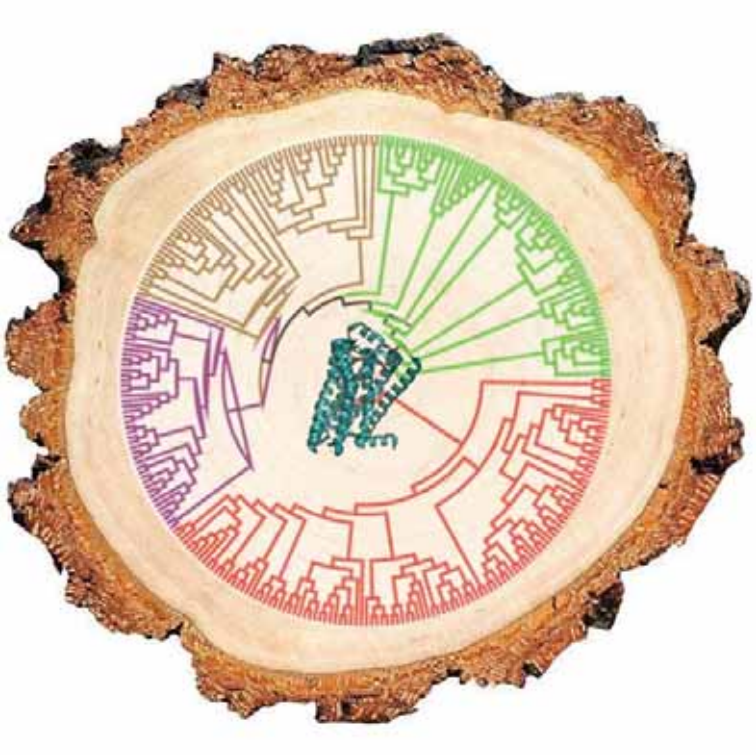
Universitat Autònoma de Barcelona

Sequence and structure-based bioinformatic tools to the characterization, clustering and modeling of G-protein-coupled receptors (GPCRs)

Santiago Rios Azuara

Doctoral Thesis

Sequence and structure-based bioinformatic tools to the characterization, clustering and modeling of G-protein-coupled receptors (GPCRs)



Santiago Rios Azuara

Laboratory of Computational Medicine,

Universitat Autònoma de Barcelona

Doctoral thesis
PhD program in Biochemistry, Molecular Biology and Biomedicine.

Sequence and structure-based bioinformatic tools to the characterization, clustering and modelling of G-protein-coupled receptors (GPCRs)

Report submitted by **Santiago Rios Azuara** to obtain the degree of doctor in Biochemistry, Molecular Biology and Biomedicine.

Directors: Gianluigi Caltabiano, Angel Gonzalez Wong

Tutor: Mireia Duñach Masjuan

Laboratory of Computational Medicine (LMC),
Medicine Faculty,
Universitat Autònoma de Barcelona (UAB)
Bellaterra, 2017

Abstract

In this work, we developed new bioinformatic tools for the study of G-protein-coupled receptors (GPCRs). The pharmacological importance of these receptors motivates the development of alternative methods to assist their classification, pharmacological identification and comparative modeling. Based on the recent advances in GPCRs crystalization, a new multiple sequence alignment strategy that incorporates irregularities observed on the receptor structures was proposed. The developed structure-based sequence alignment was used to update the GPCRs classification with significant advantages compared to previous studies. The recent structural data was also used to improve the analysis of the orthosteric binding site through the classification of the receptors in function of the ligand binding similarity. In specific, as part of this thesis we have developed a novel substitution matrix specifically derived from GPCRs (GPCRtm) and the web application (GPCR-Browser) that permits easier comparison of receptor sequence within subfamilies.

Acknowledgements

During the last years, I had the opportunity to participate at the research group under the leadership of Prof. Leonardo Pardo and Prof. Mercedes Campillo. I want as well to thank the rest of LMC group mates. With them, I enjoyed many good moments and I spent many other significant experiences, especially Arnau for “others moments”. On my future, I will feel awkward when I will do the lunch out of the lab. In particular, I want to thank my thesis directors, Dr. Gianluigi Caltabiano and Dr. Angel Gonzalez for their collaboration. I learned many thinks with them; this project would not be possible without you.

It is important to mention Prof. Mousseau and the rest of the colleagues in Montréal, Louis, Mickael and Daniel. Thanks for your company during my Canadian experience. I also want to thank my chemist teacher at the high school, Encarna, who introduced me in research. Finally, I want to thank my family for their support.

1 Introduction

The great advances in the genome sequencing, as well as in X-Ray structure determination techniques produced during the last decade, have brought to the scientific community a large amount of data about the sequences and structures of thousands of proteins. This information can effectively be used for medical and biological research with the development of adequate tools for their analysis and interpretation. In this regard, computational techniques may help us reach this goal. Bioinformatics methods are among the most powerful technologies available in life sciences. The use of statistical analysis on protein sequences and structures gives us a better understanding of their biological function, physiological roles and evolutionary relationships. There is a vast literature on successful applications of bioinformatic methodologies in the solution of biological problems, and what is more, there is an increasing need to integrate all the new knowledge in more accurate computational tools.

The main aim of this work is the development of bioinformatics techniques to characterize and study proteins with therapeutic importance. We focused our attention on the G protein-coupled receptors (GPCRs), one of the most relevant proteins families in drug discovery, with a well-recognized importance in clinical medicine. These receptors are essential in cell physiology, and their malfunction is commonly translated into pathological outcomes.

1.1 Biological and pharmacological importance of G-protein-coupled receptors

Cells are able to detect chemical signals present in their external environments by means of different classes of plasma membrane proteins, being the superfamily of G-protein-coupled receptors (GPCRs) one of the largest and most studied [1]. The origin of the GPCRs is presumed ancestral on account of their presence in most eukaryotic organisms including insects and plants. These heptahelical membrane-spanning receptors are highly diversified in mammalian genomes with current estimates of about one thousand genes, depending on the species [2].

GPCRs are involved in the translations of several endogenous and exogenous signals in cellular responses, modulating physiological processes as diverse as neurotransmission, cellular metabolism, inflammatory and immune responses, secretion, differentiation and vision among others [3]. These receptors are activated by a vast chemical diversity of natural and synthetic ligands including biogenic amines, neuropeptides, phospholipids, glycoproteins, nucleosides, nucleotides, amino acids, polypeptide hormones, odorants, ions and photons [4]. Considering the vast amount of cellular processes regulated by the GPCRs system, its wide tissue distribution and accessibility from the extracellular environment [5], it constitutes an attractive pharmaceutical target and accounts for around 30% of current drugs in market [6].

1.2 The GPCR signal transduction inside the cell

GPCRs control the activity of enzymes, ion channels and transport of vesicles via intracellular signalling cascade (Figure 1). Actions of GPCRs are driven inside the cell through heterotrimeric guanine nucleotide-binding proteins (G-proteins), ß-arrestins and G-protein-coupled receptor kinases (GRKs) among others. The activation of the G-protein (the most common secondary messenger) triggers the GTP-GDP exchange associated with the Gα-subunit and leads the βγ-dimer dissociation from Gα.

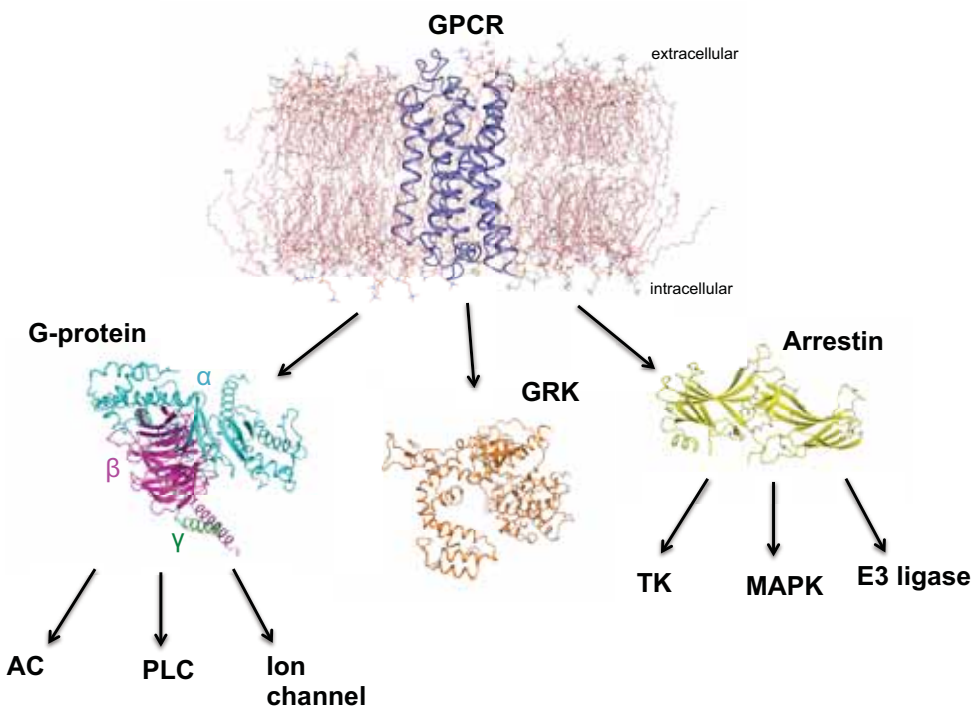


Figure 1. Scheme of the principal components of the GPCRs signal transduction machinery (GPCR coloured in purple). The G-protein, which primary effectors are adenylate cyclase (AC), phospholipase C (PLC) and ion channels, is coloured in cyan (α-subunit), magenta (ß-subunit) and green (γ-subunit). ß-arrestin, which primary effectors are tyrosine-protein kinase (TK), MAP kinase (MAPK) and E3 ubiquitin ligase (E3 ligase) is coloured in yellow. G-protein-coupled receptor kinase (GRK) is coloured in orange.

Subsequently, G-protein sub-units modulate the activity of diverse effectors like ion channels or enzymes [5] and the uncoupled GPCRs became substrates for G-protein-coupled receptor kinases (GRKs), which phosphorylate the intracellular part of the GPCRs. Phosphorylated GPCRs increase their binding affinity to ß-arrestin molecules. The GPCR - ß-arrestin complex drives the membrane protein desensitization and sequestration via clathrin-coated pits endocytosis [7]. ß-arrestins also function as G-protein independent signal transducers for TK, MAPK and E3 ubiquitin ligases effectors [8].

1.3 GPCRs classification

Early attempts to classify GPCRs had been based on phylogenetic relationships, the chemical nature of their ligands, pharmacological properties or by the design of fingerprints that encodes motifs on the seven transmembrane domains (7 TM) [9-11]. According to sequence-based approaches, the GPCR superfamily is organized on five to seven classes. Class A Rhodopsin-like, which account for over 90% of all GPCRs, class B Secretin-like, class C Metabotropic glutamate receptors, class D Pheromone receptors, class E cAMP receptors and the class F Frizzled/Smoothened receptors. They constitute the A-F system, comprising known GPCRs from both vertebrate and invertebrate species [9]. In humans, there are more than 800 GPCR genes [10]. They are divided into five large families: Glutamate, Rhodopsin, Adhesion, Frizzled/Taste2 and Secretin receptors (GRAFS) (Figure 2).

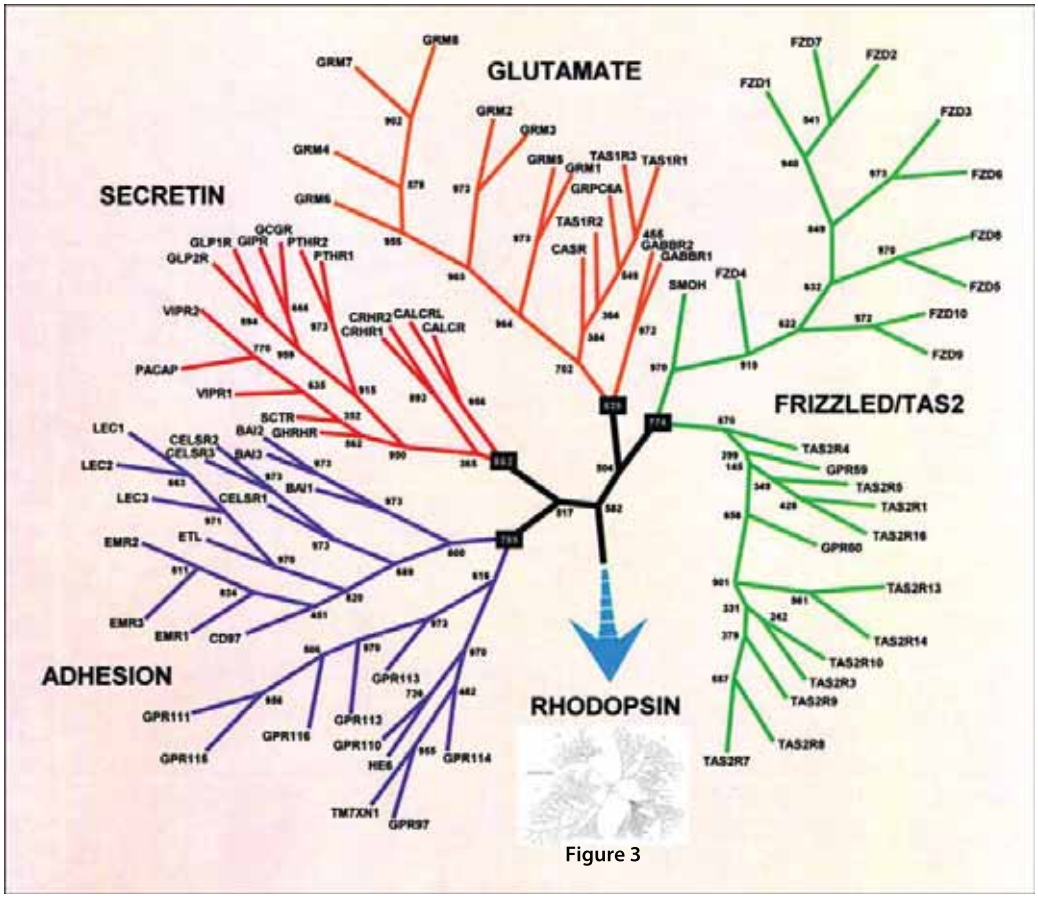


Figure 3

Figure 2. A phylogenetic relationship of the GPCR classes A, B, C and F (GRAFS classification) adapted from [10]. Rhodopsin family classification (lower inset) is described in detail in the Figure 3.

1.3.1 Class A GPCRs

The Rhodopsin or class A GPCR family has undergone a large evolutionary success in the bilateria species. This family is the biggest and the most studied of all. According to the GRAFS classification system [10], class A GPCRs are subdivided into four main branches α, β, δ and γ, and 13 sub-branches: olfactory, aminergic, peptide, chemokine-like, purine-like, somatostatin/opioid/galanin, opsin-like, glycoprotein binding,

prostaglandin, MECA (melanocortin, endoglin, adenosine, cannabinoid), MRG receptors, melatonin and melanocyte-concentrating hormone receptors. 460 out of the 710 class A GPCRs are olfactory receptors (Figure 3).

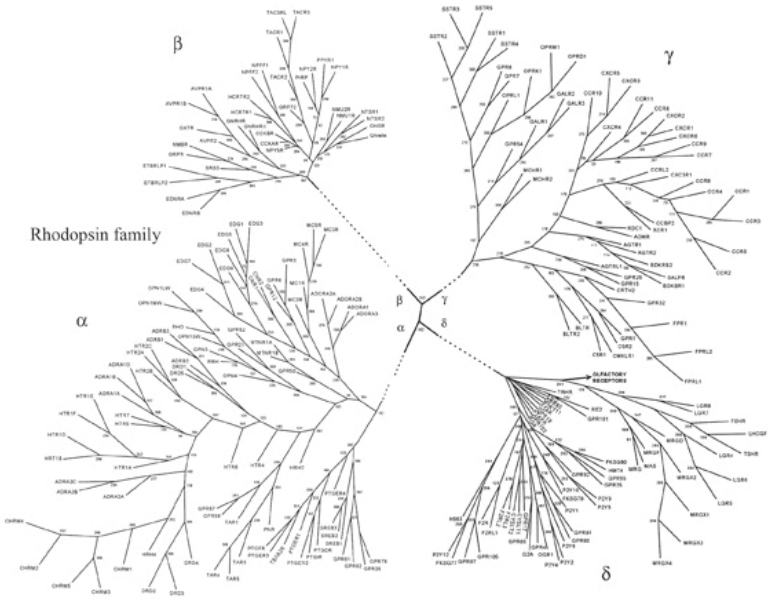


Figure 3. Phylogenetic relationships of the class A receptors according to the GRAFS classification. Adapted from [10].

1.3.2 Other GPCR classes

The class B is divided into Secretin and Adhesion receptors (15 and 24 members respectively). Class C, with 15 receptors, comprises the metabotropic Glutamate (mGlus), γ-aminobutyric acid B-type ($GABA_B$), calcium-sensing (CaS), taste 1 (TAS1) receptors, and several orphan receptors. Class F contains the Frizzled/Taste2 receptors (24 in total). The

other two classes (class D corresponds to Fungal mating pheromone and class E to cyclic AMP receptors), which are not present on humans.

1.4 Structural insights on GPCRs

Crystal structures of several GPCRs reveal an overall transmembrane fold preserved across the whole superfamily (Figure 4). These receptors display a highly conserved molecular architecture composed by seven α-helical transmembrane segments (7TM), which span the cell membrane, connected to each other by three extracellular loops (ECL1-3) and three intracellular loops (ICL1-3). In addition, an eighth α-helix (H8) is usually found at the beginning of the intracellular C-terminal, lying parallel to the membrane plane.

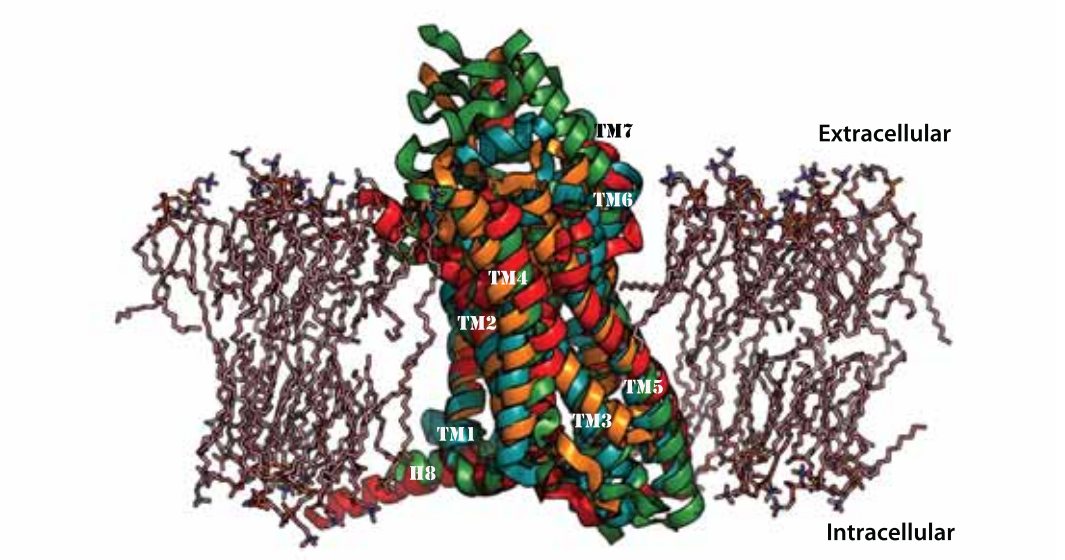


Figure 4. Comparison of the TM segments in the crystal structures from GPCRs of class A/Rhodopsin family (protein name: ADRB2, PDBid: 2RH1 [12] in cyan), class B/Secretin (GLR, 4L6R [13] in red), class C/Glutamate (GRM1, 4OR2 [14] in orange) and class F/Frizzled (SMO, 4JKV [15] in green). A lipid bilayer (in pink) is included on the representation.

In general, the interaction with ligands is produced on a binding pocket lying in the extracellular side of the receptors. Specific domains on N-terminus of several GPCR families would mediate these interactions, whereas in the majority of Class A receptors their effector molecules bind a cavity formed by the N-terminus, ECL2 and TM helices.

1.4.1 Class A GPCRs

One of the most important limitations in the study of GPCRs is the low similarity between their sequences, which in many cases is below the twilight region significant for homology detection [16]. Major sequence and structural divergences occur in the non-transmembrane regions, mostly in N- and C- terminus, ECL2 and ICL3. Interestingly, despite very low sequence conservation among class A GPCRs, they exist highly conserved residues, at least one per helix: N in TM1 (present in 98% of the sequences), D in TM2 (93%), R in TM3 (95%), W in TM4 (96%), P in TM5 (76%), P in TM6 (98%), and P in TM7 (93%) (Table A1 in Appendix). These residues have been used by Ballesteros and Weinstein to define a general numbering scheme consisting of two digits: the first (1 through 7) corresponds to the TM segment in which the amino acid of interest is located; the second pinpoints the position relative to the most conserved residue in the helix, arbitrarily assigned to 50 [17]. Significantly, the position of these highly conserved amino acids in each helix is the same in the superimposition of the currently available crystal structures (Figure 5). Thus, this finding validates their use as reference points in TM sequence alignments and for the construction of homology models of GPCRs with unknown structure [18].

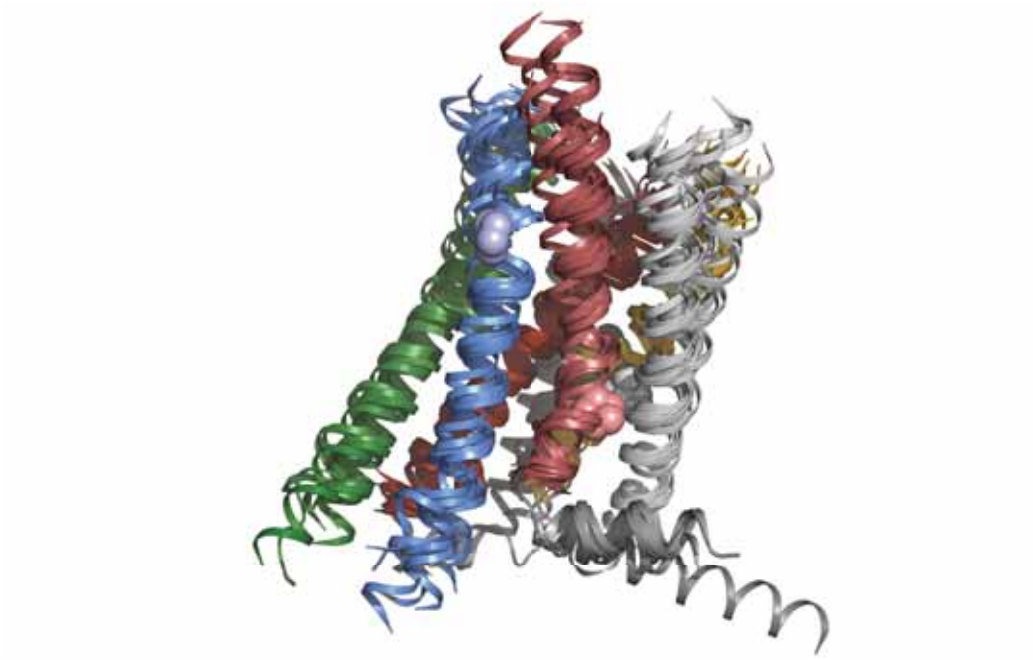


Figure 5. Comparison of the TM segments in the crystal structures of class A GPCRs: OPSD (PDBid: 1GZM), ADRB2 (2RH1), DRD3 (3PBL), S1PR1 (3V2Y), PAR1 (3VW7), NTR1 (4BUO), P2Y12 (4PXZ), OX2R (4S0V), EDNRB (5GLH), AA2AR (5IU4) and CCR9 (5LWE). The color code of the TM helices is 1 in white, 2 in yellow, 3 in red, 4 in black, 5 in green, 6 in dark blue, 7 in light red, and C-terminal in grey. The highly conserved N1.50, D2.50, R3.50, W4.50, P5.50, P6.50, P7.50 are shown in spheres.

Class A GPCRs 3D structures reveal different spatial conformations of the N-terminus and ECL2 that maintain the binding site rather accessible from the extracellular environment. Thus, each receptor subfamily has probably developed, during evolution, a specific N-terminus/ECL2 complex to adjust the structural characteristics of its cognate ligands, and to modulate the ligand binding/unbinding events [19-21]. Analysis of crystal structures shows that ligand binding mostly occurs in a major crevice located on the upper site of TMs 3, 5, 6 and 7. In addition, a minor cavity also exists on

the TMs 1, 2, 3 and 7 (Figure 6). This minor binding pocket is usually associated with ligand selectivity [22].

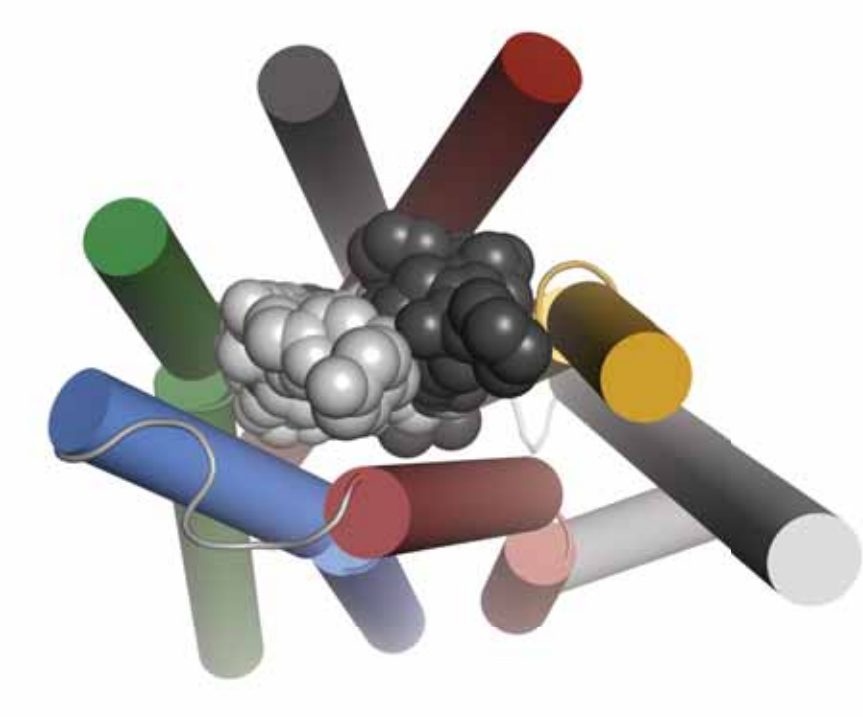


Figure 6. The ligand-binding cavities in class A GPCRs. Superimposition of ligands co-crystallized with the receptors ADBR2 (PDBid: 2RH1, 3PDS, 3D4S), ADRB1 (2VT4, 2Y00, 2Y01, 2Y02, 2Y03), AA2AR (2YD0, 2YDV, 3EML, 3PWH), CXCR4 (3ODU), DRD3 (3PBL), HRH1 (3RZE), ACM2 (3UON), ACM3 (4DAJ), S1PR1 (3V2Y) and OPRK (4DJH), all showed over ADRB2 crystal structure (PDBid: 2RH1) in cylinders helices. Ligands populating the major and the minor crevices are colored in grey and black VdW spheres respectively. The color code of the helices is the same than Figure 5.

Structural alignment of class A GPCRs has revealed changes in the α-helical scaffold in the form of tight and wide turns at some TM helices. These distortions are related to residue insertion or deletion events (known as *indels*) accumulated during the evolution and challenge the established

paradigm of avoiding gaps in alignments of the TM regions [23]. *Gonzalez et al* [24] show how sequence gaps of one or two amino acids size has to be introduced in TM2 and in TM5 sequences for some class A GPCRs in order to reflect their spatial superposition in crystal structures. These structural anomalies in TMs 2 and 5 could have played an important role in the diversification and evolutionary success of GPCRs given that a part from amino acid substitutions, indels are among the most common events in protein evolution [25]. These findings have direct implications in sequence alignments and in homology modeling as well as in phylogenetic reconstruction. Nonetheless to-date most numbering schemes of Class A GPCRs do not accurately reflect amino acid sequence position with their tertiary structure location.

1.4.2 Other GPCR classes

To date, 19 structures of the transmembrane domain of receptors from classes B, C and F have been determined in complex with ligands of varied pharmacology. The Class B Secretin receptors interact with endogenous ligands with the extracellular domain (with 100-150 residues) and part of the extracellular regions (ECL1-3) [26-28]. Whereas Adhesion receptors are characterized by the presence of an extracellular GPCR-Autoproteolysis-INducing (GAIN) domain located immediately N-terminal to the 7TM [29]. Class C GPCRs generally form dimers or higher order oligomers, with extracellular domains composed by Venus Flytrap modules (VFTM), cysteine-rich domains (CRD) and heptahelical domains (HD) [14, 30, 31]. Finally, the Class F activated by the lipoprotein WNT, contain a extracellular cysteine-rich domain (CRD), which binds

endogenous ligands, and a linker domain [15, 32] (Figure 7, Table A2 in Appendix).

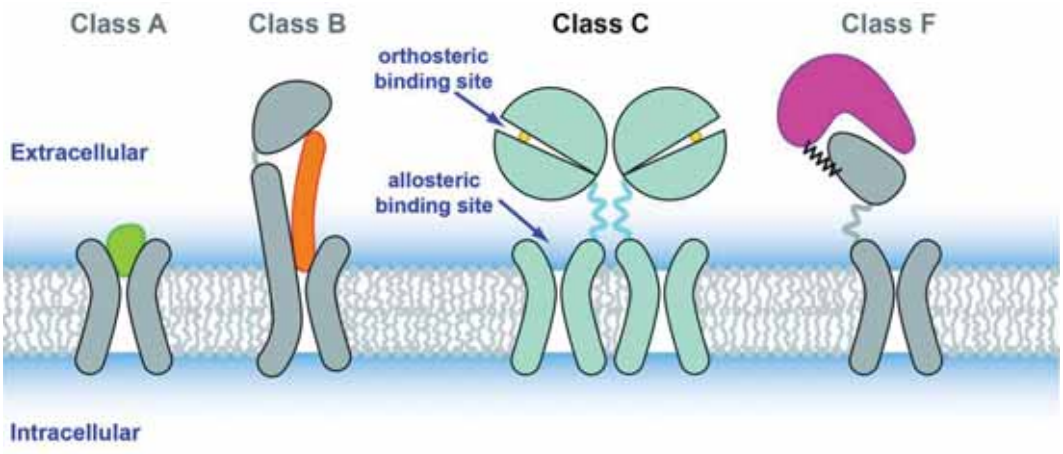


Figure 7. Representation of the structures and ligand binding regions of Class A, B, C and F GPCRs. Endogenous ligands bind transmembrane (TM) cavity in class A. In the secretin class B receptors, endogenous peptides bind on the N-terminal and extracellular loops (ECLs) regions. Class C makes dimers and ligands binds on the long N-terminal named Venus flytrap domain (VFD). In class F, ligand binds the cysteine-rich domain (CRD) in the N-terminal. Figure taken from [14].

Alternatively, small molecules could bind different allosteric sites in these receptors, modulating, in many cases the pharmacological properties of orthosteric ligands [33]. (Figure 8).

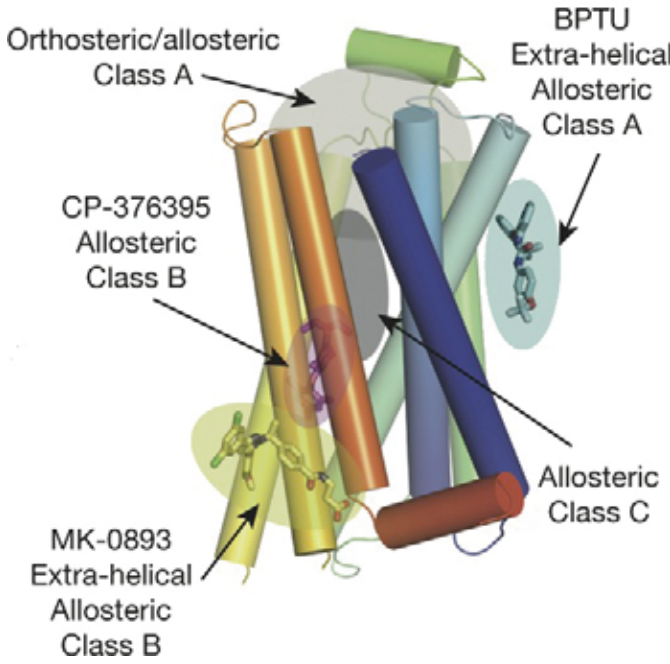


Figure 8. Schematic overview of known binding positions of class A, B and C GPCR ligands. Figure taken from [34].

1.5 Molecular switches in GPCRs

Recent advances in experimental crystallization techniques [35, 36] had allowed the crystallization of GPCRs in different conformational states. Analysis of the atomic-level information retrieved for these receptors in complex with agonists, antagonists, inverse agonists, and accessory proteins unveiled differences among states of the activation process [37]. However, despite the enormous chemical diversity of their ligands, all GPCRs shared a common activation mechanism. In class A GPCRs, this activation mechanism make use of small molecular micro-switches [38] (Figure 9), which are highly conserved across the family.

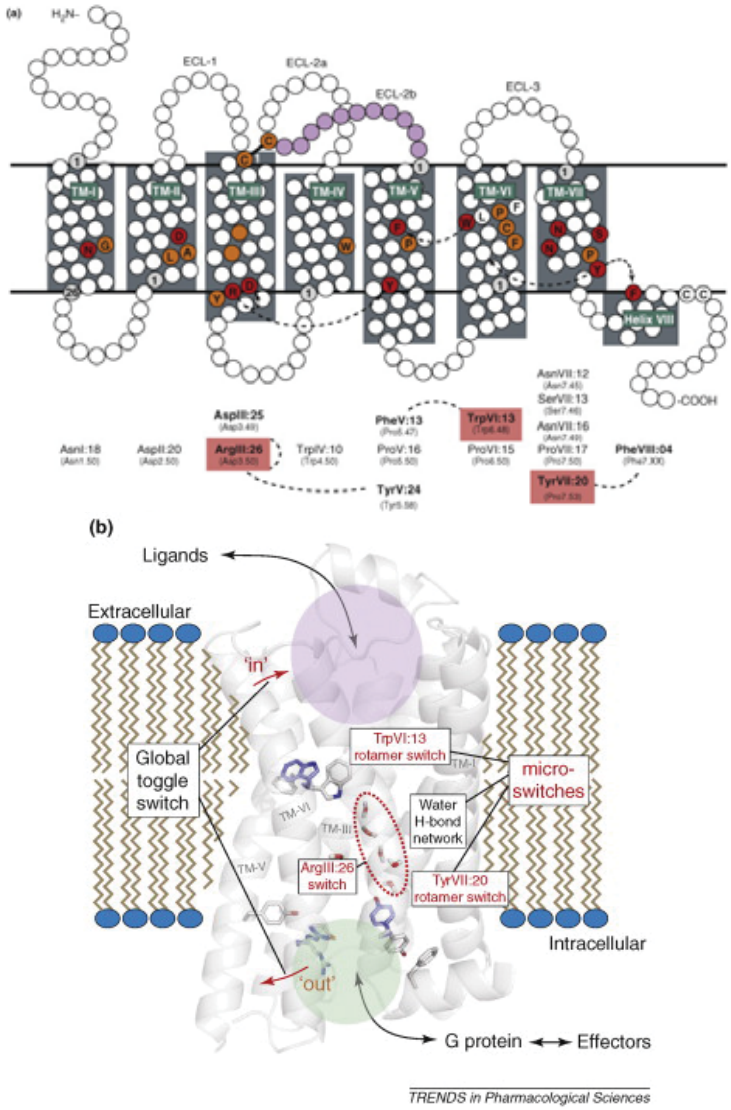


Figure 9. Micro-switches of class A GPCRs. (a) Conserved residues are indicated with orange and red (highly conserved) symbols in a serpentine model of a 7TM receptor and are listed below using both the structural generic numbering system and the mathematical generic system (in brackets below). (b) The 7TM global toggle switch with micro-switches and extra- and intracellular ligand-binding pockets, showed over ADRB2 structure. Figures taken from [38].

1.6 Clustering methods for GPCRs

Conventional strategies for identifying and classifying proteins involve similarity searches looking for common biological functions or using sequence database search tools (as example: BLAST). Unfortunately, biological data are often unknown and methods based on pairwise alignment only appreciate generic similarities between sequences. Then, alternative techniques, as sequence profiles alignment or phylogenetic analyses among others, have been developed in order to overcome such problems. The most widely accepted GPCRs classifications are listed below.

1.6.1 Pharmacological property approaches

Traditional GPCR's classification is based on receptor pharmacology, which is used as reference by all computer-generated classification. As previously mentioned, GPCRDb [11] and NC-IUPHAR [39] are the most common databases that label and classify receptors based on their endogenous ligand and pharmacological properties. However, similar ligands do not always bind to similar receptors, and similar receptors do not always recognize similar compounds. As example, all LPARs binds the same endogenous ligand [40] despite a mere 25% of identity, while other receptors (APJ and AGTRs) with higher identity, do not always share the same binding partners [41]. These observations reflect a complex evolutionary background with GPCR sequences converging or diverging before they come up with their current functional profile.

1.6.2 Phylogenetic analysis approaches

The phylogenetic analysis was originally based on phenotypes distinction. Currently, the new technologies permit more reliable classifications using the sequence alignments of DNA, RNA, proteins or non-biological data. The methodology to build the sequence classification is chosen in order to optimize the repartition of character states and the result is displayed as a phylogenetic tree or dendrogram (the graphic representation of the computed results). Most of the classification methods can be classified as i) distance- or ii) character-based methods. i) Distance (or algorithm) methods convert sequence data into a distance matrix. Distance values are stated from an evolution model algorithm and they reflect the number of differences between each pair of sequences. The tree is then constructed from these distance values by progressive clustering. ii) Character-based (or tree-searching) methods search the branch and bound tree topology that better fit the set of taxa [42]. In practice, both distance based and tree-searching methods are combined. For example, an initial tree may be estimated by distance-method Neighbour Joining (NJ) and subsequently, the maximum likelihood (ML) method maximizes the likelihood of the tree topology parameters from a given data [43].

Possibly the first and surely the benchmark of all GPCRs classifications was presented in 2003 by Fredriksson *et al.* (GRAFS classification, Figure 3), who classified GPCRs using fingerprints motifs and phylogenetic algorithms. Of the 800 human sequences classified as GPCRs, the 241 non-olfactory class A receptors were clustered in four main groups (α, ß, γ, δ) with thirteen sub-branches. α-branch clusters prostaglandins (or

prostanoid), amine (5-Hydroxytryptamine, acetylcholine, dopamine, histamine, adrenergic and trace amine receptors), opsin, melatonin and MECA receptors (Melanocortin, Lysophospholipid LPAR1-3 and S1PR, Cannabinoid, Adenosine and the orphan GPR3, GPR6 and GPR12 receptors). ß-branch includes peptide receptors: vasopresin, endothelin, bombesin, cholecystokinin, ghrelin, gonadotrophin-releasing hormone, orexin, neurotensin, neuromedin U, oxytocin, neuropeptide FF, neuropeptide Y, tachykinin, prolactin-releasing peptide and the orphan GPR83 receptors. In γ-branch, receptors are sub-grouped on the three clusters: SOG receptor cluster (somatostatin, opioid and galanin receptors), melanin-concentrating hormone “MCH” receptors and chemokine receptors: ACKRs, CCRs, CXCRs, XCR1, angiotensin II, bradykinin, chemerin, complement C5a peptide, formylpeptide, leukotriene B4, relaxin-3 and the orphan GPR1, GPR15, GPR25, GPR32 receptors. The last branch (δ) clusters MAS-related, glycoprotein (hormone receptors) and purin receptors (P2Y, Lysophospholipid LPAR4-6, platelet-activating factor, N-arachidonyl glycine, complement anaphylatoxin chemotactic, proteinase-activated, cysteinyl leukotriene, oxoglutarate, succinate, hydroxycarboxylic acid, thyrotropin-releasing hormone, ovarian cancer, QRFP, RPE-retinal and the orphan P2Y10, GPR17, GP132, GP174, GPR35, GPR55, GPR4, GP171, GPR82, GPR161 and GPR101 receptors). Olfactory and other 7TM receptors are separately grouped.

More recently, the GRAFS classification was updated [2] and new classifications have been proposed. Chabbert and colleagues trace a molecular evolution path driven by specific residues at TM2 [44], TM5 and the WXFG motif at the ECL1 [45]. From these evolution markers, class A

receptors are proposed to cluster in four groups (G0, G1, G2 and G3). G0 includes peptide, opsin and melatonin receptors. The G1 includes somatostatin/opioid, chemotactic and purinergic receptors. The G2 is composed by amine and adenosine receptors and the G3 include leucine-rich repeat, melanocortin, S1P, cannabinoid, prostaglandin and MAS-related receptors. Finally, Kakarala *et al.* developed a sequence-structure based phylogeny to identify potential ligand association for class A orphans [46].

1.6.3 Profile and pattern approaches

Many methods classify protein sequences using learning statistical models obtained from the various protein classes. Profile hidden Markov models (profile HMMs) are one of the most common, building statistics from the sequence alignment consensus. However, such profiles fail when query proteins lack significant similarity on the database sequences. Then, more accurate HMM profiles are made for classification performance. T-HMM method [47] clusters GPCRs in function of a phylogenetic tree-based profile hidden Markov model. PRED-GPCR [48] application pretends to progress the method with an exhaustive discrimination of the low selective and sensitive HMM profiles.

In 2002, Lapinsh et al. [49] developed an alignment-independent method for GPCRs classification according to the chemical properties of the amino acids. These types of Support Vector Machines (SVM) methods extract physicochemical properties of the proteins. In GDS classification [50], primary amino acid sequence are described by 26 physicochemical

properties as hydrophobicity, bulky and polarity among others. In SVMtree [51], Karchin *et al.* combine HMM and SVM methods to produce a hierarchical multi-class SVM classification. In GPCRpred [52] and GPCRsclass [53], another SVM method determines fixed-length vectors from the dipeptide composition of the proteins.

1.6.4 Chemogenomic of the TM binding cavity approaches

The particular architecture of class A GPCRs, where ligands directly contact the transmembrane bundle residues, makes feasible the classification of these receptors only using the physicochemical properties of the binding cavity residues. In 2001, Jacoby created one of the first chemogenomic strategies where monoamine receptors were clustered in base of different ligands and the related binding site [54]. The lack of the rest of the sequences in these type of analysis permits a more ligand addressed classification and helps the ligand discovery research.

In 2006, Surgand and colleagues analyzed the residues of the ligand-binding site for all GPCRs by phylogenetic classification scheme. The residues subset was obtained considering amino acids of 30 discontinuous positions supposedly involved in ligand-binding [55]. These 30 critical positions were selected by the X-ray structure of the bovin rhodopsin receptor (PDB code: 1F88 [56]). The clustering of the ligand-binding database permits an easy detection of similarity and selectivity between different ligand-receptor interactions. According to this classification, class A receptors were grouped in 19 clusters: Prostanoids, glycoproteins, MAS-related, SREB, opsins, lipids, peptides, melatonin, vasopeptides,

adenosine, amines, melanocortins, brain-gut peptides, acids, chemokines, opiates, chemoattractants and purine receptors, and three more clusters for non-class A receptors (Figure 10).

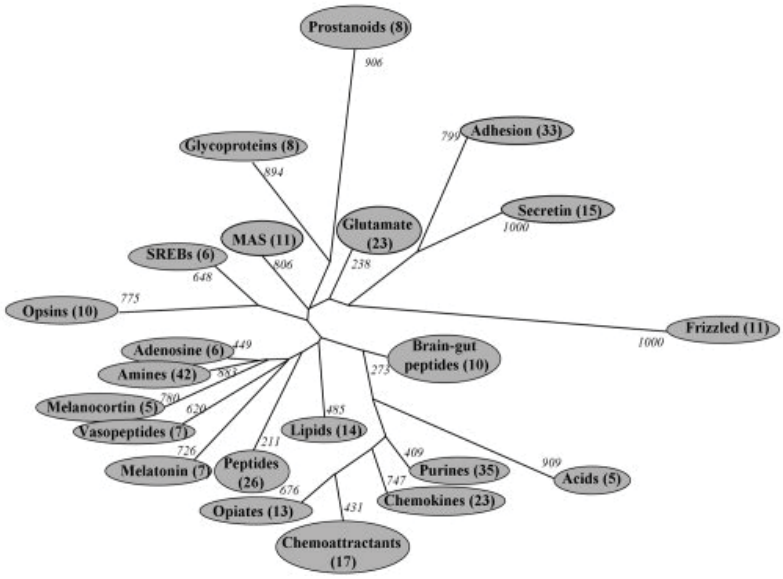


Figure 10. TM cavity-derived phylogenetic tree for human GPCRs. Taken from [55].

Thanks to the increasing number of GPCR 3D structures, more details about their ligand-binding interactions become available, allowing a more accurate classification. In 2009, a new reference set for class A GPCR binding pocket (with seven crystal structures available at that moment) defined 44 positions important in ligand binding [57]. These methods has been compared against sequence phylogenetic analysis of TM segments [58] and against classifications shemes obtained by similarity ensemble approach (SEA) [59].

In 2016, Ngo *et al* created the "*pocketome*" database in order to classify class A GPCRs by a combination of all receptor-ligand interaction pattern

extracted from crystal structures [60]. This new methodology helps the identification of related properties between receptors on different sub-families.

1.6.5 Protein-ligand fingerprint approaches

Protein-ligand fingerprint methods introduce novel low-dimensionality fingerprint encoding both ligand and receptor physicochemical properties which is suitable to mine protein−ligand chemogenomic space. Whereas ligand properties have been represented by standard chemical descriptors, protein cavities are encoded bit strings describing pharmacophoric properties of a definite number of binding site residues. This concept has been applied to G protein-coupled receptors with a homogeneous cavity description [61-64].

1.7 Computational tools for the study of GPCR

The importance of the GPCRs in cellular physiology has inspired the development of numerous computational tools and databases for their study over the years. GPCRdb [65] and Pocketome [60] are web servers that manage high quality curated sequence and structural GPCRs information. PRED-GPCR [48], GPCRpred [52], GPCRsclass [53] and SEVENS [66] predict protein classification from a query sequence. On the other hand, FoldGPCR [67], GOMoDo [68], GPCR-ModSim [69] servers allow online homology modeling, docking and molecular dynamic simulations.

Despite all these available tools, a lot of work is still needed. Currently, there are an important number of unclassified receptors (nearly one hundred [40], mostly in class A sub-family) which encourages the development of new tools to assist the identification and deorphanization of GPCRs. In addition, an improvement of the current classifications for the GPCR superfamily could be feasible taking into account the new knowledge obtained from structural data.

In this work, we propose an improved methodology for the clustering of class A GPCRs, where the new structural information derived from X-Ray crystallography studies are considered in the construction of multiple sequence alignments (MSAs) for this protein family. As a result of this approach, we propose modifications to the existing GPCRs classifications systems. In addition, this information has been used in the development of computational tools to compare GPCRs sequences, conduct similarity searches and propose structural templates for homology modeling studies.

2 Objectives

The aim of this project is to make use of the state-of-the-art information about the sequence-structure relationships in GPCRs to develop bioinformatic tools that assist in classification, pharmacological identification and comparative modeling within this receptor family.

To achieve these goals, we proposed to develop:

- A substitution matrix specific for GPCR sequences. This tool would be useful for sequence alignment, BLAST searches and phylogenetic analysis.
- A web application server to integrate structural information derived from GPCRs in the comparison of the sequences from TM regions, ligand-binding sites and for template selection in comparative modelling studies.

These new generated bioinformatic tools could be of interest in projects related to phylogenetic studies, comparative modeling, molecular docking and other pharmacological applications in GPCRs.

3 Methods

3.1 Amino acid sequence retrieval and alignment of the GPCR class A family

3.1.1 Sequence datasets employed

3.1.1.1 Class A GPCRs

Class A GPCR protein sequences from different biological sources were obtained from Uniprot [70]. The initial dataset was extended with the inclusion of 314 sequences functional human olfactory GPCR repertoire [71]. The final dataset is composed by 1019 sequences.

3.1.1.2 Human class A non-olfactory receptors

Human class A GPCR sequences were obtained from the UniProt database using the following syntax: as field name organism (OS) “Human [9606]”, as family and domains: “G protein coupled receptor 1 family”. In this set, odorant receptors were excluded with the field “NOT name: olfactory. The acquired 297 reviewed entries, were manually revised and sequences without seven TM domains were removed. The final dataset is composed by 292 sequences.

3.1.2 Post-processing of multiple sequence alignments sets according to structural information

UniProt and GPCRdb annotations were used to identify TM segments and boundaries of the TM helices were defined according to the available crystal structures of class A GPCRs (Table A1 in Appendix) [21, 72]. Sequences corresponding to TMs 1– 7 were aligned using the Win32 version of ClustalW 2.1 [73]. ClustalW was used with a gap open/extention penalty value of 40/0.1. The resulting alignment was manually curated according to the consensus signatures of class A GPCR: GN1.50, LAxxD2.50, DR3.50Y, W4.50, P5.50, Y5.58, CWxP6.50, NP7.50xxY [37], including the ECL1 WxFG motif [74] and the highly conserved cysteines in TM3 and ECL2 involved in a disulfide bridge in most receptors [75]. The disulphide bond between TM3 and ECL2 was considered as formed when both cysteine (in position 3.25 and another cysteine in ECL2) were detected and then, the cysteine at ECL2 was aligned at position 45.50 (GPCRDB numbering scheme is used at the ECL2 region [76]). Finally, gaps were inserted in the TM2 and 5 according to previous studies [24]. This resulted in two multiple sequence alignments (MSA) of 292 and 1019 non-redundant TM GPCR sequences (see Figure A1 and A2 in Appendix).

3.2 Construction of a GPCR amino acid substitution matrix

3.2.1 Construction of GPCRtm

The alignment of the TM regions of the 1019 GPCR class A sequences database was used to generate a substitution matrix representing changes

on this protein family using an implementation of the methodology described by Henikoff et al. [77]. In this regard, extracellular and intracellular regions are removed and the corresponding TM segments (1-7), which consist of multiple alignments of short regions (< 40 amino acids), were treated as sequence blocks. As initial step, a transition count (frequency) table was computed to determine the total number of amino acid transitions pairs from each column of the alignment. After the transition count table was completed, observed and expected probability of transition were computed for each pair. The observed probability (O) for the amino acid pair (i,j) is the total number of transitions observed (from the frequency table) divided by the total number of transitions for the entire alignment.

$$O_{ij} = f_{ij} / \sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij}$$

The expected probability (e) of occurrence for each (i,j) pair was calculated from the observed probabilities for the pair.

For a single residue:

$$p_i = O_{ii} + \sum_{i \neq j} O_{ij}/_2$$

for an (i,j) pair:

$$e_{ij} = p_i p_j + p_j p_i = 2 p_i p_j \quad \text{for} \quad i \neq j$$

when i = j,

$$e_{ij} = p_i p_j = p_i^2$$

Using the expected (e) and observed (O) probabilities of transitions, the substitution values were calculated from the odds ratio matrix, as the logarithm of odds, where each entry is obtained according to:

$$S_{ij} = 2\log_2\left(O_{ij}/e_{ij}\right)$$

The scaling factor of 2 is taken from Henikoff et al. [77] in order to facilitate comparisons. In the final 20 × 20 amino acid matrix, substitutions values where rounded to the nearest integer value. In addition, we calculate the average mutual information per amino acid pair or relative entropy (H) according to:

$$H = \sum_{i=1}^{20}\sum_{j=1}^{i} O_{ij} \times S_{ij}$$

3.2.2 Evaluation of the GPCRtm in database searching and pairwise alignments

One hundred random sequences from different GPCR subfamilies, including the four main groups α, β, δ and γ [10], were used as queries in BLASTP searches executed with the AB-BLAST software (http://blast.advbiocomp.com/) against the pdbaa database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/). Parameters to the customized gapped alignment score system for the GPCRtm were computed with the ALP program [78] (see Table S3). All BLASTP results were conducted with a gap existence = 15 and a gap extension = 2 scoring parameters, except for the BLOSUM62 matrix (gap existence = 11 and a gap extension = 1, default parameters). Matched comparisons of GPCRtm against JTTtm,

PHAT, BLOSUM62 and BLOSUM45 matrices were calculated with the IBM SPSS Statistics for Macintosh, Version 22.0 using the exact McNemar 2-tailed tests (p-values). Pairwise sequence alignments were generated with the MAFFT (L-INS-i) software using default parameters [79, 80].

3.3 Development of the clustering methods for GPCRs

3.3.1 Clustering GPCRs according to TM regions

The human class A non-olfactory GPCR sequence dataset, was used to build an unrooted tree using PhyML software 20120412 version [81] (see results). To claim confident alignment regions, only TM blocks were used. As mentioned earlier, two methodologies were combined in order to obtain a better phylogenetic tree. Five starting trees were built using the maximum parsimony method (pilot tests had shown better results than distance-based methods), these were optimized with Subtree Pruning and Regrafting (SPR) [82] and Nearest Neighbor Interchange (NNI) algorithms [83]. Maximum likelihood method compares the different trees, parameters and models and computes the most probable hypothesis. The amino acid equilibrium frequencies were obtained from the GPCRtm substitution matrix [24]. Bootstrap method “aLRT” based on SH and Chi-square criteria [84, 85] were used to support branch nodes (Figure A3 in Appendix). The obtained phylogenetic tree was rendered by FigTree 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) for an easier visualization.

3.3.2 Clustering GPCRs according to ligand-binding site residues

40 amino acids involved in ligand-receptor interactions were selected based on the analysis of 111 crystal structures of class A GPCRs ligand-receptor complexes deposited in Protein Data Bank (Table A1 in Appendix). Residues within a distance ≤ 5 Å of the ligand were selected in every crystal structure and annotated according to Weinstein-Ballesteros scheme. In order to decrease the bias of the available crystallographic information only positions observed in at least two crystal structures from different receptors were included in the consensus binding pocket sequence database (Table A4 in Appendix). Finally, residues of the selected positions were extracted from the human class A dataset of 292 sequences to construct ligand-binding site residue alignment (Figure A7 in Appendix). All 40-amino-acid-long sequences were compared using a similarity matrix and were visually expound using heatmap representation. Similarity scores for every receptor pairwise were weighted by GPCRtm substitution matrix [86] and normalized using equation 1.

Equation 1 $$S_{ij}^{norm} = \frac{S_{ij}}{\sqrt{S_{ii*} S_{jj}}}$$

Where S_{ij} is the initial similarity value for every pairwise receptor.

Similarity values were converted to distance by maximum metric function and clustered by average algorithm. Heatmap plots were made using the gplots library of the R software (http://CRAN.R-project.org/package=gplots). Color intervals were manually adjusted according to: strong dissimilarities (below the value -0.25) colored in dark

blue, weak dissimilarities (between -0.25 and 0) colored in white, weak similarities (between 0 and 0.15) in light blue, medium similarities (between 0.15 and 0.25) in yellow, strong similarities (between 0.25 and 0.5) in orange and very strong similarities (between 0.5 and 1) in red.

3.4 Design and implementation of the GPCR Browser web application

The GPCR-Browser (http://lmc.uab.cat/gpcr-browser/) is an easy-to-use web server built to show and take advantage of all the information retrieved in this study. Web-accessible tools were implemented in python programming language and the web page interface was developed with php code. Users can either 1) input a Uniprot protein codes or 2) the fasta sequences of the class A GPCR they are interested in. The validity of the input is always checked. In the former by validating the input code with our pre-compiled class A receptor-name list. In case a fasta sequence is pasted as input, a blast research against the “full human” GPCR database is run in order to confirms that the input sequence belongs to class A receptor (blast search has an e-value lower than 10^{-30}). Once input is checked and considered valid, GPCR-Browser run three independent analysis by mean of four python scripts and it displays the results on a user-friendly interface. The following paragraph explains the detail of each analysis.

1) The phylogenetic tree obtained by the TM database is parsed using Phylo library of the Biopython open source tools [87]. Then, the

sub-branch tree with the selected input (from protein code or blast result) is shown as a jpg picture.

2) A ligand-binding cavity representation is created showing the residues of the input sequence considered as part of the binding site. Residues are extracted from the fasta sequence by a previous alignment and a posteriori binding position identification. The protein sequence is aligned against the GPCR database with MAFFT program version 7.215 [80] using "Align full length sequences to an MSA", a gap penalty value of 5, and the GPCRtm substitution matrix are implemented by default [86]. Otherwise, residues are extracted from our database on the protein code option. Similarity scores are calculated as mentioned in 3.7 and then the twenty best similarity values are selected and sorted.
3) Finally, GPCR-Browser build a phylogenetic tree between the selected target and all receptors with known crystal structures. Sequence of all available crystal structures class A GCPR and user query are used to run a phylogenetic reconstruction using PhyML software 20120412 version with default parameters. For protein code as input, the protein sequence is extracted from our database otherwise, if fasta sequence is introduced as input, it is aligned to crystal sequences using MAFFT (as previously explained).

The GPCR-Browser tool is updated every six month. The sequence database is checked from Uniprot and the crystal structure list is renovated from pocketome website. The Uniprot and pocketome web servers are parsed by python scripts.

4 Results and Discussion

4.1 Use of structural knowledge in the improvement of sequence alignments of GPCRs

Multiple sequence alignments (MSAs) constitute an essential tool for the study of protein families. In GPCRs, the quality of an MSA could be enhanced using the available structural information (currently, 210 PDB entries for 43 unique GPCRs are available).

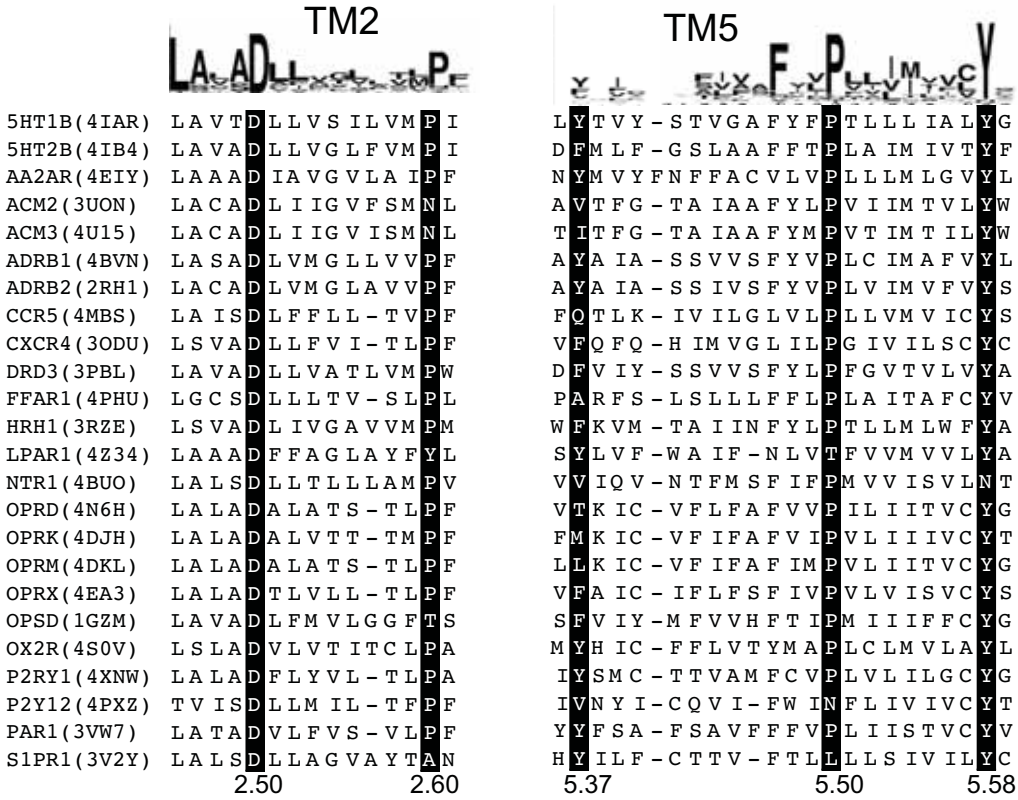


Figure 11. Multiple sequence alignment of TM2 and TM5 regions of crystalized class A GPCRs. The conserved residues are highlighted in black. Sequence gaps were included to amend spatial correspondences observed in crystal structures. Sequence logos are built with weblogo (http://weblogo.berkeley.edu/logo.cgi).

Structural alignment of GPCRs X-ray crystal structures unveiled the presence of insertion and deletion events leading to the inclusion of sequence gaps on TM helices MSAs. Figure 11 shows modifications in the TM2 and TM5 MSA based on structural differences observed in AA2AR, CCR5, CXCR4, FFAR1, LPAR1, OPRs, P2RY1, P2Y12, PAR1 and S1PR1 receptors with regard to the rest of Class A receptors. Such changes were implemented in order to fit amino acids on the same spatial location on the 3D structures.

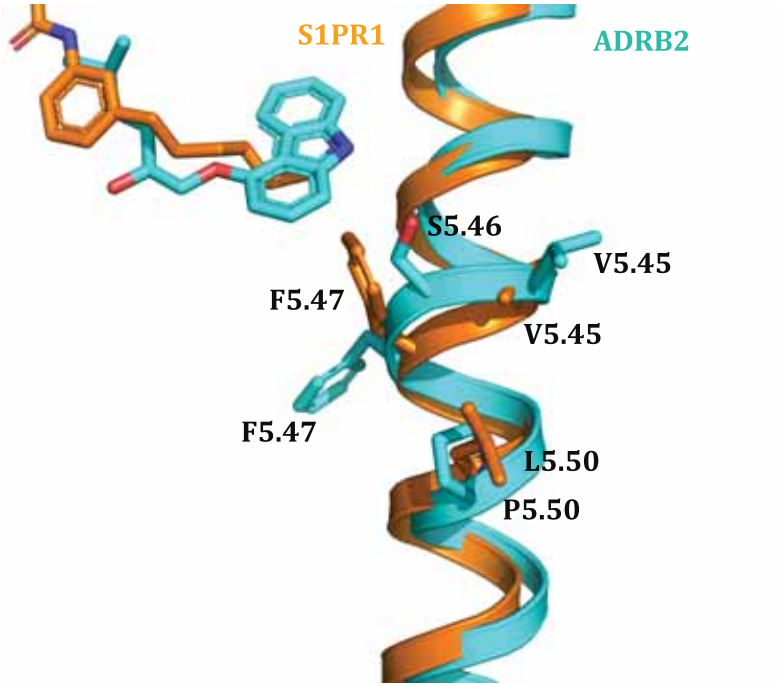


Figure 12. Representation of the ligand-binding site of the co-crystalized ADBR2 with carazolol (PDBid: 2RH1 in cyan) and S1PR1 with a phosphonic acid ligand (PDBid: 3V2Y in orange). Receptors are superposed and only the TM5 helix is showed. Residues on positions 5.45, 5.46,5.57 and 5.50 are highlighted. Ligands are shown as sticks.

As example, structural superposition shows a non-correspondence of a serine residue at position 5.46 in ADRB2 with regard the S1PR1 receptor structure (Figure 12). It is important to mention that these changes in the TMs alignment has an impact on the Ballesteros-Weinstein numbering as well as on the MSAs statistics, and highlight specific compositional bias observed in some members of the family.

4.2 Construction of an aminoacid substitution matrix for the Class A GPCRs

Protein sequence alignments and database search methods use standard scoring matrices calculated from amino acid substitution frequencies in general sets of proteins [88]. These general-purpose matrices are not optimal to align accurately sequences with marked compositional biases, such as hydrophobic transmembrane regions found in GPCRs [89]. Amino acid substitution matrices are obtained by the application of statistical methods on sequence alignments of evolutionarily related proteins. In this regard, it is known that the evolutionary selective pressure that governs the conservation and relative mutability of amino acids varies among protein families [90]. Therefore, the application of a standard matrix for the alignment of a determinate protein family could give inaccurate results, particularly if the amino acid composition differs from those used for the matrix construction.

Specific substitutions matrices for certain families of proteins are continuously developing [91-93]. These, in many cases have proven to be more effective than the standard matrices in recognizing evolutionary

relationships between the proteins of interest [94, 95]. To take account of the compositional bias and the modifications in the GPCR MSAs as a consequence of the structural analysis. A curated alignment of more than one thousand membrane spanning sequences of the rhodopsin class from different organisms were used for the construction of an amino acid substitution matrix dedicated to the study of GPCRs. This matrix was built using an approach similar to the one employed for the construction of the BLOSUM series of matrices [77].

4.2.1 Amino acid compositional bias in the class A GPCRs

The average amino acid composition of the TM regions of the class A family was compared with amino acid frequencies derived from other studies (Table 1). As expected, the fraction of hydrophobic residues in the membrane spanning regions of GPCRs is similar to other TM protein matrices (JTTtm and PHDhtm) and is higher than in general proteins (BLOSUM62, and Swiss-Prot). Leucine is the most common occurring residue followed by valine and isoleucine. Nonetheless, there are differences in the amino acid composition of GPCRs. This is the case for charged and polar residues, with the exception of serine and threonine that behave similar in all datasets. The accumulated percentage for the R, K, H, D, E, N, and Q amino acids in the GPCRtm dataset (19.6 %) is in between JTTtm (9.5 %) and PHDhtm (9.9 %) datasets and BLOSUM62 (32.3 %) and Swiss-Prot (33.8 %) datasets. In addition, TM regions of the rhodopsin family are also characterized for a lower frequency of glycine (4.6 %) and a higher frequency of cysteine (3.6 %) residues relative to the other datasets. Given such differences in amino acid composition, we

presume that general protein matrices such as the BLOSUM series and TM-derived protein matrices may not perform accurately in the alignment of the TM regions of GPCRs.

Aminoacid	GPCRtm	JTTtm [96]	PHDhtm [97]	BLOSUM62 [77]	Swiss-Prot [98]
Ala (A)	8.0	10.5	8.8	7.4	8.3
Cys (C)	3.6	2.2	2.6	2.5	1.4
Asp (D)	2.1	0.9	1.4	5.4	5.5
Glu (E)	1.9	1.0	1.0	5.4	6.7
Phe (F)	7.3	7.7	9.3	4.7	3.9
Gly (G)	4.6	7.6	5.7	7.4	7.0
His (H)	2.1	1.7	1.1	2.6	2.3
Ile (I)	8.1	11.9	11.0	6.8	5.9
Lys (K)	3.4	1.1	0.9	5.8	5.8
Leu (L)	14.1	16.3	16.0	9.9	9.7
Met (M)	3.1	3.3	4.1	2.8	2.4
Asn (N)	3.4	1.8	2.2	4.5	4.1
Pro (P)	3.8	2.6	3.2	3.9	4.7
Gln (Q)	2.2	1.4	1.2	3.4	3.9
Arg (R)	4.5	1.6	2.1	5.2	5.5
Ser (S)	6.8	5.7	6.5	5.7	6.6
Thr (T)	5.6	5.2	5.3	5.1	5.3
Val (V)	9.2	11.9	11.0	7.3	6.9
Trp (W)	1.9	2.2	1.9	1.3	1.1
Tr (Y)	4.3	3.2	4.7	3.2	2.9

Table 1. Amino acid composition of substitution matrices and the Swiss-Prot database (%).

4.2.2 Development of the GPCRtm matrix

In Figure 13, the substitution matrix developed for the TM regions of class A GPCRs is shown. Unlike BLOSUM matrices, built from sequence blocks of a variety of biological sources, we employ sequences of only GPCRs that accounts for the compositional bias in this family of receptors. Inspecting the diagonal elements of the matrix, we can estimate the mutability potential of each residue. Hydrophobic residues (V, L, I, A, F) display the highest level of relative mutability (corresponding to low values on the matrix, ≤ 2), whereas charged and polar residues are in general less mutable. Polar serine and threonine residues are special cases, displaying similar values than hydrophobic residues. These two amino acids, unlike other polar or charged residues, do not destabilize TM helices, as their hydrogen bonding potential can be satisfied by interacting with the carbonyl oxygen in the preceding turn of the same helix [99]. In contrast, N, D, R, W and P amino acids display the lowest level of relative mutability (corresponding to high values on the matrix, ≥ 7). All these residues display a high conservation pattern in at least one of TM helices of class A GPCRs [17, 75]: N in TM 1 (present in 98 % of the sequences), D in TM 2 (93 %), R in TM 3 (95 %), W in TM 4 (96 %) and P in TMs 5 (76 %), 6 (98 %) and 7 (93 %). Significantly, the position of these highly conserved amino acids in each helix is the same in the superimposition of the currently available crystal structures [100]. Positively (K, R, and H) and negatively (D, E) charged residues are easily interchangeable with each other. This could be due to a selection pressure to adapt the binding cavity of the TM bundle to the different chemical features of the ligands that, in many cases, display strong electrostatic properties (discussed below).

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

Figure 13. The G protein-coupled receptor transmembrane substitution matrix (GPCRtm).

4.2.2.1 Functional similarities of amino acids in GPCRtm. Comparison with other substitution matrices

The GPCRtm (relative entropy, H= 0.6540) displays intermediate properties between matrices derived from general TM data sets (JTTtm, H= 0.5599 and PHAT, H= 0.5550) and for water-soluble globular proteins (BLOSUM62, H= 0.6979). A comparison of GPCRtm with other matrices is shown in Figure 14. In GPCRtm, charged and polar amino acids (K, R, H, D, E, N and Q) interchange with higher frequencies than in BLOSUM62 and lower than in JTTtm. In general, there is an intermediate performance of GPCRtm between general TM-derived and globular protein matrices

with regard to the majority of charged and polar residues, which suggest a distinctive role of these amino acids in GPCRs.

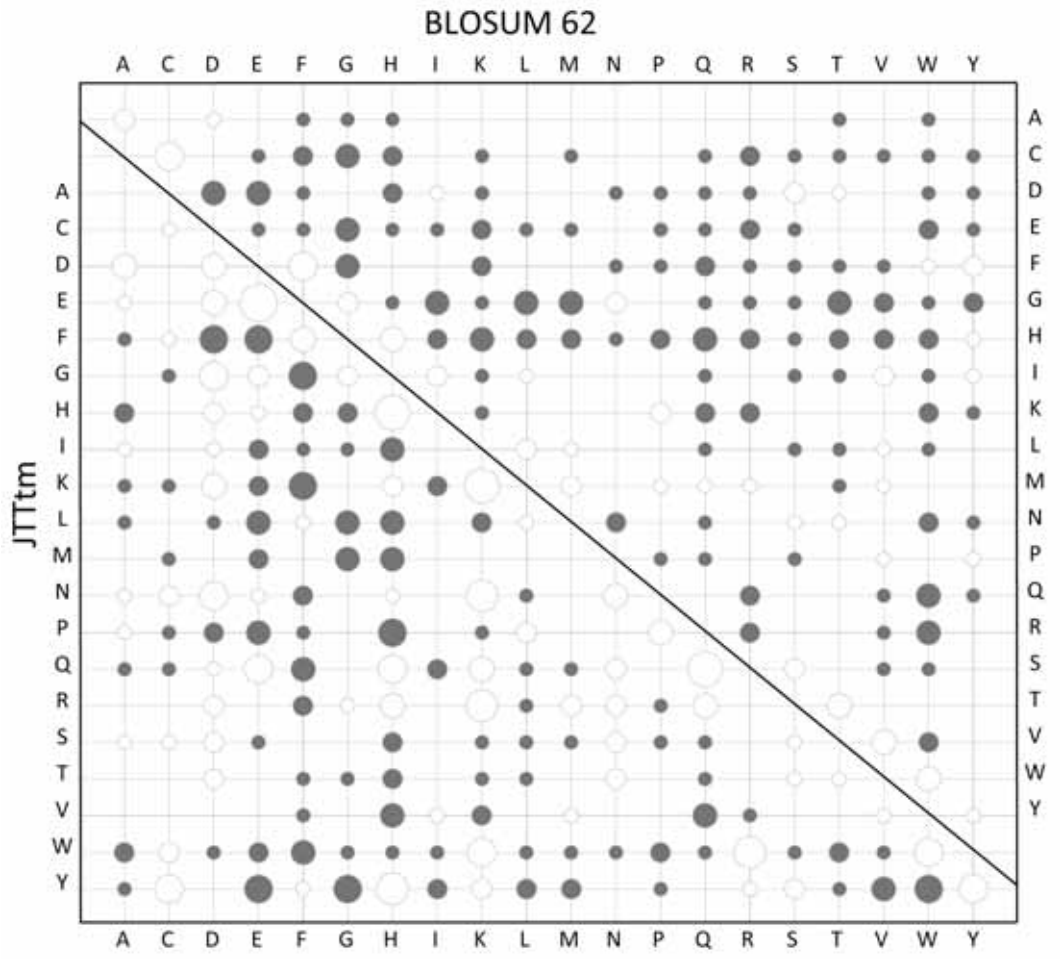


Figure 14. Bubble chart of the difference matrix obtained by subtracting from GPCRtm the JTTtm (lower) and BLOSUM62 (upper) substitution matrices. Positive and negatives values are showed in grey and white circles respectively. Bubbles are scaled according to the absolute value of the difference (numerical values are available in Figure A3 in Appendix).

One of the most important aspects of substitution matrices is amino acid grouping based on their chemical properties. These similarities could be easily visualized through the construction of dendrograms and multi-

dimensional projections to account for the correspondence of amino acids in the matrix (Figure 15). Clearly, clustering of residues in GPCRtm, JTTtm and BLOSUM62 follow similar patterns, but with significant differences. The cluster of hydrophobic residues (I, V, L, M) is closer to the cluster of small amino acids (A, S, T) in all cases. However, GPCRtm differs from other matrices in that phenylalanine is grouped with hydrophobic amino acids (the I, V, L, M, F cluster), whereas in BLOSUM62 is grouped with the aromatic tyrosine and in JTTtm with cysteine. Similarly, glycine is clustered together with the other small amino acids (A, S, T), in contrast to other matrices in which is grouped alone. Histidine clusters with positively charged and polar amino acids in GPCRtm and JTTtm, in contrast to BLOSUM62. This residue is grouped with glutamine in GPCRtm and JTTtm, probably due to its hydrogen bond donor/acceptor properties, whereas in BLOSUM62 is grouped with phenylalanine and tyrosine probably due to its aromaticity. GPCRtm clusters tryptophan and tyrosine together, preserving aromaticity and hydrogen bond capacity, whereas in the other matrices tryptophan is unaccompanied. The negatively charged aspartate and glutamate form one group in GPCRtm and JTTtm, while in BLOSUM62 aspartate pairs with asparagine and glutamate with glutamine. In this regard, positive (K, R) and negative (D, E) residues are grouped at closer distance in BLOSUM62. In contrast, positive and negative residues are distant in GPCRtm and JTTtm. Interestingly, the distance between branches containing opposite charged residues in GPCRtm is larger than in JTTtm, suggesting than the sign of the charge is apparently more conserved in the GPCR TM sequences than in a general set of TM proteins.

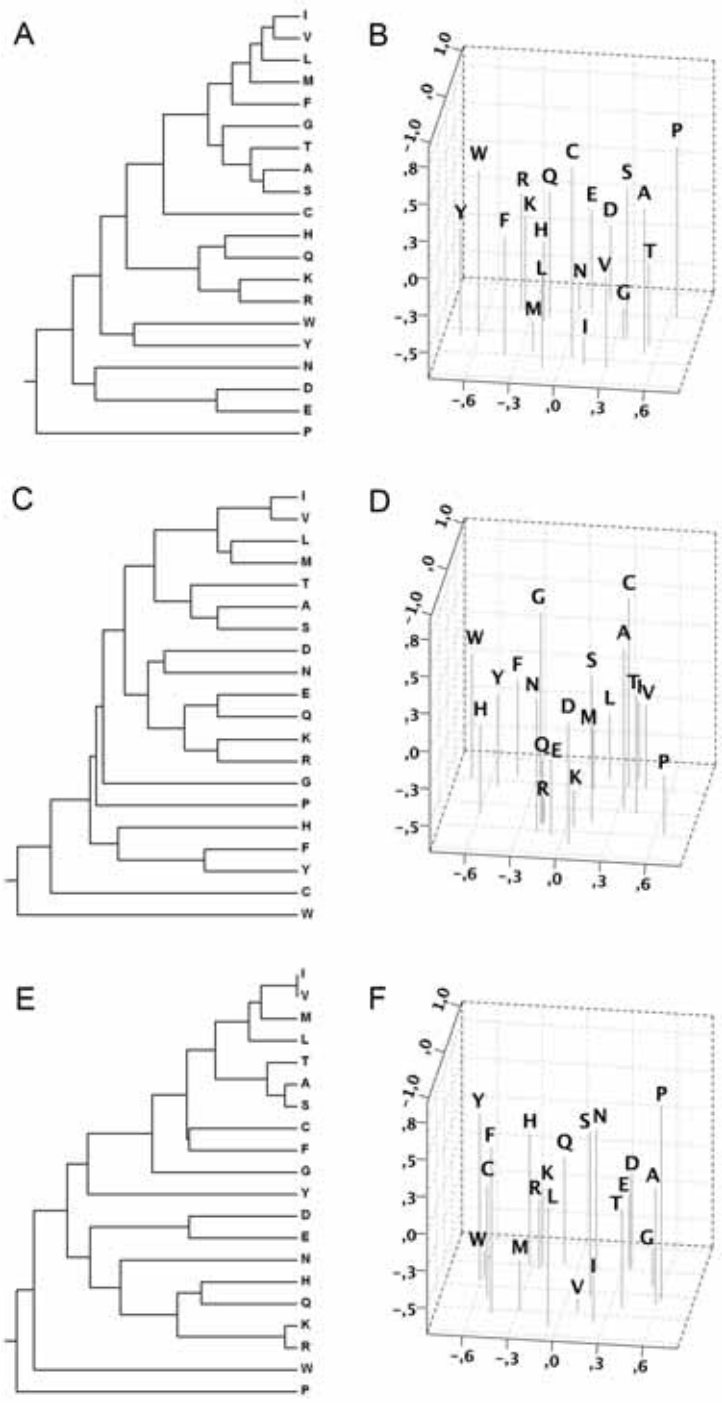


Figure 15. Unweight pair groups mean analysis dendrograms (left) and multi-dimensional scaling projections (right) of the GPCRtm a, b; the JTTtm c, d and the BLOSUM62 e, f substitution matrices

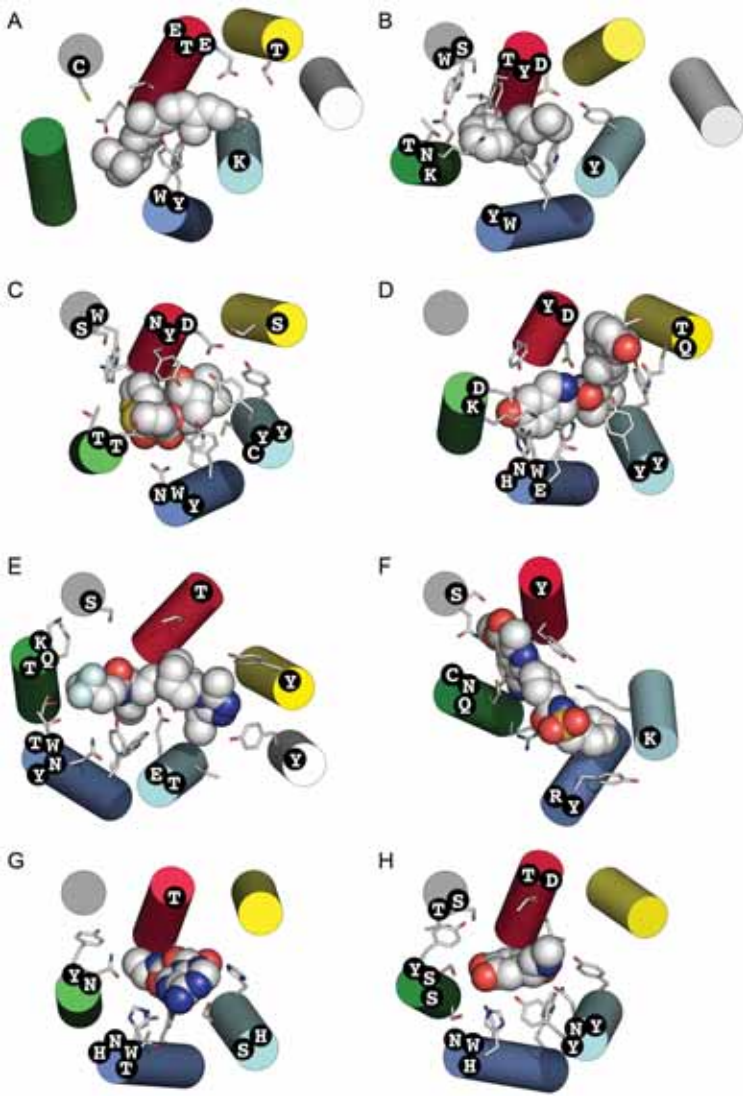


Figure 16. Diversity of ligand binding interactions involved polar and charge residues in the TM region of the rhodopsin family of GPCRs. The crystal structures corresponding to: **a** Rhodopsin (PDBid: 1U19), **b** Histamine HRH1 (3RZE), **c** Muscarinic MC3R (4DAJ), **d** Opioid OPRK (4DJH), **e** Chemokine CCR5 (4MBS), **f** Purinergic P2Y12 (4NTJ), **g** Adenosine AA2AR (2YDV) and **h** Adrenergic ADRB2 (4LDO). Polar and charged residues of the receptors at 4 Å distance of ligands (in vdW spheres) are displayed as sticks and named in the corresponding helices (circular labels). The color code of the helices is: TM1 (light grey), TM2 (yellow), TM3 (red), TM4 (grey), TM5 (green), TM6 (darkblue) and TM7 (cyan). All structures are oriented with the TM4 perpendicular to the plane.

Overall, the results show that GPCRtm prioritized the reactivity properties of the amino acids over their bulkiness. In this way, hydrophobic residues (including phenylalanine), which are key in TM regions, are clustered together. On the other side, the hydrogen bond capacity and electronic properties of the amino acids tend to be maintained in GPCR sequences. Thus, the H/Q, K/R, E/D/N and W/Y pairs together. These residues contribute largely to the diversity of interactions between ligands and the 7TM bundle as can be observed in the 3D structures of ligand-receptor complexes in some members of the rhodopsin family on Figure 16. In this respect, GPCRs are distinguished from most TM proteins for their ability to interact with a diverse variety of chemical entities.

4.2.2.2 Evaluation of the GPCRtm matrix

The GPCRtm was tested on sequence similarity searches and pairwise alignments. The results of GPCRtm were compared with commonly used amino acid exchange matrices, the JTTtm and PHAT transmembrane matrices and the general-purpose BLOSUM45 and BLOSUM62 matrices. At high sequence identity values (above the twilight zone) all matrices behave similarly. However, as sequence identity falls below 40 %, significant differences emerged. Table 2 shows a comparison among the different substitution models in BLASTP database searches for one hundred GPCR queries against the PDB database [101]. As observed in the table, the GPCRtm matrix performs better than other matrices. The second best performance was achieved by the closely related PHAT matrix, followed by the BLOSUM62, BLOSUM45 and JTTtm matrices, respectively.

Test matrix	No. of queries GPCRtm better	No. of queries GPCRtm worst	No. of queries GPCRtm the same	*p*-value
JTTtm	21	0	79	<0.001**
PHAT	8	0	92	0.008*
BLOSUM62	9	1	90	0.021*
BLOSUM45	12	1	87	0.003*

Table 2. Comparative analysis of the GPCRtm performance regarding general-purpose substitution matrices in BLASTP searches of one hundred GPCR protein queries against the PDB database. *p-values were calculated by McNemar's test (* significant differences at α = 0.05, ** significant differences at α = 0.001)*

Criteria for the performance evaluation were based on the recognition of the closest homologue with known three-dimensional structure for a determinate query, according to the well-established GPCR classification systems [10, 11]. Table 3 illustrates an example for the adrenergic receptor (ADR) subfamily of GPCRs. ADRs interact with the endogenous catecholamines adrenaline and noradrenaline and constitute essential regulators of central and peripheral metabolic functions [102]. These receptors are classified into three main groups: the α1-, α2- and β-adrenoceptors. Only two members (β1- or ADRB1 and β2- or ADRB2) have been solved by X-Ray crystallography, constituting the reference structures for the adrenoceptors subfamily [103]. According to the results shown in Table 3, the GPCRtm matrix performs better than general-purpose matrices in BLASTP searches, resolving a receptor of the same subfamily (ADRB1 or ADRB2) as a first hit for searches involved the nine ADR subtypes as queries. On the other hand, in some instances (at lower identities) the standard matrices deliver as best hit a receptor of a different GPCR subfamily.

Query Receptor	GPCRtm				JTT				PHAT				BLOSUM62				BLOSUM45			
	First Hit	Id. (%)	Score (bits)	E-value	First Hit	Id. (%)	Score (bits)	E-value	First Hit	Id. (%)	Score (bits)	E-value	First Hit	Id. (%)	Score (bits)	E-value	First Hit	Id. (%)	Score (bits)	E-value
ADA1A	ADRB1 (2VT4)	34	140.9	$1.0e^{-55}$	ACM2 (4MQS)	30	140.0	$2.3e^{-55}$	ADRB2 (3KJ6)	36	226.3	$8.8e^{-62}$	5HT1B (4IAQ)	35	133.9	$7.5e^{-57}$	5HT1B (4IAQ)	35	140.9	$5.0e^{-59}$
ADA1B	ADRB1 (2VT4)	35	139.4	$3.5e^{-56}$	ADRB1 (2VT4)	34	133.9	$4.8e^{-59}$	ADRB2 (3KJ6)	34	215.3	$3.5e^{-58}$	ADRB1 (2Y00)	35	139.3	$1.1e^{-55}$	ADRB1 (2Y00)	35	141.2	$4.6e^{-56}$
ADA1D	ADRB1 (2VT4)	35	154.3	$3.3e^{-58}$	ADRB1 (2VT4)	35	133.9	$4.8e^{-56}$	ADRB1 (2VT4)	36	156.6	$1.6e^{-60}$	ADRB1 (2VT4)	36	150.4	$1.0e^{-57}$	ADRB1 (2VT4)	36	150.3	$9.1e^{-58}$
ADA2A	ADRB1 (2VT4)	40	130.4	$6.0e^{-50}$	ACM2 (4MQS)	26	126.8	$4.2e^{-49}$	ADRB2 (3D4S)	29	163.5	$8.8e^{-53}$	5HT1B (4IAR)	39	144.3	$5.3e^{-56}$	5HT1B (4IAQ)	41	147.0	$1.0e^{-57}$
ADA2B	ADRB2 (2R4S)	30	128.9	$8.1e^{-47}$	DRD3 (3PBL)	30	195.8	$4.5e^{-53}$	DRD3 (3PBL)	31	215.9	$1.3e^{-56}$	5HT1B (4IAR)	36	135.4	$8.0e^{-55}$	5HT1B (4IAR)	36	142.9	$2.1e^{-58}$
ADA2C	ADRB1 (2VT4)	35	118.5	$1.2e^{-49}$	ADRB1 (2VT4)	34	108.1	$5.1e^{-49}$	ADRB1 (2VT4)	37	130.2	$2.8e^{-53}$	5HT1B (4IAR)	35	134.7	$6.8e^{-56}$	5HT1B (4IAR)	34	139.4	$1.7e^{-58}$
ADRB1	ADRB1 (2Y00)	77	308.0	$2.6e^{-135}$	ADRB1 (3KJ6)	57	241.1	$1.5e^{-99}$	ADRB1 (2Y00)	77	338.1	$1.1e^{-148}$	ADRB1 (2Y00)	77	317.4	$2.5e^{-130}$	ADRB1 (2Y00)	77	319.6	$2.9e^{-132}$
ADRB2	ADRB2 (2R4R)	99	696.1	$5.1e^{-204}$	ADRB2 (2R4R)	99	624.7	$3.4e^{-182}$	ADRB2 (2R4R)	99	791.0	$2.3e^{-232}$	ADRB2 (2R4R)	99	686.4	$1.2e^{-200}$	ADRB2 (2R4R)	99	678.6	$2.1e^{-198}$
ADRB3	ADRB1 (2Y00)	53	201.6	$2.3e^{-86}$	ADRB1 (2Y00)	53	185.2	$3.7e^{-78}$	ADRB1 (2Y00)	56	220.0	$2.4e^{-95}$	ADRB1 (2Y00)	56	215.7	$6.6e^{-89}$	ADRB1 (2Y00)	56	217.6	$2.4e^{-90}$

Table 3. Results of BLASTP database searches using nine human adrenergic receptor subtypes as queries against the Protein Data Bank. The table displays only the first hit (lower E-value) of each search (IUPAC name of the receptor and PDBid code in parenthesis) followed by the sequence identity values in the aligned regions and the corresponding bit scores for the GPCRtm and general substitution matrices.

One of the best ways to test alignment accuracies is to compare the results with structure-based information derived from three-dimensional structural data. In this regard, the GPCRtm matrix was tested on pairwise sequence alignments of class A GPCR whose structures are known.

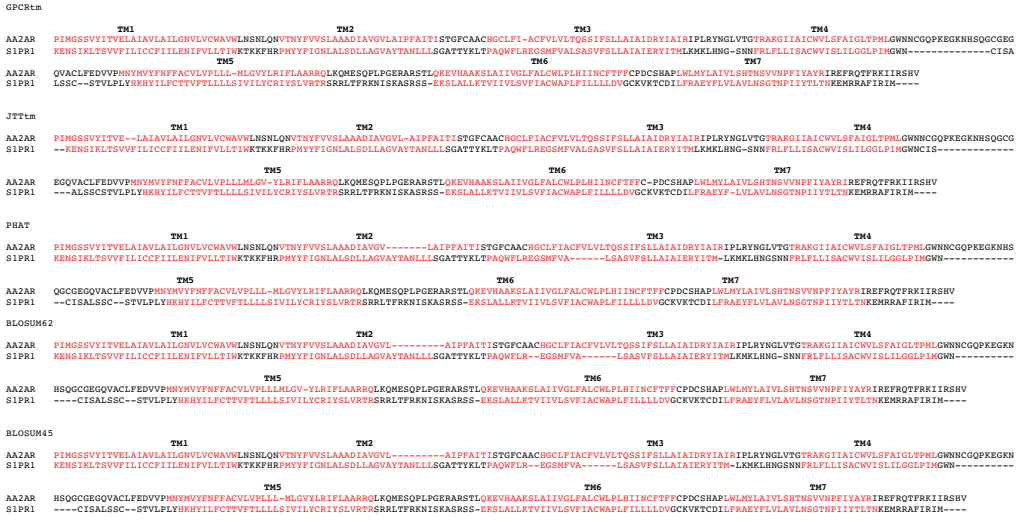


Figure 17. Example of pairwise alignments of the adenosine AA2AR and sphingosine-1-phosphate S1PR1 amino acid sequences using: GPCRtm (**a**), JTTtm (**b**), PHAT (**c**), BLOSUM62 (**d**) and BLOSUM45 (**e**) substitution matrices. Transmembrane regions TM 1 to 7 appear outlined in red according on the crystallographic 3D structural data for each receptor (PDBid: 3EML and 3V2Y). Pairwise sequence alignments were done with MAFFT program [80].

Figure 17 shows the result of the alignment between the adenosine A2A receptor (AA2AR) and sphingosine-1-phosphate receptor 1 (S1PR1) using dif- ferent substitution matrices. Both receptors are members of the MECA receptor cluster of the rhodopsin family [10] with known three-dimensional structures [104, 105]. In this example, the resulting alignments denote the accuracy of the GPCRtm to correctly align the TM helices of both receptors, whereas generalized matrices fails to correctly align some of

the TM regions. According to these results, the GPCRtm matrix improve the detection of closest homologues and produce accurate alignments in the TM regions of GPCRs, even at low sequence identities. This is particularly relevant in the development of homology models for structure-based drug discovery, which in many cases are generated from low sequence identity alignments due to the limited number of GPCRs crystallographic structural templates [24].

4.2.3 Conclusions and perspectives

The developed GPCRtm is evolutionary consistent with amino acid frequencies and actual changes occurring within the GPCR protein family. Analysis of the matrix reveals the differences between GPCRs and other membrane proteins and proteins in general. This is evidenced by distinctive frequencies of polar and charged residues and a prevalence of reactivity over size in the contribution of the conservation pattern. These observations stress the relatively high importance of charged and polar amino acids in this family of receptors with regard to other membrane proteins, possibly due to their versatility in ligand interaction. In this regard, this matrix could assist in evolutionary studies, improving the classification and increasing the accuracy of phylogenetic reconstruction for members of this family of membrane receptors. The GPCRtm, besides important from a theoretical point of view, has been successfully used in sequence alignment and database search of class A GPCRs.

4.3 Clustering of class A GPCRs using structural derived information

4.3.1 Clustering based on phylogenetic reconstruction of TM domains

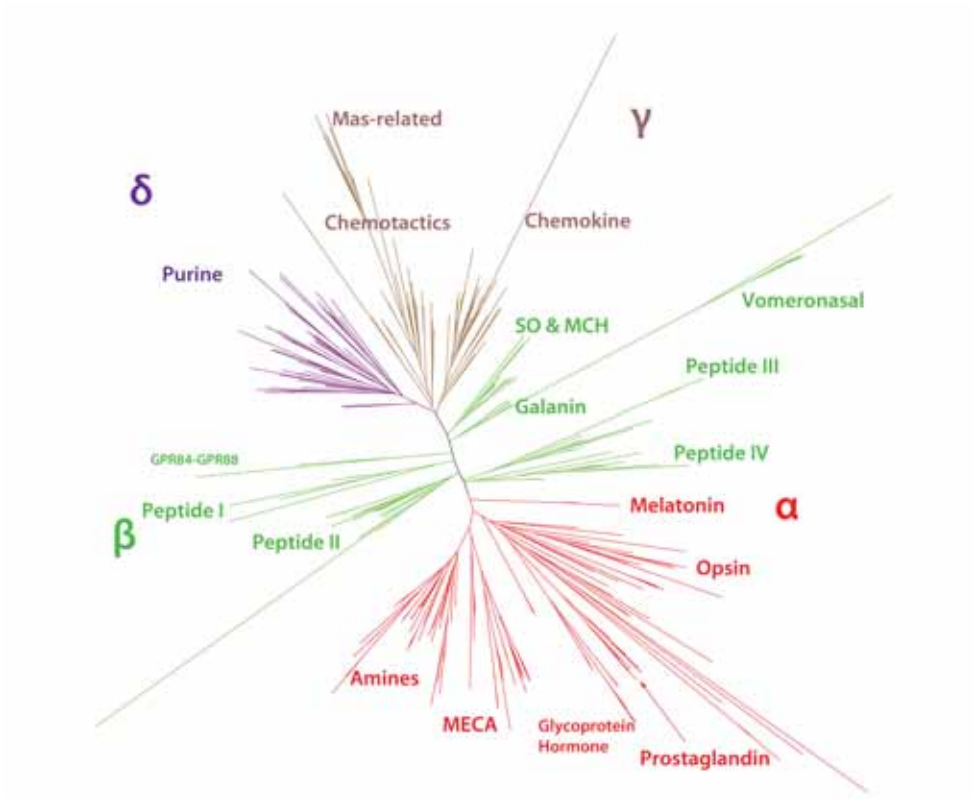


Figure 18. Schematic representation of the phylogenetic tree of the Human Class A GPCRs (see methods). Receptors clusters are colored according to GRAFS classification wherever it is possible, and each cluster is represented with different colors: α-cluster is colored in red, ß in green, γ in brown and δ in purple. Names from each cluster (Table 1) are included. The tree dendrogram supported with aLRT values is available in Figure A3 in Appendix.

A MSA of the seven transmembrane domains of the 292 sequences human class A GPCR dataset, updated with the structural derived

information, was used to generate an unrooted phylogenetic tree using the Maximum Likelihood method.

Clusters	Receptors sub-families
Amines	5-HTs, ACMs, ADRs, DRDs, HRHs and TAARs
MECA	MCRs, LPARs (1-3), S1PRs, CNRs, AARs, GP119, GPR12, GPR3 and GPR6
Prostaglandin	P2Rs except PD2R2
Opsin	OPRs, OPSB, OPSG, OPSD, OPSR, OPSX
Melatonin	MTR1s
Glycoprotein Hormone	FSHR, LSHR, TSHR, RXFPs and LGRs
Orphans in α-cluster and GPBAR	GPBAR, GP135, GPR52, GPR21, GPR22, GPR27, GP173, GPR85, GP176, GP101, GP161, GPR61, GPR62, GPR75, GPR26, GPR78, GP160, GP149, GP153, GP162, GPR45 and GPR63
Peptide I	NMBR, GRPR, BRS3, EDNRs, ETBR2, GPR37 and GP151
Peptide II	NKRs, NPYRs, NPFFs, PRLHR, OXRs, PKRs, QRFP, GP148 and GPR83
Peptide III	MTLR, GHSR, TRFR, NTRs, NMURs, GP139, GP142 and GPR39
Peptide IV	GNRHR, OXYR, VRs, NPSR1, GASR, CCKAR, GP150 and GPR19
GPR84 and GPR88	GPR84 and GPR88
SO & MCH	SSRs, OPRs, NPBWs, MCHRs and UR2R
Galanin and Vomeronasal	GALRs, KISSR and VN1Rs
Chemokine	CCRs, CXCRs, CX3C1, CCRL2, XCR1, ACKRs, BKRBs, AGTRs and GP182
Chemotactics and Mas-related	GPER1, FPRs, RL3Rs, CML1, C3AR, C5ARs, LT4Rs, APJ, PD2R2, MAS, MAS1L, MRGs, GP152, GPR32, GP146, GPR33, GPR1, GPR15 and GPR25
Purine	CLTRs, OGR1, PSYR, PTAFR, P2RYs, P2Ys, LPARs (4-6), SUCR1, OXGR1, HCARs, OXER1, PARs, FFARs, GPR4, GP132, GPR35, GPR55, GP174, GP183, GPR31, GPR20, GPR42, GPRR17, GP141, GPR82, GPR87, GP171, GPR34 and GPR18

Table 5. Clusters of the human class A GPCRs. Family names are obtained from NC-IUPHAR [40] and branch clusters names are adapted from GRAFS classification [10].

Figure 18 shows the resulting topology. The resulting distribution of the taxa shows that receptors are clustered in 15 main sub-families. Table 5 reports a summary of the classification we obtained with the phylogenetic analysis on TM domain MSAs. For an easier comparison, receptors are tagged and grouped with the same nomenclature as GRAFS classification system wherever it is possible. Despite a general similarity with other class A GPCRs' classifications, there are some important differences. The new classification shows the α-cluster (red in the figure 18), which, similarly to GRAFS, include amines, MECA, prostaglandin, opsin and melatonin receptors. However, in contrast to previous classifications, this cluster also includes the glycoprotein hormone receptors (GPBAR) and several orphan receptors. Interestingly, sequence analysis of these receptors reveals two common patterns: 95% of the sequences share either a proline at position 4.60 or a double P4.59-P4.60 motif (only observed in 3 receptors outside the α-cluster), suggesting a common evolutionary trace among them.

The central cluster (green in Figure 18) is composed by seven branches. Peptide receptors clusterize in four of them (Peptide I to IV clusters), in contrast to then main ß-cluster in GRAFS classification. The new additions include a branch tagged as "SO & MCH", composed by somatostatin, opioid, melanin-concentrating hormone, neuropeptide B/W and urotensin receptors, a branch that clusterize galanin and vomeronasal receptors and finally, two orphan receptors (GPR84 and GPR88) (Figure 18). These new additions are supported by the evidence that most of aforementioned receptors also bind peptides (59 out of 71 receptors).

Chemokines and chemotactics receptors are grouped in two branches tagged in the γ–cluster (ochre in Figure 18). Mas-related receptors (grouped with purine receptors in GRAFS system) clusterize also with chemotactics. All these receptors shared an E/DRC motif, same as the complement peptide and formylpeptide receptors.

The δ branch (purple in Figure 18) is composed by purine and several orphan receptors. A close analysis of the sequences of members in this cluster reveals a shared CFXP motif in TM6 and DPXXY's in TM7, in contrast to the common class A pattern CWXP / NPXXY. These two motifs are close spatially located, which suggest a specific co-evolution. In addition, receptors in γ– and δ–cluster (except mas-related receptors) share a gap at the end of TM2. This feature is also observed in somatostatin and opioid receptors, both closely located to chemokine and purine receptors. The presence of a gap in these closely related receptors endorses the hypothesis of an evolutionary indel event, as also suggested by previous studies [44].

4.3.2 Clustering based on ligand binding pocket residues

Because of the great pharmacological interest on the discovery of new GPCRs active compounds, we attempted to investigate the similarity of Human class A GPCRs solely based on residues involved in ligand-binding. Using the new information obtained for the recent available crystal structures, the definition of a generic ligand-binding site for class A GPCRs could be improved in comparison with previous studies [55, 57].

4.3.2.1 Definition of a generic ligand binding for class A GPCRs

Based on class A GPCR structural alignments, 40 positions in close contact with co-crystalized ligands were selected (Table 6). These positions were located in the TM region MSAs, composing a generic ligand-binding site dataset. As a consequence of the inclusion of sequence gaps in the TM2 and TM5 alignments (see section 4.1) some of the positions were shifted in order to fit the structural alignment.

The structural analysis also revealed that in most of the GPCR structures (Table A1 in Appendix), the second amino acid following the conserved cysteine in ECL2 (position 45.52) points to the binding site cavity. Taking into account the importance of the ECL2 in the ligand interactions [20], this position was included in the ligand-binding site definition (Figure A7 in Appendix).

TM1	TM2	TM3	TM4	ECL2*	TM5	TM6	TM7
1.35	2.53	3.28	4.56	45.52	5.37	6.48	7.32
1.39	2.57	3.29	4.57		5.38	6.51	7.35
	2.60	3.30	4.60		5.42	6.52	7.36
	2.61	3.32			5.43	6.54	7.38
	2.64	3.33			5.46	6.55	7.39
	2.65	3.36			5.47	6.58	7.40
		3.37					7.42
		3.40					7.43

Table 6. The 40 positions selected for the binding site definition; positions are numbered following Ballesteros & Weinstein nomenclature. *GPCRDb numbering scheme [76].

4.3.2.2 Improvements of the generic ligand binding site definition

The new released GPCRs structures had increased the knowledge of ligand-receptor interaction complexes and revealed backbone irregularities on TM regions affecting the structural alignment of GPCRs. Both novelties alter the definition of the ligand-binding positions compared with previous binding pocket data sets [55, 57]. As examples, the inclusion of position 2.60 on the generic class A GPCR binding site definition enables the selection of a glutamic residue important for ligand binding in melanocortin receptors [106]. Moreover, indels added on TM2 in the OPRX receptor modify the true location of an asparagine residue to position 2.61, which has been proven critical for in complex with a peptide mimetic (Figure 19 A). Last but not least, the inclusion of position 45.52 incorporate features of the extracellular loops that are known to modulate ligand binding in GPCRs [20]. As example, a phenylalanine residue in the ECL2 of AA2AR is key for the interaction with ligands (Figure 19 B).

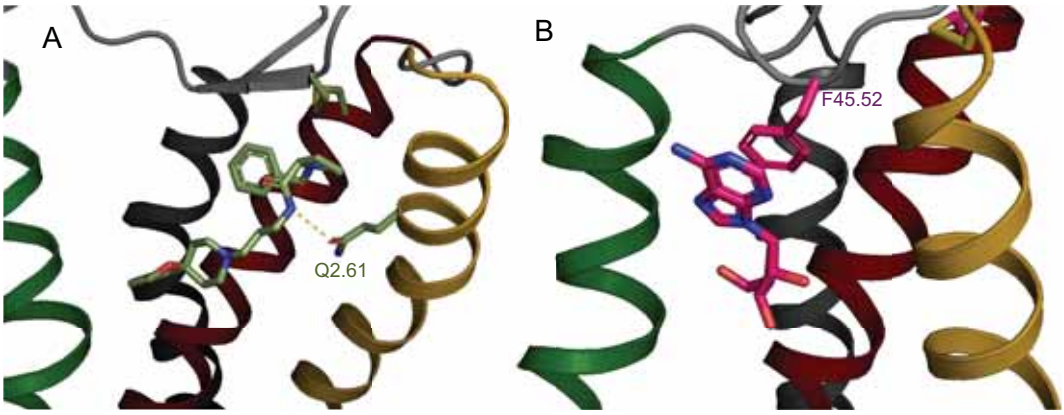


Figure 19. Crystal structure of the OPRX (PDB: 4EA3) (a) and AA2AR (PDB: 2YDO) (b). The hydrogen bond between Q107$^{2.61}$ and the amide nitrogen of the ligand and the π-stacking interaction between the exocyclic adenosine and F168$^{45.52}$ are highlighted.

4.3.2.3 Analysis of the ligand-binding site residues

Analysis of the MSAs corresponding to binding site amino acids shows some positions have similar residues in most receptors (Table 7), thus being likely part of a shared mechanism in ligand binding and/or activation mechanism, as conserved amino acids at specific location usually correspond to conserved function. For example, F5.47 and W6.48 are observed in more than 60% of the receptors. Indeed, as mentioned earlier in the introduction, they are part of one of the molecular switches. The residue at position 3.40 is hydrophobic in 80,8% of the receptors and structural data also show its importance in the activation mechanism [37].

Conserved positions	Predominant amino acids	Percentage (%)
3.40	I, L, V	75.3
4.56	I, L, V	56.1
4.57	G, A, S	63.7
5.47	F	64.7
6.48	W	67.1
6.51	F / Y	65.7
7.42	G, A, S	73.0

Table 7. Conserved residues (with more than 50%) on the binding site database. Isoleucine (I), Leucine (L) and Valine are considered as hydrophobic residues. Glycine (G), Alanine (A) and Serine (S) are considered as small residues.

Binding site analysis also unveiled that receptors with similar ligands conserve key residues. Prostaglandin receptors share an arginine at position 7.40 and experimental evidences suggest its interaction with the carboxylate moiety of the prostanoid ligands [107]. Opsin receptors share

a lysine at position 7.43 and a negative charge at position 3.28, both residues are necessary on the covalent binding with the retinal ligand [56]. Lysophospholipid receptors share a R3.28 and a Q/E3.29, both positions bind the charged head of their endogenous ligands [105, 108]. Adenosine receptors share a E1.39 and M5.38. Methionine residue can interact with aromatic moieties of the ligands while glutamic acid likely interacts with internal waters that stabilize the ligand binding [109, 110].

Structural and experimental studies on GPCRs reveals that compounds usually interact with different receptors if similar residues are located at the same position. For example, all amine receptors contain an aspartic amino acid at position 3.32 that binds the amine moieties of both agonist and antagonist ligands. They also share the highly conserved Y7.43 that stabilizes the ionic interaction into aspartic residues and amine moieties [12, 111-116]. The same pattern is found in opioid receptors although they are no closely related [117-120]. Another relationship between receptors on different pharmacological families is observed between chemokine and chemotactic receptors. The conserved W2.61 binds aromatic moieties of the ligand in chemokine [121-123] and in angiotensin receptors [124]. These types of non-standard relationships motivate a more complete comparison between class A GPCRs.

4.3.2.4 Clustering by similarity matrix of the binding site residues

The large number of GPCRs prevents a receptor-per-receptor comparison between every 40-amino-acid-long sequence. On the same way as TM sequence database, a phylogenetic tree was proposed to infer the

relationship between ligand-binding site data. Unluckily, obtained results were not robust. Bootstrap methodology typically assesses the reliability of phylogenetic trees producing multiple versions of the original alignment by extraction and duplication of columns. Unfortunately, this methodology us unsuited for databases with small number of columns (bootstrapping sampling typically lost 50% of the data for each alignment replicate) [57]. In order to perform a suitable and valuable analysis of the binding site, we thus created a similarity matrix of the dataset. Similarity values are calculated using relative mutability values of the GPCRtm substitution matrix. Then, lower and negative similarity values for pairwise receptor comparisons mean that residues on the same position are hardly interchangeable. Opposite, highly positive values mean that residues on the same position are important and conserved. Similarity values were converted in distances among receptors (see methods) and represented it by means of heat maps (Figure 20).

In order to compare the similarities values between receptors binding sites, a six color-coded scheme (darkblue for values below -0.25, white for negative values close to 0, light blue for weak positive values, yellow for medium similarity values, orange for strong similarity values and red for very strong similarity values) was used. A similarity matrix and consequent distance-heat map was also produced using the full human TM dataset, which results are shown in Figure A6 in Appendix. The latter heat map present an average values of similarity strongly higher than the binding site (much more yellowish colors indicating positive values), which confirms our hypothesis that binding site have to present a chemistry (residues involved in it) much more diverse and selective than the full TM

sequences, which share the activation mechanism (millions of ligand triggers de coupling of solely 16 G-protein), as shown by the conserved molecular switches. Indeed calculated binding site similarity values are positive in only about 50% of the matrix, versus the 97% of positive similarity values observed in full TM dataset.

Figure 20 shows two main blocks (1 and 2 highlighted with box colored in purple and red) where receptors on the clusters are distantly involved. Our results show different agglomeration of the binding site cluster with respect of the full TM database. Indeed, according to binding site's similarity, aminergic receptors cluster with opioid and somatostatin receptors. Amine and opioid receptors share a common specific moiety on the ligand and a common residue on the receptor (D3.32 mentioned in section 4.4.1.2) while histamine and somatostatin receptors are also related since the same compound (astemizole) regulates both [125]. Similar relations were found in receptors, which share key residues in their binding site as melanin-concentrating hormone, neuropeptide B/W and urotensin-2 receptors, and GPR26, GPR78 orphan receptors. Similarly, receptors clustered in block 2, as chemokine, chemotactic and purine receptors (highlighted with a purple colour box) all share the presence of key aromatic residues on some position in the binding site. As example Y1.39 and Y3.37 are present in more than 60% on this cluster. This information could be helpful on the development of drugs for these receptors.

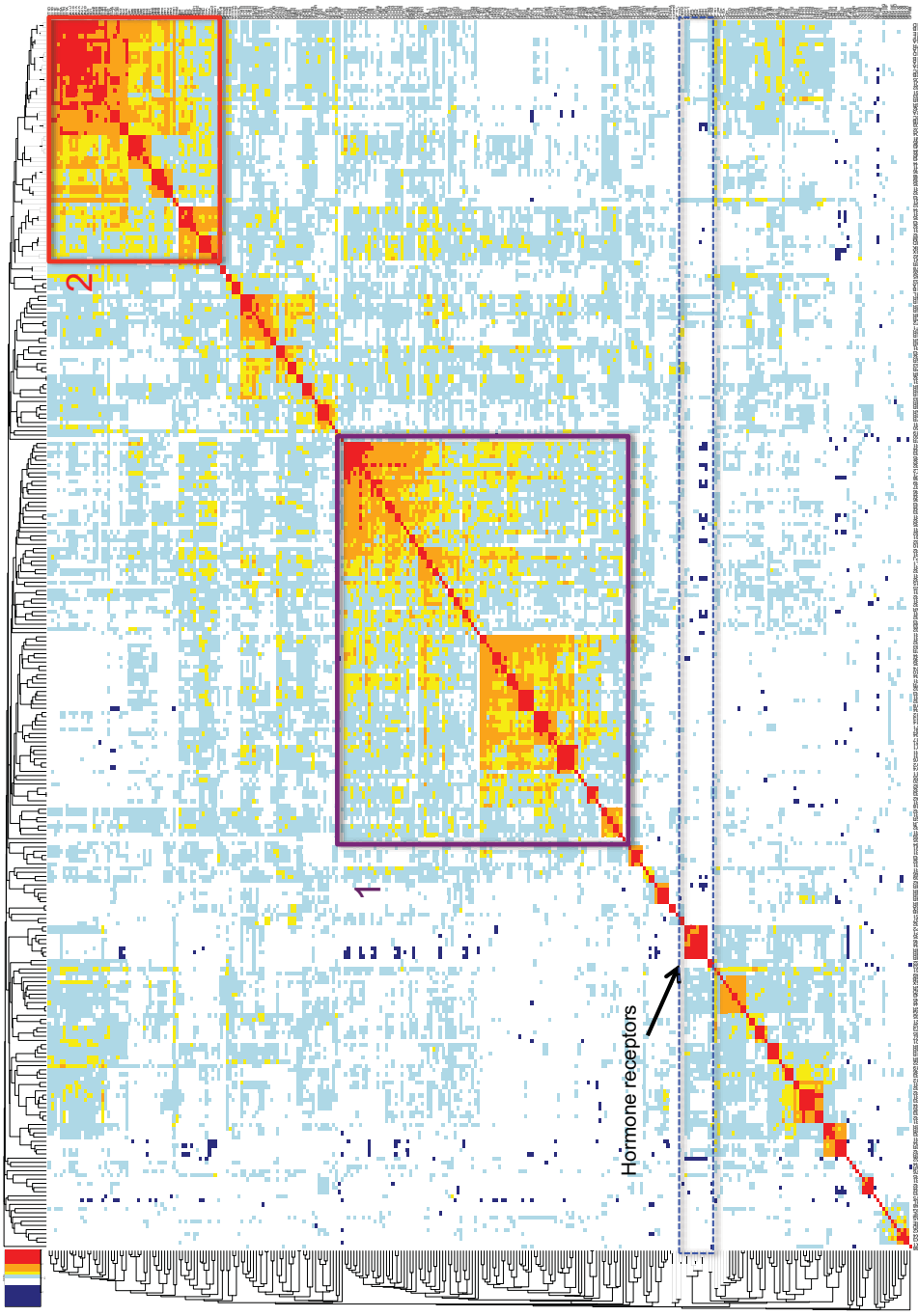


Figure 20. Heat map of the binding site subset made by the similarity matrix. Strong dissimilarities (below the value -0.25) are colored in darkblue, weak dissimilarities (between -0.25 and 0) are colored in white, weak similarities (between 0 and 0.15) in light blue, medium similarities (between 0.15 and 0.25) in yellow, strong similarities (between 0.25 and 0.5) in orange and very strong similarities (between 0.5 and 1) in red.

Distance-based heatmap, calculated by similarity values, are also useful to unveil key divergences in receptors' binding sites of the same block. In amine receptors (detail in Figure 21), the higher ligand binding similarity is observed in the sub-cluster of the adrenergic, dopaminergic and serotoninergic receptors (see sub-block 2a of the figure 21). This sub-block is characterized by the presence of block-specific SSS or STA motif at positions 5.42, 5.43 and 5.46 and aromatic residues at position 6.51 and 6.52 [126]. Another significant detail involves hormone peptide receptors, which show negative distances with respect to most other receptors (highlighted in a dark blue dotted box). The significant dissimilarities could be related to the lack of highly conserved residues as F5.47 and W6.48.

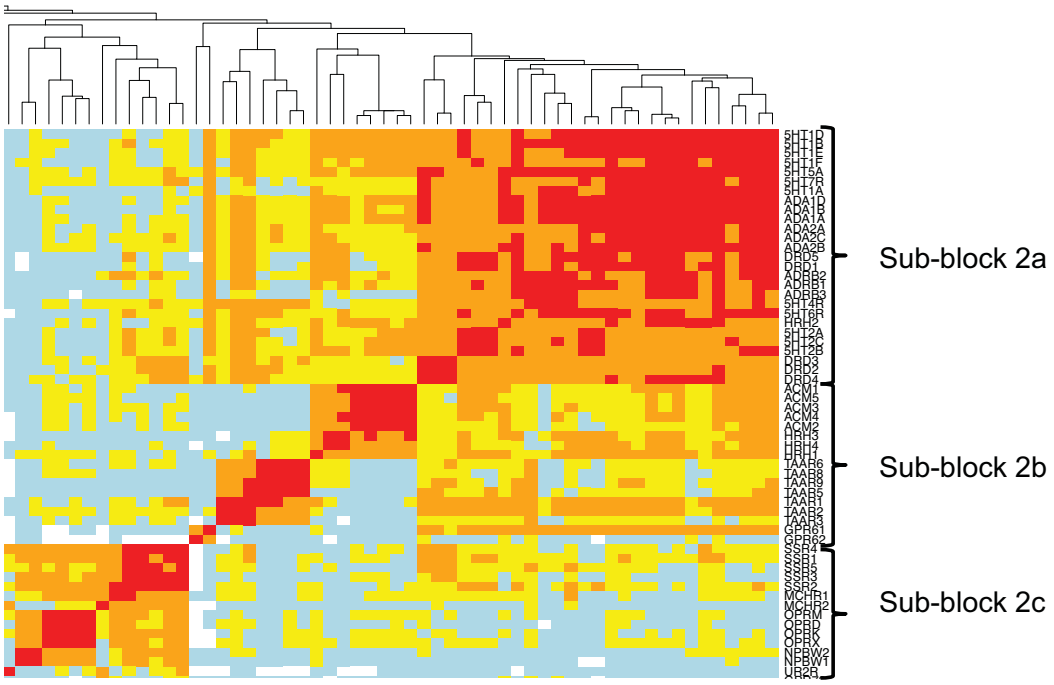


Figure 21. The heatmap fragment corresponding to amine, Somatostatin and opioid receptors (sub-block 2, colored in red in Figure 20)

4.3.3 Conclusion and Perspectives

The recent advances in GPCR crystallization resulted in a larger increase in the number of crystal structures released in the last years. Such growing pool of structures provided novel structural information that needs to be incorporated in sequence alignment. Using this information we curated a set of GPCRs MSAs according to structural features observed the TM domains. These alignments were employed to develop a phylogenetic classification for class A GPCRs, which was compared with other classification systems. As a result of this study we observe changes on the topology of the generated trees with regard to other classifications. These differences were associated in the majority of cases with the presence of conserved sequence and structural motifs. In view of the obtained results, we conclude that the proposed methodology improves the current GPCR classification and is very useful in orphan receptor research. On the other hand, the chemogenomic analysis generated for the ligand-binding site, taking advantage of the new released GPCR crystals, allowed the construction of a distance matrix among binding sites and the subsequent generation of heatmaps, which enable detailed comparison among receptors. Noteworthy, the binding site similarity values reflect strong relations between some GPCRs that were not evident by standard sequence comparison methods. These new predicted associations could have impact in drug discovery studies.

4.4 Development of computational tools for the study of class A GPCRs

4.4.1 GPCR Browser

Due to the important amount of information about Class A GPCRs their classification and sequence analysis is not ease to handle. Despite various bioinformatics web servers are available for the study of GPCRs, most of these tools generally require an important knowledge of the GPCRs in order to be effectively used. In order to make available in an easy interface all the information developed in this project we developed a new, easy and publicly available tool: “GPCR browser” (http://lmc.uab.cat/gpcr-browser/).

4.4.1.1 Content and utility

In this web server, full sequence and binding site classifications are stored and a web interface enable users to find the related information for a query receptor. Using as input the uniprot protein code or the fasta sequence of a GPCR as input, the web application displays a detailed information about it and produce several tables and analysis graphs. The displayed information is intended to be showed on a user-friendly interface (Figure 21-25).

Figure 22. GPCR Browser main page.

Once introduced the input, the server generates four different sections:

i) *TM sequence classification*: It shows the closest resembling branch of protein according to our pre-calculated phylogenetic tree. This is very useful to observe the closest similar proteins to users' input. The phylogenetic tree is browsable, adjusting the number of branch the user wants to show.

ii) *Ligand-binding site representation*: This figure shows the binding site of the users' query and their relative position in Ballesteros-Numbering scheme numeration, in the binding site.

iii) *Ligand-binding site classification*: Most similar receptors' binding sites are listed, in order of similarity. A link to IUPHAR web page of their known ligand in is provided.

iv) *Crystal structure template research*: a phylogenetic tree is generated using the input receptor and receptors with known 3D

structure. This tool is intended to assist in the template selection for homology modelling. A graph is generated showing phylogenetic distances between crystallized GPCR and the user's query. The closest crystal in the tree is the best suited to build an homology model of the query receptor.

4.4.1.2 TM sequence classification

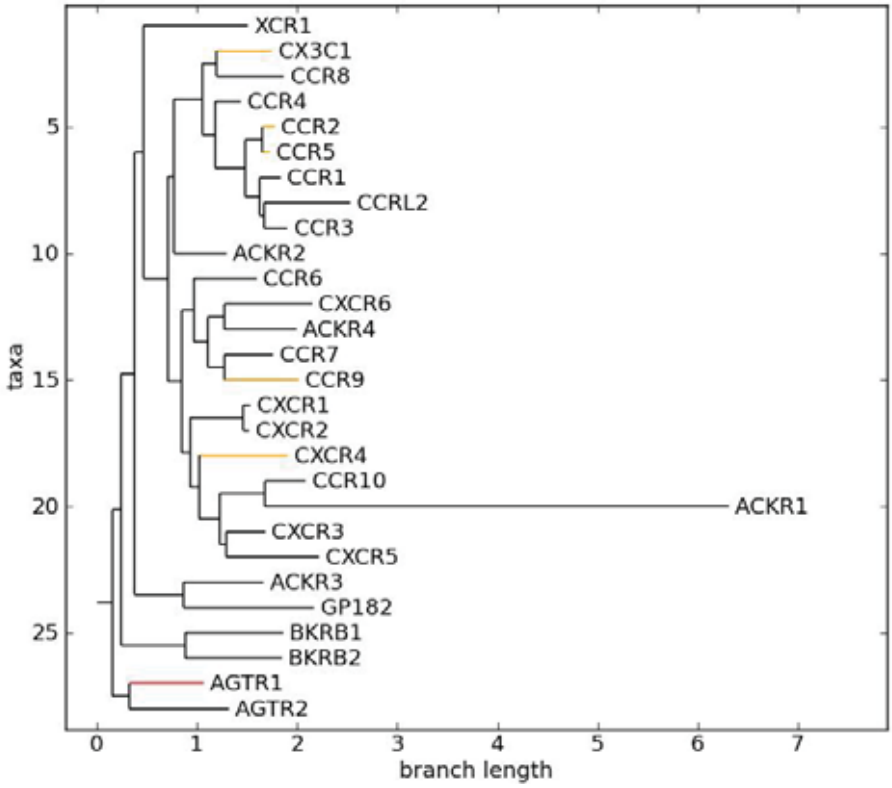


Figure 23. Tree plot of the AGTR1 (labeled in red). The number of receptors corresponds to the taxa value and the branch length is proportional to the sequence comparison distance. Receptors with crystal structure are tagged in orange.

As shown in Figure 23 users can search the submitted receptor on the TM sequence classification and check the closest GPCRs. For example, the sub-branch of the AGTR1 reflects that angiotensin receptors are closely related to the chemokine receptors that supposedly share similar characteristics with AGTR1. Users can zoom in and out trees branches, adjusting the number of branches by the input protein they want to observe. Branches of receptors with known crystal structure are shown in orange.

4.4.1.3 Ligand-binding site representation

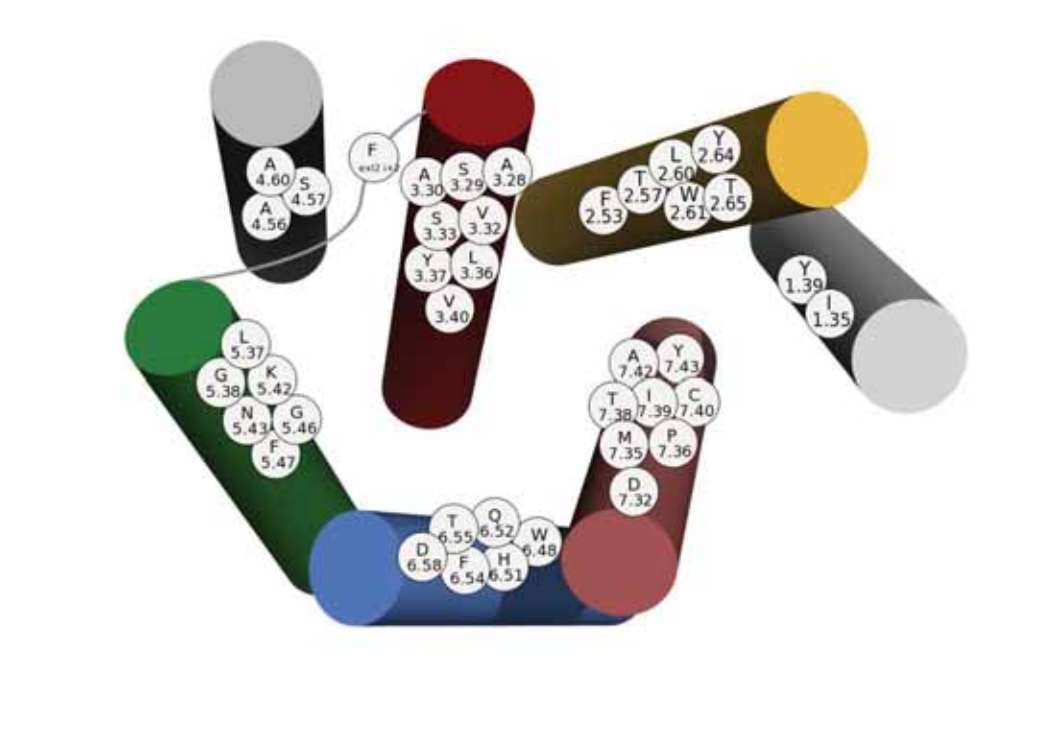


Figure 24. Binding pocket residues scheme for AGTR1. Binding site residues of the selected receptor are showed on the corresponding position.

The identification of residues directly involved in the ligand-receptor interaction is not straightforward, and less so is for users unskilled in computational structural biology. For this reason, a schematic representation of the binding cavity was developed in the GPCR-browser. As example, residues located at the generic binding site in the AGTR1 are showed in Figure 24. This information could be contrasted with the experimental data, confirming the importance of amino acids Y87$^{2.64}$, V108$^{3.32}$, K199$^{5.42}$, H256$^{6.51}$ and I288$^{7.39}$ in ligand binding for this receptor.

4.4.1.4 Ligand-binding site classification

Protein name	IUPHAR Familly	Similarity score
ETBR2	Class A Orphans	0.79
NMBR	Bombesin receptors	0.21
GRPR	Bombesin receptors	0.21
OX2R	Orexin receptors	0.2
OX1R	Orexin receptors	0.19
BRS3	Bombesin receptors,Class A Orphans	0.18
QRFPR	QRFP receptor	0.14
EDNRB	Endothelin receptors	0.1
NPY2R	Neuropeptide Y receptors	0.1
NPY4R	Neuropeptide Y receptors	0.1
CCR3	Chemokine receptors	0.1
EDNRA	Endothelin receptors	0.09
GP151	Class A Orphans	0.09
GNRHR	Gonadotrophin-releasing hormone receptors	0.08
NPY5R	Neuropeptide Y receptors	0.08
NPY1R	Neuropeptide Y receptors	0.08
GHSR	Ghrelin receptor	0.07
PRLHR	Prolactin-releasing peptide receptor	0.06
CCRL2	Chemokine receptors	0.06
CCR1	Chemokine receptors	0.06

Table 8. List of the receptors with most similar binding site of the selected query GPR37.

This tool is created to help to find new potential ligands for a determinate receptor by analogy of binding site similarities with other receptors with known ligands. It could be very useful in the orphan receptors research. Based on the idea that similar binding sites can bind similar ligands, when conducting a drug discovery project on orphan receptors, a good starting point is the close relation of receptors with similar ligand-binding pockets. As example of this approach, the orphan receptor GPR37 is proposed to have similar ligand preferences as other peptide receptors like ETBR2, bombesin, orexin and the QRFP receptors (Table 8) as has been also described by other authors [60].

4.4.1.5 Phylogenetic-based template selection tool for homology modelling

In order to speed-up the process of computational drug design, the existence of a 3D structure of the target protein is of vital importance. There exist various strategies in order to construct 3-D models for protein with unreleased structure: *ab-initio* methods [127, 128], distance-geometry based-methods [129], and homology modeling among others. Homology modeling is based on the concept proteins with similar primary sequences, share similar tertiary structures. It has been shown that identity ranges as low as 30% may be used in order to build robust models [16]. Homology modelling thus consists on constructing a structural model of a target protein from its primary sequence alignment with a known experimental three-dimensional structure used as reference [130]. Thus, the more closely related the template is to the query, the better the quality of the homology model produced. Most tools generally use a BLAST search in order to decide the most similar sequence. Here we developed the first

tool, to our knowledge, to choose a template for a class A GPCRs, based on the phylogenetic similarity between the query protein and all existing class A GPCRs crystals. The significant increase of available GPCR structures made it possible to apply sequence-structure based phylogeny methods in order to improve the alignment accuracy and consequently the models produced [131]. We thus implemented a tool to compute a phylogenetic tree for a selected query receptor and the GPCRs with available 3D coordinates, in order to assist the selection of adequate templates for comparative modeling purposes (Figure 25).

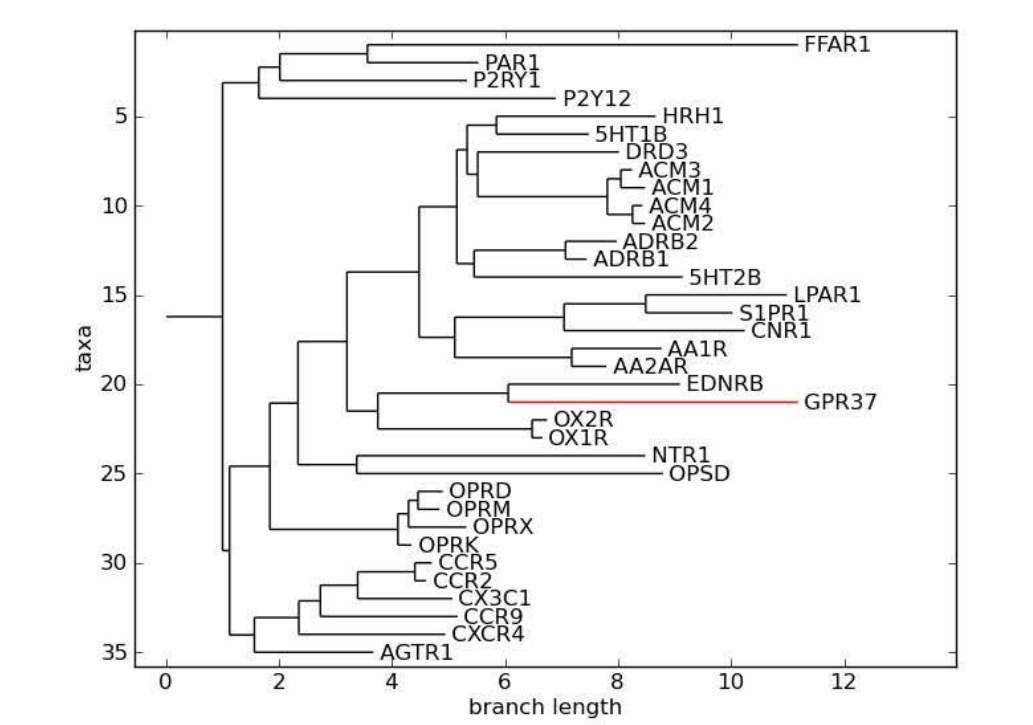


Figure 25. Example of the phylogenetic tree reconstruction of the template research using GPR37 as query. The number of receptors corresponds to the taxa value and the branch length is proportional to the sequence comparison distance.

Continuing the earlier example, the closer GPCR templates to the orphan GPR37 are listed (Table 9).

Protein name	Number of nodes
EDNRB	1
OX1R	3
OX2R	3
OPSD	5
NTR1	5

Table 9. List of the closer receptors to the query ordered by the number of nodes between both.

4.4.2 Applications in pharmacological studies

The GPCR-Browser template selection strategy was successfully applied in the development of an homology model for Angiotensin II Type 1 Receptor (AGTR1). Figure 26 shows a phylogenetic tree of the human AT1R and all class A GPCRs with known structure. It can be seen in the figure that AT1R is located in a branch that includes opioid receptors (OPRX, OPRK, OPRD, and OPRM), the protease-activated receptor (PAR1), the neurotensin receptor (NTR1), and the chemokine CXCR4 receptor, with the latter being clearly the most closely related receptor. Using this information, an AGTR1 tridimensional model was built using the CXCR4 receptor as template. The obtained structure was employed to identify the ligand binding determinants in this receptor, and in combination with a previously generated pharmacophore model, in a molecular docking study of AGTR1 ligands. This model was also successfully used in molecular dynamics (MD) simulations to estimate the affinity of a set of sartan ligands using the linear interaction energy (LIE)

method. In this way, combining a pharmacophore model with binding free energy calculations obtained from the MD simulations on the developed models a molecular mechanism by which sartans interact with AGTR1 was proposed [132].

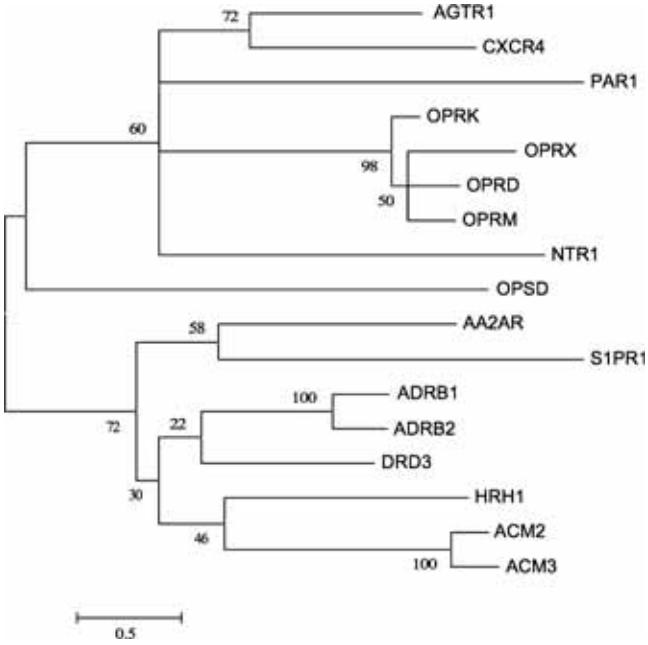


Figure 26. Phylogenetic tree for all class A GPCRs with known structure plus the human AGTR1. Taken from [132]

This result is an example of how the repertoire of currently available structural templates for GPCRs in combination with the precise knowledge on helix irregularities within the TM domains can be successfully used to develop molecular models that are useful in the understanding of experimental results. The quality of the models here presented is supported not only for its ability to explain previous experimental results on side-chain substitutions within the binding pocket but also by the

agreement between theoretically calculated and experimentally determined ligand binding free energies. The insights provided for the characterization of the mechanism by which sartans bind to AGTR1 may be useful in the design of more potent and selective compounds.

4.4.3 Applications in other web-developed bioinformatics resources

The information derived from structural knowledge of GPCRs and their impact in sequence alignments are useful to associate common features between receptors at different levels. This information could be effectively applied in conservation, covariance and correlation studies for this family of receptors. In this regard, the MSAs derived from this work were implemented in the G-protein-coupled receptors – Sequence Analysis and Statistics (GPCR-SAS) web server (http://lmc.uab.cat/gpcrsas/) developed in the Laboratory of Computational Medicine.

The GPCR-SAS compute conservation analyses on GPCR sequences, as well as covariance and correlation studies of residues located at determined positions, providing a set of tools to detect and quantify highly conserved residues or sequence motifs, identify correlations in mutations and give statistical information of such correlations in sequence alignments and to classify the results according to abundance within specific GPCR subfamilies.

4.4.4 Conclusion and perspectives

The developed GPCR-Browser web application allows to find key information from GPCR TM sequences and ligand-binding sites in order to

improve their classification, identify ligand-binding preferences and to develop comparative models. This tool is implemented in an easy user interface to allow non-expert users to investigate relations between GPCRs at the sequence and structural level. Moreover, this web application is of great help in drug discovery research, through the implementation of several tools for the analysis of the binding pocket and template selection in comparative modeling of selected pharmacological targets.

GPCR-browser is freely accessible at http://lmc.uab.cat/gpcr-browser/. Its design and implementation permits automatic updates of the available crystal structures in the Protein Data Bank, as well as the incorporation of new sequence releases from Uniprot.

5 Summary of the novelties derived from this work

The low similarity between GPCR sequences makes difficult their study. Using the new structural information provided by crystallographic data on several members of this protein family, we developed structural-informed MSAs with the inclusion of gaps in TM regions, reflecting evolutive changes observed within receptors. Using this information, we developed the following bioinformatics:

- An amino acid substitution matrix was built for the class A GPCRs. This matrix was successfully tested in sequence alignments and database searches providing improved results compared to other matrices (see section 4.2).

- A clustering method was developed for the classification and chemogenomic characterization of GPCRs (see section 4.3). This was implemented in a web application that assists the comparison between receptors and helps in the selection of templates for homology modeling. We have successfully used this tool in a pharmacological study on the AGTR1. (see section 4.4)

References

1. Bockaert J, Pin JP: **Molecular tinkering of G protein-coupled receptors: an evolutionary success**. *Embo J* 1999, **18**(7):1723-1729.
2. Fredriksson R, Schiöth HB: **The repertoire of G-protein-coupled receptors in fully sequenced genomes**. *Mol Pharmacol* 2005, **67**(5):1414-1425.
3. Smit MJ, Vischer HF, Bakker RA, Jongejan A, Timmerman H, Pardo L, Leurs R: **Pharmacogenomic and structural analysis of constitutive g protein-coupled receptor activity**. *Annu Rev Pharmacol Toxicol* 2007, **47**:53-87.
4. Lagerström MC, Schiöth HB: **Structural diversity of G protein-coupled receptors and significance for drug discovery**. *Nat Rev Drug Discov* 2008, **7**(4):339-357.
5. Wettschureck N, Offermanns S: **Mammalian G proteins and their cell type specific functions**. *Physiological reviews* 2005, **85**(4):1159-1204.
6. Imming P, Sinning C, Meyer A: **Drugs, their targets and the nature and number of drug targets**. *Nat Rev Drug Discov* 2006, **5**(10):821-834.
7. Laporte SA, Oakley RH, Zhang J, Holt JA, Ferguson SS, Caron MG, Barak LS: **The beta2-adrenergic receptor/betaarrestin complex recruits the clathrin adaptor AP-2 during endocytosis**. *Proc Natl Acad Sci U S A* 1999, **96**(7):3712-3717.
8. Luttrell LM, Lefkowitz RJ: **The role of beta-arrestins in the termination and transduction of G-protein-coupled receptor signals**. *J Cell Sci* 2002, **115**(Pt 3):455-465.
9. Attwood T, Findlay J: **Fingerprinting G-protein-coupled receptors**. *Protein engineering* 1994, **7**(2):195-203.
10. Fredriksson R, Lagerstrom MC, Lundin LG, Schiöth HB: **The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints**. *Mol Pharmacol* 2003, **63**(6):1256-1272.
11. Kolakowski Jr LF: **GCRDb: a G-protein-coupled receptor database**. *Receptors & channels* 1993, **2**(1):1-7.
12. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK *et al*: **High-resolution crystal structure of an engineered human beta2-**

adrenergic G protein-coupled receptor. *Science* 2007, **318**(5854):1258-1265.

13. Siu FY, He M, de Graaf C, Han GW, Yang D, Zhang Z, Zhou C, Xu Q, Wacker D, Joseph JS *et al*: **Structure of the human glucagon class B G-protein-coupled receptor**. *Nature* 2013, **499**(7459):444-449.
14. Wu H, Wang C, Gregory KJ, Han GW, Cho HP, Xia Y, Niswender CM, Katritch V, Meiler J, Cherezov V *et al*: **Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator**. *Science* 2014, **344**(6179):58-64.
15. Wang C, Wu H, Katritch V, Han GW, Huang XP, Liu W, Siu FY, Roth BL, Cherezov V, Stevens RC: **Structure of the human smoothened receptor bound to an antitumour agent**. *Nature* 2013, **497**(7449):338-343.
16. Olivella M, Gonzalez A, Pardo L, Deupi X: **Relation between sequence and structure in membrane proteins**. *Bioinformatics* 2013, **29**(13):1589-1592.
17. Ballesteros JA, Weinstein H: **[19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors**. *Methods in neurosciences* 1995, **25**:366-428.
18. de Graaf C, Rognan D: **Customizing G Protein-coupled receptor models for structure-based virtual screening**. *Curr Pharm Des* 2009, **15**(35):4026-4048.
19. Dror RO, Pan AC, Arlow DH, Borhani DW, Maragakis P, Shan Y, Xu H, Shaw DE: **Pathway and mechanism of drug binding to G-protein-coupled receptors**. *Proceedings of the National Academy of Sciences* 2011, **108**(32):13118-13123.
20. Peeters M, Van Westen G, Li Q, IJzerman A: **Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation**. *Trends Pharmacol Sci* 2011, **32**(1):35-42.
21. Venkatakrishnan AJ, Deupi X, Lebon G, Tate CG, Schertler GF, Babu MM: **Molecular signatures of G-protein-coupled receptors**. *Nature* 2013, **494**(7436):185-194.
22. Rosenkilde MM, Benned-Jensen T, Frimurer TM, Schwartz TW: **The minor binding pocket: a major player in 7TM receptor activation**. *Trends Pharmacol Sci* 2010, **31**(12):567-574.
23. de la Chaux N, Messer PW, Arndt PF: **DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage**. *BMC Evol Biol* 2007, **7**:191.

24. Gonzalez A, Cordomi A, Caltabiano G, Pardo L: **Impact of helix irregularities on sequence alignment and homology modeling of G protein-coupled receptors**. *Chembiochem* 2012, **13**(10):1393-1399.
25. Pascarella S, Argos P: **Analysis of insertions/deletions in protein structures**. *Journal of molecular biology* 1992, **224**(2):461-471.
26. Barwell J, Gingell JJ, Watkins HA, Archbold JK, Poyner DR, Hay DL: **Calcitonin and calcitonin receptor-like receptors: common themes with family B GPCRs?** *Br J Pharmacol* 2012, **166**(1):51-65.
27. Koth CM, Murray JM, Mukund S, Madjidi A, Minn A, Clarke HJ, Wong T, Chiang V, Luis E, Estevez A *et al*: **Molecular basis for negative regulation of the glucagon receptor**. *Proc Natl Acad Sci U S A* 2012, **109**(36):14393-14398.
28. Zhang H, Qiao A, Yang D, Yang L, Dai A, de Graaf C, Reedtz-Runge S, Dharmarajan V, Han GW, Grant TD *et al*: **Structure of the full-length glucagon class B G-protein-coupled receptor**. *Nature* 2017, **546**(7657):259-264.
29. Salzman GS, Ackerman SD, Ding C, Koide A, Leon K, Luo R, Stoveken HM, Fernandez CG, Tall GG, Piao X *et al*: **Structural Basis for Regulation of GPR56/ADGRG1 by Its Alternatively Spliced Extracellular Domains**. *Neuron* 2016, **91**(6):1292-1304.
30. Christopher JA, Aves SJ, Bennett KA, Dore AS, Errey JC, Jazayeri A, Marshall FH, Okrasa K, Serrano-Vega MJ, Tehan BG *et al*: **Fragment and Structure-Based Drug Discovery for a Class C GPCR: Discovery of the mGlu5 Negative Allosteric Modulator HTL14242 (3-Chloro-5-[6-(5-fluoropyridin-2-yl)pyrimidin-4-yl]benzonitrile)**. *J Med Chem* 2015, **58**(16):6653-6664.
31. Dore AS, Okrasa K, Patel JC, Serrano-Vega M, Bennett K, Cooke RM, Errey JC, Jazayeri A, Khan S, Tehan B *et al*: **Structure of class C GPCR metabotropic glutamate receptor 5 transmembrane domain**. *Nature* 2014, **511**(7511):557-562.
32. Zhang X, Zhao F, Wu Y, Yang J, Han GW, Zhao S, Ishchenko A, Ye L, Lin X, Ding K *et al*: **Crystal structure of a multi-domain human smoothened receptor in complex with a super stabilizing ligand**. *Nat Commun* 2017, **8**:15383.
33. May LT, Leach K, Sexton PM, Christopoulos A: **Allosteric modulation of G protein-coupled receptors**. *Annu Rev Pharmacol Toxicol* 2007, **47**:1-51.
34. Jazayeri A, Dore AS, Lamb D, Krishnamurthy H, Southall SM, Baig AH, Bortolato A, Koglin M, Robertson NJ, Errey JC *et al*: **Extra-helical**

binding site of a glucagon receptor antagonist. *Nature* 2016, **533**(7602):274-277.
35. Day PW, Rasmussen SG, Parnot C, Fung JJ, Masood A, Kobilka TS, Yao XJ, Choi HJ, Weis WI, Rohrer DK *et al*: **A monoclonal antibody for G protein-coupled receptor crystallography**. *Nat Methods* 2007, **4**(11):927-929.
36. Serrano-Vega MJ, Magnani F, Shibata Y, Tate CG: **Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form**. *Proc Natl Acad Sci U S A* 2008, **105**(3):877-882.
37. Tehan BG, Bortolato A, Blaney FE, Weir MP, Mason JS: **Unifying family A GPCR theories of activation**. *Pharmacol Ther* 2014, **143**(1):51-60.
38. Nygaard R, Frimurer TM, Holst B, Rosenkilde MM, Schwartz TW: **Ligand binding and micro-switches in 7TM receptor structures**. *Trends Pharmacol Sci* 2009, **30**(5):249-259.
39. Southan C, Sharman JL, Benson HE, Faccenda E, Pawson AJ, Alexander SP, Buneman OP, Davenport AP, McGrath JC, Peters JA *et al*: **The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands**. *Nucleic Acids Res* 2016, **44**(D1):D1054-1068.
40. Alexander SP, Davenport AP, Kelly E, Marrion N, Peters JA, Benson HE, Faccenda E, Pawson AJ, Sharman JL, Southan C: **The Concise Guide to PHARMACOLOGY 2015/16: G protein‐coupled receptors**. *Br J Pharmacol* 2015, **172**(24):5744-5869.
41. Lee DK, Lança AJ, Cheng R, Nguyen T, Ji XD, Gobeil F, Chemtob S, George SR, O'Dowd BF: **Agonist-independent nuclear localization of the Apelin, angiotensin AT1, and bradykinin B2 receptors**. *Journal of Biological Chemistry* 2004, **279**(9):7901-7908.
42. Sleator RD: **Phylogenetics**. *Arch Microbiol* 2011, **193**(4):235-239.
43. McCormack GP, Clewley JP: **The application of molecular phylogenetics to the analysis of viral genome diversity and evolution**. *Rev Med Virol* 2002, **12**(4):221-238.
44. Deville J, Rey J, Chabbert M: **An indel in transmembrane helix 2 helps to trace the molecular evolution of class A G-protein-coupled receptors**. *J Mol Evol* 2009, **68**(5):475-489.
45. Pele J, Abdi H, Moreau M, Thybert D, Chabbert M: **Multidimensional scaling reveals the main evolutionary pathways of class A G-protein-coupled receptors**. *PLoS One* 2011, **6**(4):e19094.
46. Kakarala KK, Jamil K: **Sequence-structure based phylogeny of GPCR Class A Rhodopsin receptors**. *Mol Phylogenet Evol* 2014, **74**:66-96.

47. Qian B, Soyer OS, Neubig RR, Goldstein RA: **Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs**. *FEBS Lett* 2003, **554**(1-2):95-99.
48. Papasaikas PK, Bagos PG, Litou ZI, Hamodrakas SJ: **A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models**. *SAR QSAR Environ Res* 2003, **14**(5-6):413-420.
49. Lapinsh M, Gutcaits A, Prusis P, Post C, Lundstedt T, Wikberg JE: **Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences**. *Protein Sci* 2002, **11**(4):795-805.
50. Davies MN, Secker A, Freitas AA, Mendao M, Timmis J, Flower DR: **On the hierarchical classification of G protein-coupled receptors**. *Bioinformatics* 2007, **23**(23):3113-3118.
51. Karchin R, Karplus K, Haussler D: **Classifying G-protein coupled receptors with support vector machines**. *Bioinformatics* 2002, **18**(1):147-159.
52. Bhasin M, Raghava GP: **GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W383-389.
53. Bhasin M, Raghava GP: **GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W143-147.
54. Jacoby E: **A Novel Chemogenomics Knowledge‐Based Ligand Design Strategy—Application to G Protein‐Coupled Receptors**. *Quantitative Structure‐Activity Relationships* 2001, **20**(2):115-123.
55. Surgand JS, Rodrigo J, Kellenberger E, Rognan D: **A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors**. *Proteins* 2006, **62**(2):509-538.
56. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE *et al*: **Crystal structure of rhodopsin: A G protein-coupled receptor**. *Science* 2000, **289**(5480):739-745.
57. Gloriam DE, Foord SM, Blaney FE, Garland SL: **Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design**. *J Med Chem* 2009, **52**(14):4429-4442.
58. Kratochwil NA, Malherbe P, Lindemann L, Ebeling M, Hoener MC, Muhlemann A, Porter RH, Stahl M, Gerber PR: **An automated system for the analysis of G protein-coupled receptor transmembrane**

binding pockets: alignment, receptor-based pharmacophores, and their application. *J Chem Inf Model* 2005, **45**(5):1324-1336.
59. Lin H, Sassano MF, Roth BL, Shoichet BK: **A pharmacological organization of G protein-coupled receptors**. *Nat Methods* 2013, **10**(2):140-146.
60. Ngo T, Ilatovskiy AV, Stewart AG, Coleman JL, McRobb FM, Riek RP, Graham RM, Abagyan R, Kufareva I, Smith NJ: **Orphan receptor ligand discovery by pickpocketing pharmacological neighbors**. *Nat Chem Biol* 2017, **13**(2):235-242.
61. Bock JR, Gough DA: **Virtual screen for ligands of orphan G protein-coupled receptors**. *J Chem Inf Model* 2005, **45**(5):1402-1414.
62. Frimurer TM, Ulven T, Elling CE, Gerlach LO, Kostenis E, Hogberg T: **A physicogenetic method to assign ligand-binding relationships between 7TM receptors**. *Bioorg Med Chem Lett* 2005, **15**(16):3707-3712.
63. Jacob L, Hoffmann B, Stoven V, Vert JP: **Virtual screening of GPCRs: an in silico chemogenomics approach**. *BMC Bioinformatics* 2008, **9**:363.
64. Weill N, Rognan D: **Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands**. *J Chem Inf Model* 2009, **49**(4):1049-1062.
65. Isberg V, Vroling B, van der Kant R, Li K, Vriend G, Gloriam D: **GPCRDB: an information system for G protein-coupled receptors**. *Nucleic Acids Res* 2014, **42**(Database issue):D422-425.
66. Ono Y, Fujibuchi W, Suwa M: **Automatic gene collection system for genome-scale overview of G-protein coupled receptors in eukaryotes**. *Gene* 2005, **364**:63-73.
67. Michino M, Chen J, Stevens RC, Brooks CL, 3rd: **FoldGPCR: structure prediction protocol for the transmembrane domain of G protein-coupled receptors from class A**. *Proteins* 2010, **78**(10):2189-2201.
68. Sandal M, Duy TP, Cona M, Zung H, Carloni P, Musiani F, Giorgetti A: **GOMoDo: A GPCRs online modeling and docking webserver**. *PLoS One* 2013, **8**(9):e74092.
69. Esguerra M, Siretskiy A, Bello X, Sallander J, Gutierrez-de-Teran H: **GPCR-ModSim: A comprehensive web based solution for modeling G-protein coupled receptors**. *Nucleic Acids Res* 2016, **44**(W1):W455-462.
70. **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2017, **45**(D1):D158-D169.

71. Zozulya S, Echeverri F, Nguyen T: **The human olfactory receptor repertoire**. *Genome Biol* 2001, **2**(6):RESEARCH0018.
72. Topiol S, Sabio M: **X-ray structure breakthroughs in the GPCR transmembrane region**. *Biochem Pharmacol* 2009, **78**(1):11-20.
73. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948.
74. Klco JM, Nikiforovich GV, Baranski TJ: **Genetic analysis of the first and third extracellular loops of the C5a receptor reveals an essential WXFG motif in the first loop**. *J Biol Chem* 2006, **281**(17):12010-12019.
75. Mirzadegan T, Benko G, Filipek S, Palczewski K: **Sequence analyses of G-protein-coupled receptors: similarities to rhodopsin**. *Biochemistry* 2003, **42**(10):2759-2767.
76. Isberg V, de Graaf C, Bortolato A, Cherezov V, Katritch V, Marshall FH, Mordalski S, Pin JP, Stevens RC, Vriend G *et al*: **Generic GPCR residue numbers - aligning topology maps while minding the gaps**. *Trends Pharmacol Sci* 2015, **36**(1):22-31.
77. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proc Natl Acad Sci U S A* 1992, **89**(22):10915-10919.
78. Sheetlin S, Park Y, Spouge JL: **The Gumbel pre-factor k for gapped local alignment can be estimated from simulations of global alignment**. *Nucleic Acids Res* 2005, **33**(15):4987-4994.
79. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic Acids Res* 2002, **30**(14):3059-3066.
80. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability**. *Mol Biol Evol* 2013, **30**(4):772-780.
81. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0**. *Syst Biol* 2010, **59**(3):307-321.
82. Hordijk W, Gascuel O: **Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood**. *Bioinformatics* 2005, **21**(24):4338-4347.
83. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood**. *Syst Biol* 2003, **52**(5):696-704.

84. Anisimova M, Gascuel O: **Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative**. *Syst Biol* 2006, **55**(4):539-552.
85. Goldman N, Anderson JP, Rodrigo AG: **Likelihood-based tests of topologies in phylogenetics**. *Syst Biol* 2000, **49**(4):652-670.
86. Rios S, Fernandez MF, Caltabiano G, Campillo M, Pardo L, Gonzalez A: **GPCRtm: An amino acid substitution matrix for the transmembrane region of class A G Protein-Coupled Receptors**. *BMC Bioinformatics* 2015, **16**:206.
87. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B *et al*: **Biopython: freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics* 2009, **25**(11):1422-1423.
88. Altschul SF: **Amino acid substitution matrices from an information theoretic perspective**. *Journal of molecular biology* 1991, **219**(3):555-565.
89. Yu YK, Altschul SF: **The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions**. *Bioinformatics* 2005, **21**(7):902-911.
90. Yang Z, Nielsen R, Goldman N, Pedersen AM: **Codon-substitution models for heterogeneous selection pressure at amino acid sites**. *Genetics* 2000, **155**(1):431-449.
91. Lemaitre C, Barre A, Citti C, Tardy F, Thiaucourt F, Sirand-Pugnet P, Thebault P: **A novel substitution matrix fitted to the compositional bias in Mollicutes improves the prediction of homologous relationships**. *BMC Bioinformatics* 2011, **12**:457.
92. Ng PC, Henikoff JG, Henikoff S: **PHAT: a transmembrane-specific substitution matrix**. *Bioinformatics* 2000, **16**(9):760-766.
93. Sutormin RA, Rakhmaninova AB, Gelfand MS: **BATMAS30: amino acid substitution matrix for alignment of bacterial transporters**. *Proteins* 2003, **51**(1):85-95.
94. Brick K, Pizzi E: **A novel series of compositionally biased substitution matrices for comparing Plasmodium proteins**. *BMC Bioinformatics* 2008, **9**:236.
95. Coronado JE, Attie O, Epstein SL, Qiu WG, Lipke PN: **Composition-modified matrices improve identification of homologs of saccharomyces cerevisiae low-complexity glycoproteins**. *Eukaryot Cell* 2006, **5**(4):628-637.
96. Jones DT, Taylor WR, Thornton JM: **A mutation data matrix for transmembrane proteins**. *FEBS Lett* 1994, **339**(3):269-275.

97. Ng PC, Henikoff JG, Henikoff S: **PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane**. *Bioinformatics* 2000, **16**(9):760-766.
98. **UniProtKB/Swiss-Prot protein knowledgebase release statistics Oct-29, 2014**.
99. Deupi X, Olivella M, Sanz A, Dolker N, Campillo M, Pardo L: **Influence of the g- conformation of Ser and Thr on the structure of transmembrane helices**. *J Struct Biol* 2010, **169**(1):116-123.
100. Gonzalez A, Cordomi A, Matsoukas M, Zachmann J, Pardo L: **Modeling of G protein-coupled receptors using crystal structures: from monomers to signaling complexes**. *Adv Exp Med Biol* 2014, **796**:15-33.
101. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based archival file for macromolecular structures**. *Journal of molecular biology* 1977, **112**(3):535-542.
102. Bylund DB, Eikenberg DC, Hieble JP, Langer SZ, Lefkowitz RJ, Minneman KP, Molinoff PB, Ruffolo RR, Trendelenburg U: **International Union of Pharmacology nomenclature of adrenoceptors**. *Pharmacological reviews* 1994, **46**(2):121-136.
103. Soriano-Ursúa MA, Trujillo-Ferrara JG, Correa-Basurto J, Vilar S: **Recent structural advances of β1 and β2 adrenoceptors yield keys for ligand recognition and drug design**. *J Med Chem* 2013, **56**(21):8207-8223.
104. Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien EY, Lane JR, Ijzerman AP, Stevens RC: **The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist**. *Science* 2008, **322**(5905):1211-1217.
105. Hanson MA, Roth CB, Jo E, Griffith MT, Scott FL, Reinhart G, Desale H, Clemons B, Cahalan SM, Schuerer SC *et al*: **Crystal structure of a lipid G protein-coupled receptor**. *Science* 2012, **335**(6070):851-855.
106. Srinivasan S, Santiago P, Lubrano C, Vaisse C, Conklin BR: **Engineering the melanocortin-4 receptor to control constitutive and ligand-mediated G(S) signaling in vivo**. *PLoS One* 2007, **2**(7):e668.
107. Narumiya S, Sugimoto Y, Ushikubi F: **Prostanoid receptors: structures, properties, and functions**. *Physiological reviews* 1999, **79**(4):1193-1226.
108. Chrencik JE, Roth CB, Terakado M, Kurata H, Omi R, Kihara Y, Warshaviak D, Nakade S, Asmar-Rovira G, Mileni M *et al*: **Crystal**

Structure of Antagonist Bound Human Lysophosphatidic Acid Receptor 1. *Cell* 2015, **161**(7):1633-1643.

109. Barbhaiya H, McClain R, Ijzerman A, Rivkees SA: **Site-directed mutagenesis of the human A1 adenosine receptor: influences of acidic and hydroxy residues in the first four transmembrane domains on ligand binding**. *Mol Pharmacol* 1996, **50**(6):1635-1642.
110. Lebon G, Warne T, Edwards PC, Bennett K, Langmead CJ, Leslie AG, Tate CG: **Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation**. *Nature* 2011, **474**(7352):521-525.
111. Chien EY, Liu W, Zhao Q, Katritch V, Han GW, Hanson MA, Shi L, Newman AH, Javitch JA, Cherezov V *et al*: **Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist**. *Science* 2010, **330**(6007):1091-1095.
112. Haga K, Kruse AC, Asada H, Yurugi-Kobayashi T, Shiroishi M, Zhang C, Weis WI, Okada T, Kobilka BK, Haga T *et al*: **Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist**. *Nature* 2012, **482**(7386):547-551.
113. Kruse AC, Hu J, Pan AC, Arlow DH, Rosenbaum DM, Rosemond E, Green HF, Liu T, Chae PS, Dror RO *et al*: **Structure and dynamics of the M3 muscarinic acetylcholine receptor**. *Nature* 2012, **482**(7386):552-556.
114. Wacker D, Wang C, Katritch V, Han GW, Huang XP, Vardy E, McCorvy JD, Jiang Y, Chu M, Siu FY *et al*: **Structural features for functional selectivity at serotonin receptors**. *Science* 2013, **340**(6132):615-619.
115. Wang C, Jiang Y, Ma J, Wu H, Wacker D, Katritch V, Han GW, Liu W, Huang XP, Vardy E *et al*: **Structural basis for molecular recognition at serotonin receptors**. *Science* 2013, **340**(6132):610-614.
116. Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AG, Tate CG, Schertler GF: **Structure of a beta1-adrenergic G-protein-coupled receptor**. *Nature* 2008, **454**(7203):486-491.
117. Granier S, Manglik A, Kruse AC, Kobilka TS, Thian FS, Weis WI, Kobilka BK: **Structure of the delta-opioid receptor bound to naltrindole**. *Nature* 2012, **485**(7398):400-404.
118. Manglik A, Kruse AC, Kobilka TS, Thian FS, Mathiesen JM, Sunahara RK, Pardo L, Weis WI, Kobilka BK, Granier S: **Crystal structure of the micro-opioid receptor bound to a morphinan antagonist**. *Nature* 2012, **485**(7398):321-326.

119. Wu H, Wacker D, Mileni M, Katritch V, Han GW, Vardy E, Liu W, Thompson AA, Huang XP, Carroll FI *et al*: **Structure of the human kappa-opioid receptor in complex with JDTic**. *Nature* 2012, **485**(7398):327-332.
120. Thompson AA, Liu W, Chun E, Katritch V, Wu H, Vardy E, Huang XP, Trapella C, Guerrini R, Calo G *et al*: **Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic**. *Nature* 2012, **485**(7398):395-399.
121. Tan Q, Zhu Y, Li J, Chen Z, Han GW, Kufareva I, Li T, Ma L, Fenalti G, Zhang W *et al*: **Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex**. *Science* 2013, **341**(6152):1387-1390.
122. Wu B, Chien EY, Mol CD, Fenalti G, Liu W, Katritch V, Abagyan R, Brooun A, Wells P, Bi FC *et al*: **Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists**. *Science* 2010, **330**(6007):1066-1071.
123. Zheng Y, Qin L, Zacarias NV, de Vries H, Han GW, Gustavsson M, Dabros M, Zhao C, Cherney RJ, Carter P *et al*: **Structure of CC chemokine receptor 2 with orthosteric and allosteric antagonists**. *Nature* 2016, **540**(7633):458-461.
124. Zhang H, Han GW, Batyuk A, Ishchenko A, White KL, Patel N, Sadybekov A, Zamlynny B, Rudd MT, Hollenstein K *et al*: **Structural basis for selectivity and diversity in angiotensin II receptors**. *Nature* 2017, **544**(7650):327-332.
125. Guba W, Green LG, Martin RE, Roche O, Kratochwil N, Mauser H, Bissantz C, Christ A, Stahl M: **From astemizole to a novel hit series of small-molecule somatostatin 5 receptor antagonists via GPCR affinity profiling**. *J Med Chem* 2007, **50**(25):6295-6298.
126. Ishiguro M: **Ligand-binding modes in cationic biogenic amine receptors**. *Chembiochem* 2004, **5**(9):1210-1219.
127. Shacham S, Topf M, Avisar N, Glaser F, Marantz Y, Bar-Haim S, Noiman S, Naor Z, Becker OM: **Modeling the 3D structure of GPCRs from sequence**. *Med Res Rev* 2001, **21**(5):472-483.
128. Vaidehi N, Floriano WB, Trabanino R, Hall SE, Freddolino P, Choi EJ, Zamanakos G, Goddard WA, 3rd: **Prediction of structure and function of G protein-coupled receptors**. *Proc Natl Acad Sci U S A* 2002, **99**(20):12622-12627.
129. Lomize AL, Pogozheva ID, Mosberg HI: **Structural organization of G-protein-coupled receptors**. *J Comput Aided Mol Des* 1999, **13**(4):325-353.

130. Webb B, Sali A: **Protein structure modeling with MODELLER**. *Protein Structure Prediction* 2014:1-15.
131. Cvicek V, Goddard WA, 3rd, Abrol R: **Structure-Based Sequence Alignment of the Transmembrane Domains of All Human GPCRs: Phylogenetic, Structural and Functional Implications**. *PLoS Comput Biol* 2016, **12**(3):e1004805.
132. Matsoukas MT, Cordomi A, Rios S, Pardo L, Tselios T: **Ligand binding determinants for angiotensin II type 1 receptor from computer simulations**. *J Chem Inf Model* 2013, **53**(11):2874-2883.

Appendix

List of Publications

1. Matsoukas MT, Cordomi A, Rios S, Pardo L, Tselios T: **Ligand binding determinants for angiotensin II type 1 receptor from computer simulations**. *J Chem Inf Model* 2013, **53**(11):2874-2883.

2. Rios S, Fernandez MF, Caltabiano G, Campillo M, Pardo L, Gonzalez A: **GPCRtm: An amino acid substitution matrix for the transmembrane region of class A G Protein-Coupled Receptors**. *BMC Bioinformatics* 2015, **16**:206.

3. Rios S, Caltabiano G, Gonzalez A, Pardo L: **GPCR-Browser web server: Classification of the multiple sequence alignment of the G-protein-coupled receptors.** *Manuscript under preparation*

4. Gomez-Tamayo JC, Olivella M, Rios S, Hoogstraat M, Gonzalez A, Deupi X, Campillo M, Pardo L, Cordomi A: **The GPCR-SAS web server: G Protein-Coupled Receptors Sequences Analysis and Statistics.** *Manuscript under preparation*

Figures and tables

Protein code	Protein name (co-crystallized ligands: IAG: Inverse Agonist; AGO: Agonist; ANT: Antagonist, APO: no ligand)	PDB Code
5HT1B	5-hydroxytryptamine receptor 5-HT_{1B} bound to:ergotamine, dihidroergotamine	4IAQ, 4IAR
5HT2B	5-hydroxytryptamine receptor 5-HT_{2B} bound to:ergotamine	4IB4, 4NC3
AA2AR	A_{2A} adenosine receptor bound to: ZM241385, XAC, caffeine$_{AGO}$	3EML, 3PWH, 3REY, 3RFM, 3VG9, 3VGA, 4EIY
	A_{2A} adenosine receptor bound to: T4G. T4E$_{ANT}$	3UZA, 3UZC
	A_{2A} adenosine receptor bound to: adenosine, NECA, UK-4342097, CGS21680$_{AGO}$	2YDO, 2YDV, 3QAK, 4UG2, 4UHR
ACM2	Muscarinic M_2 receptor bound to QNB$_{ANT}$	3UON
	Muscarinic M_2 receptor bound to iperoxo$_{AGO}$ in complex with a G-protein mimetic	4MQS, 4MQT
ACM3	Muscarinic M_3 receptor bound to tiotropium,NMS$_{ANT}$	4DAJ, 4U14, 4U15, 4U16
ADRB1	β_2-adrenergic receptor bound to: cryanopindolol, carazolol, iodocyannopindolol, 4-(piperazin-1- yl)-1H-indole, 4-methyl-2-(piperazin-1-yl) quinoline, carvedilol$_{ANT}$	2VT4, 2YCW, 2YCX, 2YCY, 2YCZ, 3ZPQ, 3ZPR, 4AMJ, 4BVN
	β_2-adrenergic receptor bound to: dobutamine, carmoterol, isoproterenol, salbutamol, bucindolol $_{AGO}$	2Y00, 2Y01, 2Y02, 2Y03, 2Y04, 4AMI
ADRB2	β_2-adrenergic receptor bound to: carazolol, timolol, ICI-118551$_{IAG}$	2RH1, 3D4S, 3NY8, 3NY9
	β_2-adrenergic receptor bound to: alprenolol$_{ANT}$	3NYA
	β_2-adrenergic receptor bound to: alprenolol, BI-167107, procaterol, adrenaline, hydroxybenzylisoproterenol, covalent noradrenaline analog $_{AGO}$	3POG, 3PDS, 3SN6, 4LDO, 4LDE, 4LDL, 4QKX
	β_2-adrenergic receptor in complex to G-protein	3SN6
AGTR1	Angiotensin AT_1 receptor bound to ZD7155$_{ANT}$	4YAY
CCR5	C-C chemokine receptor type 5 bound to maraviroc$_{ANT}$	4MBS
CXCR4	Chemokine CXCR4 receptor bound to: Itit, CVX15, vMIP-II$_{ANT}$	3ODU, 3OE0, 3OE6, 3OE8, 3OE9, 4RWS
DRD3	Dopamine D_3 receptor in complex with eticlopride$_{ANT}$	3PBL
FFAR1	Free fatty acid receptors bound to TAK-875$_{ANT}$	4PHU
HRH1	Histamine H_1 receptor in complex with doxepin$_{IAG}$	3RZE
LPAR1	Lysophospholipid (LPA) receptors type 1 bound to ONO9780307, ONO-9910539, ONO-3080573$_{ANT}$	4Z34, 4Z35, 4Z36
NTR1	Neurotensin NTS_1 receptor in complex with neurotesin$_{AGO}$	3ZEV, 4BUO, 4BWB, 4GRV
OPRD	δ - Opioid receptor in complex with naltrindole, DIPP$_{ANT}$	4EJ4, 4N6H, 4RWA, 4RWD
OPRM	μ - Opioid receptor in complex with beta-funaltrexamine$_{ANT}$	4DKL
OPRK	κ - Opioid receptor in complex with JDTic$_{ANT}$	4DJH
OPRX	NOP Opioid receptor in complex with peptide mimetic C-24$_{ANT}$	4EA3
OPSD	Bovine and squid rhodopsin bound cis-retinalIAG	1F88, 1GZM, 1HZX, 1L9H, 1U19, 2G87, 2HPY, 2I35, 2J4Y, 2PED, 3C9L, 3C9M, 3OAX, 3PXO, 2Z73, 2ZIY, 3AYN
	Bovine rhodopsin bound trans-retinal$_{AGO}$	2X72, 3PQR, 4A4M
	opsin$_{APO}$	3CAP
	Metarhodopsin II$_{AGO}$	1LN6, 3PXO
OX2R	Orexin OX_1 receptor bound to suvorexant$_{ANT}$	4S0V
P2RY1	Purinergic $P2Y_1$ receptor in complex with BPTU, MRS2500$_{ANT}$	4XNV, 4XNW
P2Y12	Purinergic $P2Y_{12}$ receptor in complex with AZD1283$_{ANT}$	4NTJ
	Purinergic $P2Y_{12}$ receptor in complex with 2MeSADP, 2MeSATP$_{AGO}$	4PXZ, 4PY0
PAR1	Proteinase-activated receptor type 1 in complex with vorapaxar$_{ANT}$	3VW7
S1PR1	Sphingosine 1-phosphate receptor in complex with sphingolipid mimic$_{ANT}$	3V2W, 3V2Y
US28	chemokine viral US28 receptor in complex with CX3CL1$_{AGO}$	4XT1, 4XT3

Table A1. List of the class A GPCRs crystal structures used to define the generic ligand-binding site.

	Protein code	Protein name	PDB Code
Class B	CRFR1	Corticotropin-releasing factor receptor type 1	4K5Y, 4Z9G
	GLR	GLucagon receptor	4L6R, 5EE7, 5XF1, 5XEZ
	GLP1R	Glucagon-like peptide receptor	5VEX, 5VEW
Class C	GRM1	Metabotropic glutamate receptor type 1	4OR2
	GRM5	Metabotropic glutamate receptor type 5	4OO9, 5CGC, 5CGD
Class F	SMO	Smoothened homolog receptor	4JKV, 4N4W, 4QIM, 4QIN, 4O9R, 5V57, 5V56

Table A2. List of the crystal structures of class B, C and F GPCRs

Gap penalty parameters	λ	κ	*H*
Q=9 R=2	0.3218	0.6727	0.2013
Q=10 R=4	0.3420	0.6374	0.2470
Q=11 R=1	0.3121	0.6943	0.1819

Table A3. Gumbel distribution statistical parameters λ and κ and the relative entropy H for gapped local alignment scores operating at different gap penalties parameters.

Protein code	135	139	253	257	258	260	261	264	265	326	327	328	329	330	332	333	334	335	336	337	340	456	457	458	459	460	461	535	537	538	542	543	546	547	644	648	651	652	654	655	658	732	734	735	736	738	739	740	742	743
5HT1B_HUMAN	0	0	0	0	0	0	0	2	0	0	0	2	2	0	2	2	0	0	2	2	2	2	0	0	0	0	0	0	0	0	2	1	2	0	0	2	2	2	0	2	2	2	0	2	2	0	2	0	0	2
5HT2B_HUMAN	0	0	0	0	0	0	0	1	0	0	0	3	3	0	3	3	0	0	3	3	1	2	0	0	0	0	0	1	2	2	3	3	3	0	0	3	3	3	0	3	3	3	0	3	3	0	3	0	0	3
AA2AR_HUMAN	1	0	0	1	0	0	10	8	5	0	0	1	1	1	12	14	0	0	6	5	5	0	0	0	0	0	0	8	0	15	12	0	4	1	0	13	15	14	1	15	1	8	0	15	7	0	15	0	8	9
ACM2_HUMAN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	0	0	3	3	3	0	3	0	0	0	0	0	0	1	1	1	3	3	0	3	3	3	0	1	0	0	0	0	0	0	3	0	3	3
ACM3_RAT	0	0	3	0	0	0	0	0	0	0	0	0	0	0	9	9	0	0	9	9	8	0	9	0	0	0	0	0	0	7	8	9	9	9	0	9	9	9	0	8	0	0	0	0	1	9	0	9	9
ADRB1_MELGA	0	0	0	0	0	0	5	9	6	0	0	27	25	0	31	31	0	0	31	21	0	0	0	0	0	0	0	0	31	25	31	31	31	0	0	27	31	31	0	31	0	0	0	6	10	0	31	7	0	31
ADRB2_HUMAN	0	0	4	0	0	0	2	5	2	0	0	12	10	0	12	12	0	0	12	11	0	0	0	0	0	0	0	0	11	9	12	12	12	0	0	12	12	12	0	12	0	3	0	12	6	0	12	2	0	12
AGTR1_HUMAN	1	1	1	0	1	0	1	1	1	0	0	1	1	0	1	1	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1
CCR5_HUMAN	1	2	0	0	0	0	0	2	2	0	0	0	0	2	0	2	2	0	0	2	0	0	0	0	0	0	1	0	0	2	2	2	0	0	0	0	2	2	0	0	2	0	0	0	2	2	0	2	2	0	2
CXCR4_HUMAN	0	6	0	0	0	3	9	9	8	0	0	9	10	0	9	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0	0	7	0	0	1	2	2	0	2	4	0	10	0	0	0
DRD3_HUMAN	0	0	0	0	0	0	0	2	0	0	0	0	2	2	0	2	2	0	0	2	2	0	0	0	0	0	0	0	0	0	2	2	2	2	0	0	2	2	2	0	2	0	0	0	2	0	0	2	0	0	2
FFAR1_HUMAN	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	1	0	0	0	1	0	0	1	1	0	1	1	0	1	1	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0
HRH1_HUMAN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	1	0	0	1
LPAR1_HUMAN	3	0	0	3	0	3	3	3	3	3	0	0	2	3	0	3	3	0	0	3	0	0	0	0	0	0	0	0	0	1	3	3	0	0	0	0	3	3	3	3	3	0	0	0	3	3	3	3	0	3	0
NTR1_RAT	0	0	0	0	0	0	0	0	5	9	0	0	0	0	9	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	7	0	0	0	0	0	0	0	0	9	7	9	9	0	9	0	0	0	0	0
OPRD_HUMAN	0	0	0	0	0	0	0	4	0	0	0	0	0	4	0	6	6	0	0	6	0	0	0	0	0	0	0	0	0	1	6	6	0	0	0	0	6	6	6	0	6	6	0	0	6	0	1	6	0	2	6
OPRK_HUMAN	0	0	2	2	0	0	2	2	0	0	0	2	2	0	2	2	0	2	2	2	0	0	0	0	0	0	0	0	0	2	2	2	0	0	0	0	2	2	2	0	2	0	0	0	2	0	0	2	0	2	2
OPRM_MOUSE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	0	1	1	0	0	1	0	0	1	0	1	1
OPRX_HUMAN	0	0	0	0	0	0	2	2	0	0	0	2	2	0	2	2	0	0	2	2	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0	2	2	2	2	0	2	0	0	0	0	2	0	2	0	0	2
OPSD	0	0	11	7	0	0	0	0	0	0	1	31	29	0	36	36	1	2	33	36	31	2	0	1	0	0	0	0	0	2	4	36	35	29	35	28	36	36	35	0	9	0	0	0	0	0	0	34	16	21	36
OX2R_HUMAN	0	0	0	1	0	1	1	1	0	0	0	1	1	0	1	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	1	0	1	1
P2RY1_HUMAN	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	2	2	0	2	0	0	0
P2Y12_HUMAN	0	0	0	0	0	0	0	0	0	2	0	0	3	3	0	3	3	0	0	3	0	3	3	0	0	3	0	0	3	3	3	1	0	0	0	0	0	1	0	1	3	3	2	1	3	1	1	0	0	0	0
PAR1_HUMAN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0
S1PR1_HUMAN	0	0	0	1	2	1	2	1	2	0	0	2	2	0	2	2	0	0	2	0	0	0	0	0	0	0	0	0	0	2	2	2	0	2	0	2	2	2	2	0	2	0	0	0	0	2	0	2	0	0	2
US28_HCMVA	2	2	0	0	0	0	2	2	2	0	0	0	2	0	2	2	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	1	0	2	2	0	2	0	0	2
present receptors	5	4	5	6	2	4	14	16	9	2	2	14	21	3	22	24	2	2	19	15	7	6	5	2	2	9	1	2	11	21	20	11	11	6	2	18	22	16	5	24	10	10	2	18	14	4	22	4	9	19
bs position	135	139	253	257		260	261	264	265			328	329	330	332	333			336	337	340	456	457			460			537	538	542	543	546	547		648	651	652	654	655	658	732		735	736	738	739	740	742	743

Table A4. Count of the residues within a distance ≤ 5 Å of the ligand in every crystal structure and annotated according to Weinstein-Ballesteros scheme. Only positions observed in at least two crystal structures from different receptors were included in the consensus binding pocket sequence database (last row).

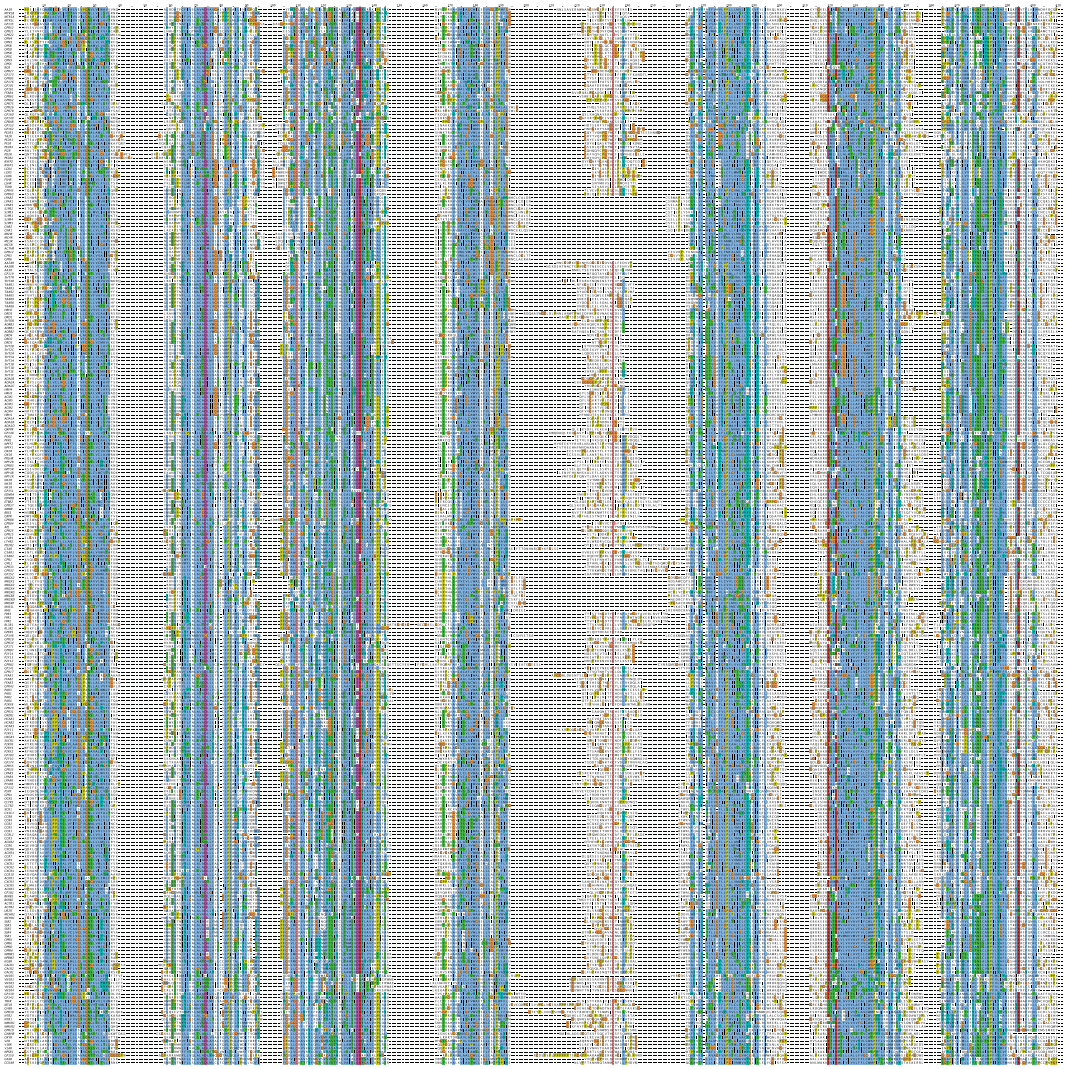


Figure A1. Sequence alignment of class A GPCR. N- and C- terminus are avoided. Conserved amino acids are colored as clustalx format. The alignment can be downloaded from http://lmc.uab.cat/gpcr-browser/.

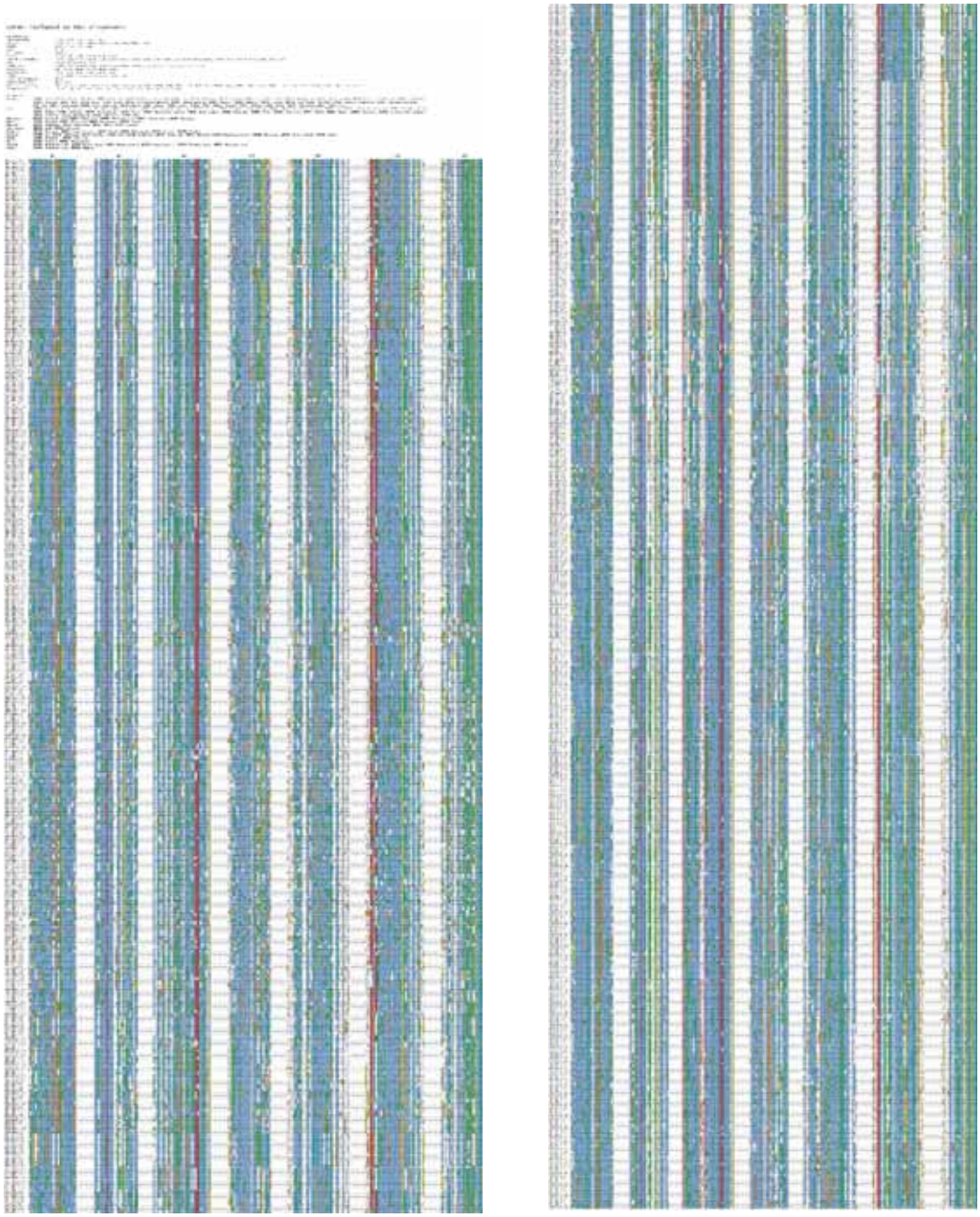


Figure A2. Compilation of subfamilies, principal clades and sequence alignment of class A GPCR transmembrane regions (TM1 to 7) used to generate the GPCRtm substitution matrix. Conserved amino acids are colored as clustalx format. The alignment can be downloaded from http://lmc.uab.cat/gpcr-browser/.

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

Figure A3. Difference matrix obtained by subtracting from the GPCRtm the JTTtm (Lower) and the BLOSUM62 substitution matrices (Upper)

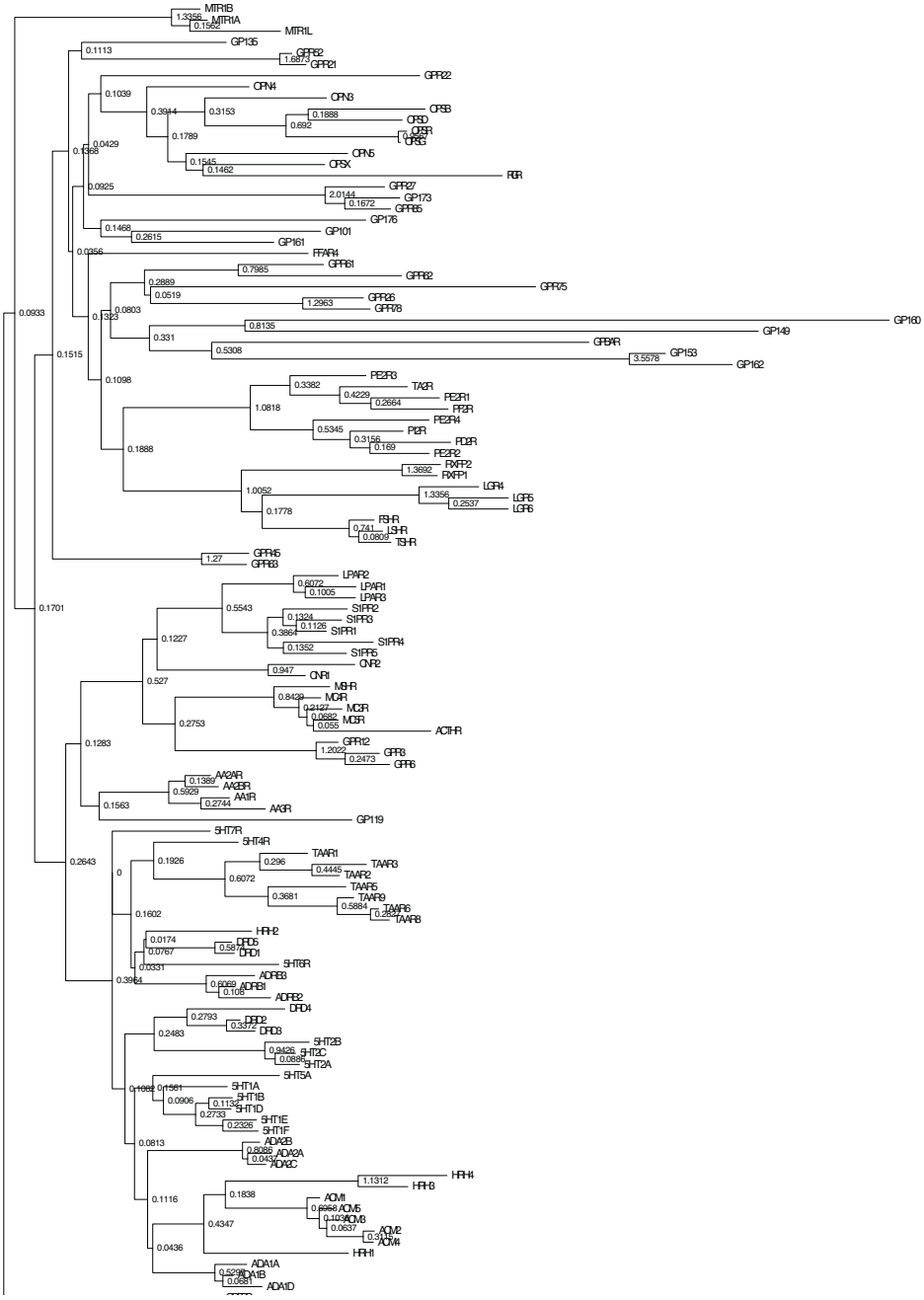


Figure A4. (continue in next page)

Figure A4. Phylogenetic tree of the Human class A receptors from TM sequences. The aLRT support values are shown on each node.

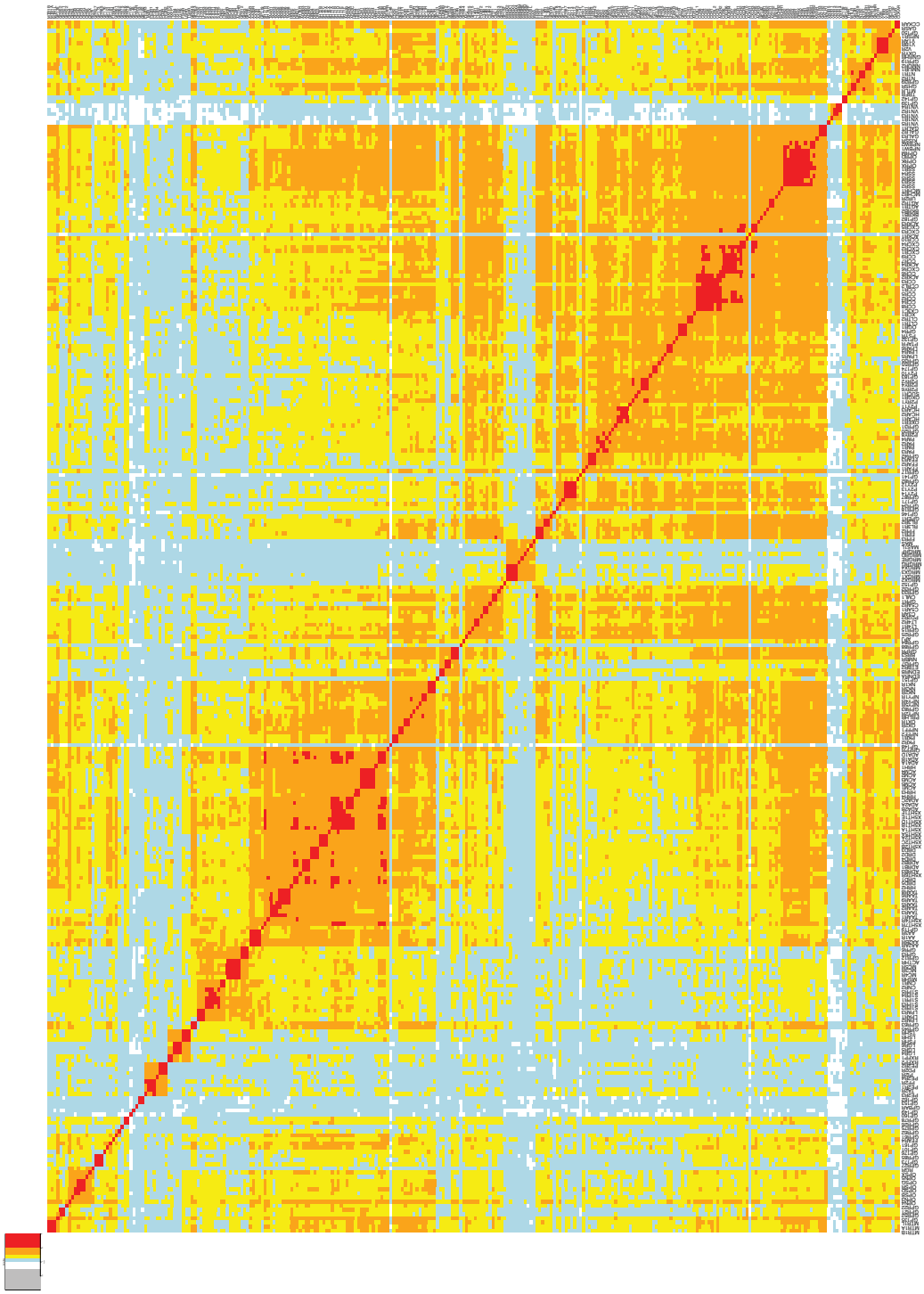


Figure A5. Heat map of the TM sequences made. Similarity values and color scheme are the same as the binding site heat map. Sequences are order as the phylogenetic tree of TM regions.

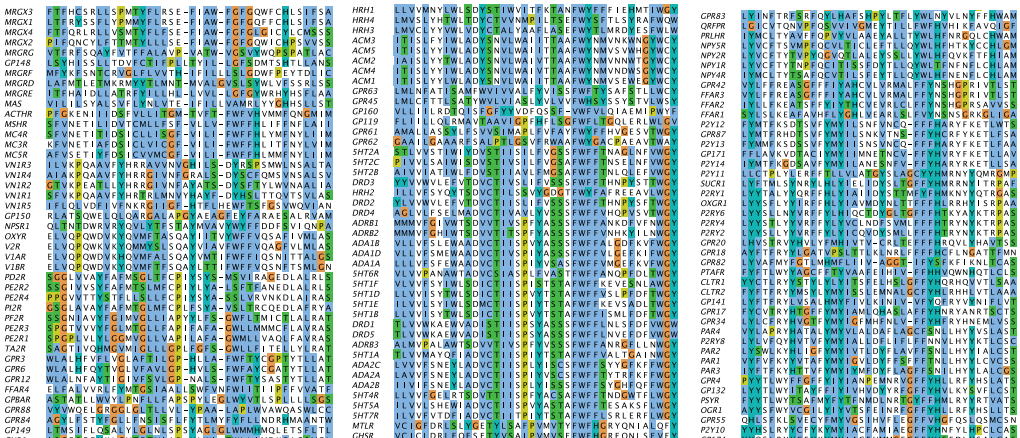


Figure A6. Ligand-binding site database for class A GPCRs (see methods, section 3.3.2). The 40-length-residue dataset was split in three parts.

List of abbreviations

5-HTs: 5-Hydroxytryptamine receptors
AARs: Adenosine receptors
AC: Adenylate cyclase
ACKRs: Atypical chemokine receptors
ACMs: Acetylcholine (muscarinic) receptors
ACTHR: Adrenocorticotropic hormone receptor
ADRs: Adrenoceptors
AGTRs: Angiotensin receptors
APJ: Apelin receptor
BKRBs: Bradykinin receptors
BRS3: Bombesin receptor subtype-3
C3AR: C3a anaphylatoxin chemotactic receptor
C5ARs: C5a anaphylatoxin chemotactic receptors
cAMP: cyclic Adenosine monophosphate
CaS: Calcium-sensing
CCRL2: C-C chemokine like receptors
CCKAR: Cholecystokinin receptor type A
CCRs: CC chemokine receptors
CXCRs: CXC chemokine receptors
CX3C1: CX3C chemokine receptor 1
CML1: Chemerin receptor
CNRs: Cannabinoid receptors
CLTRs: Cysteinyl leukotriene receptors
CRD: Cysteine-rich domain

CRFRs: Corticotropin-releasing factor receptors

DNA: Deoxyribonucleic acid

DRDs: Dopamine receptors

ECL: Extracellular loop

EDNRs: Endothelin receptors

ETBR2: Prosaposin receptor GPR37L1

FFARs: Free fatty acid receptors

FPRs: Formylpeptide receptors

FSHR: Follitropin receptor

$GABA_B$: γ-aminobutyric acid B-type

GALRs: Galanin receptors

GAIN: GPCR-autoproteolysis-inducing

GASR: Gastrin receptor

GDP: Guanidine diphosphate

GHSR: Ghrelin receptor

GLR: Glucagon receptor

GLP1R: Glucagon-like peptide 1 receptor

GNRHR: Gonadotrophin-releasing hormone receptor

GPBAR: Bile acid receptor

GPCR: G-protein-coupled receptor

GPER1: G-protein-coupled estrogen receptor

GPRXX, GPXXX: orphan GPCR

GRAFS: Glutamate, Rhodopsin, Adhesion, Frizzled/Taste 2 and Secretin

GRK: G-protein-coupled receptor kinase

GRMs: Metabotropic glutamate receptors

GRPR: Gastrin-releasin peptide receptor

GTP: Guanidine triphosphate

H8: Helix 8
HCARs: Hydroxycarboxylic acid receptors
HHM: Hidden Markov Model
HRHs: Histamine receptors
ICL: Intracellular loop
KISSR: Kisspeptin receptor
LGRs: Leucine-rich repeat-containing receptor
LIE: Linear interaction energy
LPARs: Lysophospholipid acid receptors
LSHR: Lutropin-choriogonadotropic hormone receptor
LT4Rs: Leukotriene B4 receptors
MAPK: Mitogen-activated protein kinase
MAS: Proto-oncogen Mas
MAS1L: Mas-related G-protein-coupled receptor MRG
MCHRs: Melanin-concentrating hormone receptors
MCRs: Melanocortin receptors
MECA: Melanocortin, endoglin, adenosine and cannabinoid
mGlu: metabotropic Glutamate
MSHR: Melanocyte-stimulating hormone receptor
MRGRs: Mas-related G-protein-coupled receptors
MSA: Multiple Sequence Alignment
MTLR: Motilin receptor
MTR1s: Melatonin receptors
NJ: Neighbour Joining
NKRs: Tachykinin receptors
MD: Molecular Dynamics
ML: Maximum Likelihood

NC-IUPHAR: International Union of basic and clinical Pharmacology committee on receptor Nomenclature and Drug Classification

NMBR: Neuromedin B receptor

NMURs: Neuromedin U receptors

NNI: Nearest Neighbor Interchange

NPBWs: Neuropeptide BW receptors

NPFFs: Neuropeptide FF receptors

NPSR1: Neuropeptide S receptor

NPYRs: Neuropeptide Y receptors

NTRs: Neurotensin rceptors

OGR1: Ovarian cancer G-protein coupled receptor

OPNs: opsin receptors

OPRs: Opioid receptors

OPSB: Short-wave-sensitive opsin receptor

OPSD: Rhodopsin

OPSG: Medium-wave-sensitive opsin receptor

OPSR: Long-wave-sensitive opsin receptor

OPSX: Visual pigment-like receptor peropsin

OXER1: Oxoeicosanoid receptor 1

OXGR1: 2-oxoglutarate receptor type 1

OXRs: Orexin receptors

OXYR: Oxytocin receptor

P2Rs: Prostanoid receptors

P2RYs, P2Ys: P2Y purinoceptors

PARs: Proteinase-activated receptors

PDB: Protein Data Bank

PKRs: Prokineticin receptors

PLC: Phospholipase C

PRLHR: Prolactin-releasing peptide receptor

PSYR: Psychosine receptor

PTAFR: Platelet-activating factor receptor

QRFPR: Pyroglutaminated RFamide peptide receptor

RGR: RPE-retinal G-protein-coupled receptor

RL3Rs: Relaxin-3 receptors

RNA: Ribonucleic acid

RXFPs: Relaxin family peptide receptors

S1PRs: Sphingosine 1-phosphate receptors

SEA: Sequence Ensemble Approach

SMO: Smoothened homolog

SPR: Subtree Pruning and Regrafting

SSRs: Somatostatin receptors

SUCR1: Succinate receptor 1

SVM: Support Vector Machines

TA2R: Thromboxane A2 receptor

TAARs: Trace amine-associated receptors

TAS1: Taste 1 receptors

TK: Tyrosine-protein kinase

TM: Transmembrane

TRFR: Thyrotropin-releasing hormone receptors

TSHR: Tyrotropin receptor

UR2R: Urotensin-2 receptor

VFTM: Venus Flytrap modules

VN1Rs: Vomeronasal type-1 receptors

VRs: Vasopressin receptors

XCR1: XC chemokine receptor 1