U SCIENCE TECH
FACULTAT DE CIÈNCIES
I TECNOLOGIA
UVIC-UCC

Treball de Fi de Grau

Prediction and punctuation of the effect of non-synonymous single nucleotide polymorphisms on the structure and function of membrane proteins.

Iker Reina Fuente

Grau en biotecnologia
Tutors: Mireia Olivella
Arnau Cordomi
Vic, 09 de 2016

Index

ABSTRACT

Title: Prediction and punctuation of the effect of non-synonymous single nucleotide polymorphisms on the structure and function of membrane proteins.
Key words: Single nucleotide polymorphism (SNP), transmembrane (TM), protein, mutation, UniProt, python, MYSQL, missense, SNPtmDB, bootstrap.
Author: Iker Reina Fuente
Tutors: Dr. Arnau Cordomí (UVic – UAB) i Dra. Mireia Olivella (UVic - UAB)
Data: September 2016

Membrane proteins are targets for over 60 % of currently marketed drugs, as they are involved in a wide variety of diseases. During these last years, many single nucleotide polymorphisms (SNP) have been identified in the transmembrane region of alpha membrane proteins.

The aim of this study is to develop a web server able to calculate, for all SNPs identified in the transmembrane region of membrane proteins, parameters that can be useful to predict if a SNP can affect the structure and/or function of a membrane protein, resulting in a pathological mutation. These parameters include entropy, amino acid frequencies and substitution score based on sequence alignments.

In order to analyse the capacity of the web server to predict the effect of a SNP in the structure and function of the protein, the entropy, amino acid frequencies and substitution scores have been compared for pathological and non-pathological SNPs.

The results show significative differences between the parameters of both groups, indicating that the information provided by the web server can be used to predict the effect of SNP on membrane proteins.

INTRODUCTION

A genome is the complete set of nucleic acid sequence—Deoxyribonucleic acid in humans (DNA)—that contains all of the information needed to develop and maintain that organism. The genome consists coding regions (genes) and noncoding regions. In humans, a copy of the entire genome—more than 3 billion DNA base pairs—is contained in all nucleated cells. DNA consists of two biopolymer strands coiled around each other to form a double helix of four different nucleotides (adenine, cytosine, guanine and thymine, A, C, G and T respectively).[1]

Genes control different characteristics such as eye colour, height or susceptibility to specific diseases. The development of new technologies has made genome sequencing dramatically easy and cheap, and the number of complete genome sequences is growing rapidly. This permitted to identify many genetic differences between the genome of different humans. New sequencing technologies, such as massive parallel sequencing, have also opened up the prospect of personal genome sequencing as a diagnostic tool. This diagnostic includes genetic illness and predisposition to an illness.[2]

Genetic Variations

Genomic information suffers alterations in its nucleotide sequence. Mutations appear when this change in the DNA sorts the error-prone repair and remains in the genome. Damage in DNA is constantly present, it can be caused by external processes or DNA replication damage[3].

There are differences between somatic and germinal cell mutations. If a somatic mutation occurs in a single cell in a developing somatic tissue, all cells descended from the mutated cells will be mutated. A germinal mutation occurs in the germline, a special tissue that is set aside in the course of development to form sex cells. If a mutant sex cell participates in fertilization, then the mutation will be passed on to the next generation. An individual of perfectly normal phenotype and of normal ancestry can harbour undetected mutant sex cells[4]. Depending on their specific localization in the genome, mutations may or not have altered function or expression[5].

SNPs

A single nucleotide polymorphism (abbreviated as SNP; Figure 1), is a genetic variation present to some appreciable degree within a population (e.g. >1%) in a single nucleotide that occurs at a specific position in the genome[6]. For example, base A may appear in most individuals at a specific base position in the human genome, but in a minority of individuals, the position may be occupied by base C. Thus, we will say that there is a SNP at this specific base position, and the two possible nucleotide variations —C or A— will be the alleles for this base position.

Single-nucleotide polymorphisms may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions (regions between genes)[7]. SNPs that falls in the protein coding region of a gene can be classified into three different kinds of SNP depending on whether it results in a change in the amino acid sequence or not. Silent mutations code for the same amino acid, missense mutations code for a different amino acid nonsense mutation code for a stop codon (see Figure 1)[8]. SNPs that are not in protein-coding regions may still affect gene splicing, transcription factor binding, messenger RNA degradation, or the sequence of non-coding RNA. Gene expression affected by this type of SNP is referred to as an eSNP (expression SNP) and may be upstream or downstream from the gene[9].

Thus, each SNP is different and may or not have functional consequences depending on its position or its amino acid change. SNPs underlie differences in our susceptibility to disease and may be associated with a wide range of diseases. Some of these SNP been associated to different kind of genetic diseases[10], but the effect of most SNPs in pathologies or in predisposition to suffer certain pathologies is still unknown.

In the context of the growing number of SNP recently identified, there is a need to develop tools able to predict the effect of SNPs in the structure and function of proteins and the association to pathologies.

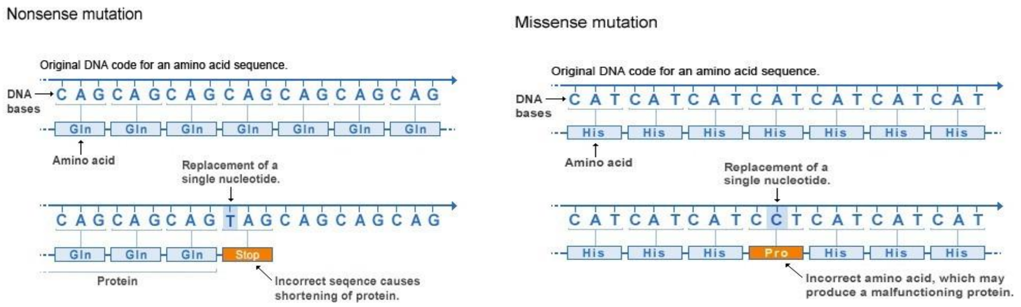

Nonsense mutation
Original DNA code for an amino acid sequence.
DNA bases
C A G C A G C A G C A G C A G C A G C A G
Gln Gln Gln Gln Gln Gln Gln
Amino acid
Replacement of a single nucleotide.
C A G C A G C A G T A G C A G C A G C A G
Gln Gln Gln Stop
Protein
Incorrect seqence causes shortening of protein.
Missense mutation
Original DNA code for an amino acid sequence.
DNA bases
C A T C A T C A T C A T C A T C A T C A T
His His His His His His His
Amino acid
Replacement of a single nucleotide.
C A T C A T C A T C C T C A T C A T C A T
His His His Pro His His His
Incorrect amino acid, which may produce a malfunctioning protein.

Figure 1. *Nonsense mutations introduce a codon stop, while missense mutations result in a change in the amino acid. (Taken from Natural library of medicine)*

Web servers available to predict the effect of SNPs

During the last years, the number of sequenced human genomes have exponentially increased, and so has grown the number of reported new identified SNPs. Many different tools have been developed in order to study the effect of SNP that produce amino acid changes in regions encoding proteins. Usually, an amino acid change with similar size and physico-chemical properties (e.g. substitution of leucine by valine, as in SNP rs13166360) has mild effect, and vice-versa. Moreover, if a SNP disrupts secondary structure elements (e.g. substitution of proline by another residue in an alpha helix) such mutation usually may affect whole protein structure and function.

Using those simple and many other machine learning derived rules a group of programs for the prediction of SNP effect was developed. Some examples are: SIFT[11], SNAP2[12], SuSPect[13], PolyPhen-2[14], PredictSNP[15] and Variant Effect Predictor from the Ensembl project[16]. Most of these web servers have limited predictive power and were designed to be employed with globular proteins, which feature a wide diversity of folds and primary structures.

Membrane proteins

Membrane proteins represent over 25 % of all proteins in sequenced genomes and mediate the interaction of the cell with its surroundings, including selective molecular transport, signalling, respiration and motility[17]. Because of their accessibility from the extracellular environment, membrane proteins (see Figure 2), they are the targets for over 60 % of currently marketed drugs. Due to the difficulty in over-expressing, purifying and crystallizing them, only 2 % of the structures deposited in Protein Data Bank are membrane proteins. Membrane proteins display specific features that differ from those of water-soluble ones, due to their different environment[18]. For instance, the number of folds that membrane proteins can adopt is limited to only α-helical bundles and β-barrels, due to the physical constraints imposed by the lipid bilayer. Consequently TM regions have specific distributions of amino acids (mostly hydrophobic)[19].

Given the lack of experimental structural information, computational tools specific for membrane proteins have become highly valuable. Thus developing tools for studying TM regions would help, for instance, studying how drugs target membrane proteins (see Figure 2).

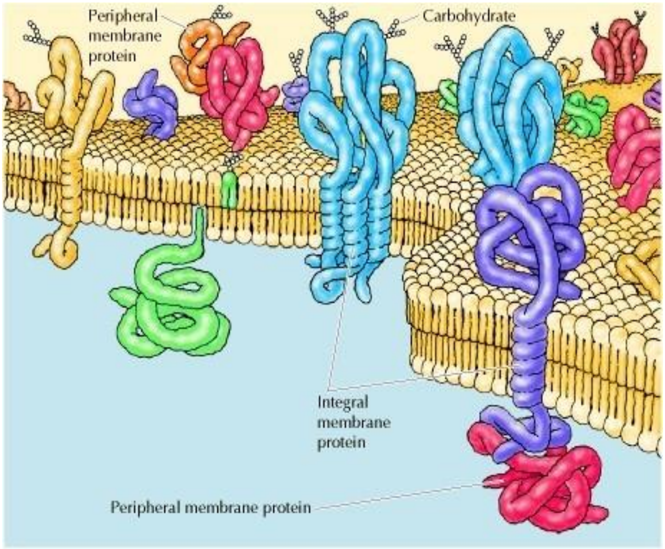
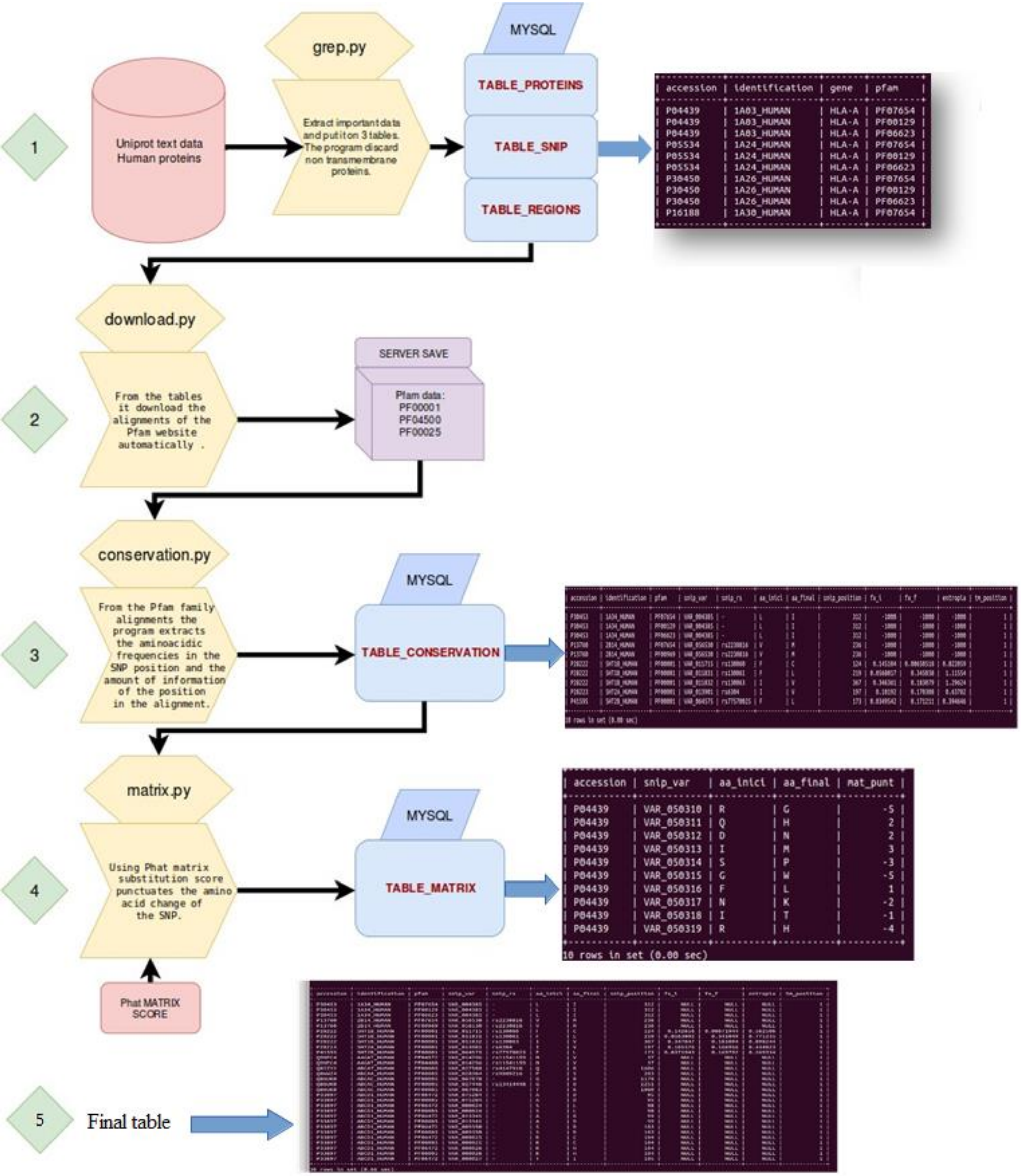

Peripheral membrane protein
Carbohydrate
Integral membrane protein
Peripheral membrane protein

Figure 2. *Cartoon representation of a membrane with different types of membrane proteins embedded (taken from http://oregonstate.edu/instruction/bi314/summer09/membranes.html)*

METHODS

SNPtmDB web server consists of both a database of SNPs located on the TM domains of alpha membrane proteins and a tool that computes quantitative parameters that evaluate the amino acidic change. The server has the ability to systematically survey sequences of TM regions and provide the users parameters such as frequencies (considering the same amino acid on the same family of proteins in a sequence alignment), information content (of the alignment), as well as punctuation of the amino acid change (using a substitution matrix). These parameters can be useful to classify SNP as pathological or non-pathological. SNPtmDB relies on Python programs in combination with a MySQL database and Bootstrap web-page interface.



1
Uniprot text data
Human proteins
grep.py
Extract important data and put it on 3 tables. The program discard non transmembrane proteins.
MYSQL
TABLE_PROTEINS
TABLE_SNP
TABLE_REGIONS
download.py
2
From the tables it download the alignments of the Pfam website automatically .
SERVER SAVE
Pfam data:
PF00001
PF04500
PF00025
conservation.py
3
From the Pfam family alignments the program extracts the aminoacidic frequencies in the SNP position and the amount of information of the position in the alignment.
MYSQL
TABLE_CONSERVATION
matrix.py
4
Using Phat matrix substitution score punctuates the amino acid change of the SNP.
MYSQL
TABLE_MATRIX
Phat MATRIX SCORE
5
Final table

```
+-----------+----------------+--------+---------+
| accession | identification | gene   | pfam    |
+-----------+----------------+--------+---------+
| P04439    | 1A03_HUMAN     | HLA-A  | PF07654 |
| P04439    | 1A03_HUMAN     | HLA-A  | PF00129 |
| P04439    | 1A03_HUMAN     | HLA-A  | PF06623 |
| P05534    | 1A24_HUMAN     | HLA-A  | PF07654 |
| P05534    | 1A24_HUMAN     | HLA-A  | PF00129 |
| P05534    | 1A24_HUMAN     | HLA-A  | PF06623 |
| P30450    | 1A26_HUMAN     | HLA-A  | PF07654 |
| P30450    | 1A26_HUMAN     | HLA-A  | PF00129 |
| P30450    | 1A26_HUMAN     | HLA-A  | PF06623 |
| P16188    | 1A30_HUMAN     | HLA-A  | PF07654 |
+-----------+----------------+--------+---------+
```

```
+-----------+----------------+---------+------------+-------------+---------+----------+---------------+-----------+-------------+-----------+-------------+
| accession | identification | pfam    | snip_var   | snip_rs     | aa_inici | aa_final | snip_position | fx_i      | fx_f        | entropia  | tm_position |
+-----------+----------------+---------+------------+-------------+---------+----------+---------------+-----------+-------------+-----------+-------------+
| P30453    | 1A34_HUMAN     | PF07654 | VAR_004385 | -           | L        | I        |           312 |     -1000 |       -1000 |     -1000 |           1 |
| P30453    | 1A34_HUMAN     | PF00129 | VAR_004385 | -           | L        | I        |           312 |     -1000 |       -1000 |     -1000 |           1 |
| P30453    | 1A34_HUMAN     | PF06623 | VAR_004385 | -           | L        | I        |           312 |     -1000 |       -1000 |     -1000 |           1 |
| P13760    | 2B14_HUMAN     | PF07654 | VAR_050530 | rs2230816   | V        | M        |           236 |     -1000 |       -1000 |     -1000 |           1 |
| P13760    | 2B14_HUMAN     | PF00969 | VAR_050530 | rs2230816   | V        | M        |           236 |     -1000 |       -1000 |     -1000 |           1 |
| P28222    | 5HT1B_HUMAN    | PF00001 | VAR_011715 | rs130060    | F        | C        |           124 |  0.145104 | 0.000505510 |  0.822859 |           1 |
| P28222    | 5HT1B_HUMAN    | PF00001 | VAR_011831 | rs130061    | F        | L        |           219 | 0.0560057 |    0.345838 |   1.11554 |           1 |
| P28222    | 5HT1B_HUMAN    | PF00001 | VAR_011832 | rs130063    | I        | V        |           367 |  0.346361 |    0.183879 |   1.29624 |           1 |
| P28223    | 5HT2A_HUMAN    | PF00001 | VAR_013901 | rs6304      | I        | V        |           197 |   0.10192 |    0.170308 |   0.43782 |           1 |
| P41595    | 5HT2B_HUMAN    | PF00001 | VAR_064575 | rs775708825 | F        | L        |           173 | 0.0349542 |    0.171211 |  0.394046 |           1 |
+-----------+----------------+---------+------------+-------------+---------+----------+---------------+-----------+-------------+-----------+-------------+
10 rows in set (0.00 sec)
```

```
+-----------+------------+----------+----------+----------+
| accession | snip_var   | aa_inici | aa_final | mat_punt |
+-----------+------------+----------+----------+----------+
| P04439    | VAR_050310 | R        | G        |       -5 |
| P04439    | VAR_050311 | Q        | H        |        2 |
| P04439    | VAR_050312 | D        | N        |        2 |
| P04439    | VAR_050313 | I        | M        |        3 |
| P04439    | VAR_050314 | S        | P        |       -3 |
| P04439    | VAR_050315 | G        | W        |       -5 |
| P04439    | VAR_050316 | F        | L        |        1 |
| P04439    | VAR_050317 | N        | K        |       -2 |
| P04439    | VAR_050318 | I        | T        |       -1 |
| P04439    | VAR_050319 | R        | H        |       -4 |
+-----------+------------+----------+----------+----------+
10 rows in set (0.00 sec)
```

```
| accession | identification | pfam | snip_var | snip_rs | aa_inici | aa_final | snip_position | fx_i | fx_f | entropia | tm_position |
```

Figure 3. *Schematic representation of how the dataset of SNPtmDB was produced. "Black boxes show examples of the MySQL tables generated." - The color-code employed in this figure is red: external input, yellow: programs, blue: MySQL tables names.*

Construction of the database of SNP in transmembrane segments

All human proteins are manually downloaded (in text format) from the UniProt web-server (see Figure 3). The program 'grep.py' reads the text and extracts SNP located in the TM region of these membrane proteins. The Uniprot accession, the amino acid sequence, Pfam family ID[20], amino acid change and position and pdb code was downloaded for each SNP. ‘table_SNP’, ‘table_regions’ and ‘table_proteins’ contains these data.

Multiple sequence alignments of the transmembrane domains

Using the database described in the previous section, the program ‘download.py’ automatically downloads the Pfam alignments and the files containing the information. Only Pfam accessions that appear in the TM proteins are downloaded (Figure 3).

Frequency of each position

The program 'conservation.py' combines the database with the Pfam-downloaded files to compute the frequency for an amino acid at each position of the alignment. Knowing the exact position of the SNP, ‘conservation.py’ looks for each amino acid in the column alignment and calculates frequencies of the reference amino acid (from the major allele) and the amino acid codified by the missense mutation. It also calculates the information content of the position alignment (see Figure 3). The relative frequencies of the amino acids are obtained as:

$$Fa(i) = na(i)/n(i)$$

Where *na(i)* is the number of sequences in which position *i* is occupied by amino acid *a*, and *n(i)* is the total number of aligned sequences in which position *i* is present (no gap at this position). Data is saved in 'table_conservation'.

Information content

The information content *I(X)* corresponds to the reduction in uncertainty (Entropy H) that occurs in the alignment observed:

$$I(X) = H(before) – H(after)$$

Where *H(before)* and *H(after)* are the order of a system measured by its entropy, before and after the alignment and it can be used in particular for measuring sequence variability, as was proposed for example by Shenkin et al. (1991) [21] and has been implemented in a number of studies.

Entropy for a position *i* is maximal if all 20 amino acids at this position have equal frequencies. We use entropy with the reverse sign defined on position-specific frequencies *fa(i)* to estimate the conservation index. Entropy does not take into account possible bias in amino acid composition or similarities among amino acids.

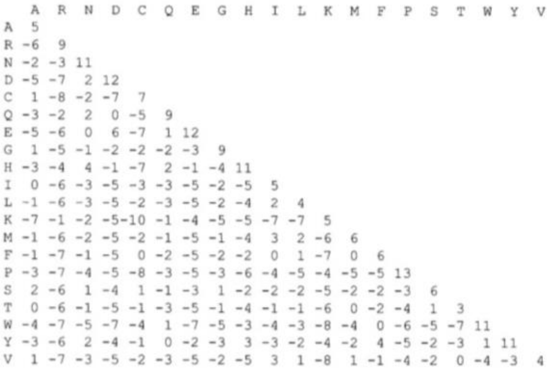
Entropy is calculated as a measure of the average uncertainty of a random variable. If *X* is a variable that can take *k* values *xi*, entropy *X* is defined as:

$$H(X) = -\sum_{i=1}^{k} p(x_i)\, log\, p(x_i)$$

Instead of saying that the entropy before the alignment is random (saying that all 20 amino acids at this position have equal frequencies) we calculate it from all the TM regions of the membrane proteins. So using the value of the initial entropy, to know the information content, we calculate the entropy of the column alignment and solve the information content equation. The data is saved in 'table_conservation'.

Substitution Scores

The program 'matrix.py' requests the amino acid change of the SNP and employs the PHAT75/73 (TM-specific) substitution matrix[22] to produce a substitution score (see Figure 4). Data is saved in 'table_matrix'.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5																			
R	-6	9																		
N	-2	-3	11																	
D	-5	-7	2	12																
C	1	-8	-2	-7	7															
Q	-3	-2	2	0	-5	9														
E	-5	-6	0	6	-7	1	12													
G	1	-5	-1	-2	-2	-2	-3	9												
H	-3	-4	4	-1	-7	2	-1	-4	11											
I	0	-6	-3	-5	-3	-3	-5	-2	-5	5										
L	-1	-6	-3	-5	-2	-3	-5	-2	-4	2	4									
K	-7	-1	-2	-5	-10	-1	-4	-5	-5	-7	-7	5								
M	-1	-6	-2	-5	-2	-1	-5	-1	-4	3	2	-6	6							
F	-1	-7	-1	-5	0	-2	-5	-2	-2	0	1	-7	0	6						
P	-3	-7	-4	-5	-8	-3	-5	-3	-6	-4	-5	-4	-5	-5	13					
S	2	-6	1	-4	1	-1	-3	1	-2	-2	-2	-5	-2	-2	-3	6				
T	0	-6	-1	-5	-1	-3	-5	-1	-4	-1	-1	-6	0	-2	-4	1	3			
W	-4	-7	-5	-7	-4	1	-7	-5	-3	-4	-3	-8	-4	0	-6	-5	-7	11		
Y	-3	-6	2	-4	-1	0	-2	-3	3	-3	-2	-4	-2	4	-5	-2	-3	1	11	
V	1	-7	-3	-5	-2	-3	-5	-2	-5	3	1	-8	1	-1	-4	-2	0	-4	-3	4

Figure 4. *The PHAT 75/73 matrix (H=0.5605) constructed from PHDhtm 75 (H = 0.5007) target values and Persson–Argos 73 background frequencies (H = 0.5038).*

Integration of all parameters

The different MySQL tables are linked through the Uniprot accession code. The 'final_table' combines the information obtained in the different steps: Uniprot accession code, protein ID, position in the sequence, sequence, amino acid change, aminoacid change frequencies, matrix punctuation, information content, pathological or non-pathological mutation and transmembrane or non-transmembrane position.

The web application collects the requested data from this table and return as output the information for all SNPs of the selected protein. (see Figure 3)

Statistical Analysis

In order to validate the prediction capability of the parameters computed by SNPtmDB, we addressed the comparison between pathological and non-pathological mutations. It is important, however, to take into account that not all SNPs that affect the structure and function are classified as pathological. This is the case, for instance, of SNPs that modify the binding sites of taste receptors. Although the result of such SNPs is the lack of (or decreased) taste sensation, these this does not derive into a pathology. For this reason, our non-pathological mutations contain only those mutations that appear in proteins with known pathological SNPs. The mean and the standard deviation were computed for each parameter of each group. Normality test and t-test were used in order to identify statistically differences between both groups. The statistical analysis, including box plots, was performed with scipy module of Python[23].

Database Update

The database will be regularly and automatically updated twice a year, in order to incorporate new SNP and the computed parameters.

RESULTS AND DISCUSSION

Overview of the current data

The initial set of SNPs was extracted from the UniProt database[24]. Table 1 shows an overview of main important data contained in SNPtmDB. The database currently contains 20.197 unique Homo sapiens proteins from which only 5192 proteins contained a TM domain. These proteins represent 1380 distinct families according to Pfam[20]. From the 5192 proteins with a TM domain, 1299 of them had at least a SNP that fell in the TM region. Approximately half of the SNPs in transmembrane regions are pathological. Contrastingly, approximately 30% of SNP in non-transmembrane regions are pathological (see Table 1).

	Number (amount)
Filtered proteins	5192
Total SNPs	24564
Pathogenic SNPs	9726
Transmembrane regions	4312
pathogenic SNPs and transmembrane	2265
pathogenic SNPs and non-transmembrane	7461
non-pathogenic and transmembrane	2047
non-pathogenic and non-transmembrane	12791

Table 1. *Overview of the data contained in SNPtmDB. There are 5192 human proteins containing a transmembrane domain. There are 2265 pathological SNP and 2047 non pathological SNP in the transmembrane region of human membrane proteins.*

Pathological vs non-pathological SNP

We have performed statistical analyses in order to determine if the parameters computed for pathological and non-pathological mutations reveal that the two populations can be separated. Table 2 displays average values and their standard deviations. Figure 5 displays the same information in the form of a box-plot. It can be seen in the figure that both populations overlap (due to large standard deviation). All variables display groups with a normal distribution (D'Agostino and Pearson's normal test, see Table 2). Therefore, we could perform t-tests for each parameter that revealed significant differences in all cases (see Table 3).

	Mean (punctuation)	Desvest (punctuation)	Normal test (p-value)
fi	0.246	0.219	<10e-11
fip	0.388	0.291	<10e-59
infCont	0.737	0.537	<10e-06
infContp	0.986	0.657	<10e-24
matrixScore	0.209	2.614	<10e-07
matrixScorep	-2.131	2.998	<10e-84

Table 2. *Mean, standard deviation and p-value of the normal test for initial frequency (fi and fip), content of information (infCont and infContp) and substitution matrix score (matrixScore and matrix scorep) for pathological and non pathological mutations (p indicating pathological mutations).*

	T-test (p-value)
fi	1.0577e-07
infCont	1.0713e-05
matrixScore	3.5225e-19

Table 3. *p-value for the T-test comparing initial frequency (fi), content of information (infCont), substitution matrix score (matrixScore) for pathological and non-pathological mutations. Significative differences arise between the two groups.*

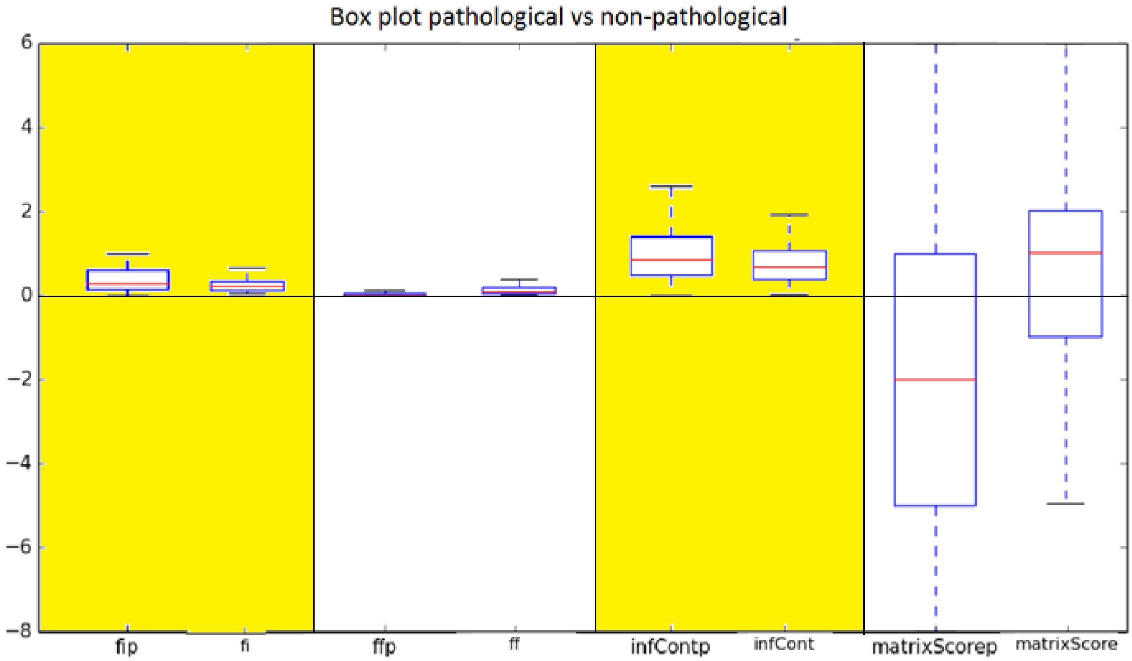


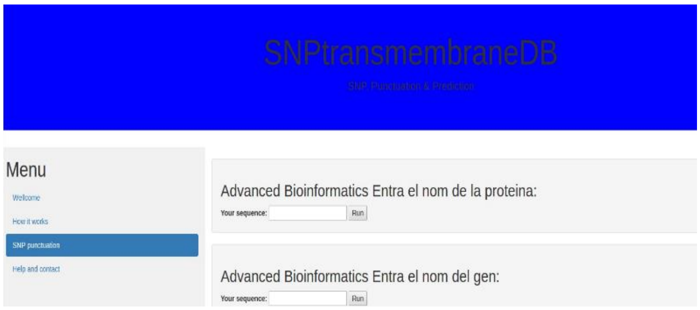
Figure 5. *Box plots for initial frequency (fi), final frequency (fp), content of information (infcont) and matrix substitution score (matrixScore) for pathological and non-pathological mutations.*

If the final frequency is greater than the initial frequency then usually is a non-pathological SNP, whatever the score or the information content values. So, if the frequency of the amino acid in the column of the alignment is high it means that it is a conserved residue and its change can affect more in a pathological way than if it is not conserved.

Matrix score values presents big differences between the two groups, the mean of pathological matrix score (Figure 5) is negative, indicating that the amino acid has a biggest effect in the SNP pathological effect.

EXAMPLE OF USAGE

Figure 6 displays the input page of the web application. The page displays two HTML-forms where the user can provide either the protein name (Uniprot accession or Uniprot ID) or the gene name.



Advanced Bioinformatics Entra el nom de la proteina:

Your sequence: Run

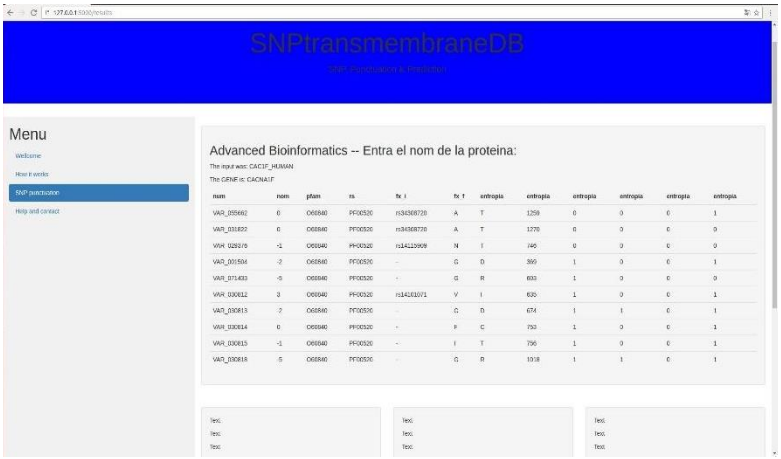
Advanced Bioinformatics Entra el nom del gen:

Your sequence: Run

Figure 6. *Input of SNPtmDB. The user can introduce the name of the protein or the name of a gene.*

A click on "Run" requests the preparation of a table with all the SNPs for such protein. The output consists on a list of SNPs(both VAR-code and rs-code, if available), the Pfam family, the score associated to the amino acid change, the reference amino acid (from the major allele) and final (from the SNP) frequencies in the alignment, the amount of information of the SNP position (see Figure 7). Each SNP is also classified as pathological (tagged as "1") or non-pathological (tagged as "0"). Pathological mutations are linked to OMIM database[10].

SNPtmDB web-page has an easy-to-use interface and returns results in less than one second.



Advanced Bioinformatics -- Entra el nom de la proteina:

The input was: CAC1F_HUMAN

The GENE is: CACNA1F

num	nom	pfam	rs	fx_i	fx_f	entropia	entropia	entropia	entropia	entropia	entropia
VAR_055662	0	O60840	PF00520	rs34308720	A	T	1259	0	0	0	1
VAR_031822	0	O60840	PF00520	rs34308720	A	T	1270	0	0	0	0
VAR_029376	-1	O60840	PF00520	rs14115909	N	T	746	0	0	0	0
VAR_001504	-2	O60840	PF00520	-	G	D	369	1	0	0	1
VAR_071433	-5	O60840	PF00520	-	G	R	603	1	0	0	0
VAR_030812	3	O60840	PF00520	rs14101071	V	I	635	1	0	0	1
VAR_030813	-2	O60840	PF00520	-	G	D	674	1	1	0	1
VAR_030814	0	O60840	PF00520	-	F	C	753	1	0	0	1
VAR_030815	-1	O60840	PF00520	-	I	T	756	1	0	0	1
VAR_030818	-5	O60840	PF00520	-	G	R	1018	1	1	0	1

Text:
Text:
Text:

Text:
Text:
Text:

Text:
Text:
Text:

Figure 7. *The Output of SNPtmDB consists on the initial and final frequency, the entropy and the substitution score for all SNP on a gene or protein.*

CONCLUSIONS

We have developed a web server database (SNPtmDB) that computes the initial and final frequencies, content of information and substitution score for all SNP located in the transmembrane regions of membrane proteins.

Comparison of these parameters between pathological and non-pathological SNP shows that there are statistical differences between the two groups.

These results suggest that these parameters can be used to predict if a mutation can be pathological or not and points SNPtmDB as a predictor of pathological mutations in membrane proteins.

BIBLIOGRAPHY

1. Brosius J. The Fragmented Gene. Ann N Y Acad *Sci. 2009;1178(1):186-193. doi:10.1111/j.1749-6632.2009.05004.x.*

2. Ridley M. Genome : *The Autobiography of a Species in 23 Chapters.; 2013. doi:10.1176/appi.ps.51.11.1457.*

3. Sharma S, Javadekar SM, Pandey M, Srivastava M, Kumari R, Raghavan SC. *Homology and enzymatic requirements of microhomology-dependent alternative end joining. Cell Death Dis. 2015;6(3):e1697. doi:10.1038/cddis.2015.58.*

4. Chen J, Miller BF, Furano A V. *Repair of naturally occurring mismatches can induce mutations in flanking DNA. 2014;3:e02001. doi:10.7554/eLife.02001.*

5. Griffiths AJF, Miller JH, Suzuki DT et al. *An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman; 2000. Somatic versus germinal mutation.* Available from: *https://www.ncbi.nlm.nih.gov/books/NBK21894/*

6. *"single nucleotide polymorphism / SNP | Learn Science at Scitable".*
Available from: *www.nature.com.*

7. Ingram, V. M. (1956). *"A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin". Nature. 178 (4537): 792–794. PMID 13369537.*

8. Hamosh A, King TM, Rosenstein BJ, et al. *Cystic fibrosis patients bearing both the common missense mutation, Gly->Asp at codon 551 and the ?F508 mutation are clinically indistinguishable from ?F508 homozygotes, except for decreased risk of meconium ileus. 1992;51(2):245-250.*
Available from: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1682672/pdf/ajhg00066-0020.pdf.*

9. Kui Zhang, Zhaohui S. Qin, *Haplotype Block Partitioning and Tag SNP Selection Using Genotype Data and Their Applications to Association Studies. Genome Res. 2004. 14: 908-916*

10. 1966-2016 Johns Hopkins University. Available from: *http://www.omim.org/ Ommim*

11. SIFT. *"SIFT Human DB updated to support GRCh37 Ensembl release 63"* Available from: *http://sift.bii.a-star.edu.sg/*

12. SAP2 "ROSTLAB all rights reserved."
Available from: *https://rostlab.org/services/snap/*

13. SuSPect "Yates CM, Filippis I, Kelley LA & Sternberg MJE (2014) SuSPect: *Enhanced prediction of single amino acid variant (SAV) phenotype using network features. Journal of Molecular Biology."*
In press. http://dx.doi.org/10.1016/j.jmb.2014.04.026
Available from: *http://www.sbg.bio.ic.ac.uk/suspect/index.html*

14. PolyPhen-2 *"predicts possible impact of an amino acid substitution"* Available from: *http://genetics.bwh.harvard.edu/pph2/*

15. predictSNP *"Consensus classifiers for prediction of disease-related mutations"* Available from: *http://loschmidt.chemi.muni.cz/predictsnp/*

16. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. *The Ensembl Variant Effect Predictor.Genome Biology Jun 6;17(1):122. (2016)* Available from: *http://www.ensembl.org/info/docs/tools/vep/index.html*

17. Agerberg L, Jonasson K, von Heijne G, Uhlen M, Berglund L: *Prediction of the human membrane proteome.*

18. Overington JP, Al-Lazikani B, Hopkins AL.: *How many drug targets are there? Nat Rev Drug Discov. 2006;5:993*

19. Perea M, Lugtenburg I, Mayol E, et al. TMalphaDB and TMbetaDB: *web servers to study the structural role of sequence motifs in $α$-helix and $β$-barrel domains of membrane proteins. BMC Bioinformatics. 2015;16(1):1-6. doi:10.1186/s12859-015-0699-5.*

20. Pfam *"The Pfam protein families database: towards a more sustainable future: R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, S.C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G.A. Salazar, J. Tate, A. Bateman"* Available from: http://pfam.xfam.org/

21. Sander and Schneider, 1991; Atchley et al.1999; Mirny and Shakhnovich, 1999; Lowry and Atchley, 2000 *"AL2CO: calculation of positional conservation in a protein sequence" - bioinformatics.oxfordjournals*

22. Pauline C. Ng, Jorja G. Henikoff, and Steven Henikoff.
PHAT: a transmembrane-specific substitution matrix. Bioinformatics (2000) 16 (9): 760-766 doi:10.1093/bioinformatics/16.9.760

23. Scipy "SciPy 0.18.0 released 2016-07-25" Available from: http://www.scipy.org/

24. UniProt "2002 – 2016 UniProt Consortium | License & Disclaimer" Available from: http://www.uniprot.org/

Acknowledgements

In this final project thesis I faced my first personal challenge.

I really want to thank my tutors, Arnau Cordomi and Mireia Olivella for helping and encouraging me to start this difficult but rewarding project.
I also want to thank them for solving all my problems and council me in the critical moments, offering me their experience and letting me to contribute in their studies.
They are going to be an exemplary professional role model to continue with my future career.

I also want to thank all the UAB research group that always made me feel part of them and help me in what was possible.

I want to mention my Uvic teachers in subjects related with boinformatics; Cristina Borralleras, Malu Calle and David Torrents that work hard to discover me the beauty in every informatics code line and enhance it in an extraordinary way.

Finally, I want to name my parents who always support me in all my live-projects.

Thanks to this project I really started to appreciate the huge possibilities that biology give and that make me love the route of my live.