Universitat de Vic
Escola Politècnica Superior

FINAL MASTER PROJECT

COMPUTATIONAL TOOLBOX TOWARDS EVOLUTIONARY DOMAIN MAPPING OF MEMBRANE PROTEINS

Alba Crespi i Boixader

Master in Omics Data Analysis

Directors: Mireia Olivella / Alex Peràlvarez

1/20/2014

Abstract

Membrane proteins account for about 20% to 30% of all proteins encoded in a typical genome. They play central roles in multiple cellular processes mediating the interaction of the cell with its surrounding. Over 60% of all drug targets contain a membrane domain. The experimental difficulties of obtaining a crystal structural severely limits our ability or understanding of membrane protein function. Computational evolutionary studies of proteins are crucial for the prediction of 3D structures. In this project, we construct a tool able to quantify the evolutionary positive selective pressure on each residue of membrane proteins through maximum likelihood phylogeny reconstruction. The conservation plot combined with a structural homology model is also a potent tool to predict those residues that have essentials roles in the structure and function of a membrane protein and can be very useful in the design of validation experiments.

Table of contents

List of tables and figures

1 Introduction

1.1 Membrane Proteins

Membrane proteins account for about 20% to 30% of all proteins encoded in a typical genome. They play central roles in multiple cellular processes mediating the interaction of the cell with its surrounding, such as the transport of nutrients and metabolites and in signalling of regulatory networks (Liang et al. 2012). Over 60% of all drug targets contain a membrane domain. One of the largest families of membrane proteins is G protein-coupled receptors (GPCRs), which are enriched in druggable target domains, and around of half of actual drugs are designed against them (Hofmann et al. 2009)(Russ & Lampel 2005).

The environment of membrane proteins is predominantly lipophilic, lacks hydrogen-bonding potential, and provides little screening of electrostatic interactions. At a primary sequence level, compared to water soluble proteins, there are significant differences in amino acid composition and the probabilities of amino acid substitutions during evolution, generally favouring residues with hydrophobic side chains, especially at the protein-lipid interface.

A major obstacle in studying membrane proteins is the difficulty in experimental determination of their three dimensional structures(Bill et al. 2011): many membrane proteins are difficult to crystallize, or are too large to be studied with NMR (Liang et al. 2012)(Pierri et al. 2010) and only represent <2% of crystal structures (Kozma et al. 2013) deposited in the Protein Data Bank. The 3D structure of membrane proteins is essential for the characterization of its molecular mechanisms and is crucial in the development of pharmacological agent targets.

1.2 Evolutionary studies

The absence of structural information severely limits our ability or understanding of membrane protein function. Computational evolutionary studies of proteins are crucial for the prediction of 3D structures (Marks et al. 2011) in order to understand their function (Pierri et al. 2010). Protein patterns and motifs are result of the selective pressure of evolution. Some residues play key roles either in structure or function (Liang et al. 2012) at specific positions. As an example, Pro^{50} of bacteriorhodopsin is essential for lipid-protein and protein-protein interaction and consequently maintain the proper folding; and Pro^{91} is basic for the functionality of the active site (Perálvarez-Marín et al. 2008). Typically, the most accurate models of protein structures are achieved through homology modelling, where a known structure is used as a template for the construction of a model of a related protein (Forrest et al. 2006) revealing the parts which are changing rapidly and those residues shaped by natural selection(Holder & Lewis 2003).

Primary sequence evolutionary conservation as a tool to identify structural conservation is limited to its use in the case of membrane proteins. The same transmembrane fold can show an overall low conservation. However, within the same family of proteins, certain residues of conserved function or conserved structural stability cannot escape evolution. These residues are suffering what is called evolutionary pressure. This evolutionary pressure is so high, that even in the case that a mutation occurs in a key region of a protein; this mutation will be compensated by other mutations to ameliorate the effect of the first one.

Evolutionary studies of the sequence of membrane protein permits to estimate those residues with a high selective pressure during evolution. The identification of these positions and the experimental validation gives insight to the structure and function of the membrane protein (Marks et al. 2011)(Grishin 2012)(Nugent & Jones 2012).

1.3 Aim

Although certain statistical studies have dealt with the effect of conservation/mutation and co-evolution in specific membrane proteins, such as ABC transporters (Gulyas-Kovacs 2012), a useful tool should be able to deal with any kind of membrane protein. The aim of this study is to construct a tool able to quantify the evolutionary pressure on each residue or on each transmembrane segment from the sequence of any non determined membrane protein. The conservation plot combined with a structural homolgy model can be a potent tool to predict those residues that have an essential role in the structure and function of a membrane protein and can be very useful in the design of validation experiments.

A user-friendly web server interface is also under development. This, will also allow the user to set the parameters which best fits to the analysis.

Box 1-1. Keywords and Definitions

- *Positive selection:* Darwinian selection fixing advantageous mutations with positive selective coefficients. The term is used interchangeably with molecular adaptation and adaptive molecular evolution
- *Algorithm:* A fixed procedure embodied in a computer program.
- *Alignment:* The process or result of matching up the nucleotide or amino acid residues of two or more biological sequences to achieve maximal levels of identity and, in the case of amino acid sequences, conservation, for the purpose of assessing the degree of similarity and the possibility of homology.
- *Identity:* The extent to which two (nucleotide or amino acid) sequences have the same residues at the same positions in an alignment, often expressed as a percentage.
- *Similarity:* The extent to which nucleotide or sequences are related. Similarity between two sequences can be expressed as percent sequence identity and/or percent positive substitutions.
- *Domain:* A discrete portion of a protein assumed to fold independently of the rest of the protein and possessing its own function.
- *E-value:* The Expectation value or Expect value represents the number of different alignments with scores equivalent to or better than S that is expected to occur in a database search by chance. The lower the E value, the more significant the score and the alignment.
- *Homology:* Similarity attributed to descent from a common ancestor. Homologous biological components (genes, proteins, structures) are called homologs.
- *Orthologs:* Homologous biological components (genes, proteins, structures) in different species that arose from a single component present in the common ancestor of the species; orthologs may or may not have a similar function.
- *Paralogs:* Homologous biological components within a single species that arose by gene duplication.
- *Motif:* A short conserved region in a protein sequence. Motifs are frequently highly conserved parts of domains.

http://www.ncbi.nlm.nih.gov/books/NBK62051/

Figure 1-1. Homologs, orthologs & paralogs

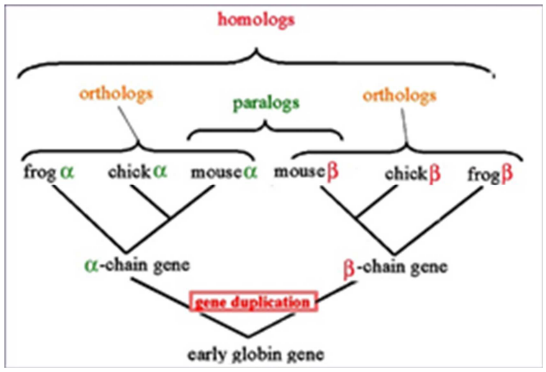
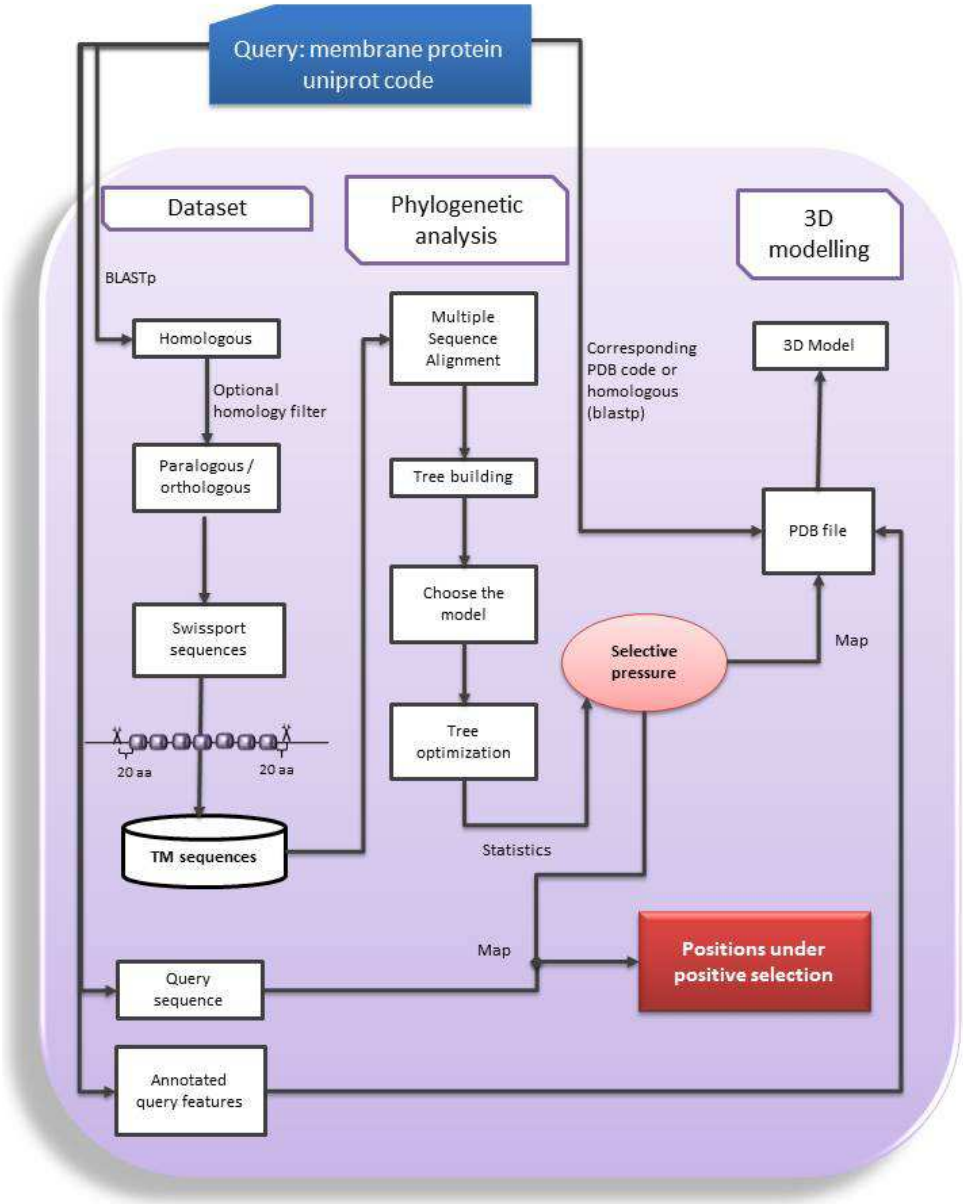


Figure 1-2. Methodology workflow



2 Methods

The methodology presented in this work follows the general procedure of evolutionary studies. Starting from the membrane protein template or query, object of the analysis, find its homologous. Next, a multiple sequence alignment is performed followed by the phylogenetic tree construction, which ends up with an assigned log likelihood value to each position. Finally, a model is presented with a colour gradient of its associated value of the corresponding or homolog crystal structural.

A pipeline has been created through python, biopython and R scripts in Linux environment. Python programming is used to the basic local alignment, to obtain the protein sequences, to perform the multiple sequence alignment, mapping scores and plotting, and the chimera-python interface for the 3D homology model. R programming is used in phylogenetic trees generation.

2.1 Basic Local Sequence Alignment of the membrane protein in order to obtain homologous

Sequence alignments searches against databases are used to find homologous proteins. Basic local alignment search (BLAST) is an algorithm which has been optimised in order to find the optimal local alignment to a query in a speed search against databases (Altschul et al. 1990) (Altschul et al. 1997).

A protein BLAST search (blastp against protein database using a protein query) is carried out using as a template a protein Uniprot code of a membrane protein in order to find its homologs. The search is done against Swissprot database (Bairoch & Apweiler 2000), and the user could specify the desired values for the following parameters: sequence identity, E-value cut-offs, query coverage and hit list size.

From the list of uniprot accession codes obtained, the corresponding full sequences are extracted through Swissprot database with their corresponding transmembrane and mutagenesis regions annotated.

Considering the construction of reliable phylogenetic tree, sequences should be neither so similar nor so divergent (Castresana 2000). Optionally, a filter for paralogous sequences can be performed. In order to obtain the transmembrane bundle, and to avoid the extracellular or intracellular regions, only the region ranged from the first to the last transmembrane domain plus 20 amino acids for each extreme are extracted of all proteins to further analyse.

2.2 Multiple Sequence Alignment of the homologous

Multiple sequence alignment (MSA) is able to detect the evolutionary relationship between membrane proteins and key functional residues (Liang et al. 2012). Many successful approaches have been designed to overcome with MSA scoring system and consequently with its accuracy. MAFFT is a global MSA tool based on the fast Fourier transform (FFT), which allows rapid detection of homologous (Katoh et al. 2002), and has been considered one of the best programs (Edgar & Batzoglou 2006).

After the extraction of the transmembrane bundle of all homologous sequences, a MAFFT multiple sequence alignment (MSA) is performed with its default parameters (Katoh & Standley 2013).

2.3 Phylogenetic tree generation

Phylogenetic analysis reveals the selective pressure among sites for a given MSA. Between the phylogenetic methods described (Massingham & Goldman 2005) (Wong et al. 2004), such as SG method for nucleotides (Suzuki & Gojobori 1999) or its modification(Suzuki 2004), character based ones are the most widely used. These include the maximum parsimony (Wu et al. 2006) and maximum-likelihood (ML)(Yang & Bielawski 2000a)(Zhang et al. 2005)(Nielsen & Yang 1998) (Tamura et al. 2011) which evaluates and maximizes the probability that the chosen evolutionary model has generated the observed data (Brinkman & Leipe 2001) (Huelsenbeck & Bollback 2001). This allows the assessment of the reliability of each amino acid position in an alignment on the basis of all other positions (Yang et al. 2000)(Yang 1998)(Holder & Lewis 2003)(Williams & Lovell 2009)(Yang & Bielawski 2000b).

Phylogenetic inference from amino acid sequence data uses mainly empirical models of amino acid replacement and is therefore dependent on those models (Abascal et al. 2005). These models encompass estimate of the instantaneous substitution rates from any amino acid to another one within time (Le & Gascuel 2008). They are used to compute substitution probabilities along phylogeny branches and thus the likelihood of the data. Several have been specifically designed for different families and subfamilies of proteins. Some of them are specifically designed for soluble proteins and often are considered not appropriate for membrane proteins such as BLOSUM or PAM (Liang et al. 2012). These models assume very similar amino acid replacement across all positions. Nevertheless, conservation of protein function and structure imposes constraints on which positions can change. This evolutionary information can be inferred by considering a fraction of amino acids to be invariable ('+I'), or assigning each site a probability to belong to given gamma rate categories ('+G') (Abascal et al. 2005).

ML method provides a better estimation of the model of replacement. Moreover, it allows the use of different models of evolution depending on the examined dataset (Keane et al. 2006). However, it has a high computational cost (Whelan & Goldman 1995)(Holder & Lewis 2003) and it doesn't work for big datasets. This method has been implemented in R through "Phangorn" package (Schliep 2011). The replacement models used in this tool are a subset of the implemented for proteins: WAG(Whelan & Goldman 2001), LG(Le & Gascuel 2008), cpREV(Adachi et al. 2000), mtArt(Abascal et al. 2007), MtZoa(Rota-Stabelli et al. 2009), mtREV24(Adachi & Hasegawa 1996).

A phylogenetic tree is generated by ML method. Firstly, an initial tree is calculated through "bionj" algorithm (Gascuel 1997), which is a method for reconstructing phylogenetic trees from a matrix of pairwise evolutionary distances, and WAG model. This data is tested to the models inferred by both invariant sites and gamma rate categories. The model which best fits under BIC (Bayesian Information Criterion) is selected to perform the study.

The tree is optimized by the parameters NNI (branch swapping method), proportion of variable size, gamma rate, and edge lengths. Log likelihood is obtained for each MSA site and for the global tree.

2.4 Positive Selective Pressure

For determining whether a concrete position of the query membrane protein is under positive selection, a FDR correction of the likelihood for multiple testing is performed as each site is tested for positive selection independently. The p-values are ranked from the lowest to the highest. A site is considered to be significant for positive selection if its adjusted p-value is smaller than the designed alpha divided by its rank (Wong et al. 2004). The pressure measure is obtained by alpha divided by the rank minus the adjusted p-value.

2.5 3D Model of the membrane protein

In this step is fundamental to dispose of the PDB file (Berman et al. 2000). If the query protein has been crystallized, the structure with maximum coverage and maximum resolution obtained by X-ray method is selected. Otherwise, the homology model will be constructed by the closest homologous protein crystal structure found by running a BLASTp against the PDB database.

The analysed sequence with its site log likelihood assigned is mapped to the original query sequence and to the corresponding or homologous protein crystal structure. In the case that two sequences are different, an alignment of the query versus the PDB sequence is previously performed (pairwise global alignment, using BLOSUM62 matrix, gap penalisation of -10 for opening an -0.5 for extending).

2.6 Conservation plot along the 3D model of the membrane protein

Chimera (Pettersen et al. 2004) python interface is used in order to obtain the conservational plot of the 3D Model of the studied protein. A gradient colour of the selective pressure measure is applied to the residues of the corresponding structure ranging from cyan to magenta, maximum to minimum correspondingly. Described mutagenesis residues are plotted in stick shape. Grey areas are the non-analysed residues.

2.7 Example of usage through bovine rhodopsin

Bovine rhodopsin (P02699) is one the most well characterized membrane proteins. It is one of the most abundant and stable membrane proteins and it was soon crystallised (Bill et al. 2011). Experimental information regarding of the role of some residues is available (Palczewski et al. 2000).

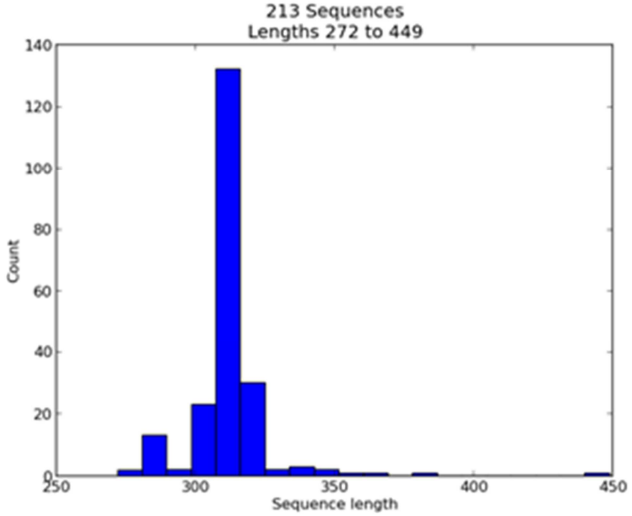
For this example we set the blastp parameters as follows: 25% sequence identity, which is in the middle of the called twilight zone(Rost 1999), ensuring to find enough similar sequence to perform an acceptable MSA and taking into account that TM structure of most membrane proteins have a strong conservation at low-sequence identity (Olivella et al. 2013); E-value cut-offs of $1x10^{-4}$, query coverage of 70% and hit list of size 10,000. Afterwards we filtered the paralogous sequences. The alpha value is set up at 0.05.

3 Results and Discussion

We identified 213 orthologous sequences. After keeping the transmembrane bundle for all of them, their lengths rage from 272 to 449. (See figure 3-3). The figure 3-4 shows a partial view of the MAFFT Multiple Sequence Alignment.

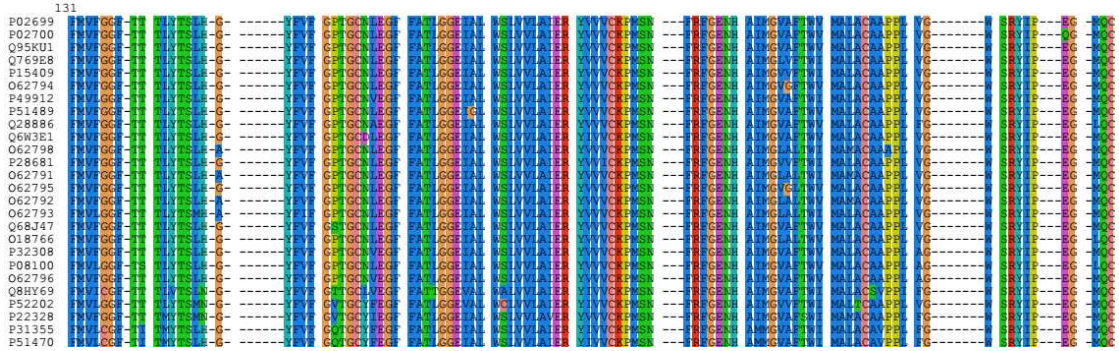
Thanks to ML methodology it is possible to evaluate the model which best fits to the data. Table 3-2 shows the evaluation of the initial constructed phylogenetic tree to determine which model best fits considering invariable sites ('+I'), and assigning each site a probability to belong to a given rate ('+G'). In this case, the model which minimizes Bayesin Information Criterion (BIC) is LG+G+I.

Figure 3-3. Frequency of Lengths analysed sequences.



This graph represents the frequencies of sequences length of the selected area for the analysis, which ranges 20 amino acids up and down from the first TM domain to the last one. There is a total of 213 sequences after filtering the paralogues which passes the blastp search criteria (25% sequence identity; E-value cutoffs of 1x10-4, cover query of 70% and hitsize list of 10,000).

Figure 3-4. Multiple Sequence Alignment



Partial view of the top MSA from position 131 by using SeaView (Gouy et al. 2010).

Table 3-1. Model Test

Model	df	logLik	AIC	BIC
WAG	631	-49146.50	99554.99	102370.18
WAG+I	632	-49039.78	99343.56	102163.21
WAG+G	632	-48081.00	97426.00	100245.64
WAG+G+I	633	-48051.74	97369.47	100193.58
LG	631	-48848.30	98958.60	101773.79
LG+I	632	-48767.16	98798.32	101617.97
LG+G	632	-47666.16	96596.32	99415.97
LG+G+I	633	-47648.51	96563.03	99387.14
cpREV	602	-49255.83	99715.66	102401.46
cpREV+I	603	-49159.01	99524.02	102214.29
cpREV+G	603	-48108.11	97422.23	100112.49
cpREV+G+I	604	-48087.02	97382.05	100076.77
mtArt	533	-51180.56	103427.12	105805.08
mtArt+I	534	-51121.19	103310.39	105692.81
mtArt+G	534	-49202.02	99472.03	101854.46
mtArt+G+I	535	-49193.75	99457.50	101844.39
MtZoa	603	-50038.74	101283.47	103973.74
MtZoa+I	604	-49980.51	101169.01	103863.74
MtZoa+G	604	-48268.15	97744.29	100439.02
MtZoa+G+I	605	-48258.16	97726.33	100425.52
mtREV24	600	-50873.63	102947.25	105624.13
mtREV24+I	601	-50776.43	102754.86	105436.21
mtREV24+G	601	-49387.93	99977.86	102659.21
mtREV24+G+I	602	-49365.41	99934.83	102620.63

Results from the test performed to establish the model that fits the best under BIC. *Df: degrees of freedom. LogLik: overall likelihood ratio. AIC: Akaike Information Criterion. BIC: Bayesian Information Criterion.*

Once the tree is optimised, a log likelihood value for each position is obtained (See appendix). After calculating the selective pressure, a total of 275 residues are under selective positive pressure, leaving 36 out. This means that the first ones are susceptible to mutations in more or less degree (see appendix for detailed information of the measure of the positive selection).

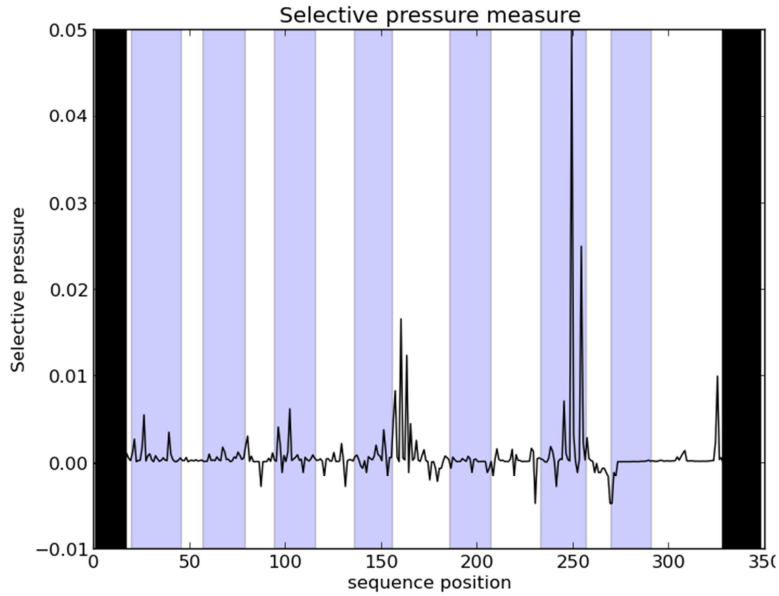
Table 3-3 shows the positions and its corresponding amino acid under positive selective pressure.

Table 3-3. Positive Selected residues

45	**G**	**81**	T	**117**	A	**153**	A	**189**	G	**227**	E	**267**	T	**305**	A
46	P	**82**	T	**118**	W	**154**	F	**190**	G	**228**	Y	**268**	L	**306**	F
47	M	**83**	Y	**119**	Y	**155**	D	**191**	V	**229**	L	**269**	L	**307**	Y
48	I	**84**	I	**120**	E	**156**	M	**192**	E	**230**	T	**270**	R	**308**	F
49	F	**85**	F	**121**	R	**157**	Y	**193**	F	**231**	T	**271**	G	**309**	I
50	Q	**86**	F	**122**	L	**158**	F	**194**	R	**232**	P	**272**	F	**310**	S
51	G	**87**	C	**123**	H	**159**	L	**195**	G	**233**	P	**273**	I	**311**	T
52	H	**88**	V	**124**	L	**160**	S	**196**	M	**234**	F	**274**	M	**312**	V
53	V	**89**	V	**125**	C	**161**	H	**197**	G	**235**	L	**275**	L	**313**	F
54	W	**90**	V	**126**	G	**162**	A	**198**	G	**236**	V	**276**	M	**314**	V
55	I	**91**	Q	**127**	G	**163**	E	**199**	F	**237**	L	**279**	K	**315**	L
56	T	**92**	L	**128**	S	**164**	Y	**200**	L	**238**	N	**280**	L	**316**	I
57	Q	**93**	Y	**129**	V	**165**	N	**201**	C	**239**	R	**281**	L	**317**	G
58	A	**94**	I	**130**	G	**166**	V	**202**	V	**240**	Y	**282**	W	**318**	A
59	L	**95**	Y	**131**	F	**167**	S	**203**	H	**241**	V	**283**	M	**319**	T
60	F	**96**	F	**132**	T	**168**	A	**204**	L	**242**	V	**284**	I	**320**	I
61	M	**97**	S	**133**	L	**169**	P	**205**	T	**243**	Q	**285**	E	**321**	V
62	P	**98**	V	**134**	A	**170**	F	**206**	I	**244**	F	**286**	I	**322**	S
63	G	**99**	L	**135**	N	**171**	V	**207**	V	**245**	A	**287**	Y	**323**	F
64	F	**100**	T	**136**	C	**172**	A	**208**	M	**246**	P	**288**	E	**324**	P
65	A	**101**	G	**137**	E	**173**	F	**209**	F	**247**	M	**289**	I	**325**	G
66	V	**102**	W	**138**	Q	**174**	E	**210**	Y	**248**	C	**290**	K	**326**	L
67	A	**103**	T	**139**	F	**175**	L	**211**	A	**249**	Q	**291**	A	**327**	E
68	A	**104**	W	**140**	K	**176**	R	**212**	P	**250**	P	**292**	A		
69	I	**105**	F	**141**	K	**177**	G	**213**	M	**251**	Y	**293**	R		
70	E	**106**	P	**142**	I	**178**	Q	**214**	V	**256**	H	**294**	E		
71	C	**107**	T	**143**	Q	**179**	L	**215**	T	**257**	L	**295**	Q		
72	I	**108**	F	**144**	C	**180**	N	**216**	Y	**258**	T	**296**	Q		
73	N	**109**	A	**145**	H	**181**	V	**219**	S	**259**	V	**297**	I		
74	P	**110**	S	**146**	S	**182**	N	**220**	A	**260**	V	**298**	L		
75	G	**111**	I	**147**	L	**183**	E	**221**	P	**261**	V	**299**	P		
76	F	**112**	V	**148**	M	**184**	P	**222**	S	**262**	A	**300**	N		
77	N	**113**	K	**149**	P	**185**	F	**223**	A	**263**	T	**301**	T		
78	D	**114**	D	**150**	M	**186**	G	**224**	A	**264**	A	**302**	A		
79	Y	**115**	Y	**151**	I	**187**	L	**225**	K	**265**	T	**303**	E		
80	G	**116**	E	**152**	I	**188**	V	**226**	L	**266**	A	**304**	I		

Next figure 3-6 shows the selective pressure for each position of the query.

Figure 3-6. Position Selective pressure

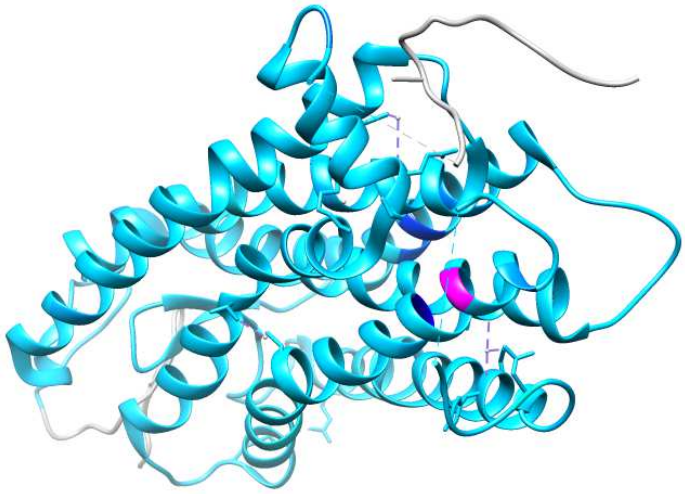


The plot shows the calculated selecteve pressure measure for each position from the P02699 peptide. Black areas are the extracellular and cytoplasmatic domains (correspondingly) which has not been taken into account for obtaning the score. Gray areas corresponds to the TM domains. Positives values are under positive selective pressure.

In general traits selective pressure is very close to 0. Some values are clearly positives whereas just a few are negative. Regarding the positive coincide with positive results, those values needs to be studied carefully for the user.

Once the selective pressure is calculated, a 3D model is set. In this case the template is the pdb code 1F88, a crystal structure obtained by X-ray. In this example, the protein entirely coincides with the original query. In other cases, an alignment is performed between the 2 sequences and then the selective pressure measure is mapped. As an output a chimera session is generated.

Figure 3-7. 3D model



1F88 crystal structure. A gradient colour of the selective pressure measure is applied to the residues of the corresponding structure ranging from cyan to magenta, maximum to minimum correspondingly. Described mutagenesis residues are plotted in stick shape. Grey areas are the non-analysed residues.

4 Conclusions

The results obtained through this tool need to be validated by the experimental scientist at the laboratory and through bibliography and experiments.

We generate a tool able to identify form a given membrane protein, which positions are under positive selective pressure. It is able to give hints to the experimental researchers on which positions of a membrane protein can or/and cannot mutate in order to validate the results.

5 References

Abascal, F., Posada, D. & Zardoya, R., 2007. MtArt: a new model of amino acid replacement for Arthropoda. *Molecular biology and evolution*, 24(1), pp.1–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17043087 [Accessed January 14, 2014].

Abascal, F., Zardoya, R. & Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)*, 21(9), pp.2104–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15647292 [Accessed January 12, 2014].

Adachi, J. et al., 2000. Plastid Genome Phylogeny and a Model of Amino Acid Substitution for Proteins Encoded by Chloroplast DNA. *Journal of molecular biology*, 50, pp.348–358.

Adachi, J. & Hasegawa, M., 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of molecular evolution*, 42(4), pp.459–468.

Altschul, S.F. et al., 1990. Basic Local Alignment Search Tool. *Journal of molecular biology*, (215), pp.403–410.

Altschul, S.F. et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17), pp.3389–402. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rendertype=abstract.

Bairoch, a & Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research*, 28(1), pp.45–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102476&tool=pmcentrez&rendertype=abstract.

Berman, H.M. et al., 2000. The Protein Data Bank. *Nucleic acids research*, 28(1), pp.235–42. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102472&tool=pmcentrez&rendertype=abstract.

Bill, R.M. et al., 2011. Overcoming barriers to membrane protein structure determination. *Nature biotechnology*, 29(4), pp.335–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21478852 [Accessed December 17, 2013].

Brinkman, F.S. & Leipe, D.D., 2001. *Phylogenetic analysis.*, Available at: http://www.ncbi.nlm.nih.gov/pubmed/11449731.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution*, 17(4), pp.540–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10742046.

Edgar, R.C. & Batzoglou, S., 2006. Multiple sequence alignment. *Current Opinion in Structural Biology*, 16, pp.368–373. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16679011.

Forrest, L.R., Tang, C.L. & Honig, B., 2006. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophysical journal*, 91(2), pp.508–17. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1483079&tool=pmcentrez&rendertype=abstract [Accessed October 7, 2013].

Gascuel, O., 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular biology and evolution*, 14(7), pp.685–95. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9254330.

Gouy, M., Guindon, S. & Gascuel, O., 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution*, 27(2), pp.221–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19854763 [Accessed January 10, 2014].

Grishin, N. V, 2012. Membrane protein structure predictions for exploration. *Cell*, 149(7), pp.1424–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3688449&tool=pmcentrez&rendertype=abstract [Accessed January 12, 2014].

Gulyas-Kovacs, A., 2012. Integrated Analysis of Residue Coevolution and Protein Structure in ABC Transporters. *PloS one*, 7(5).

Hofmann, K.P. et al., 2009. A G protein-coupled receptor at work: the rhodopsin model. *Trends in biochemical sciences*, 34(11), pp.540–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19836958 [Accessed January 13, 2014].

Holder, M. & Lewis, P.O., 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews. Genetics*, 4(4), pp.275–84. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12671658 [Accessed November 6, 2013].

Huelsenbeck, J.P. & Bollback, J.P., 2001. Empirical and hierarchical Bayesian estimation of ancestral states. *Systematic biology*, 50(3), pp.351–66. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12116580.

Katoh, K. et al., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14), pp.3059–66. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=135756&tool=pmcentrez&rendertype=abstract.

Katoh, K. & Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4), pp.772–80. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3603318&tool=pmcentrez&rendertype=abstract [Accessed January 9, 2014].

Keane, T.M. et al., 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC evolutionary biology*, 6, p.29. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1435933&tool=pmcentrez&rendertype=abstract [Accessed January 12, 2014].

Kozma, D., Simon, I. & Tusnády, G.E., 2013. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic acids research*, 41(Database issue), pp.D524–9. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531219&tool=pmcentrez&rendertype=abstract [Accessed November 27, 2013].

Le, S.Q. & Gascuel, O., 2008. An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7), pp.1307–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18367465 [Accessed January 12, 2014].

Liang, J. et al., 2012. Computational studies of membrane proteins: models and predictions for biological understanding. *Biochimica et biophysica acta*, 1818(4), pp.927–41. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3737997&tool=pmcentrez&rendertype=abstract [Accessed September 20, 2013].

Marks, D.S. et al., 2011. Protein 3D structure computed from evolutionary sequence variation. *PloS one*, 6(12), p.e28766. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3233603&tool=pmcentrez&rendertype=abstract [Accessed January 17, 2014].

Massingham, T. & Goldman, N., 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics*, 169(3), pp.1753–62. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1449526&tool=pmcentrez&rendertype=abstract [Accessed November 22, 2013].

Nielsen, R. & Yang, Z., 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3), pp.929–36. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1460041&tool=pmcentrez&rendertype=abstract.

Nugent, T. & Jones, D.T., 2012. Membrane protein structural bioinformatics. *Journal of structural biology*, 179(3), pp.327–37. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22075226 [Accessed November 7, 2013].

Olivella, M. et al., 2013. Relation between sequence and structure in membrane proteins. *Bioinformatics (Oxford, England)*, 29(13), pp.1589–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23677941 [Accessed November 25, 2013].

Palczewski, K. et al., 2000. Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science*, 289(5480), pp.739–745. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.289.5480.739 [Accessed January 9, 2014].

Perálvarez-Marín, A. et al., 2008. Influence of proline on the thermostability of the active site and membrane arrangement of transmembrane proteins. *Biophysical journal*, 95(9), pp.4384–95. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2567942&tool=pmcentrez&rendertype=abstract [Accessed January 9, 2014].

Pettersen, E.F. et al., 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13), pp.1605–12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15264254 [Accessed December 12, 2013].

Pierri, C.L., Parisi, G. & Porcelli, V., 2010. Computational approaches for protein function prediction: a combined strategy from multiple sequence alignment to molecular docking-based virtual screening. *Biochimica et biophysica acta*, 1804(9), pp.1695–712. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20433957 [Accessed October 2, 2013].

Rost, B., 1999. Twilight zone of protein sequence alignments. *Protein engineering*, 12(2), pp.85–94. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10195279.

Rota-Stabelli, O., Yang, Z. & Telford, M.J., 2009. MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies. *Molecular phylogenetics and evolution*, 52(1), pp.268–72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19475713 [Accessed January 19, 2014].

Russ, A.P. & Lampel, S., 2005. The druggable genome: an update. *Drug discovery today*, 10(23-24), pp.1607–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16376820.

Schliep, K.P., 2011. phangorn: phylogenetic analysis in R. *Bioinformatics (Oxford, England)*, 27(4), pp.592–3. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3035803&tool=pmcentrez&rendertype=abstract.

Suzuki, Y., 2004. New Methods for Detecting Positive Selection at Single Amino Acid Sites Yoshiyuki. *Journal of Molecular Evolution*, 59(1), pp.11–19.

Suzuki, Y. & Gojobori, T., 1999. A method for detecting positive selection at single amino acid sites. *Molecular biology and evolution*, 16(10), pp.1315–28. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10563013.

Tamura, K. et al., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, 28(10), pp.2731–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3203626&tool=pmcentrez&rendertype=abstract [Accessed September 16, 2013].

Whelan, S. & Goldman, N., 1995. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. , pp.691–699.

Whelan, S. & Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5), pp.691–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11319253.

Williams, S.G. & Lovell, S.C., 2009. The effect of sequence evolution on protein structural divergence. *Molecular biology and evolution*, 26(5), pp.1055–65. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19193735 [Accessed January 15, 2014].

Wong, W.S.W. et al., 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, 168(2), pp.1041–51. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1448811&tool=pmcentrez&rendertype=abstract [Accessed December 11, 2013].

Wu, C.H. et al., 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic acids research*, 34(Database issue), pp.D187–91. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347523&tool=pmcentrez&rendertype=abstract [Accessed November 15, 2013].

Yang, Z., 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular biology and evolution*, 15(5), pp.568–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9580986.

Yang, Z. & Bielawski, J., 2000a. Statistical methods for detecting molecular adaptation. *Trends in ecology & evolution*, 15(12), pp.496–503. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11114436.

Yang, Z. & Bielawski, J., 2000b. Statistical methods for detecting molecular adaptation. *Trends in ecology & evolution*, 15(12), pp.496–503. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11114436.

Yang, Z., Swanson, W.J. & Vacquier, V.D., 2000. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Molecular biology and evolution*, 17(10), pp.1446–55. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11018152.

Zhang, J., Nielsen, R. & Yang, Z., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution*, 22(12), pp.2472–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16107592 [Accessed December 12, 2013].

6 Appendix

Proteins which pass the blastp criteria (E-value cutoffs of 1x10-4, cover query of 70% and hitsize list of 10,000) and filtered by paraloguous. A total of 213 are obtained.

Ref name	Uniprot code	Organism
OPSD_BOVIN	P02699	Bos taurus
OPSD_SHEEP	P02700	Ovis aries
OPSD_FELCA	Q95KU1	Felis catus
OPSD_OTOCR	Q769E8	Otolemur crassicaudatus
OPSD_MOUSE	P15409	Mus musculus
OPSD_PHOVI	O62794	Phoca vitulina
OPSD_RABIT	P49912	Oryctolagus cuniculus
OPSD_RAT	P51489	Rattus norvegicus
OPSD_MACFA	Q28886	Macaca fascicularis
OPSD_CALPD	Q6W3E1	Caluromys philander
OPSD_TURTR	O62798	Tursiops truncatus
OPSD_CRIGR	P28681	Cricetulus griseus
OPSD_DELDE	O62791	Delphinus delphis
OPSD_PAGGO	O62795	Pagophilus groenlandicus
OPSD_GLOME	O62792	Globicephala melas
OPSD_MESBI	O62793	Mesoplodon bidens
OPSD_LOXAF	Q68J47	Loxodonta africana
OPSD_PIG	O18766	Sus scrofa
OPSD_CANFA	P32308	Canis familiaris
OPSD_HUMAN	P08100	Homo sapiens
OPSD_TRIMA	O62796	Trichechus manatus
OPSD_SMICR	Q8HY69	Sminthopsis crassicaudata
OPSD_ALLMI	P52202	Alligator mississippiensis
OPSD_CHICK	P22328	Gallus gallus
OPSD_RANPI	P31355	Rana pipiens
OPSD_LITCT	P51470	Lithobates catesbeiana
OPSD_RANTE	P56516	Rana temporaria
OPSD_AMBTI	Q90245	Ambystoma tigrinum
OPSD_BUFMA	P56515	Bufo marinus
OPSD2_ANGAN	Q90215	Anguilla anguilla
OPSD_BUFBU	P56514	Bufo bufo
OPSD_XENLA	P29403	Xenopus laevis
OPSD_DANRE	P35359	Danio rerio
OPSD_ANOCA	P41591	Anolis carolinensis
OPSD_SCYCA	O93459	Scyliorhinus canicula
OPSB_CONCO	O13227	Conger conger
OPSD_LITMO	Q9YH00	Lithognathus mormyrus
OPSD_DIPVU	Q9YH04	Diplodus vulgaris
OPSD_GALML	O93441	Galeus melastomus
OPSD_SPAAU	Q9YH02	Sparus aurata
OPSD_SARPI	Q9YGZ0	Sardina pilchardus
OPSD_DIPAN	Q9YH05	Diplodus annularis
OPSD_SARSL	Q9YH03	Sarpa salpa
OPSD_LEUER	P79863	Leucoraja erinacea
OPSD_MUGCE	Q9YGZ9	Mugil cephalus
OPSD_CYPCA	P51488	Cyprinus carpio
OPSD_LIZAU	Q9YGZ6	Liza aurata
OPSD_LIZSA	Q9YGZ7	Liza saliens
OPSD_TETNG	Q9DGG4	Tetraodon nigroviridis
OPSD_DICLA	Q9YGZ4	Dicentrarchus labrax
OPSD_CARAU	P32309	Carassius auratus
OPSD_ASTFA	P41590	Astyanax fasciatus
OPSD_CHELB	Q9YGZ8	Chelon labrosus
OPSD_LAMJA	P22671	Lampetra japonica
OPSD_ZEUFA	O42604	Zeus faber
OPSD_ORYLA	P87369	Oryzias latipes
OPSD_SOLSO	Q9YGZ5	Solea solea
OPSD_MULSU	Q9YH01	Mullus surmuletus
OPSD_POERE	P79848	Poecilia reticulata
OPSD_ATHBO	Q9YGZ1	Atherina boyeri
OPSD_SALPV	Q9YGZ3	Salaria pavo
OPSD_GOBNI	Q9YGZ2	Gobius niger
OPSD_ZOSOP	Q9YGY9	Zosterisessor ophiocephalus
OPSD_GAMAF	P79756	Gambusia affinis
OPSD_NEOSA	P79812	Neoniphon sammara
OPSD_SARDI	P79898	Sargocentron diadema
OPSD_MYRVI	P79807	Myripristis violacea
OPSD_NEOAR	P79808	Neoniphon argenteus
OPSD_MYRBE	P79798	Myripristis berndti
OPSD_PETMA	Q98980	Petromyzon marinus
OPSD_SARMI	P79901	Sargocentron microstoma
OPSD_SARXA	P79914	Sargocentron xantherythrum
OPSD_SARPU	P79902	Sargocentron punctatissimum
OPSD_SARTI	P79911	Sargocentron tiere
OPSD_POMMI	P35403	Pomatoschistus minutus
OPSD_SARSP	P79903	Sargocentron spiniferum
OPSD_NEOAU	P79809	Neoniphon aurolineatus
OPSB_GECGE	P35357	Gecko gecko
OPSD_ICTPU	O42268	Ictalurus punctatus
OPSD_COTIN	O42330	Cottocomephorus inermis.
OPSD_ABYKO	O42294	Abyssocottus korotneffi
OPSD_PROJE	O42451	Procottus jettelesi.
OPSD_BATMU	O42300	Batrachocottus multiradiatus.
OPSD_COTBO	O42307	Cottinella boulengeri.
OPSD_COMDY	O42327	Comephorus dybowskii.
OPSD_BATNI	O42301	Batrachocottus nikolskii
OPSD_PARKN	O42452	Paracottus kneri.
OPSD_LIMBE	O42427	Limnocottus bergianus.
OPSD_LIMPA	O42431	Limnocottus pallidus.
OPSD_LEOKE	Q90373	Leocottus kesslerii
OPSD_TAUBU	O42466	Taurulus bubalis
OPSD_COTGR	O42328	Cottocomephorus grewingki.
OPSP_COLLI	P51476	Columba livia
OPSUV_MELUD	O57605	Melopsittacus undulatus
OPSB_SAIBB	O13092	Saimiri boliviensis boliviensis
OPSL_CALJA	P34989	Callithrix jacchus
OPSR_CAPHI	Q95170	Capra hircus
OPSG_CAVPO	Q9R024	Cavia porcellus
OPSG_SCICA	O35478	Sciurus carolinensis
OPSO_RUTRU	Q7T3Q7	Rutilus rutilus
OPSR_HORSE	O18912	Equus caballus
OPSG_ODOVI	O18911	Odocoileus virginianus virginianus
OPSO_SALSA	O13018	Salmo salar
OPN4_PODSI	Q4U4D2	Podarcis sicula
OPN4_PHOSU	Q5XXP2	Phodopus sungorus
OPN4B_GADMO	Q804Q2	Gadus morhua
OPSD1_MIZYE	O15973	Mizuhopecten yessoensis
OPS6_DROME	O01668	Drosophila melanogaster
OPSD_LOLFO	P24603	Loligo forbesi
OPSC2_HEMSA	Q25158	Hemigrapsus sanguineus
OPSD_SEPOF	O16005	Sepia officinalis
OPSD_ALLSU	Q17094	Alloteuthis subulata
OPSD_ENTDO	P09241	Enteroctopus dofleini
OPSO_LIMPO	P35361	Limulus polyphemus
OPSD_PROML	O16020	Procambarus milleri

OPSD_TODPA	P31356	Todarodes pacificus
OPSD_CATBO	Q17296	Cataglyphis bombycina
OPS1_CALVI	P22269	Calliphora vicina
OPS1_DROPS	P28678	Drosophila pseudoobscura pseudoobscura
OPSD_CAMAT	Q17292	Camponotus atriceps
OPSD_CAMSC	O16018	Cambarellus shufeldtii
OPN4_BRABE	Q4R1I4	Branchiostoma belcheri
OPS1_SCHGR	Q94741	Schistocerca gregaria
OPSD_SPHSP	P35362	Sphodromantis sp.
OPSD_ORCVI	O16019	Orconectes virilis
OPSD_PROCL	P35356	Procambarus clarkii
OPSD_CAMLU	O16017	Cambarus ludovicianus
OPS1_MANSE	O02464	Manduca sexta
OPSCE_BOMMO	Q95YI3	Bombyx mori
OPSD_PROOR	O18485	Procambarus orcinus
OPSD_APIME	Q17053	Apis mellifera
OPSD_CAMHU	O18312	Cambarus hubrichti
OPSD_ORCAU	O18481	Orconectes australis
OPSD_CAMMA	O18315	Cambarus maculatus
OPSD_PROSE	O18486	Procambarus seminolae
NK2R_MESAU	P51144	Mesocricetus auratus
OPRM_MACMU	Q9MYW9	Macaca mulatta
OPS4_DROVI	P17646	Drosophila virilis
NPR11_CAEEL	Q18179	Caenorhabditis elegans.
GPR54_ORENI	Q6BD04	Oreochromis niloticus
OPRM_PANTR	Q5IS39	Pan troglodytes
NK1R_MERUN	Q5DUB1	Meriones unguiculatus
APJ_XENTR	Q4VA82	Xenopus tropicalis
GNRR2_CLAGA	O42329	Clarias gariepinus
CCR5_CERSO	Q9BGN6	Cercopithecus solatus
CCR5_CERLH	Q9XT76	Cercopithecus lhoesti
GR101_LYMST	P46023	Lymnaea stagnalis
GHSR_MUSPF	A5A4L1	Mustela putorius furo
CCR5_ERYPA	Q95ND0	Erythrocebus patas
CCR5_GORGO	P56439	Gorilla gorilla gorilla
CCR5_PONPY	O97881	Pongo pygmaeus
CCR5_HYLSY	Q95NC5	Hylobates syndactylus
CCR5_TRAJO	Q95NC6	Trachypithecus johnii
CCR5_NASLA	Q95NC7	Nasalis larvatus
CCR5_TRAFR	O97878	Trachypithecus francoisi
CCR5_TRAPH	O97879	Trachypithecus phayrei
CCR5_COLPO	Q95NC8	Colobus polykomos
CCR5_PYGBI	O97880	Pygathrix bieti
OAR1_LOCMI	Q25321	Locusta migratoria
CCR5_RHIAV	O97962	Rhinopithecus avunculus
CCR5_MACAR	O97975	Macaca arctoides
CCR5_PYGNE	O97882	Pygathrix nemaeus
EDNRB_COTJA	Q90328	Coturnix coturnix japonica
CCR5_CERTA	Q95NE8	Cercopithecus tantalus
CCR5_CALMO	Q95NC2	Callicebus moloch
CCR5_CERPY	Q9TV42	Cercopithecus pygerythrus
CCR5_LOPAT	P61755	Lophocebus aterrimus
CCR5_CERGA	Q9TV49	Cercocebus galeritus

CCR5_CHLSB	Q9TV43	Chlorocebus sabaeus
CCR5_MIOTA	Q95NC3	Miopithecus talapoin
CCR5_NOMLE	O97883	Nomascus leucogenys
CCR5_HYLML	Q95NC0	Hylobates moloch
CCR5_SAISC	Q8HZT9	Saimiri sciureus
CCR5_CERAT	O62743	Cercocebus atys
CCR5_MANSP	Q95ND1	Mandrillus sphinx
CCR5_CERNS	Q9TV45	Cercopithecus nictitans
BRS4_BOMOR	P47751	Bombina orientalis
CCR5_THEGE	Q95NC1	Theropithecus gelada
GPR18_AMPAM	Q93127	Amphibalanus amphitrite
CCR3_CHLAE	P56492	Chlorocebus aethiops
FSHR_CAIMO	Q7ZTV5	Cairina moschata
CCR5_MANLE	Q95ND2	Mandrillus leucophaeus
CCR5_CALHU	Q6WN98	Callithrix humeralifera
CCR5_CERCP	Q9TV47	Cercopithecus cephus
BKRB1_TUPMI	Q8HZP1	Tupaia minor
CCR5_ATEGE	Q95NC4	Ateles geoffroyi
CCR5_CERAS	Q9TV48	Cercopithecus ascanius
ITR_CATCO	Q90334	Catostomus commersonii
V1AR_MICMA	Q9WTV8	Microtus montanus
CNR1B_TAKRU	Q98895	Takifugu rubripes
V1AR_MICOH	Q9WTV9	Microtus ochrogaster
GLHR_ANTEL	P35409	Anthopleura elegantissima
FSHR_MACEU	Q6YNB6	Macropus eugenii
CTR2_OCTVU	Q5WA50	Octopus vulgaris
ADRB1_MELGA	P07700	Meleagris gallopavo
CCR5_ALOSE	Q95NC9	Alouatta seniculus
ADRB2_TSCTR	Q4KWL2	Tscherskia triton
FSHR_EQUAS	Q95179	Equus asinus
C3AR_ONCMY	Q2WED0	Oncorhynchus mykiss
PAR1_CRILO	Q00991	Cricetulus longicaudatus
VGE1_EHV2	Q89609	Equine herpesvirus 2
CNR1_TARGR	Q9PUI7	Taricha granulosa
CXCR4_TUPCH	Q7YS92	Tupaia chinensis
GPER1_MICUN	B0F9W3	Micropogonias undulatus
CXCR4_PAPAN	P56491	Papio anubis
CNR1_TAEGU	P56971	Taeniopygia guttata
CNR1_RANES	Q333S9	Rana esculenta
GPR85_PONAB	Q5RBG7	Pongo abelii
VK02_SWPVK	Q08520	Swinepox virus
CXCR6_MACNE	O19024	Macaca nemestrina
V027_FOWPN	Q9J5H4	Fowlpox virus
DRD1_DIDVI	P42288	Didelphis virginiana
US28_HCMVM	F5HF62	Human cytomegalovirus