golgi apparatus and/or Endoplasmic reticulum or they can be transported but unable to form gap junctions. Lastly, mutant proteins can form channels but in limited number (Bortolozzi). These mutations can be caused due to mutations on the GJB1 gene that either affects the production or function of the protein.

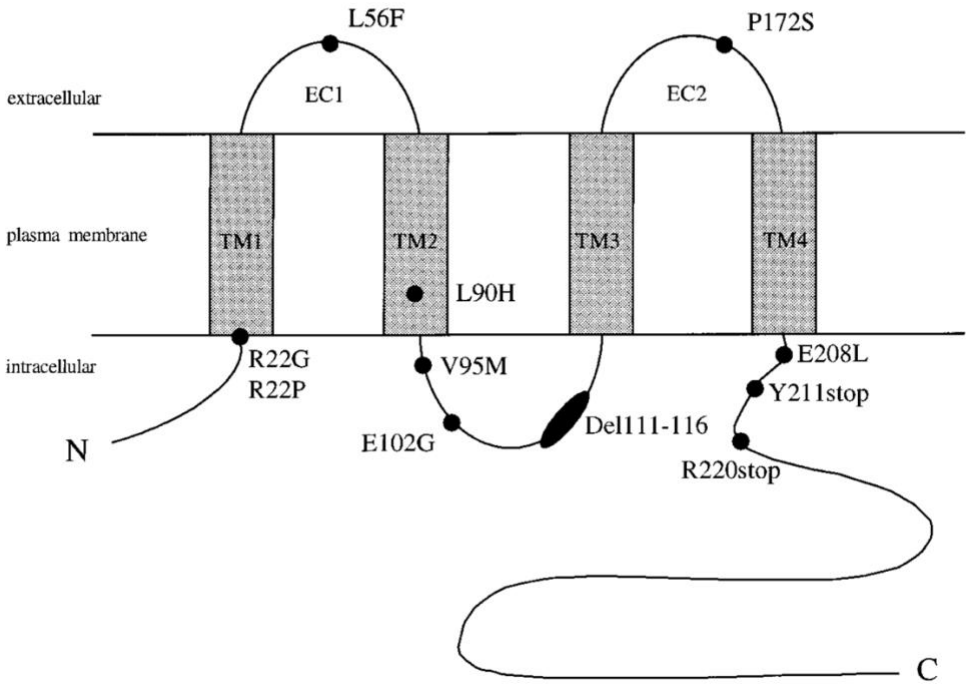

Figure 1: Topological regions of the Gap Junction Beta 1 protein

In Figure 1, the topological regions, structure and some reported mutations, in the paper it was taken from, of the GJB1 protein are shown (Sohl et al.). According to UniProt from amino acids 1-22 is the first topological cytoplasmic region, 23-45 is the first transmembrane region, 46-75 is the first extracellular topological region, 76-95 is the second transmembrane region, 96-130 is the second topological cytoplasmic region, 131-153 is the third transmembrane region, 154-191 is the second extracellular region, 192-214 is the fourth transmembrane region and 215-283 is the cytoplasmic tail. All of the transmembrane regions are helical. According to SMART some important domains of the protein are the CNX (connexin) domain which is between amino acids 42-75, transmembrane region two 76-95, Connexin_CCC (cysteine rich domain) from 145-212 and low complexity domain 215-224.

In order to determine the most conserved domain or possibly to find a new one Shannon Entropy test was conducted using a python script. According to Shannon Entropy the H value ranges from 0 to 4.22. Positions where H is greater than two are considered variable while positions where H is smaller than two are considered conserved and positions where H is between zero and one are considered highly conserved ("Sequence Variability Server"). Shannon Entropy test was applied to the multiple sequence alignment found from NCBI's MSAviewer. The average Shannon Entropy scores for the known topological regions were calculated and the most conserved topological region was observed. A plot was made using R studio ggplot.

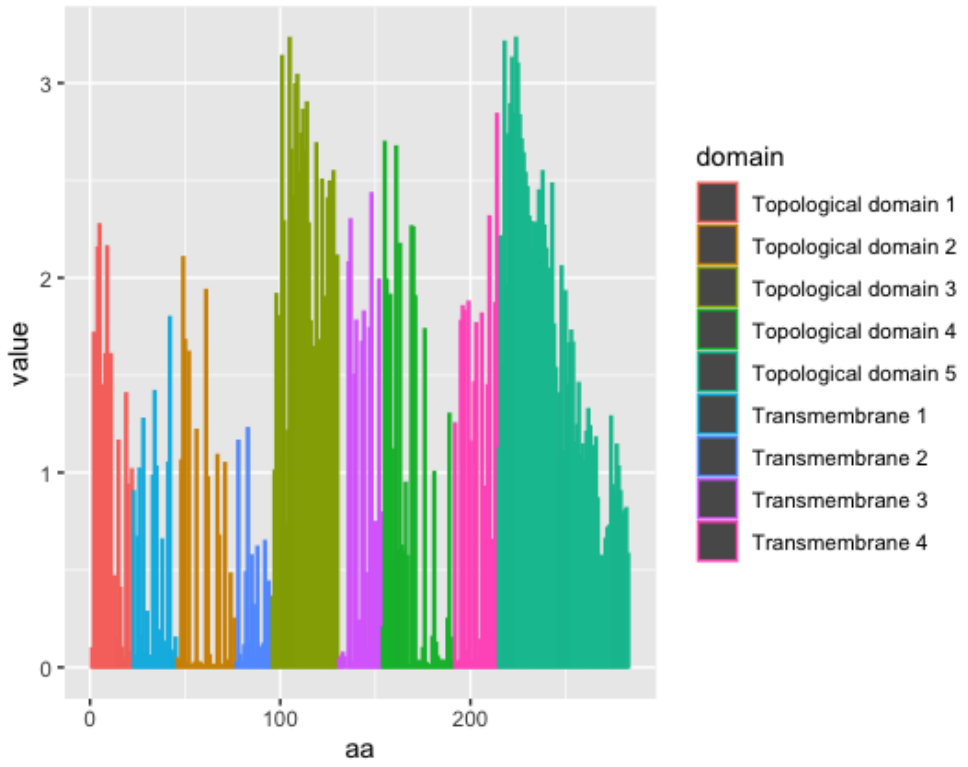

Figure 2: Distribution of Shannon Entropy Scores

Figure 2 shows the distribution of entropy scores for amino acid residues. The colors indicate the known regions of the GJB1 protein and their borders. According to the python script the average entropy scores for the known regions are as follows; the first topological cytoplasmic region had an average score of 0.954, the first transmembrane region had an average score of 0.534, the first extracellular topological region had an average score of 0.469, the second transmembrane region had an average score of 0.347, the second topological cytoplasmic region had an average score of 1.948, the third transmembrane region had an average score of 0.885, the second extracellular region had an average score of 0.746, the fourth transmembrane region had an average score of 1.100 and the cytoplasmic tail had an average score of 1.677.

According to the results from Shannon entropy the most conserved region on the protein was the second transmembrane region followed closely by the first extracellular region. The literature shows that the second transmembrane region is in fact a domain of the GJB1 protein and since the first extracellular region is located between the 46th and 75th residues it corresponds to the CNX (connexin) domain ("Domain Architecture Analysis"). It can be inferred that since these two regions are highly conserved domains of the protein, a mutation corresponding to these domains could possibly be pathogenic. Since the second transmembrane region was conserved the most, a more detailed look to the entropy of the region between the 75th and 96th residues are provided in the following figure.

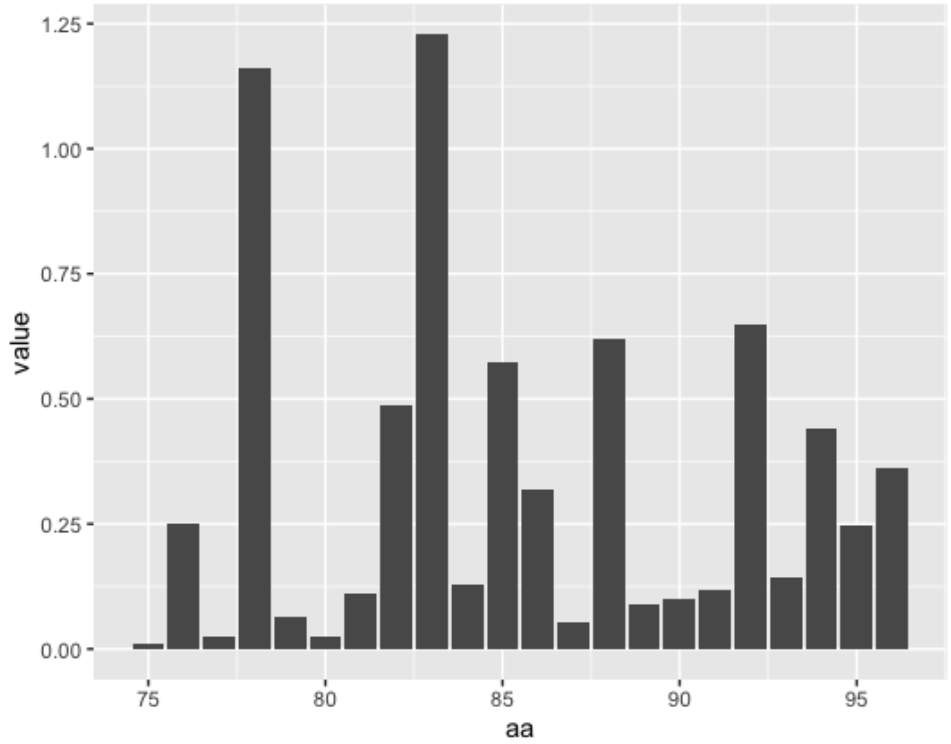

Figure 3: Distribution of Shannon Entropy Scores in the Second Transmembrane Domain

In Figure 3 the entropy scores of the second transmembrane domain are shown by residue. The results show that the 75th, 77th and 80th residues were highly conserved. The results align with the known variants of these residues that cause CMTX1 according to UniProt. The mutations Arg75Gln, Arg75Pro, Arg75Trp, Trp77Ser and Gln80Arg are observed in patients with CMTX1 (UniProt, "Gap Junction Beta-1 Protein."). According to the Transporter Classification Database, the Leu89Pro mutation in this domain hinders the transportation of GJB1 and gap junctions cannot be formed.

In addition to the highly conserved domains and residues found by Shannon Entropy, the 135th residue had an entropy score of zero and although the 136th residue had an entropy score of 2.07, making it a variable region, there was a Val136Ala mutation seen in CMTX1 (UniProt, "Gap Junction Beta-1 Protein."). There was a Tyr135Cys mutation reported for CMTX1 in UniProt; however, there was no publication linked to it. There were also no mutations reported relating to the 96th residue. Using the known mutations as controls mutations were modeled using the Meta-predictor of disease-causing variants tool.

Sequence File: 1-HOMO.seq

Mutation	PANTHER	PhD-SNP	SIFT	SNAP	Meta-SNP	RI	Profile		
R75P	Disease	Disease	Disease	Disease	Disease	8	F[R]=99%	F[P]=0%	Nali=400
	0.820	0.952	0.000	0.850	0.923				
R75Q	Disease	Disease	Disease	Disease	Disease	7	F[R]=99%	F[Q]=0%	Nali=400
	0.659	0.911	0.000	0.805	0.838				
R75W	Disease	Disease	Disease	Disease	Disease	8	F[R]=99%	F[W]=0%	Nali=400
	0.933	0.948	0.000	0.825	0.897				
S85C	Disease	Disease	Neutral	Disease	Disease	2	F[S]=65%	F[C]=4%	Nali=403
	0.579	0.727	0.060	0.585	0.601				
A96V	Disease	Neutral	Neutral	Neutral	Disease	1	F[A]=32%	F[V]=15%	Nali=392
	0.621	0.299	0.580	0.245	0.564				
Y135A	Disease	Disease	Disease	Disease	Disease	9	F[Y]=100%	F[A]=0%	Nali=377
	0.997	0.942	0.000	0.770	0.925				
Y135C	Disease	Disease	Disease	Disease	Disease	9	F[Y]=100%	F[C]=0%	Nali=377
	0.999	0.967	0.000	0.815	0.961				
Y135E	Disease	Disease	Disease	Disease	Disease	9	F[Y]=100%	F[E]=0%	Nali=377
	0.998	0.966	0.000	0.830	0.950				
Y135K	Disease	Disease	Disease	Disease	Disease	9	F[Y]=100%	F[K]=0%	Nali=377
	0.998	0.963	0.000	0.830	0.952				
Y135N	Disease	Disease	Disease	Disease	Disease	9	F[Y]=100%	F[N]=0%	Nali=377
	0.998	0.963	0.000	0.790	0.952				
Y135Q	Disease	Disease	Disease	Disease	Disease	9	F[Y]=100%	F[Q]=0%	Nali=377
	0.998	0.958	0.000	0.790	0.951				
Y135S	Disease	Disease	Disease	Disease	Disease	9	F[Y]=100%	F[S]=0%	Nali=377
	0.997	0.959	0.000	0.770	0.927				
Y135T	Disease	Disease	Disease	Disease	Disease	9	F[Y]=100%	F[T]=0%	Nali=377
	0.998	0.944	0.000	0.785	0.949				
V136A	Neutral	Disease	Neutral	Disease	Disease	2	F[V]=44%	F[A]=2%	Nali=378
	0.307	0.804	0.050	0.520	0.592				
V136F	Neutral	Disease	Disease	Disease	Disease	6	F[V]=44%	F[F]=1%	Nali=378
	0.375	0.938	0.000	0.630	0.786				
V136P	Disease	Disease	Disease	Disease	Disease	8	F[V]=44%	F[P]=0%	Nali=378
	0.657	0.966	0.000	0.725	0.892				

Mutation: WT+POS+NEW
  WT: Residue in wild-type protein
  POS: Residue position
  NEW: New residue after mutation
Prediction:
  Neutral: Neutral variants
  Disease: Disease causing variants
Outputs: Value reported under each prediction
  PANTHER: Between 0 and 1. (If >0.5 mutation is predicted Disease)
  PhD-SNP: Between 0 and 1. (If >0.5 mutation is predicted Disease)
  SIFT: Positive Value (If >0.05 mutation is predicted Neutral)
  SNAP: Output normailzed between 0 and 1 (If >0.5 mutation is predicted Disease)
  Meta-SNP: Between 0 and 1. (If >0.5 mutation is predicted Disease)
RI: Reliability Index between 0 an 10
F[X]: Frequency of residue X in the sequence profile
Nali: Number of aligned sequences in the mutated site

Figure 4: Mutation Modelling

Different mutations to the 96th, 135th and 136th residues were modeled. The mutation Ala96Val, both of which are non-polar amino acids, showed some disease probability but the likelihood was low when compared to the rest. All mutations modeled on the 135th residue, which were from polar amino acid to polar amino acid, showed high probability for disease. It can be inferred that the 135th residue is important for a functioning GJB1 protein. Various mutations were also modelled on the 136th residue however the probability for disease was low when compared to the rest.

Following the analysis of conserved domains and residues a phylogenetic tree was constructed in order to understand the evolutionary relationship between the homologs of the GJB1 protein. Since the alignment obtained from MSAviewer NCBI Blast was not a commonly used alignment for making phylogenetic trees an alignment using Mafft were made. In the multiple sequence alignment some patterns, in the way the gaps were inserted, were observed. These observations were counted using a custom Python script. Approximately 880 sequences had a large gap in the beginning, 80 sequences had a short gap in the beginning and the remaining sequences did not have a particular pattern. In Figure 5 the sequences are colored with magenta, red and orange respectively. The species colored in red tended to cluster together. Anser cygnoides domesticus was determined as the outgroup.

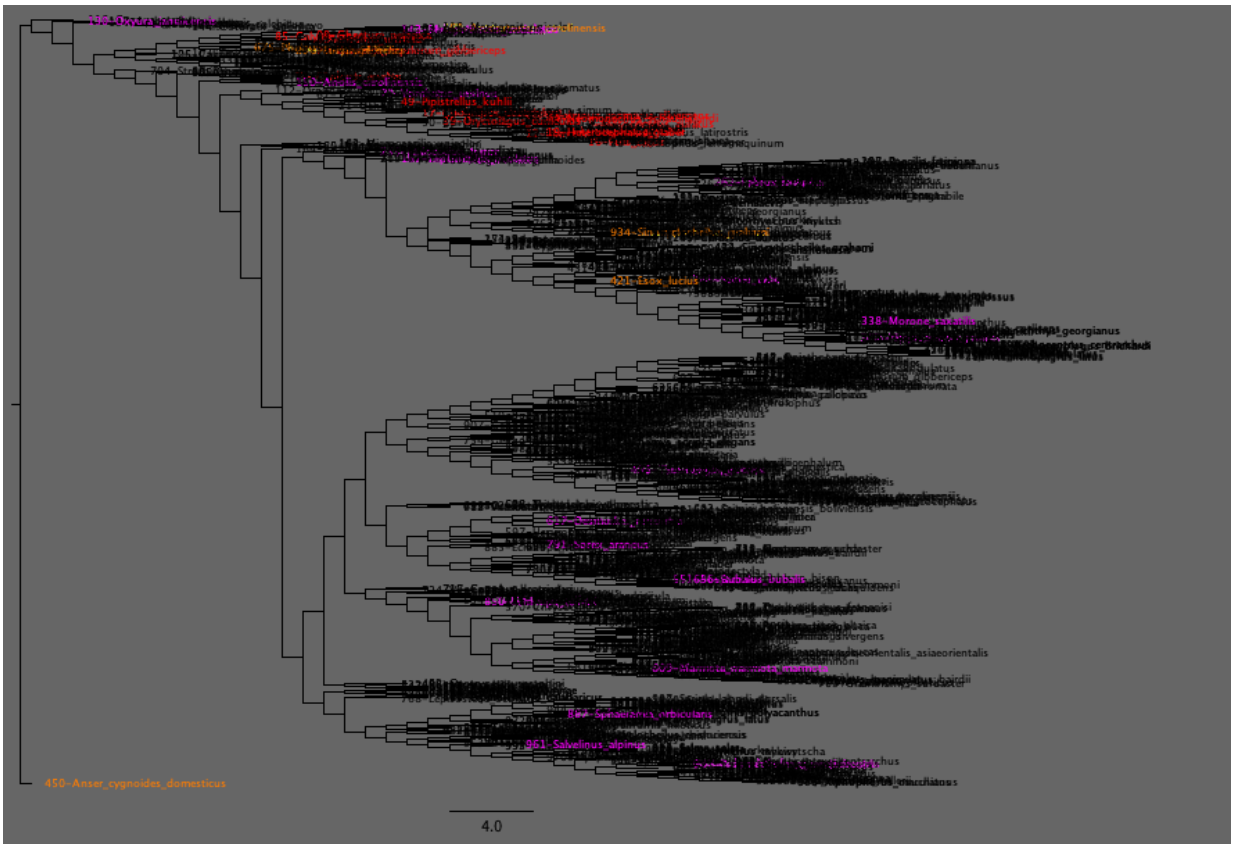

Figure 5: Maximum Likelihood Phylogenetic Tree

Since the phylogenetic tree in Figure 5 has approximately 1000 species it is quite hard to understand the more detailed evolutionary relationships. In order to address this problem four species from mammals, birds, lizards, turtles and amphibians that had the GJB1 protein were randomly selected and a smaller phylogenetic tree was constructed.

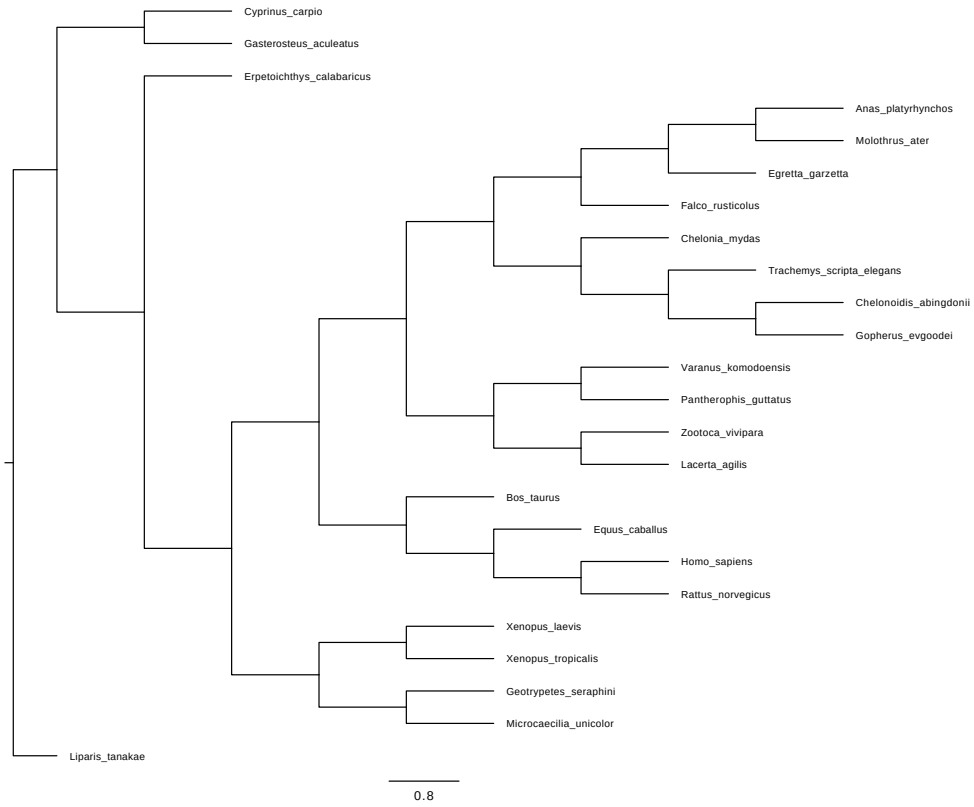

Figure 6: Smaller Maximum Likelihood Tree

It can be seen from Figure 6 that Liparis tanakae is the outgroup and species from the same taxonomical origin tended to cluster together. For instance, Bos Taurus, Equus caballus, Homo sapiens and Rattus norvegicus are all mammals and they are located on the same node.

3. Discussion:

In order to fulfill the purpose of the study the most conserved domain of the Gap Junction Beta 1 protein was found through an analysis on known topological regions using Shannon Entropy calculation. The results indicated that the second transmembrane region was the most conserved and this finding aligned with existing knowledge that the second transmembrane region was a domain of the protein. Certain mutations were modeled according to the most conserved residues of the most conserved region and those that were highly conserved when compared to the rest. Within the conserved domain the results showed that the 75^{th}, 77^{th} and 80^{th} residues were highly conserved and many mutations for these domains were reported in UniProt database. Residue 135 was also highly conserved; however, there were not many reports on how it related to the Charcot Marie Tooth Disease X dominant 1. Certain mutations were modeled on the 135^{th} residue and even polar amino acid to polar amino acid mutations indicated disease. Additionally, phylogenetic trees constructed by the Maximum Likelihood method were used in order to understand the evolutionary history of the homologs of the GJB1 protein. The alignment used indicated that approximately 80 sequences had a possible gap

deletion that kept them clustered closer to the root and less differentiation occurred in the GJB1 of those species. The bigger tree showed that there was a deletion or duplication from the ancestral protein that formed the outgroup, and the remaining species were on the same node. The smaller tree of vertebrates randomly selected from different groups showed a similar type of initial deletion or duplication where the outgroup was on a separate node and the remaining species clustered around one node. In the smaller tree organisms that were taxonomically close tended to form nodes together. Further investigation can be performed into the relation of the 135th residue in the GJB1 protein and the CMTX1 disease.

4. Materials and Methods:

The general aim of the project was to study a protein family connected to a Mendelian disease. In order to determine such a protein, Mendelian diseases were found from the OMIM database. Studying a relatively less common disease was a group decision so diseases that are not commonly known were looked into. After the Charcot Marie Tooth X dominant 1 disease was found and decided upon. OMIM entries on CMTX1 were analyzed. Once it was determined that the mutations on the Gap Junction Beta 1 gene caused the problem that led the disease the GJB1 protein and its family was researched.

Initially the protein sequence for GJB1 was obtained from the UniProt database. The topology of the protein was also available on UniProt and that information was used in the following steps. The homologs of the protein sequence were found using National Center for Biotechnology Information database's Blast tool's refseq protein database. On the NCBI database there was a link called MSAviewer which showed the alignments made by the search query of the protein sequence. The aligned sequences were available for download and the multiple sequence alignment was obtained in fasta format from the database. After the sequences were obtained initially a phylogenetic tree was formed using the Neighbor Joining algorithm on MegaX; however, that tree showed some duplicates of the species. In order to make sure that there were no sequence duplicates, a python script that removed sequences which had the same species name and sequence were removed from the alignment. Additionally, the headers of the fasta file obtained from NCBI were quite long, the headers were reduced to species names for clarity.

After editing the alignment file for clarity, the conservation of the topological regions of the protein were calculated according to Shannon Entropy using a phyton script. After the most conserved topological region was found the results were compared to the known domains of the GBJ1 protein which were acquired from the SMART database. The Shannon Entropy of the domain was calculated, and the most conserved residue was observed. The results were graphed using R studio ggplot. Some predicted mutations were modeled using the Meta-predictor of disease-causing variants at biofold.org.

In order to observe the evolutionary history of the homologs of the GJB1 protein a phylogenetic tree was constructed using the Maximum Likelihood method using MegaX. In order to construct the tree in Mega an alignment was made using Mafft. In order to find the outgroup a python script that formed a position probability matrix and calculated the worst match was used. The tree was re-rooted according to the outgroup identified by

the python script. The tree was later colored according to some patterns identified in the alignment. Since there were close to 1000 samples, random samples from different taxonomic groups were taken from GenBank and a smaller tree was construct for a more in-depth analysis.

5. References:

Abrams, Charles K. "GJB1 Disorders: Charcot Marie Tooth Neuropathy (CMT1X) and Central Nervous System Phenotypes." *GeneReviews® [Internet].*, U.S. National Library of Medicine, 20 Feb. 2020, www.ncbi.nlm.nih.gov/books/NBK1374/.

"BLAST: Basic Local Alignment Search Tool." *National Center for Biotechnology Information*, U.S. National Library of Medicine, blast.ncbi.nlm.nih.gov/Blast.cgi.

Bortolozzi, Mario. "What's the Function of Connexin 32 in the Peripheral Nervous System?" *Frontiers in Molecular Neuroscience*, Frontiers Media S.A., 10 July 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6048289/.

Capriotti, Emidio. "Meta-SNP - Meta-Predictor of Disease Causing Variants." *SNP*, snps.biofold.org/meta-snp/.

"Domain Architecture Analysis." *SMART*, smart.embl-heidelberg.de/smart/show_motifs.pl?ID=P08034.

"GenBank Overview." *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/genbank/.

McKusick, Victor A. "CHARCOT-MARIE-TOOTH DISEASE, X-LINKED DOMINANT, 1; CMTX1." Edited by Cassandra L. Kniffin, *OMIM*, 18 Mar. 2014, www.omim.org/entry/302800.

McKusick, Victor A. "GAP JUNCTION PROTEIN, BETA-1; GJB1." Edited by George E. Tiller, *OMIM*, www.omim.org/entry/304040.

NCBI Multiple Sequence Alignment Viewer 1.18.1. www.ncbi.nlm.nih.gov/projects/msaviewer/.

Sequence Variability Server. imed.med.ucm.es/Tools/svs_help.html.

Sohl, G. and K. Willecke. "Gap Junctions and the Connexin Protein Family." *Cardiovascular Research*, vol. 62, no. 2, 2004, pp. 228-232, doi:10.1016/j.cardiores.2003.11.013.

"Transporter Classification Database." *TCDB " SEARCH*, www.tcdb.org/search/result.php?acc=P08034.

UniProt ConsortiumEuropean Bioinformatics InstituteProtein Information ResourceSIB Swiss Institute of Bioinformatics. “Gap Junction Beta-1 Protein.” *UniProt ConsortiumEuropean Bioinformatics InstituteProtein Information ResourceSIB Swiss Institute of Bioinformatics*, 2 Dec. 2020, www.uniprot.org/uniprot/P08034.