FACULTAT
DE CIÈNCIES, TECNOLOGIA
I ENGINYERIES
UVIC | UVIC·UCC

Treball de Fi de Grau Experimental

PATHOGENICITY OF HOMOLOGOUS MUTATIONS IN MEMBRANE PROTEINS

SERGI SOLDEVILA GÁLVEZ

Grau en Biotecnologia

Tutor/a: Mireia Olivella García

Co-tutor: Arnau Cordomí Montoya

Vic, Juny de 2022

Acknowledgement

This research would not have been possible without the constant help, supervision, and correction of my tutor, Mireia Olivella García. I am really thankful for her patience and guidance through all the course. In addition, I would like to thank Arnau Cordomí Montoya, my co-tutor, for his support. Last but not least, I would like to mention the support of Adrián García Recio, who has been crucial in the optimization and development of scripts, especially when it was hard to find a solution. All three have been my mentors, helping me in all phases of the project, even when I was stuck.

Finally, I find it necessary to mention the support received from Uvic. The university has helped me develop the basic knowledge needed to start my career in bioinformatics and be able to have the essential skills to do this research.

Resum

Títol: Patogenicitat de mutacions homòlogues en proteïnes de membrana
Autora: Sergi Soldevila
Co-Tutors: Dra. Mireia Olivella García (UVic) i Dr. Arnau Cordomí Montoya (UPF)
Data: Juny de 2022
Paraules clau: *Proteïnes de membrana, regions transmembrana, Via Glutamatèrgica, mutacions, patogenicitat*

Les variacions en les proteïnes de membrana poden conduir a conseqüències patògenes o no patògenes. És de vital rellevància estudiar aquelles variacions que poden acabar donant patogenicitat, per tal de descobrir com identificar correlacions entre mutacions similars i procedir a la detecció precoç de causes patogèniques o possibles tractaments. En aquest article s'han utilitzat múltiples bases de dades i eines computacionals per estudiar la patogenicitat de mutacions homòlogues en proteïnes de membrana. Els resultats finals han demostrat la hipòtesis inicial on es creia que les proteïnes de membrana de la regió transmembrana estan molt conservades, juntament amb la seva patogenicitat.

Addicionalment, en aquest estudi s'han identificat 397 gens dins de la via glutamatèrgica a partir d'1.233 codis GO. La investigació ha estudiat i classificat els gens més patògens implicats en la sinapsi glutamatèrgica. Els descobriments previs relacionats amb el tractament amb L-serine i la classificació permet que els pacients amb la via glutamatèrgica afectada podran ser tractats amb aquest fàrmac i estudiar la seva millora.

Summary

Title: *Pathogenicity of homologous mutations in membrane proteins*
Author: Sergi Soldevila Gálvez
Supervisor: Dra. Mireia Olivella García (UVic) and Dr. Arnau Cordomí Montoya (UPF)
Date: June de 2022
Keywords: *Membrane proteins, transmembrane regions, Glutamatergic synapse, mutations, pathogenicity*

Variations in membrane proteins can lead to pathogenic or non-pathogenic consequences. It is vitally important to study those variations that may end up conferring pathogenicity, in order to discover how to identify correlations between similar mutations and to proceed to the early detection of pathogenic causes or possible treatments. Multiple databases and computational tools have been used in this paper to study the pathogenicity of homologous mutations in membrane proteins. The final results have shown the initial hypothesis that membrane proteins in the transmembrane region were thought to be highly conserved, along with their pathogenicity.

Additionally, in this study, 397 genes were identified within the glutamatergic pathway from 1,233 GO codes. Research has studied and classified the most pathogenic genes involved in glutamatergic synapse. Previous findings related to L-serine treatment and classification allow patients with the affected glutamatergic pathway to be treated with this drug and to study its improvement.

Table of Contents

Table list

Figure list

1. Introduction

1.1. Technology advances in identifying new sequence variants

Between 1977 and 1978, the bacteriophage Φ-X174 was sequenced (1), and from then, the number of determined sequences has increased exponentially. From the year 2000, several sequencing projects began to appear in different organisms, including the sequencing of the human genome. Among others, this data is analysed to identify genes that codify for proteins, mutations, regulatory sequences, and also allows comparative genomics between genes and species.

In order to process this amount of information, the most effective and efficient way of doing it is by using programming and statistics methods. Additionally, the methods of sequence analysis have not ceased to be optimized, and the number of professionals specialized in analysis, and the amount of data has not stopped increasing (2). The cost of genome sequencing has decreased over the last decade, automatizing the processes of sequence analysis and improving methodologies. These advances done in sequencing technologies have been a revolution in the research of genetic variation, and have allowed exploring in a more efficient way the basis of human disorders, enabling to study huge databases of pathogenic and non-pathogenic variants (3).

One of the most used techniques of sequencing is the Next Generation Sequencing (NGS). This technique grants the possibility of rapidly increasing the knowledge about genes and mutations (4). Consequently, rates in diagnosis and in genetic etiology (a gene abnormality that is inherited from a parent at conception) have improved (5). This improvement, combined with the sequencing techniques, has permitted the discovery of recurrent *de novo* mutations, and the development of new databases which allow comparing the effects of mutations in "normal genes", copy number variations, mosaicism, and many others) (6). With NGS, it is now possible to sequence a huge quantity of ADN and to perform exome sequencing, that covers between 1 percent and 2 percent of the genome, depending on the species (7). Exome sequencing allows us to identify variations in the protein-coding region, specifically in the 5'UTR and in the 3'UTR region. As most of the mutations that cause disease occur in exons, this method of sequencing is efficient for the identification of possible disease-causing mutations (8).

Recently, owing to whole-genome and exome sequencing, it has been discovered that the rare Mendelian diseases caused by missense mutations are more frequent than expected, affecting millions of patients globally (9). Mendelian diseases happen when germline mutations (specific mutations in single genes) are inherited. Some examples of Mendelian diseases are Cystic fibrosis, sickle cell disease and Duchenne muscular dystrophy (10,11). Nowadays, the identification of mutations leading to human genetic diseases has increased, and the research is every day more accurate. It is relevant to study the variations that lead to pathologies and through DNA analysis, it is possible to determine the mutational spectrum for a specific disease. What is more, international databases are now available worldwide, which makes it easier to work with more data at the same time without having to restart something that is already studied (12). Understanding the relationship between genetic variation and functional implications in proteins is critical for deciphering genomic data and finding disease-causing variations. Integrating protein function knowledge with genome annotation can help understand genetic diversity in complicated biological processes more quickly. To demonstrate the power of combining protein annotations with genome annotations for functional interpretation of homologous variants. Data is used to process mapping UniProtKB human sequences and positional annotations, such as transmembrane regions, Pfam codes, and other annotations, to the human genome (GRCh38) (13).

It has been demonstrated that *de novo* mutations are one big cause of genetic disorders, such as intellectual disability and autism. Neutral mutations may simply spread as a result of genetic drift, whereas mutations that rely on a phenotype spread quickly through a population. However, harmful mutations that result in harmful traits before or during the reproductive phase are subjected to a "purifying selection" and are prevented from spreading through the population. *De novo* mutations are genetically distinct from inherited variants because they are the result of mutagenic processes that occur between generations. At the population level, loss or acquisition of traits drives species evolution, whereas at the individual level, loss or acquisition of traits can result in disease. For these reasons, it is vitally important to identify the mutations *de novo*, or not, and its possible pathogenic effects (14).

1.2. Classification of sequence variations

A mutation is any change in the DNA sequence of an organism. A germline mutation is a type of mutation that occurs in the eggs or sperms. Otherwise, a somatic mutation is given in any other cell of the body different from germline cells (15). Mutations in the DNA sequence are a consequence of substitution, insertion, or deletion of base pairs which are usually harmless. Due to the lethal or disease potential of mutations,

evolution has developed mechanisms to repair them. There are different DNA repairing mechanisms such as mismatch repair, where the exonuclease corrects when there's a single base insertion or deletion, and Nucleotide Excision Repair (NER) which forms a protein complex where the DNA is damaged in order to repair it. There are also specialized enzymes which repair the DNA directly and the repairing process which occurs during the recombination process. As said before, mutations can arise by errors during the replication process, cellular division, mutagenic agents or even a viral infection. Normally, changes in the DNA sequences occur due to endogenous (Reactive oxygen species (ROS) produced from normal metabolic pathways) or exogenous (exposure to X-rays, radiation, toxins, chemicals, viruses, …) DNA damage (16, 17).

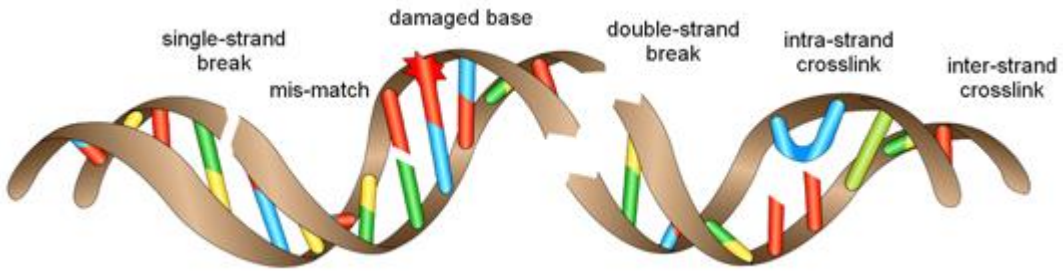

Figure 1: Doble-stranded damaged DNA from Davis A, Tinker A v., Friedlander M. "Platinum resistant" ovarian cancer: what is it, who to treat and how to measure benefit? Gynecol Oncol [Internet]. 2014 [cited 2022 Jun 6];133(3):624–31. Available from: https://pubmed.ncbi.nlm.nih.gov/24607285/

Mutations can be classified into large scale mutations and small scale mutations.

i. **Large scale mutations** (18)

Mutations on large scale implies that the effect is higher, larger genetic material is affected at once. This, as a consequence, is said that have an impact on chromosome changes. These can be classified as:

a. Duplications or Amplifications

When a chromosomal segment has been duplicated, so it is repeated.

b. Inversions

When there's a change in the DNA direction of a chromosome.

c. Deletions

When a chromosomal fragment has been lost (19).

d. Insertions

A chromosomal change that involves the insertion of one or more nucleotides into a DNA sequence.

e. Translocations

A chromosome splits and the shattered portions (usually two) reattach to other different chromosomes.

Simple translocations

One-way transfer.

Reciprocal translocations

Two-way transfer.

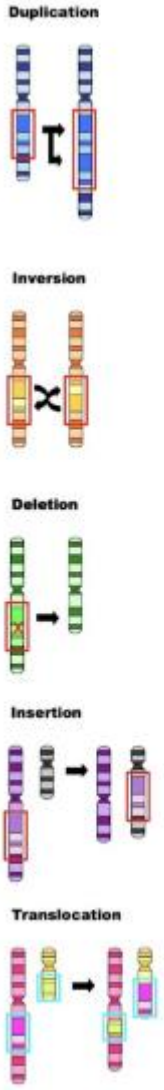

Figure 2: Large scale mutations representation from Peter J. Russell. iGenetics: : a molecular approach. Pearson Education Inc as BC, editor. 2010.

ii. Small scale mutations (18)

Small scale mutations imply that one or few nucleotides within a gene are affected. These are the result of base substitution, base addition, base deletion or base insertion. These can be classified as:

a. Synonymous substitutions

DNA change which does not affect the protein sequence.

1. **Silent** (20)

The silent mutations are the ones that result in a codon that codes for the same or a different amino acid without suffering any change in the function of the protein. In other words, the amino acid sequence that is encoded by the gene, in this case, is not changed, and that is why this type of mutation is called silent because it is not noticed in the amino acid sequence itself.

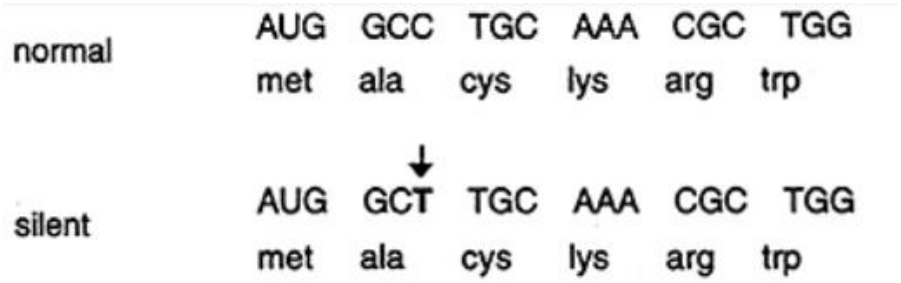

Figure 3: Silent mutation representation from Dr. Noel Sturm. DNA Mutation and Repair [Internet]. 2019 [cited 2022 Jun 6]. Available from: http://www2.csudh.edu/nsturm/CHEMXL153/DNAMutationRepair.htm

b. Nonsynonymous substitutions

DNA change does affect the protein sequence.

1. Base Substitutions

Base substitutions, also called point mutations, are a single base substitution and are the most frequent, the most common.

These can also be classified as the molecular consequence in the protein:

i. Missense mutations (20, 21)

In this case, the mutation generates a codon that specifies a different amino acid and leads to another polypeptide sequence. This type of mutation replaces one nucleotide of the DNA codon with another nucleotide and resulting in an amino acid change.

Conservative: in this type of missense mutation, the resultant amino acid conserves a similar function and shape to the amino acid that has been replaced.

Non-conservative: in contrast to the conservative mutations, non-conservative mutations result in a completely different amino acid. In this type of mutation, the amino acid is altered in functionality and shape, and they do not conserve any of the original characteristics (22).

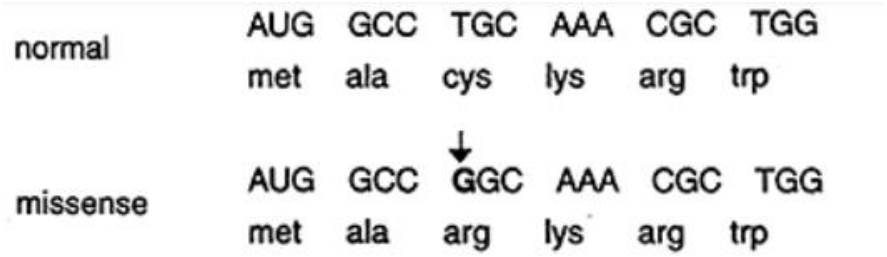

Figure 4: Missense mutation representation from Dr. Noel Sturm. DNA Mutation and Repair [Internet]. 2019 [cited 2022 Jun 6]. Available from: http://www2.csudh.edu/nsturm/CHEMXL153/DNAMutationRepair.htm

ii. Truncation mutations

When talking about nonsense mutations, we are referring to a kind of nucleotide point mutation, i.e., a substitution, which results in a stop codon. A nonsense codon is a mutation that ends in the truncation of the protein, a non-functional protein (20, 23).

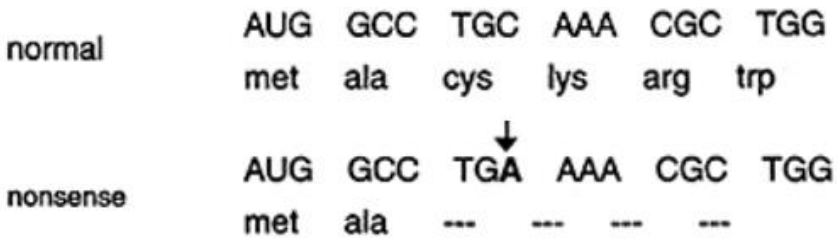

Figure 5: Truncation mutation representation (from Dr. Noel Sturm. DNA Mutation and Repair [Internet]. 2019 [cited 2022 Jun 6]. Available from: http://www2.csudh.edu/nsturm/CHEMXL153/DNAMutationRepair.htm)

iii. Deletion and Insertions (20)

One or more base pairs (bp) are lost or inserted from the DNA, possibly leading to frameshifts. The insertion or

deletion changes the sequences and usually ends with a premature stop codon.

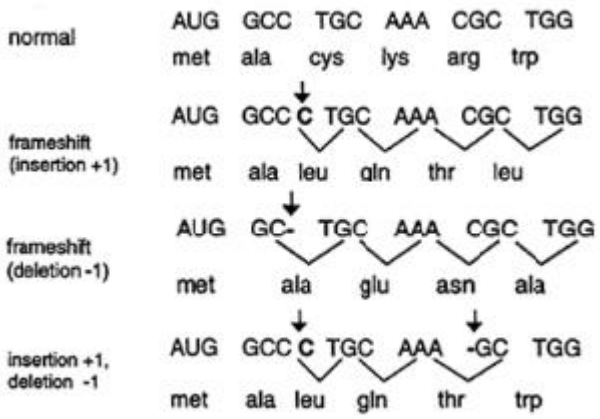

Figure 6: Insertion/Deletion mutation representation from Dr. Noel Sturm. DNA Mutation and Repair [Internet]. 2019 [cited 2022 Jun 6]. Available from: http://www2.csudh.edu/nsturm/CHEMXL153/DNAMutationRepair.htm

1.3. Classification of sequence variants into disease-causing or neutral

This massive amount of data that is generated from genome sequencing produces tons of newly identified mutations. The stratification of these mutations into disease-causing, neutral or disease-related is vital for the diagnosis of diseases and in order to define a strategic therapy (12,24).

A relevant research focus are the variants from a single nucleotide that lead to amino acid substitutions at the protein level. These mutations, called missense mutations, are associated with more than half of inherited diseases known up to date. A big amount of computational methods have been developed to identify those missense mutations that are potentially pathogenic. Each method has a different approach to studying missense mutations, although they share methodological approximations. For example, some important aspects to consider are evolutionary conservation, changes in physicochemical properties of amino acids, biological function, known disease association and protein structure (25). Although these predictors are useful and allow us to distinguish between disease-causing and neutral variants, the prediction power is still limited, making it difficult to rely totally on the final diagnosis.

Due to the fast-enlarging of new genomic variants identified and the need of understanding the relationship between the phenotype and the genotype, it is

requested to develop new computational tools to understand these variants and which of these are responsible for causing a disease between candidate variants. There are multiple predictors available such as SIFT (26), MutationTaster (27) and Polyphen-2 (28) based on the evolutive conservation and the impact that is expected on the structure and function of amino acids from sequence data. To do so, they use parameters based on evolutionary conservation and physicochemical properties of these amino acids from the data sequence (29). Apart from those predictors, not long ago, TMSNP the latest and most advanced prediction method has been released. This is a database and a predictive tool for the pathogenicity of protein-membrane variants with higher predictive ability than other mentioned tools. Despite improving the predictive power, it only works for membrane proteins.

1.4. The human proteome

Proteins are the major key regulators of function in biology, therefore a thorough understanding of their structure and characteristics is essential for basic and translational research. Proteins are big biomolecules that are made up of one or more long chains of amino acids. These, play a variety of roles in organisms, including catalyzing metabolic events, DNA replication, acting as receptors, giving cells and organisms structure, and moving materials from one place to another. Proteins differ primarily in their amino acid sequence, which is governed by their genes' nucleotide sequence and usually culminates in protein folding into a specific 3D structure that dictates its activity (30).

Proteins are classified as globular proteins and membrane proteins. One of the most frequent protein kinds is globular proteins, which are spherical proteins and are relatively water-soluble (forming colloids in water). Because there are many alternative architectures that can fold into a roughly spherical shape, there are several-fold classes of globular proteins. Globular proteins can act as enzymes, messengers, transporters, regulators, and more (31).

1.5. Membrane proteins

Membrane proteins (MP) are proteins that are found in biomembranes or are able to interact with them. The interaction between the interior of the cell and its environment is mediated by these kinds of proteins, which represent 25% of the whole *Homo sapiens* proteome. Membrane proteins are classified into different families such as ionic

channels, enzymes, and receptors. All of these families play an important role in cellular functions. Integral proteins are a permanent part of membranes that can either permeate it or connect with one side or the other. The cell membrane is transiently linked with peripheral membrane proteins. As a result of the large variety of functions developed by these proteins, when disrupted, they tend to a range of diseases. Identification of membrane proteins with abnormal properties can lead to the result of discovering new therapeutic targets, which is something that pharmaceutics can take advantage of (32). Previous studies have shown that about 50% of membrane proteins are the pharmacological target of various (33). It is also estimated that 90% of membrane proteins have mutations that are pathogenic. These mutations can affect some functions that are necessary to the correct functioning of the proteins, such as folding or stability (34).The current bioinformatics tools are the key to understanding the function of the membrane proteins. However, membrane proteins are more related to disease than globular proteins. In contrast to the globular proteins, fewer available structures are found in the databases, due to experimental limitations in the process of X-ray crystallography. For this reason, the current tools are based on the prediction of globular proteins, meaning that it is necessary to implement tools for the membrane proteins, so they are studied as well (35).

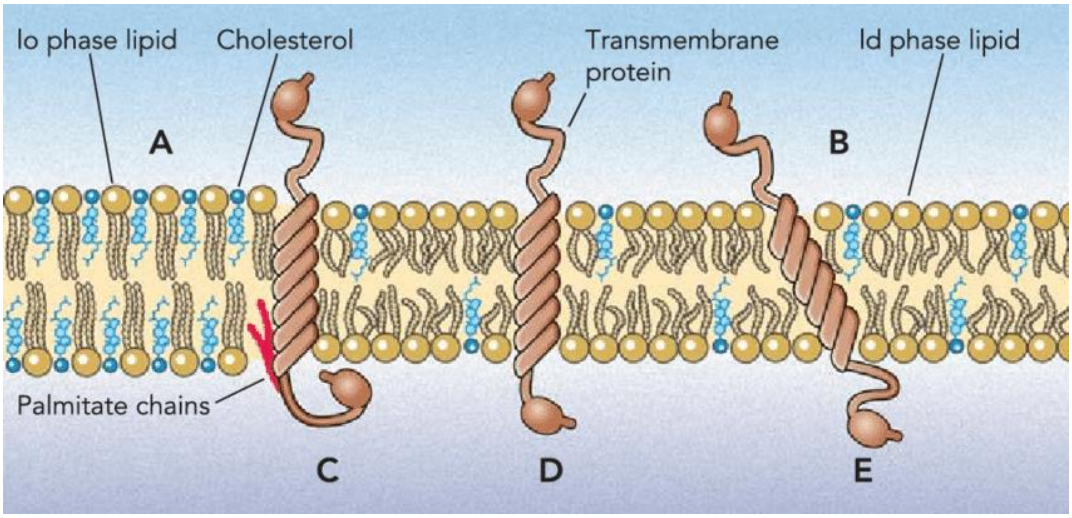

Figure 7: Types of membrane proteins and its schematic representation from Brown DA. Lipid rafts, detergent-resistant membranes, and raft targeting signals. Physiology (Bethesda) [Internet]. 2006 Dec [cited 2022 Jun 6];21(6):430–9. Available from: https://pubmed.ncbi.nlm.nih.gov/17119156/

One of the main differences between the membrane proteins and the globular proteins is their environment, and they also differ in the transmembrane (TM) region. This global difference is the cause of other dissimilarities, seen in the amino acid sequence, secondary structure and conformations. This contrasts between both types of proteins, evidence that the membrane proteins require less sequence identity than the globular ones to maintain their folding (36,37). By the other side, the sequence in the

transmembrane segments of membrane proteins is more conserved than the globular domains of membrane proteins or than globular proteins. Although membrane proteins are more robust to sequence variation, membrane proteins are more pathogenic than globular proteins due to their essential role in the cell.

Overall, the specific features of membrane proteins: high sequence conservation in the transmembrane segments, more robust to amino acid change and more structure conservation, combined with their relation to diseases make them an ideal group of proteins to test new methodological approximations to discern between neutral and disease-causing variants.

1.6. Glutamatergic synapse

Glutamate is the most excitatory neurotransmitter in the central nervous system (CNS). Multiple glutamate transporters, as well as ionotropic and metabotropic receptors, are involved in central and peripheral glutamate signalling (38–40).

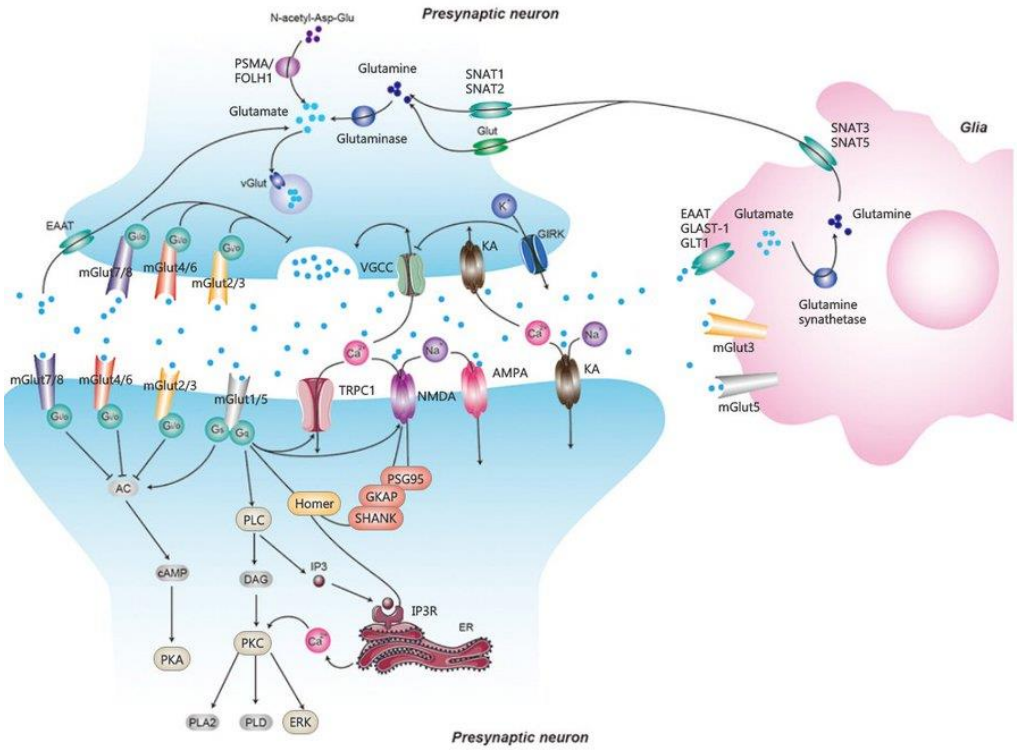

Figure 8: Schematic representation of a glutamatergic synapse indicating the main proteins involved and their functional pathways from Javitt DC. Glutamate as a therapeutic target in psychiatric disorders. Mol Psychiatry [Internet]. 2004 Nov [cited 2022 Jun 6];9(11):984–97. Available from: https://pubmed.ncbi.nlm.nih.gov/15278097/

In the mammalian CNS, glutamate is a key signalling molecule as well as a primary excitatory neurotransmitter (42). It regulates a wide range of processes through its receptors, which are found on both neuronal and non-neuronal cells. It is released as a neurotransmitter into the synaptic cleft and initiates the propagation of action potentials under normal physiological conditions (43).

Apart from its vital involvement in CNS, glutamate also contributes to autocrine and paracrine signalling in peripheral tissues such as the bone, pancreas, pineal glands, etc (44). Glutamate is also crucial in peripherally mediated pain signalling to the central nervous system (45). Glutamate release, absorption, metabolism, and signalling are all closely regulated activities, which partially explains glutamate's extensive significance in essential central and peripheral processes. Aetiologically, disturbances in these pathways are frequently linked to central disorders (46,47).

Glutamate dysregulation has been well-studied in schizophrenia, fragile X syndrome, and epilepsy, among other psychiatric, neurodevelopmental, and neurodegenerative illnesses (48). The growing importance of glutamate signalling in various illnesses, including depression and anxiety, has prompted new hypotheses about glutamate dysregulation (49). Psychiatric and neurodegenerative illnesses, on the other hand, are complicated disease states that are likely the result of several interconnected processes (50).

Due to advances in sequence technology, exome sequencing is increasing on children with neurodevelopment disorders, leading to the identification of pathogenic variants in genes involved in the glutamatergic synapse, which are usually classified as rare diseases. Unfortunately, scarce treatments are available for these rare diseases. The recent discovery of L-serine to treat GRIN related disorders (51), a disorder that mainly results in a hypofunctionality of NMDA (*N-methyl-D-aspartate*) receptors and a reduction of glutamatergic synapse, has opened the door to extend L-serine treatment to other glutamatergic synapse genes related diseases that also result in a decrease of the glutamatergic synapse (52,53).

2. Objectives

1. The high sequence identity in the TM segments of membrane protein, combined with its robustness to sequence variation and the involvement of membrane proteins in disease, make them ideal to study and develop new methods able to classify sequence variants into disease-causing or neutral. The aim of this study is to inspect if the pathogenicity of sequence variants in the TM segments of membrane proteins can be extrapolated between homologous variants.

2. The increasing number of sequence variants involved in genes of the glutamatergic synapse, that result in neurodevelopmental disorders, urges finding therapeutic strategies. The aim of this study is to obtain an overall picture of the pathogenesis of this group of genes and to identify the most pathogenic genes that result in a decay of the glutamatergic synapse.

3. Methods

3.1. Public Biological Databases used

UniProt (Universal Protein Resource) is a free repository whose aim is to provide a comprehensive, high-quality, and widely accessible database of protein sequence and functional information to the scientific community. This repository is globally used because it contains manually reviewed and annotated proteins from different species (54,55).

ClinVar is a public database with the particularity that the data collected allows the interpretation of human (Homo sapiens) variation and observed health status. It provides open access to data about the relationship between phenotypes and medical important variants (56).

Pfam is a database of protein families and domains that is often used to examine new genomes and metagenomes, as well as to guide experimental work on individual proteins and systems. A representative set of sequences is provided by a seed alignment for each Pfam family. The HMMER tool automatically creates a profile hidden Markov model (HMM) from the seed alignment and searches it against the pfamseq sequence database. The HMM is used to align all regions within a curated umbral (57).

The Genome Aggregation Database (GnomAD) is a database that was released relatively early (in 2016), and it is freely available to the rest of biomedical and scientific community. This database harmonizes and aggregates the data retrieved from exome and genome large-scale sequencing projects. Despite its new release, GnomAD has demonstrated that it is an invaluable genetic resource, assisting in the discovery and interpretation of disease-causing variations as well as possible treatment targets. This database contains data from 6 global and 8 subcontinental ancestors, giving sample diversity (58,59).

3.2. Data preparation

3.2.1. Sequence annotations of the human proteome

Considering this reliability in the data provided by UniProt, all transmembrane proteins belonging to *Homo sapiens* organism were retrieved from this database with the certainty of having accurate, rich and consistent functional information of proteins. For each of the 5.202 human transmembrane reviewed proteins or UniProtKB entries, the core mandatory data has been extracted (amino acidic sequence, protein names, gene names, Pfam codes and TM ranges) (60).

3.2.2. Database of disease-causing variants

Information from all membrane proteins classified with UniProt into TM proteins were mined. All disease-causing (or pathogenic and likely pathogenic variants) related to Mendelian illnesses as described in ClinVar were extracted, starting with a dataset with 169.998 disease-causing sequence variants. This database is also based on the fact that the TM segments are highly conserved and its Pfam alignments are precise as well. This created database also contains the same molecular consequences as the neutral variants database. Knowing that, the databases finally ended with 2.991 disease-causing missense variants in the TM regions of membrane proteins also from *Homo sapiens* organisms. Before extracting the final dataset, a previous step was done in which disease-causing membrane proteins were selected, resulting on 40.111 variants (29).

This data was used to generate a list of pathogenic membrane proteins: those membrane proteins that present disease-causing variants that are related to Mendelian inheritance in genetic diseases. Non-pathogenic membrane proteins are those that do not present any sequence variation that is disease-causing and thus, sequence variants that affect the function and the folding of the protein are not disease-causing.

3.2.2. Database of neutral variants

All human neutral mutations containing missense, truncation, splice sites, UTRs and frameshift molecular consequences were retrieved from GnomAD allowing the creation of the data set containing 3.767.975 neutral variants. Only those entries from membrane proteins were kept, with this step the database was reduced by 70% until 949.840 neutral variants in membrane proteins. In order to keep only sequence variants in the TM segments, the coordinates of the TM segments from UniProt were used to filter out the sequence variants that were

not in the TM segments. This proceeded to a new reduction which was carried on when an even more restrictive filter was applied, ending with 45.337 neutral missense variants into TM regions of membrane proteins from *Homo sapiens.*

There were mutations classified in both neutral and disease-causing databases, some mutations in the GnomAD server were also identified as pathogenic or likely pathogenic and other mutations were related to recessive or complex variants (needing other third-party proteins to become a disease and being found on a healthy population, as well as the population with diseases). These variants were not included in the database. Only neutral variants in pathogenic proteins were kept, as variants in non-pathogenic proteins can affect the structure and the function of the protein without being associated or related to any disease.

3.2.4. Use Sequence Alignments to obtain equivalent positions for each variant

Pfam database was downloaded and only human alignments were kept. The Pfam alignments were used to identify the position in the Pfam sequence alignment of each missense sequence variant. The position in the Pfam sequence alignments allows finding other sequence variants that are located at the same equivalent position of the Pfam alignment for homologous proteins.

3.2.5. Identification of homologous variants in the TM segments of membrane proteins and comparison of its pathogenicity

In order to identify homologous variants, the same initial amino acid, same final amino acid, same Pfam alignment and same equivalent position were used. From the list of homologous variants, the pathogenicity of pairs of homologous variants were compared. The variants that presented a dual classification as neutral and disease-causing within a protein were filtered out, as these proteins and genes could be related to recessive inheritance. The sequence variants from these 24 variants were filtered out.

Finally, all pairs of homologous variants were classified as (I) neutral homologous variants (II) disease-causing homologous variants (III) neutral and disease-causing homologous variants.

3.3 Additional projects

3.3.1 Syngo Genes

A group of 1.233 GO codes from glutamatergic synapse are used to identify the genes involved in this pathway. The procedure is practically identical to the one used for the glutamatergic synapse. Databases from both ClinVar and GnomAD repositories were created to first identify and then quantify the number of mutations of the glutamatergic synapse pathway.

3.3.2. Genes involved in Glutamatergic synapse

The given genes involved in the glutamatergic synapse were crossed with the UniProt database and the desired information, such as gene names or UniProt accession codes, was mined. A database with pathogenic variants from ClinVar is then created, containing both pathogenic and likely pathogenic genes as well as their molecular consequences (Missense, Nonsense, Frameshift, and Splice site). Another database was created with neutral variants from GnomAD. Both databases were crossed to retrieve an ordered list by pathogenicity. This list was also crossed with the data from affected patients to prove if this L-Serine supplement would also be effective in patients with low glutamatergic synapses.

3.4. Tools

The tools used to develop this project are:

3.4.1. Python

Python is a programming language characterized by its easy comprehension. It is really useful for the development of any kind of tool. In this project, Python 3 is employed to treat the data and ending to extract the final results, with the combination of other mentioned tools. Python allows the use of different libraries, a set of precompiled instructions that can be utilized later in a program to perform certain well-defined activities. A library may also include literature, data format, message templates, objects, and variables, among other things (61,62). These are the most remarkable libraries used in this project:

- Jupyter notebook is the most recent interactive programming environment for notebooks, code, and data that is available on the web. Users can create and arrange workflows in bioinformatics, computational biology, digital media, and algorithms using its versatile interface. It is an interface, that allows the execution of code by blocks, a useful solution to find, identify and solve code issues (63).

- Selenium is an open-source tool which is used for automating tasks in web browsers or web applications. It allows, among other things, to replicate the steps a user can perform in a browser. In this project, it is used because GnomAD server does not allow access via API. Selenium is called to introduce every gene into the server and download its query (64).

3.4.2. Git Bash

Git Bash is a Windows application that gives Unix-based shell utilities as well as expertise with Git command line operations. It has been mostly used to perform verifications of Python code with synthetic and fast commands, but it also has been useful in some critical steps of the project, such as joining all downloaded GnomAD files or selecting human alignments from Pfam (65).

It is important to consider that all developed scripts can be run automatically to update the databases and be up-to-date at any desired moment. Scripts available in the following link: https://github.com/Igres17/TFG.git

4. Results and Discussion

4.1. Pathogenesis of homologous variants in membrane proteins

4.1.1. Pathogenesis of membrane proteins

The whole curated *Homo sapiens* proteome was retrieved from UniProt (accessed on 01/11/2021). The database of all human proteins consists of 20.386 proteins, belonging to 6.319 Pfam domains (see Table 1). From these proteins, only membrane proteins were kept, by selecting those proteins that contained at least one TM segment, resulting in a total of 5.202 human TM proteins and 1.485 Pfam codes.

Membrane proteins were classified as pathogenic or non-pathogenic. Pathogenic membrane proteins are those that present at least one disease-causing sequence variant related to a Mendelian genetic disease. By contrast, non-pathogenic membrane proteins are those where any disease-causing sequence variants are found. In these proteins, although mutations are able to affect the protein structure and function, this is not disease-causing. 1051 pathogenic membrane proteins, about the 20%, and 4.151 non-pathogenic membrane proteins were identified, corresponding to 706 Pfam and 779 Pfam codes, respectively. This is in contrast with previous studies that predicted that 90% of membrane proteins contained pathogenic variants (34). However, these 1051 pathogenic membrane proteins are the target of 25% of current drugs in the market (32).

The study is performed on these 1.051 membrane pathogenic proteins, as in this group of proteins the pathogenicity is related to how an amino acid change is able to affect the function and the folding of the protein due to changes in their physicochemical properties. This effect cannot be evaluated in non-pathogenic membrane proteins, because although protein folding and function may be affected in a sequence variant, theses are not disease-causing.

Table 1: Homo sapiens proteins and TM proteins, containing pathogenic or non-pathogenic mutations, and its Pfam codes

	Number of proteins	Number of Pfam codes
Human proteins	**20.386**	**6.319**
Human TM proteins	**5.202**	**1.485**
Human TM proteins with pathogenic mutations	**1.051**	**706**
Human TM proteins with non-pathogenic mutations	**4.151**	**779**

As it can be observed in the Figure 9, membrane proteins represent approximately 25% of the human proteome 20% of membrane proteins contain pathogenic mutations, whereas 80% are classified as non-pathogenic proteins, with no disease-causing variants (66).

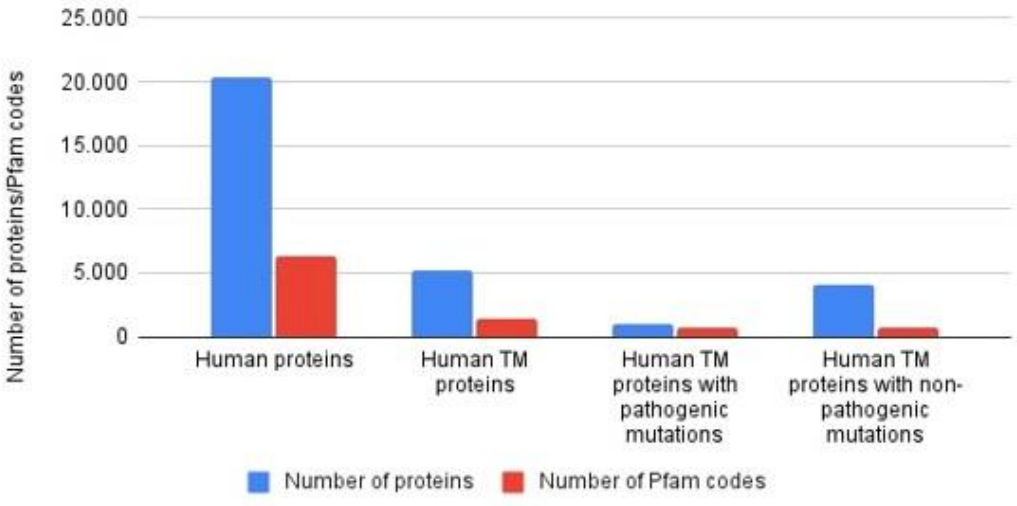

Figure 9: Number of Human proteins and Transmembrane proteins classification into pathogenic mutations and neutral mutations

4.1.2. Sequence variants in the human proteome

All neutral and disease-causing sequence variants were retrieved from GnomAD and ClinVar respectively (see Table 2). From these, a dataset containing missense, frameshifts, truncation, splice sites and UTR molecular consequences was created with a total of 3.937.973 sequence variants. Most of the dataset is represented by missense and splice sites + UTR mutations. Missense and splice sites + UTR mutations contain 1.880.489 and 1.820.077 sequence variants, 47,75% and 46,2% respectively. Thus, missense mutations are the most frequent sequence variants. This is not only for its number superiority, it is also because it must be considered that the other that has similar number is the total of two sequence variants, splice site and UTR. Sequence variants from frameshift and truncation mutations represent between 2 and 3% each of all the dataset, whereas missense and splice site + UTR sequence variants represent approximately 50%.

We can separate the previous dataset into disease-causing and neutral (see Figure 10 and Figure 11). There are some differences between neutral and disease-causing sequence variants datasets. As seen, the number of sequence variants in disease-causing between the different types of molecular consequences (Missense, Frameshift, Truncation, Splice Site + UTR) is similar. These are about 170.000 sequence variants, being the frameshift mutations the ones with more sequence variants with 57.049 and the splice sites + UTR the ones with less. These changes are a little different, and contrasting when comparing with the neutral sequence variants or with the general dataset. The neutral sequence variant group, frameshift, and truncation mutations go, practically, unnoticed. These, containing 49.708 and 89.313 sequence variants, represent a 1-2% of all neutral variants. Nevertheless, missense mutations and splice sites + UTRs are the ones with more neutral sequence variants, containing about two million neutral sequence variants each. That two million variants are about 50% of neutral sequence variants dataset so, practically, missense and splice site + UTR molecular consequences, represent the whole dataset.

If the comparatives between both datasets is performed, it can be observed that there's a difference of 3.597.977 sequences or what is the same, the disease-causing group represents a 4,3% of the whole sequence variants whereas the neutral group represents the other 95,7%. These percentages differ when analysing the data by molecular consequence individually. In missense and splice site + UTR mutations, the percentage remains similar to the percentage of the difference between databases sequences, but in frameshifts and

truncation, the numbers change. Missense neutral sequences represent a 97,5% whereas missense disease-causing sequences the rest 2,5%, in splice sites+ UTR the neutral sequences represent a 98,6% and the disease-causing a 1,4%. On the other hand, frameshift neutral sequences represent a 46,6% and the disease-causing a 54,4% and the truncation neutral is a 68,4% whereas the disease-causing is a 31,6%.

Table 2: Classification of neutral and disease-causing mutations by their molecular consequences and the region where they are found. A comparison between neutral and pathogenic variants is performed using a percentage

	Pathogenicity	Missense	Frameshift	Truncations	Splice Sites + UTR	Total
Sequence variants	Neutral	1.833.470 (97,5%)	49.708 (46,6%)	89.313 (68,4%)	1.795.484 (98,7%)	3.767.975 (95,7%)
	Disease-causing	47.025 (2,5%)	57.049 (53,4%)	41.331 (31,6%)	24.593 (1,3%)	169.998 (4,3%)
	Total	1.880.495	106.757	130.644	1.820.077	3.937.973
Sequence variants in Membrane Proteins	Neutral	439.819 (96,8%)	9.264 (45,4%)	16.566 (65,3%)	484.191 (98,8%)	949.840 (95,9%)
	Disease-causing	14.345 (3,2%)	11.157 (54,6%)	8.804 (34,7%)	5.805 (1,2%)	40.111 (4,1%)
	Total	454.173	20.421	25.370	489.996	989.951
Sequence variants in TM segments	Neutral	45.337 (93,8%)	-	-	-	-
	Disease-causing	2.991 (6,2%)	-	-	-	-
	Total	48.328	-	-	-	-

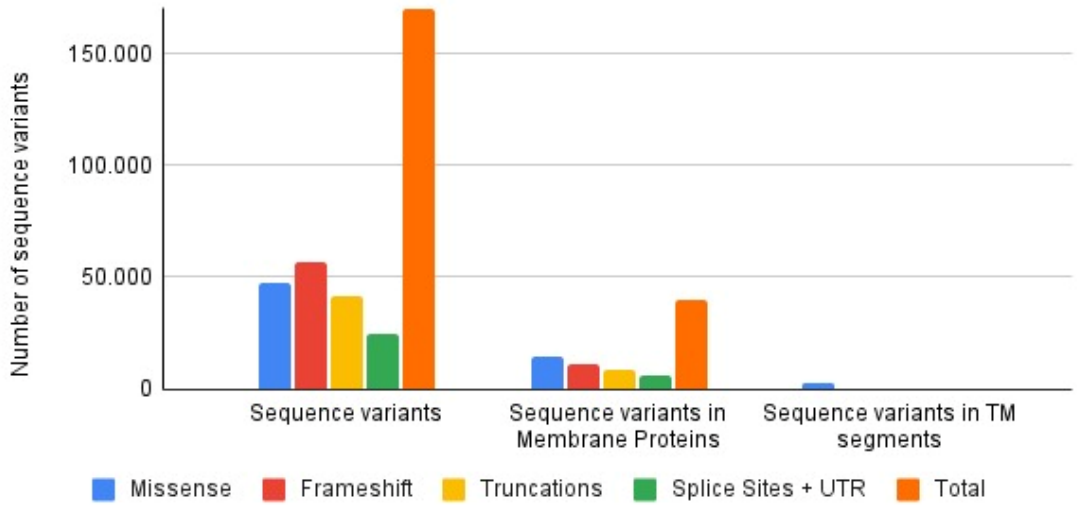

Figure 10: Disease-causing sequence variants representation (ClinVar).

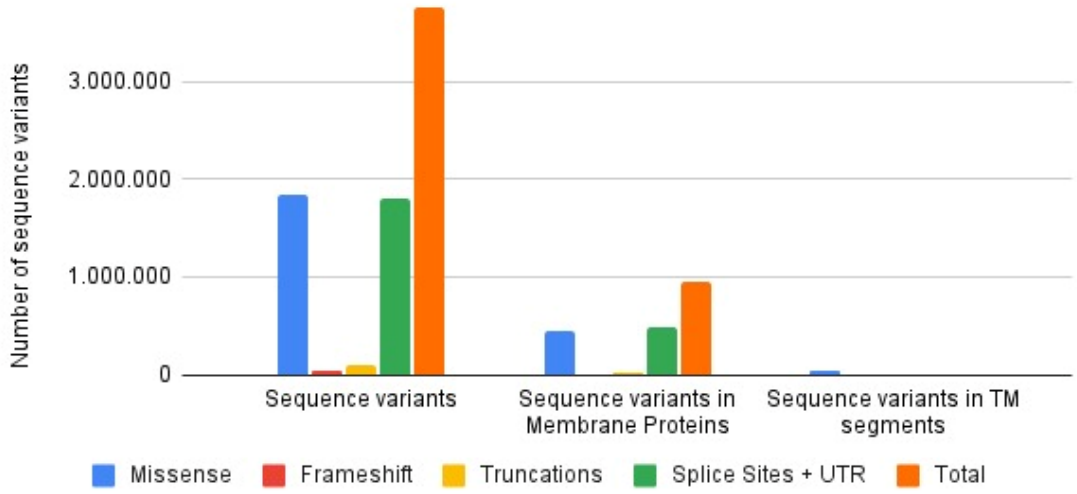

Figure 11: Neutral sequence variants representation (GnomAD).

4.1.3. Sequence variants in human membrane proteins

989.951 (out of 3.937.973) correspond to sequence variants in membrane proteins. Membrane sequences represent, approximately, 25% of all sequence variants found in *Homo sapiens*. These results show that the number of sequence variants in MP is in accordance with the number of membrane proteins and thus, membrane proteins do not seem more pathogenic than globular proteins.

The number of sequence variants in disease-causing between the different types of molecular consequences remains similar, but now represents, the MP, a 25% of all sequence variations (see Table 2). A remarkable change is that missense variants in MP are now the group with more sequences, 14.345. The neutral sequence variants remain similar, with a reduction of the number of sequences up to 75% for each molecular consequence.

These sequence variants were filtered out in order to obtain missense variants contained in the transmembrane segments. This region is highly conserved in structure and function (67) and thus presents reliable Pfam sequence alignments. The total number of sequence variants in the TM segments of membrane proteins are 48.328 missense variants in the TM segments, which represents 10% of the missense variants in membrane proteins. The database is not containing frameshift, truncation, splice sites or UTR variants because our final goal is to compare the pathogenicity of homologous variants missense mutations in transmembrane segments resulted in 45.337 neutral and 2.991 pathogenic missense mutations. 24 missense mutations were found in both datasets, as disease-causing and neutral, corresponding to 533 proteins. These proteins were deleted from the dataset since they possibly correspond to more complex diseases or recessive genetic diseases.

4.1.4. Pathogenesis of Homologous variants in the TM segments of membrane proteins

The Pfam sequence alignments allowed to identify a total of 4.726 homologous variants, corresponding to 5.249 pairs of homologous variants. A pair of variants is considered homologous if the variants present the same initial and final amino acid and the same position in the Pfam alignment. As we are using the transmembrane segments of membrane proteins that are highly conserved, the Pfam alignments are highly reliable, allowing to find equivalent positions between homologous proteins (68).

In order to assess if pathogenicity could be extrapolated between pairs of homologous variants, the pathogenicity of homologous variants was compared.

282 homologous variants, corresponding to 191 pairs were identified as disease-causing and 4.434 homologous variants, corresponding to 5.028 pairs were identified as neutral. There were no pairs of homologous variants with a different pathogenicity classification. These results indicate that the pathogenicity of homologous variants can be extrapolated for neutral variants and also for disease-causing variants in the TM segments of membrane proteins.

Thus, it seems that the high sequence and structure conservation in the TM segments of membrane proteins also implies conservation in pathogenicity. Thus, if one amino acid change is able to affect the function and the structure of a protein, the same amino acid change in the same equivalent position in homologous proteins, will produce the same effect. These findings allow increasing the dataset of disease-causing and neutral, as the pathogenesis of homologous variants in the TM segments of membrane proteins can be predicted with high reliability.

It would be interesting to check if, in addition to variants in equivalent positions, the pathogenicity of similar homologous variants can also be extrapolated. Similar homologous variants are those variants in which the final amino acids are not identical but with similar physicochemical properties (i.e. present a high score in a scoring matrix).

Also, it would be interesting to test if the pathogenesis of homologous variants can also be extrapolated beyond the non-TM segments of membrane proteins or for globular proteins, where sequence and structure is less conserved.

Table 3: Homologous mutations

	Variants	**Pairs**(69)
Neutral Homologous variants	**4.434**	**5.058**
Disease-causing homologous variants	**282**	**191**
Disease-causing and neutral homologous variants	**0**	**0**
Total number of homologous variants (pairs)	**4.726**	**5.249**

4.2. Mutations in genes involved in the Glutamatergic synapse

From 397 genes that were identified from 1.233 GO codes, sequence variants were extracted (frameshift, missense, nonsense, splice site…) from ClinVar and GnomAD to study the most pathogenic genes involved in glutamatergic synapse pathway. All these genes involved in this synapse, were analyzed and ordered, from the most pathogenic to the less pathogenic gene. It is known that, pathogenic variants in this glutamatergic synapse may develop in neurodevelopmental disorders, usually classified as rare diseases (45).

Table 4 show the first 15 most pathogenic genes found in this pathway. As previously mentioned, these are classified from most pathogenic to less pathogenic using ClinVar as the main database to perform the classification. From this, the most pathogenic gene in the synapse is the DMD which is involved in muscle dystrophy.

Scarce treatments are available for treating these rare diseases but repentantly, a treatment with L-serine, a nutraceutical compound, has been discovered to treat GRIN related disorders. A disorder characterized by NMDA receptor hypofunction and a reduction in glutamatergic synapses, has paved the way for L-serine treatment to be extended to other glutamatergic synaptic genes-related diseases characterized by a reduction in glutamatergic synapses. The results show the list of genes that present more mutations in the glutamatergic synapse. This list has now been shared with our clinical collaborators in order to start a study with patients having disease-causing mutations in glutamatergic genes in order to activate the glutamatergic synapse with L-serine.

Table 4: The most 15 pathogenic genes involved in the Glutamatergic synapse.

Gene(s)	Frameshift	Missense	Nonsense	Splice site	Total	Disease(s)/Function
DMD	379	25	515	208	1.127	Muscle dystrophy
TSC2	258	92	173	159	682	Tuberous Sclerosis 2 Protein
PTEN	260	157	118	95	630	+++ phosphatase
SCN1A	165	277	88	57	587	Dravet syndrome
MECP2	334	81	65	27	507	Rett syndrome
DICER1	213	46	146	73	478	Endoribonuclease helicase
LOC102724058|SCN1A	108	251	76	35	434	Dravet syndrome
LAMA2	124	16	171	109	420	Extracellular matrix. Musc. Dystrophy
SCN2A	60	234	48	19	361	+++
TSC1	168	10	119	59	356	Tuberous Sclerosis 1 Protein
CDKL5	145	82	72	42	341	+++ phosphatase
CDH1	118	19	66	55	258	Cadherin 1, cancer (gastric, endometrial, ovarian)
CACNA1A	69	80	71	33	253	
STXBP1	65	82	43	50	240	Syntaxin Binding Protein

						1 (MUNC18-1); NDD
SYNGAP1	107	39	56	24	226	Synaptic Ras GTPase Activating Protein 1, ID, Epileptic encephalopathy

Supplementary table 1 shows extended Table 4 with the list of the 15 most pathogenic genes involved in the glutamatergic synapse. The complete table with all related information and containing all genes, with each disease, ClinVar variants, GnomAD variants and much more is available in the following Git Hub repository: https://github.com/Igres17/TFG/tree/main/TFG%20Graphs%20%26%20Tables

In this supplementary table, the quantity of pathogenic and non-pathogenic variants available for these gens in ClinVar and GnomAD databases is recovered. The ratio between pathogenic and non-pathogenic databases has been calculated to compare their size of the sample in relation to each other. This allows to measure and express quantities by making them easier to interpret.

Supplementary table 1: List of 15 first most pathogenic genes involved in glutamatergic synapse, the number of disease-causing variants, the number of neutral variants and the corresponding disease.

Gene(s)	Total ClinVar	Total GnomAD	Totals ClinVar + GnomAD	Ratio P/NP	Disease(s)/Function
DMD	1.127	4.661	5.788	24,2	Muscle dystrophy
TSC2	682	4.451	5.133	15,3	Tuberous Sclerosis 2 Protein
PTEN	630	456	1.086	138,2	+++ phosphatase
SCN1A	587	1.798	2.385	32,6	Dravet syndrome
MECP2	507	673	1.180	75,3	Rett syndrome

DICER1	**478**	**2.056**	**2.534**	**23,2**	**Endoribonuclease helicase**
LOC102724058| SCN1A	**434**	**0**	**434**		**Dravet syndrome**
LAMA2	**420**	**4.932**	**5.352**	**8,5**	**Extracellular matrix. Musc. Dystrophy**
SCN2A	**361**	**15**	**376**	**2.406,7**	**+++**
TSC1	**356**	**1.397**	**1.753**	**25,5**	**Tuberous Sclerosis 1 Protein**
CDKL5	**341**	**889**	**1.230**	**38,4**	**+++ phosphatase**
CDH1	**258**	**1.321**	**1.579**	**19,5**	**Cadherin 1, cancer (gastric, endometrial, ovarian)**
CACNA1A	**253**	**3.358**	**3.611**	**7,5**	
STXBP1	**240**	**924**	**1.164**	**26,0**	**Syntaxin Binding Protein 1 (MUNC18-1); NDD**
SYNGAP1	**226**	**1.355**	**1.581**	**16,7**	**Synaptic Ras GTPase Activating Protein 1, ID, Epileptic encephalopathy**

5. Conclusions

Results support the initial hypothesis that because membrane proteins in the TM region are highly conserved, the pathogenicity between homologous variants (i.e. the same amino acid change in the same equivalent position is an homologous protein) is also conserved. When two similar proteins are homologous and one pathogenic mutation affects one of those two proteins, it can be extrapolated that the same mutation will be pathogenic in an equivalent position of a homologous protein. Additionally, we have found that only 20% of membrane proteins present disease-causing variants, in contrast with previous studies, although they represent 75% of drug targets.

The use of homologous variants to predict the pathogenesis of a sequence variant can help in prediction when the variant does have any information regarding pathogenesis.

It would be interesting to study if, in addition to equivalent mutations, the pathogenesis of similar homologous variants can also be extrapolated. Also to study to which degree the pathogenesis of homologous variants can also be extrapolated in the non-TM segments of membrane proteins or in globular proteins.

Since the use of L-serine as a therapy for GRIN-related illnesses was found, able to increase glutamatergic synapse, it is hypothesized that the treatment could be extended to other glutamatergic synaptic genes-related diseases with a reduction in glutamatergic synapses. We have identified the genes involved in this synapse through GO codes and identified the list of the most pathogenic genes in glutamatergic synapse in order study if L-serine treatment can be extrapolated to these diseases caused by genes involved in the glutamatergic synapse.

6. Bibliography

1. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. Nucleotide sequence of bacteriophage λ DNA. Journal of Molecular Biology. 1982 Dec 25;162(4):729–73.

2. Mount DW. Bioinformatics: Sequence and Genome Analysis . (2nd ed.). Spring Harbor Press; 2004.

3. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020 581:7809 [Internet]. 2020 May 27 [cited 2022 May 26];581(7809):434–43. Available from: https://www.nature.com/articles/s41586-020-2308-7

4. Behjati S, Tarpey PS. What is next generation sequencing? Archives of Disease in Childhood - Education and Practice [Internet]. 2013 Dec 1 [cited 2022 May 19];98(6):236–8. Available from: https://ep.bmj.com/content/98/6/236

5. Epilepsy ILA. GENETIC ETIOLOGY [Internet]. International League Against Epilepsy. [cited 2022 May 19]. Available from: https://www.epilepsydiagnosis.org/aetiology/genetic-groupoverview.html

6. Wilfert AB, Sulovari A, Turner TN, Coe BP, Eichler EE. Recurrent de novo mutations in neurodevelopmental disorders: properties and clinical implications. Genome Medicine 2017 9:1 [Internet]. 2017 Nov 27 [cited 2022 May 19];9(1):1–16. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0498-x

7. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome Sequencing: Current and Future Perspectives. G3 Genes|Genomes|Genetics [Internet]. 2015 Aug 1 [cited 2022 Jun 6];5(8):1543–50. Available from: https://academic.oup.com/g3journal/article/5/8/1543/6025359

8. Medline Plus. What are whole exome sequencing and whole genome sequencing?: MedlinePlus Genetics [Internet]. National Library of Medicine. 2021 [cited 2022 Jun 6]. Available from: https://medlineplus.gov/genetics/understanding/testing/sequencing/

9. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. The American Journal of Human Genetics. 2015 Aug 6;97(2):199–215.

10. Carvallo P. CONCEPTOS SOBRE GENÉTICA HUMANA PARA LA COMPRENSIÓN E INTERPRETACIÓN DE LAS MUTACIONES EN CÁNCER Y OTRAS PATOLOGÍAS HEREDITARIAS. Revista Médica Clínica Las Condes. 2017 Jul 1;28(4):531–7.

11. Papadimitriou S, Gazzo A, Versbraegen N, Nachtegael C, Aerts J, Moreau Y, et al. Predicting disease-causing variant combinations. Proc Natl Acad Sci U S A [Internet]. 2019 Jun 11 [cited 2022 Jun 6];116(24):11878–87. Available from: www.pnas.org/cgi/doi/10.1073/pnas.1815601116

12. Hanna N, Parfait B, Vidaud D, Vidaud M. [Mutation mechanisms and their consequences]. Medecine sciences : M/S [Internet]. 2005 [cited 2022 May 22];21(11):969–80. Available from: https://pubmed.ncbi.nlm.nih.gov/16274649/

13. McGarvey PB, Nightingale A, Luo J, Huang H, Martin MJ, Wu C, et al. UniProt genomic mapping for deciphering functional effects of missense variants. Human Mutation [Internet]. 2019 Jun 1 [cited 2022 Jun 6];40(6):694–705. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/humu.23738

14. Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. Genome Biology 2016 17:1 [Internet]. 2016 Nov 28 [cited 2022 Jun 6];17(1):1–19. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1110-1

15. National Human Genome Research Institute. Mutación | NHGRI [Internet]. National Human Genome Research Institute. [cited 2022 May 22]. Available from: https://www.genome.gov/es/genetics-glossary/Mutacion

16. CALIFORNIA STATE UNIVERSITY. DNA Mutation and Repair [Internet]. CSU. [cited 2022 May 22]. Available from: http://www2.csudh.edu/nsturm/CHEMXL153/DNAMutationRepair.htm

17. Davis A, Tinker A v., Friedlander M. "Platinum resistant" ovarian cancer: what is it, who to treat and how to measure benefit? Gynecol Oncol [Internet]. 2014 [cited 2022 Jun 6];133(3):624–31. Available from: https://pubmed.ncbi.nlm.nih.gov/24607285/

18. Peter J. Russell. iGenetics: : a molecular approach. Pearson Education Inc as BC, editor. 2010.

19. Deletion Mutation - Definition and Examples | Biology Dictionary [Internet]. [cited 2022 May 22]. Available from: https://biologydictionary.net/deletion-mutation/

20. Dr. Noel Sturm. DNA Mutation and Repair [Internet]. 2019 [cited 2022 Jun 6]. Available from: http://www2.csudh.edu/nsturm/CHEMXL153/DNAMutationRepair.htm

21. Biologyonline. Silent mutation Definition and Examples - Biology Online Dictionary [Internet]. biologyonline. [cited 2022 May 22]. Available from: https://www.biologyonline.com/dictionary/silent-mutation

22. Kryukov G v., Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. American Journal of Human Genetics. 2007;80(4):727–39.

23. Biologyonline. Nonsense mutation Definition and Examples - Biology Online Dictionary [Internet]. biologyonline. [cited 2022 May 22]. Available from: https://www.biologyonline.com/dictionary/nonsense-mutation

24. García Recio A. Bioinformatics tools for membrane proteins: from sequences to structure and function [Internet]. Uvic; 2022 [cited 2022 Jun 6]. Available from: http://dspace.uvic.cat/xmlui/handle/10854/7032

25. Niroula A, Vihinen M. Variation Interpretation Predictors: Principles, Types, Performance, and Choice. Hum Mutat [Internet]. 2016 Jun 1 [cited 2022 Jun 6];37(6):579–97. Available from: https://pubmed.ncbi.nlm.nih.gov/26987456/

26. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Research [Internet]. 2012 Jul 1 [cited 2022 May 21];40(W1):W452–7. Available from: https://academic.oup.com/nar/article/40/W1/W452/1751364

27. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nature Methods 2014 11:4 [Internet]. 2014 Mar 28 [cited 2022 May 21];11(4):361–2. Available from: https://www.nature.com/articles/nmeth.2890

28. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Current Protocols in Human Genetics [Internet]. 2013 Jan 1 [cited 2022 May 21];76(1):7.20.1-7.20.41. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/0471142905.hg0720s76

29. Garcia-Recio A, Gómez-Tamayo JC, Reina I, Campillo M, Cordomí A, Olivella M. TMSNP: a web server to predict pathogenesis of missense mutations in the transmembrane region of membrane proteins. NAR Genomics and Bioinformatics [Internet]. 2021 Jan 6 [cited 2022 May 21];3(1). Available from: https://academic.oup.com/nargab/article/3/1/lqab008/6148837

30. Murray JE, Laurieri N, Delgoda R. Proteins. Pharmacognosy: Fundamentals, Applications and Strategy. 2017 Jan 1;477–94.

31. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Research [Internet]. 2014 Jan 1 [cited 2022 Jun 6];42(Database issue):D310. Available from: /pmc/articles/PMC3964979/

32. Tan S, Hwee TT, Chung MCM. Membrane proteins and membrane proteomics. Proteomics. 2008 Oct;8(19):3924–32.

33. Almeida JG, Preto AJ, Koukos PI, Bonvin AMJJ, Moreira IS. Membrane proteins structures: A review on computational modeling tools. Biochimica et Biophysica Acta (BBA) - Biomembranes. 2017 Oct 1;1859(10):2021–39.

34. Ganesan K, Kulandaisamy A, Binny Priya S, Michael Gromiha M. HuVarBase: A human variant database with comprehensive information at gene and protein levels. PLOS ONE [Internet]. 2019 Jan 1 [cited 2022 Jun 6];14(1):e0210475. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0210475

35. Brown DA. Lipid rafts, detergent-resistant membranes, and raft targeting signals. Physiology (Bethesda) [Internet]. 2006 Dec [cited 2022 Jun 6];21(6):430–9. Available from: https://pubmed.ncbi.nlm.nih.gov/17119156/

36. Olivella M, Gonzalez A, Pardo L, Deupi X. Relation between sequence and structure in membrane proteins. Bioinformatics [Internet]. 2013 Jul 1 [cited 2022 Jun 6];29(13):1589–92. Available from: https://academic.oup.com/bioinformatics/article/29/13/1589/200225

37. Mayol E, Campillo M, Cordomí A, Olivella M. Inter-residue interactions in alpha-helical transmembrane proteins. Bioinformatics [Internet]. 2019 Aug 1 [cited 2022 Jun 6];35(15):2578–84. Available from: https://academic.oup.com/bioinformatics/article/35/15/2578/5253260

38. Sato H, Tamba M, Ishii T, Bannai S. Cloning and expression of a plasma membrane cystine/glutamate exchange transporter composed of two distinct proteins. J Biol Chem [Internet]. 1999 Apr 23 [cited 2022 Jun 6];274(17):11455–8. Available from: https://pubmed.ncbi.nlm.nih.gov/10206947/

39. Eulenburg V, Gomeza J. Neurotransmitter transporters expressed in glial cells as regulators of synapse function. Brain Res Rev [Internet]. 2010 May [cited 2022 Jun 6];63(1–2):103–12. Available from: https://pubmed.ncbi.nlm.nih.gov/20097227/

40. Kew JNC, Kemp JA. Ionotropic and metabotropic glutamate receptor structure and pharmacology. Psychopharmacology (Berl) [Internet]. 2005 Apr [cited 2022 Jun 6];179(1):4–29. Available from: https://pubmed.ncbi.nlm.nih.gov/15731895/

41. Diagnostics C. Glutamatergic Synapse Pathway - Creative Diagnostics [Internet]. Creative Diagnostics. [cited 2022 May 22]. Available from: https://www.creative-diagnostics.com/glutamatergic-synapse-pathway.htm

42. Fonnum F. Glutamate: a neurotransmitter in mammalian brain. J Neurochem [Internet]. 1984 [cited 2022 Jun 6];42(1):1–11. Available from: https://pubmed.ncbi.nlm.nih.gov/6139418/

43. Bliss TVP, Collingridge GL. A synaptic model of memory: long-term potentiation in the hippocampus. Nature 1993 361:6407 [Internet]. 1993 [cited 2022 Jun 6];361(6407):31–9. Available from: https://www.nature.com/articles/361031a0

44. Hinoi E, Takarada T, Ueshima T, Tsuchihashi Y, Yoneda Y. Glutamate signaling in peripheral tissues. Eur J Biochem [Internet]. 2004 Jan [cited 2022 Jun 6];271(1):1–13. Available from: https://pubmed.ncbi.nlm.nih.gov/14686914/

45. Moretto E, Murru L, Martano G, Sassone J, Passafaro M. Glutamatergic synapses in neurodevelopmental disorders. Progress in Neuro-Psychopharmacology and Biological Psychiatry. 2018 Jun 8;84:328–42.

46. García-Recio A, Santos-Gómez A, Soto D, Julia-Palacios N, García-Cazorla À, Altafaj X, et al. GRIN database: A unified and manually curated repertoire of GRIN variants. Hum Mutat [Internet]. 2021 Jan 1 [cited 2022 May 22];42(1):8–18. Available from: https://pubmed.ncbi.nlm.nih.gov/33252190/

47. Meldrum BS. Glutamate as a neurotransmitter in the brain: review of physiology and pathology. J Nutr [Internet]. 2000 [cited 2022 Jun 6];130(4S Suppl). Available from: https://pubmed.ncbi.nlm.nih.gov/10736372/

48. Santoro MR, Bray SM, Warren ST. Molecular mechanisms of fragile X syndrome: a twenty-year perspective. Annu Rev Pathol [Internet]. 2012 [cited 2022 Jun 6];7:219–45. Available from: https://pubmed.ncbi.nlm.nih.gov/22017584/

49. Javitt DC. Glutamate as a therapeutic target in psychiatric disorders. Mol Psychiatry [Internet]. 2004 Nov [cited 2022 Jun 6];9(11):984–97. Available from: https://pubmed.ncbi.nlm.nih.gov/15278097/

50. Moretto E, Murru L, Martano G, Sassone J, Passafaro M. Glutamatergic synapses in neurodevelopmental disorders. Progress in Neuro-Psychopharmacology and Biological Psychiatry. 2018 Jun 8;84:328–42.

51. Soto D, Olivella M, Grau C, Armstrong J, Alcon C, Gasull X, et al. L-Serine dietary supplementation is associated with clinical improvement of loss-of-function GRIN2B-related pediatric encephalopathy. Sci Signal [Internet]. 2019 Jun 18 [cited 2022 Jun 6];12(586). Available from: https://pubmed.ncbi.nlm.nih.gov/31213567/

52. Seyhan AA, Carini C. Are innovation and new technologies in precision medicine paving a new era in patients centric care? Journal of Translational Medicine [Internet]. 2019 Apr 5 [cited 2022 May 22];17(1):1–28. Available from: https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-019-1864-9

53. Miladinovic T, Nashed MG, Singh G. Overview of Glutamatergic Dysregulation in Central Pathologies. Biomolecules 2015, Vol 5, Pages 3112-3141 [Internet]. 2015 Nov 11 [cited 2022 May 22];5(4):3112–41. Available from: https://www.mdpi.com/2218-273X/5/4/3112/htm

54. Bateman A. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res [Internet]. 2019 Jan 8 [cited 2022 May 26];47(D1):D506–15. Available from: https://pubmed.ncbi.nlm.nih.gov/30395287/

55. McGarvey PB, Nightingale A, Luo J, Huang H, Martin MJ, Wu C, et al. UniProt genomic mapping for deciphering functional effects of missense variants. Human Mutation [Internet]. 2019 Jun 1 [cited 2022 May 26];40(6):694–705. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/humu.23738

56. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res [Internet]. 2014 Jan 1 [cited 2022 May 26];42(Database issue). Available from: https://pubmed.ncbi.nlm.nih.gov/24234437/

57. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Research [Internet]. 2021 Jan 8 [cited 2022 Jun 6];49(D1):D412–9. Available from: https://academic.oup.com/nar/article/49/D1/D412/5943818

58. gnomAD. gnomAD [Internet]. 2016 [cited 2022 Jun 6]. Available from: https://gnomad.broadinstitute.org/

59. Koch L. Exploring human genomic diversity with gnomAD. Nature Reviews Genetics 2020 21:8 [Internet]. 2020 Jun 2 [cited 2022 Jun 6];21(8):448–448. Available from: https://www.nature.com/articles/s41576-020-0255-7

60. Uniprot. Homo sapiens, reviewed and transmem in UniProtKB [Internet]. Uniprot. 2022 [cited 2022 Jun 6]. Available from: https://www.uniprot.org/uniprot/?query=reviewed%3Ayes+organism%3A%22Homo+sapiens+%28Human%29+%5B9606%5D%22+annotation%3A%28type%3Atransmem%29&sort=score

61. parthmanchanda81. Libraries in Python - GeeksforGeeks [Internet]. https://www.geeksforgeeks.org/. 2021 [cited 2022 Jun 6]. Available from: https://www.geeksforgeeks.org/libraries-in-python/

62. Python Software Foundation. What is Python? Executive Summary | Python.org [Internet]. Python.org. 2020 [cited 2022 Jun 6]. Available from: https://www.python.org/doc/essays/blurb/

63. The Jupyter Trademark. Project Jupyter | Home [Internet]. https://jupyter.org/. 2022 [cited 2022 Jun 6]. Available from: https://jupyter.org/

64. Neha Vaidya. Selenium Using Python: All You Need to Know | Edureka [Internet]. edureka. 2021 [cited 2022 Jun 6]. Available from: https://www.edureka.co/blog/selenium-using-python/

65. Bitbucket. Git bash: Definition, commands, & getting started | Atlassian [Internet]. atlassian.com. [cited 2022 Jun 6]. Available from: https://www.atlassian.com/git/tutorials/git-bash

66. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research [Internet]. 2016 Jan 1 [cited 2022 Jun 6];44(Database issue):D279. Available from: /pmc/articles/PMC4702930/

67. Gaddie KJ, Kirley TL. Conserved Polar Residues Stabilize Transmembrane Domains and Promote Oligomerization in Human Nucleoside Triphosphate Diphosphohydrolase 3 (NTPDase3). Biochemistry [Internet]. 2009 Oct 10 [cited 2022 Jun 7];48(40):9437. Available from: /pmc/articles/PMC2758327/

68. Pearson WR. An Introduction to Sequence Similarity ("Homology") Searching. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis . [et al] [Internet]. 2013 [cited 2022 Jun 6];0 3(SUPPL.42). Available from: /pmc/articles/PMC3820096/

69. NA. Calculadora de Combinaciones - Encuentra posibles combinaciones [Internet]. calculator-online.net/. 2020 [cited 2022 Jun 6]. Available from: https://calculator-online.net/es/combination-calculator/

70. merilutururu. Mutations: deletion, dna, duplication, en, genes, genetics, insertion, inversion, mutations, science | Glogster EDU - Interactive multimedia posters [Internet]. Glogster. 2015 [cited 2022 Jun 6]. Available from: https://edu.glogster.com/glog/mutations/28ld78b65tm?=glogpedia-source