Eda Suku

G-protein coupled receptors activation mechanism: from ligand binding to the transmission of the signal inside the cell

Ph.D. Thesis

April 27, 2019

Università degli Studi di Verona
Dipartimento di Biotecnologie

Advisor:
prof. Alejandro Giorgetti

Series N°: **xxx**

Università di Verona
Dipartimento di Biotecnologie
Strada le Grazie 15, 37134 Verona
Italy

To Mauro,
l'amore della mia vita

Abstract

G-protein coupled receptors (GPCRs) are the largest family of pharmaceutical drug targets in the human genome and are modulated by a large variety of endogenous and synthetic ligands. GPCRs activation usually depends on agonist binding (except for receptors with basal activity), which stabilizes receptor conformations and allow the requirement and activation of intracellular transducers. GPCRs are unique receptors and very well studied, since they play an important role in a great number of diseases. They interact with different type of ligands (such as light, peptides, proteins) and different partners in the intracellular part (such as G-proteins or β-arrestins). Based on homology and function GPCRs are divided in five classes: Class A or Rhodopsin, Class B1 or Secretin, Class B2 or Adhesion, Class C or Glutamate, Class F or Frizzled. What is still missing in the state of the art of these receptor, and in particular in Class A, is a global study on different binding cavities with divergent properties, with the aim to discover common binding characteristics, preserved during years of evolution. Gaining more knowledge on common features for ligand recognition shared among all the receptors may become crucial to deeply understand the mechanism used to transmit the signal into the cell. In the first step of this thesis we have used all the solved Class A receptors structures to analyze and find, if exist, a common way to transmit the signal inside the cell. We identified and validated ten positions shared between all the binding cavities and always involved in the interaction with ligands. We demonstrated that residues in these positions are conserved and have co-evolved together. In a second step, we used these positions to understand how ligands could be positioned in the binding cavities of three study cases: Muscarinic receptors, Kisspeptin receptors and the GPR3 receptor. We did not have any experimental information *a priori*. We used homology modeling and docking techniques for the first two cases, adding molecular dynamics simulations in the third case. All the predictions and suggestions from the computational point of view, turned out to be very successful. In particular for the GPR3 receptor we were able to identify and validate by alanine-scanning mutagenesis the role of three functionally relevant residues. The latter were correlated with the constitutive and agonist-stimulated adenylate cyclase activity of GPR3 receptor. Taken together, these results suggest an important role of computational structural biology and pave the way of strong collaborations between computational and experimental researches.

Sommario

I recettori accoppiati a proteine G (GPCRs) sono la piú grande famiglia di target di prodotti farmaceutici nel genoma umano e sono modulati da un'ampia varietá di ligandi endogeni e sintetici. L'attivazione dei GPCR di solito dipende dal legame con agonisti (eccetto per i recettori con attivit basale), che stabilizza le conformazioni del recettore e consente l'attivazione dei trasduttori intracellulari. I GPCR sono recettori unici e molto ben studiati, poich svolgono un ruolo importante in un gran numero di malattie. Interagiscono con diversi tipi di ligandi (come la luce, i peptidi, le proteine) e diversi partner nella parte intracellulare (ad esempio G-proteins o β-arrestins). I GPCR sono suddivisi in cinque classi: Rodopsina o classe A, Secretina o classe B1, Adesione o di classe B2, Glutammato o classe C, Frizzled o classe F. Ció che ancora manca nello stato dell'arte di questi recettori, e in particolare nella Classe A, uno studio globale su diverse cavitá di legame focalizzandosi su caratteristiche comuni per il riconoscimento di ligando, condivise tra tutti i recettori, che pu essere cruciale per la comprensione del meccanismo utilizzato per trasmettere il segnale all'interno della cellula. Nella prima fase di questa tesi abbiamo utilizzato tutte le strutture dei recettori di classe A risolti per analizzare e trovare, se esiste, un modo comune per trasmettere il segnale all'interno della cellula. Abbiamo identificato e convalidato dieci posizioni condivise tra tutte le cavit di legame e sempre coinvolte nell'interazione con i ligandi. Abbiamo dimostrato che i residui in queste posizioni sono conservati e si sono coevoluti insieme. In una seconda fase, abbiamo utilizzato queste posizioni per capire i ligandi nelle cavitá di legame di tre casi: recettori muscarinici, recettori Kisspepin e recettore GPR3. Non avevamo alcuna informazione sperimentale *a priori*. Abbiamo utilizzato tecniche di modellazione e docking per l'omologia per i primi due casi, aggiungendo simulazioni di dinamica molecolare nel terzo caso. Tutte le previsioni e i suggerimenti dal punto di vista computazionale si sono rivelati molto efficaci. In particolare per il recettore GPR3 siamo stati in grado di identificare e convalidare mediante mutagenesi di scansione alanina il ruolo di tre residui funzionalmente rilevanti. Questi ultimi erano correlati all'attivit di adenilato ciclasi costitutiva e agonista-stimolato del recettore GPR3. Presi insieme, questi risultati suggeriscono un ruolo importante della biologia strutturale computazionale e aprono la strada a forti collaborazioni tra ricerche computazionali e sperimentali.

Contents

Part I Introduction

1 Introduction . . . 3

2 G-protein coupled receptors . . . 7

3 Methods . . . 27
3.1 Homology Modeling . . . 27
3.2 Docking . . . 32
3.2.1 Docking on homology models . . . 35
3.3 Molecular dynamics simulation . . . 38
3.3.1 Molecular dynamics: an overview . . . 38
3.3.2 Force fields in MD . . . 39
3.3.3 Solvation . . . 40
3.3.4 Non-bonded Interactions . . . 41
3.3.5 Integration of the equations of motion . . . 41
3.3.6 Combined Molecular Dynamics Approches: MM/CG . . . 44

Part II Contributions

4 Results . . . 49
4.1 Common evolutionary binding mode of rhodopsin-like GPCRs . . . 49
4.1.1 Abstract . . . 49
4.1.2 Results . . . 49
4.1.3 Conclusion . . . 53
4.2 Analysis of the Muscarinic Receptors binding site . . . 55
4.2.1 Abstract . . . 55
4.2.2 Results: The variability of mouse muscarinic receptors lies in the allosteric binding site . . . 55
4.3 Analysis of the Kisspeptine Receptors binding site . . . 57
4.3.1 Abstract . . . 57
4.3.2 Results . . . 57

4.3.3 Conclusions 58
4.4 Allosteric sodium binding cavity in GPR3: a novel player in modulation of A production 60
4.4.1 Abstract 60
4.4.2 Results: Homology modeling and molecular dynamics simulations 60
4.5 Agonist Binding to Chemosensory Receptors: Receptor Activation Predictions 67
4.5.1 Abstract 67
4.5.2 Results: Receptor Activation Predictions 67
4.6 Extra Contributions 70
4.6.1 Identification of new BMP6 propeptide mutations in patients with iron overload 70
4.6.2 Abstract 70

5 Conclusions 73

6 Acknowledgments 75

Part I

Introduction

1

Introduction

The human body is a beautiful and complicated machine composed by many systems and processors. It works in a very fascinating way transforming inputs from the external world in internal signals. The maiority of the extenal stimuli are captured by a family of proteins called GPCRs (G-protein coulped receptors). It is impossible to evaluate their real importance. GPCRs are one of the most ancient and ubiquitous family of eukaryotic plasma membrane receptors. With about 850 members in the human genome [1–3] they are involved in a astonishing amount of physiological processes. Information derived from senses such as olfaction, vision and taste, and even pain or mechanoperception in some cases, is first captured by GPCRs. GPCRs are also responsable for the post-synaptic transmission of neurotransmitters. Hormones regulating behaviours as different as maternal instincts or fear response are received by GPCRs. GPCRs are thus the closest thing to a universal receptor architecture, capable to evolve the transduction of an infinite variety of chemical and (in the case of opsins) physical stimuli.

GPCRs have a long and important history. The first receptor that paved the way for styding the GPCRs was the rhodopsin receptor almost 40 years ago (1970) [4–6] (Figure 1.1).

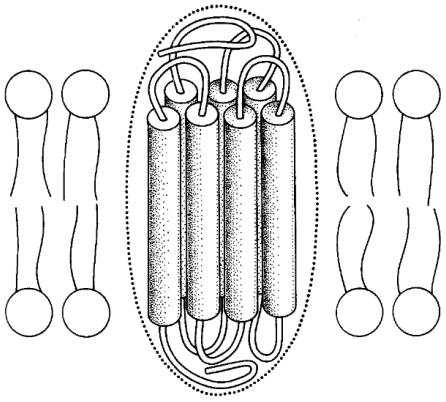

Fig. 1.1. First structural model of the bovine Rhodopsin receptor [7].

Since then we have accumulated an enormous amount of information on their biology and function. Just as an example if we search the string "GPCR" on Pubmed we retrieve almost 10.000 hits. However, a deep study on GPCRs structural biology started only in 2000, when a 2.8Å resolution structure of rhodopsin receptor was published and so far, thanks to ensemble of technological advancements, 50 GPCRs have obtained a solved structure. This boom in GPCRs crystallization included also the Nobel Prize-winning structure of the β2-adrenergic receptor in complex with its cognate heterotrimeric G-protein. Still, despite this, the solved GPCRs structures rapresent only 5% of the diversity of GPCRs in the human genome. Given the current situation, a full experimental coverage of, for example, the non-olfactory GPCRs would take more than a century. Thus, limiting our research to only the characterized and solved GPCRs structures means that we should restrict our biological understanding to a mere 5% of GPCRs. A deep comprehension of GPCRs also requires to know the structure of each receptor in each one of the multiplicity of states which is populated by GPCR molecules (since they are very dynamic proteins), and essential to their function. So far, however, only five GPCRs have been crystallized in more than one activation state (inactive, partially active or fully active). This means that the GPCRs structural biology is still in its infancy, presenting a bottleneck to the type of research questions we can answer. Thus, complementing experimental techniques like X-ray, NMR, FRET, site-directed mutagenesis with computational biology provides a valid toolkit to generate quantitative structural biological information. Homology modeling for example can predict unknown GPCRs structures starting from the known ones; docking algorithms can provide an idea of the binding modes of ligands to their receptors; molecular simulation adds the time dimension, predicting the conformational states of proteins and their complexes with ligands and finally the structure analysis can collect and analyze data from all these techniques. In this thesis, we have used all these computational biology techniques to face up the general problem of the recognition and transduction of a biological signal by GPCRs. We embraced two complementary points of view:

- the binding of a ligand (agonist, antagonist or inverse agonist) to a specific biological system, considering different GPCRs binding cavities
- the generic problem of class A GPCRs activation, using analysis of interactions within GPCRs solved structures and molecular dynamics simulations

In both cases, the aim was to provide a dynamical portrait of the GPCRs function from the binding site to the interaction with the G-protein. The study on the GPR3 receptor is an example of how the use of advanced computational techniques is necessary to answer biology-driven problems. Only static homology modeling and docking approaches are not accurate enough when the sequence identity of the target protein falls below 25%. However, knowledge-based modeling, combined with long-scale hybrid molecular dynamics, can be used as approach to obtain accurate results, as shown by the comparison with experimental data. This means that using computational biology techniques can help us to answer to different biological questions and propose new ones. In our case no crystallographic structure was available for the GPR3 receptor and sequence identity with the templates was below 23%. Nevertheless the work presented in this thesis

and published on Scientific Reports is one of the most advanced structural and mutagenesis study on GPR3 to date, explaining mechanistically the experimental data. It also allowed to suggest a new possible drug target binding cavity, thus taking advantage of a single case study to shed light on new interesting biological questions.

For receptors which structural information is instead available, computational methods can extract and systematize quantitative information, which is not readily apparent by inspecting the structures. So far the number of the structures available in both active and inactive conformations reached a good amount for statistical investigation. By using structural bioinformatics *in-house* written scripts we uncovered a novel switch involved in activation of GPCRs and we confirmed the evolutionary conservation of several other previously published switches in GPCRs. The network in which this novel switch is inserted links the GPCR binding cavity to the intracellular space, and gives a quantitative foundation to the GPCRs activation models proposed in the literature. This revealed therefore an unexpected link between the activation mechanism and the downstream signaling pathway.

2

G-protein coupled receptors

How receptors work has fascinated researchers for more than a century. The first idea regarding the existence of receptors was made by a british pharmacologist, in 1905, J. N. Langley. He wrote the following: So we may suppose that in all cells two constituents at least are to be distinguished. The chief substance which is concerned with the chief function of the cell as contraction and secretion and receptive substances which are acted upon by chemical bodies and in certain cases by nervous stimuli. The receptive substance affects or is capable of affecting the metabolism of the chief substance". Thus receptors can be considered proteins or protein complexes that allow the communication between the external of the cell with its internal environment. GPCRs were recognized as receptors only in the late 1980s, when the similarities in sequence and transduction mechanism between rhodopsin and β-adrenergic receptors had been first noticed [8]. Today we know that the G-protein coupled receptors (GPCRs) superfamily rapresents the largest group of plasma eukaryotic membrane receptors [1] (Figure 2.1). They cover an immense variety of physiological functions. Endocrine, neurological, cardiovascular and reproductive functions also depend in great part from GPCRs signal transduction [9]. Many neurotransmitters and hormones such as serotonin, endogenous opioids, glutamate, acetylcholine, histamine, melatonin, secretin, orexin, glucagon, vasopressin, oxytocin target GPCRs [9]. Sensory information is first received and transmitted by GPCRs for vision (through the opsins), olfaction (through the olfactory receptor family) and several taste senses (umami and sweet tastes via the TAS1R receptors, bitter taste via the TAS2R receptors). Given this, it is not surprising that GPCRs are a primary pharmaceutical target. They target between 30% and 40% of the marketed drugs [10, 11].

The huge role of GPCRs in human physiology derives from the evolutionary diversity in the sequence encoding the seven transmembrane domains, which form the core of the receptor common to all GPCRs. According to the GRAFS classification [12], human G-protein coupled receptors (hGPCRs) are divided in five different families, i.e.: Rhodopsin-like (or class A), Glutamate (or class C), Adhesion (or class B2), Frizzled (or class F) and Secretin (or class B1). They all share a common seven transmembrane (TM) helix bundle shape [13], [14] (see Figure 2.2):

- Class A or **Rhodopsin family**: In vertebrate genomes, the class A/Rhodopsin family constitutes the vast majority of GPCRs. In humans, class A GPCRs are by far the largest family which includes 670 members) [15] and even in case we don't consider the almost 400 olfactory receptors, class A is still dominant, with about 270 remaining members. We will later discuss the Class A in a detailed way, given its size and importance (see Figure 2.3).
- Class B1 or **Secretin family**: Class B1 is a small family (15 members in humans) [15] of peptide hormone-binding GPCRs, unique to Metazoa. They contain a hormone-binding domain located in the extracellular part; receptors belonging to this class are the glucagon receptors, the corticotropin-releasing hormone receptor or the parathyroid hormone receptors. They share between 21% and 67% of sequence identity, with a variable N-terminal region. [15] (see Figure 2.3).
- Class B2 or **Adhesion family**: Class B2 is the second largest family of GPCRs in the human genome, with 33 members. So far, it is not well studied and relatively obscure, with several members still orphaned [15]. Characterized Class B2 receptors has a large highly glycosylated N-terminal domain and bind extracellular matrix proteins. These features displays considerable variability between members. The extracellular domain self-cleaves and is functional as two non-covalently attached subunits; an internal N-terminal epitope acts as a cryptic ligand [16]. This domain can include protein domains of diverse origin, such as laminin or Ig-like domains [15] (see Figure 2.3).
- Class C or **Glutamate family**: Class C in humans includes 22 receptors with very different physiological roles. Apart from the eight metabotropic glutamate receptors (mGluR1-8), that name the family, they include a calcium-sensing

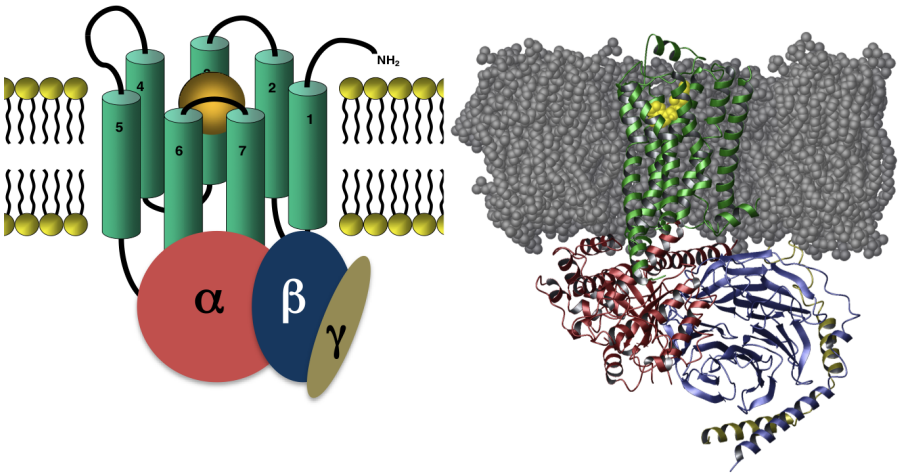

Fig. 2.1. Schematic and atomistic representations of a GPCR embedded in the membrane, bound to a ligand in the extracellular part and to the trimeric G-protein in the intracellular part. [Image from the Costanzi Research Group]

receptor, two GABA receptors and especially the TAS1R subfamily of taste receptors (3 members), which include both the umami and the sweet taste receptor [15]. Many members have at least some affinity for aminoacids. The N-terminus of Class C receptors is highly variable but the TAS1R and glutamate receptors contain a characteristic large extracellular N-terminal domain, dubbed the Venus flytrap domain, which binds ligands [15]. From different studies emerged that this domain has structural homology with bacterial aminoacid-binding proteins [17]. The N-terminal domain is often connected to the 7-TM GPCR transmembrane core by a cysteine-rich domain. Intriguingly, at least the mGluRs and TAS1Rs function as heterodimers [18, 19]: TAS1R1/R3 senses glutamate and signals the umami taste; contrarily, TAS1R2/R3 senses sugars and sweeteners, signalling the sweet taste [20] (see Figure 2.3).

- Class F or **Frizzled family**: In humans, composed of ten Frizzled receptors (FZD1-10) and the Smoothened receptor (SMO). They contain a N-terminal domain of 200-320 aminoacids, connected to the transmembrane domain by a variable linker [15]. FZDs are receptors for the family of Wnt glycoproteins [21], while SMO is a component of the SMO/Patched/Sonic Hedgehog complex [22]. Both receptors have been identified as involved in embryonic development and cancer development (see Figure 2.3).
- **TAS2Rs** (Bitter Taste receptors): The position of TAS2Rs (25 members in humans) in the GPCRs phylogenetic tree has been always debated because iniatially they were considered a divergent subfamily of the Class F [23]. However recent phylogenetic studies on the origin of GPCRs families has cast doubt on this classification, and places instead TAS2Rs as a branch evolved from the Class A/Rhodopsin family [24]. They significant lack of the N-terminal domains; all evidence indicates that they bind their ligands in the orthosteric binding cavity within the 7-TM bundle, as most class A receptors [25–27]. The ones characterized so far bind a surprisingly diverse range of chemically diverse agonists [28, 29] (see Figure 2.6).

As we have already pointed out, GPCRs are versatile proteins that regulate a diverse array of intracellular signaling cascades in response to the interaction with a ligand. As their name suggests, GPCRs mostly transduce their signal by binding and activating guanine-nucleotide binding heterotrimeric G-proteins. The heterotrimeric G-proteins are constitute of three subunits, Gα, Gβ and Gγ. Gβ and Gγ are always bound together. They are normally associated to the plasma membrane in the inactive state of a receptor [31]. Upon GPCRs activation through the interaction with an external ligand, the receptor is able at this point to activate the Gα subunit. The activation signal of GPCRs begin with a physical interaction with the heterotrimeric G-protein, as well as with the G-protein-coupled receptor kinase (GRK)-mediated phosphorylation and arrestins coupling, a G-protein indipendent way. Once it has been activated, the Gα exchanges GDP for GTP. The GTP-bound Gα dissociates from the GPCR and also from the Gβ-Gγ complex, affecting further effectors downstream (see Figure 2.4) [32, 33]. Gα subunits have an intrinsic slow GTPase activity: GTP is hydrolyzed to GDP and the Gα subunit reassociates with the Gβ-Gγ complex, restoring finally the initial state [34]. The human genome contains 17 different genes for Gα subunits divided principally in four subclasses: G$\alpha_{i/o}$, G$\alpha_{q/11}$, Gα_{s}, G$\alpha_{12/13}$. Each GPCR can bind one or more

Gα belonging to different subclasses. The β2 adrenergic receptor for example can couples both to G$\alpha_{i/o}$ or Gα_s [35]. Gα_s activates the cAMP-dependent pathway by stimulating the production of cyclic AMP (cAMP) while G$\alpha_{i/o}$ inhibits this pathway. G$\alpha_{q/11}$ stimulates the membrane-bound phospholipase Cβ starting a downstream signal transduction pathway for many hormones and finally G$\alpha_{12/13}$ are involved in Rho family GTPase signaling, controlling the cell cytoskeleton remodeling, and regulating cell migration. The Gβ-Gγcomplex can also function as an effector, regulating ion channels such as the muscarinic K+ channel, adenylate cyclases, phosphatidylinositol-kinases and phospholipase Cβ [35]. While there are 5 subtypes of Gβ subunits and 11 types of Gγ subunits, the downstream effects of most Gβ-Gγ complexes are not substantially different [33].

The maiority of GPCRs show also a certain basal activity that can be modulated by ligands with different efficacy. Full agonists induce a maximum in the signaling response, whereas partial agonists and inverse agonists promote submaximal signaling or decrease the basal activity, respectively. For most GPCRs, subsequent agonist-dependent phosphorylation of the receptor C-terminus or intracellular loops by GRKs promotes arrestin recruitment, which, in turn, induces receptor desensitization by sterically blocking additional G-protein coupling and stimulating receptor internalization. The arrestins family includes four members: β-arrestin$_{1-4}$. β-arrestin$_1$ and β-arrestin$_4$ are also called visual arrestins due to their expression in the retina [36]. Initially they were called "arrestins" because they were found to stop the GPCRs signal transduction. Arrestins however are able to promote signalling by theirselves starting an alternative way of the transduction of the signal [37]. Indeed biased signaling by GPCRs means that some ligands preferably activate either G-proteins or β-arrestins pathway to transduce

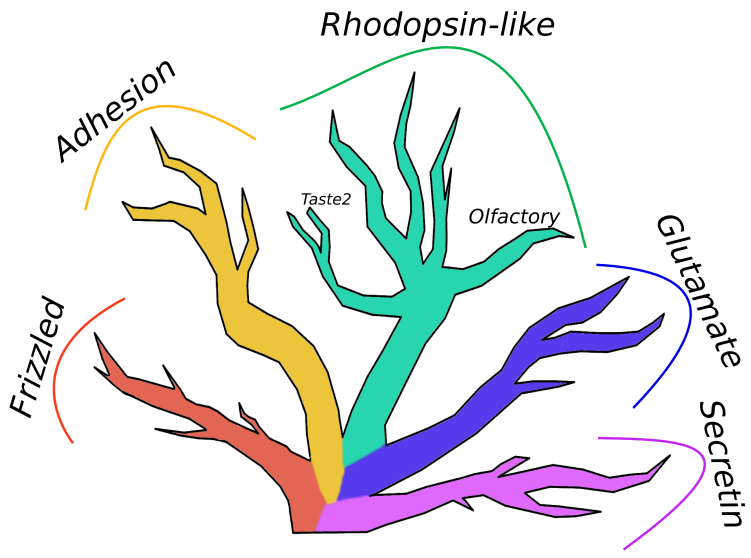

Fig. 2.2. The classical human GPCR tree, according to the standard GRAFS classification.

the signal. They selectively activate certain receptor-associated pathways at the expense of others (see Figure 2.5).

How GPCRs evolved? GPCRs are the results of an early eukaryotic innovation. The origin of GPCRs is still very much debated and they must have evolved before or around the time of the last common eukaryotic ancestor [40]. Plants are one of the few major eukaryotic clades with a poor GPCR repertoire, even if at least one probable GPCR and several further candidates have been identified in Arabidopsis [34]. GPCRs share visible basic structural similarity (a structurally similar 7-transmembrane alpha helical core) with prokaryotic proteins such as proteo-, bacterio- and halorhodopsins [40]. The sequence divergence between prokaryotic rhodopsins and eukaryotic GPCRs is too high to confirm a true evolutionary relationship; nonetheless structural lines of evidence, e.g. in the activation network might point at a common origin [41]. It has been proposed that exon shuffling has led to the emergence of GPCRs from prokaryotic 7-TM proteins, due to the observation that the highest sequence similarity is found between non-homologus helices [40, 42]. The low sequence identity between GPCRs, often below than 30% even between members within the same family, makes phylogenetic studies difficult. However, by using several sequence alignment approaches, a

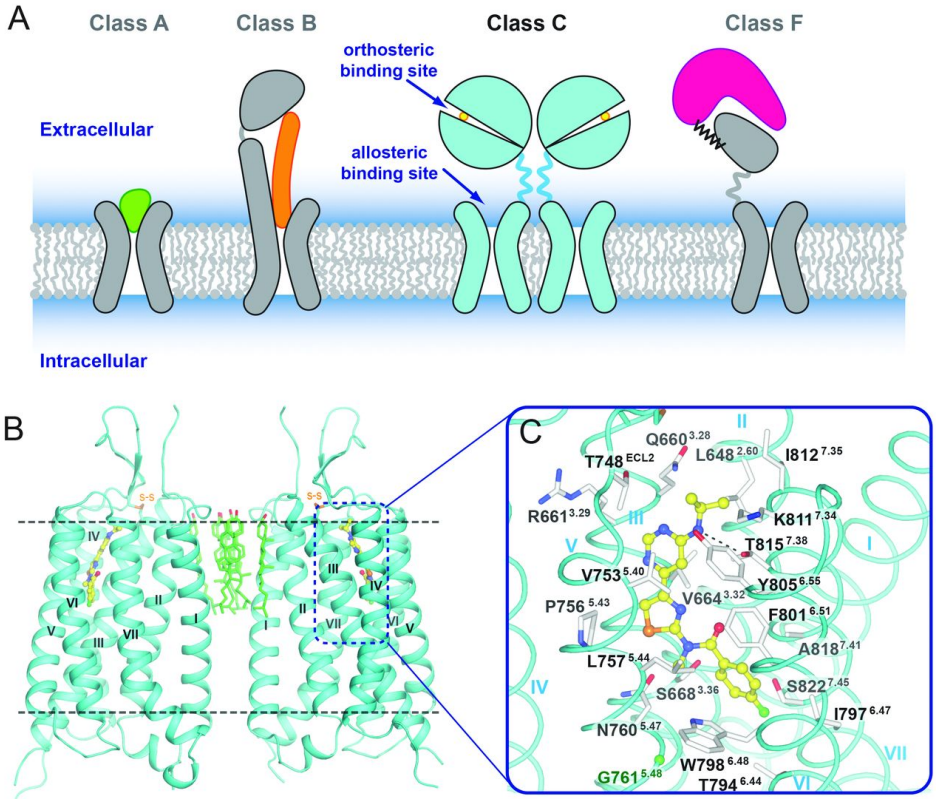

Fig. 2.3. General structure of Class A GPCR with detailed functional regions [30].

suggestion of phylogeny of GPCRs has been identified [43]. According to it, originally have been existed two classes, the cAMP and the Class C, that are the two earliest diverging families. The cAMP family then branched into the Class A, Class F and Class B2. From Class B2, early branched B1. TAS2Rs branch and diverge from the Class A family, within the vertebrate lineage [43]. A further study further refined the picture of GPCR evolution within the context of eukaryotic evolution [24]. The common ancestor of eukaryotes would already have included both cAMP family receptors and Class C receptors in its genome. The Archaeplastida lineage, including plants, retained only the ancestral cAMP family; Alveolata retain both cAMP and class C. In unikonts GPCR diverge ulteriorly, with Class F and B2 evolving from the cAMP receptor family. The Class A split later from the cAMP

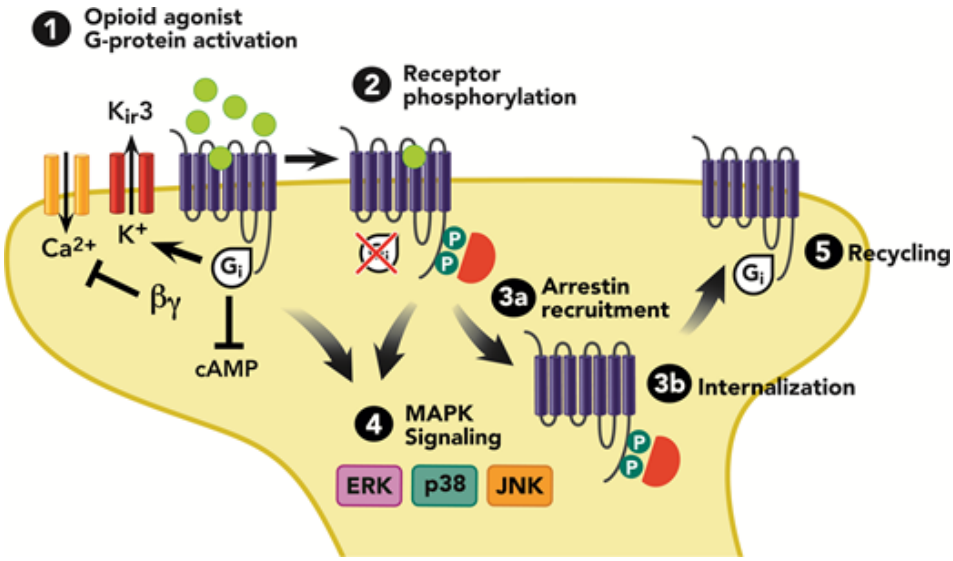

Fig. 2.4. G-protein Signaling Cycle [38]

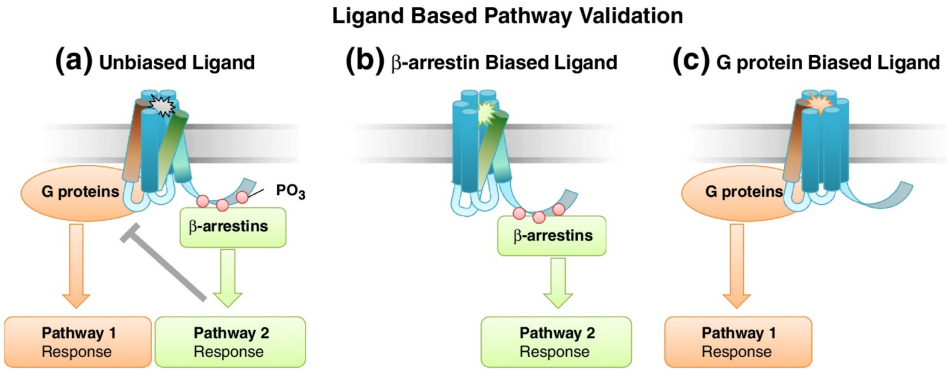

Fig. 2.5. GPCR signaling in the presence of a biased ligand [39]

receptors but still well before animal evolution, in the early opisthokont lineage; while family B1 diverges from B2 much later, after the split between Cnidaria and Bilateria [24] (see Figure 2.7).

Let's us now introduce and discuss in a detailed way the GPCRs Class A. The GPCRs Class A, also known as the "Rhodopsin-like family", is the largest subfamily of GPCRs, which includes hormones, neurotransmitters, and light receptors and accounts for around 80% of GPCRs. It plays lots of key physiological roles in human physiology with its members and due to this it is the most studied and characterized GPCRs family. In humans, it includes about 271 non-olfactory members and almost between 340 and 460 olfactory members [23, 44]. The non-olfactory receptors are principally grouped in four major subfamilies:

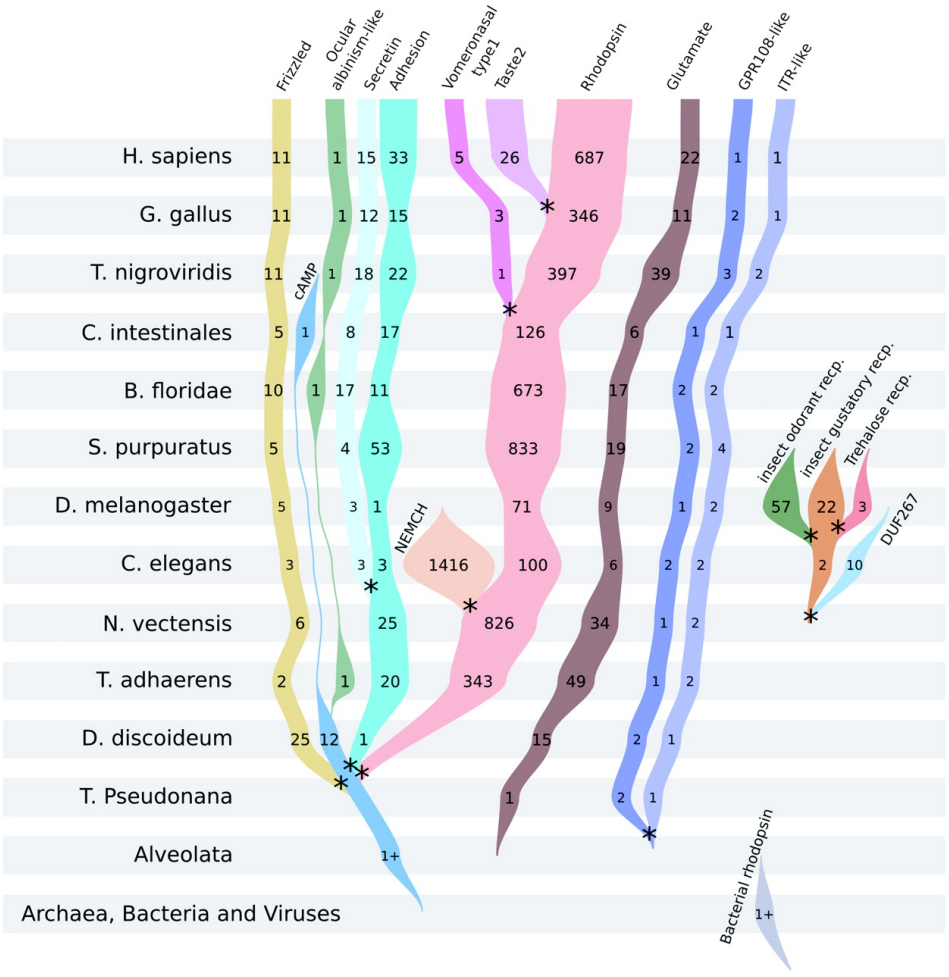

Fig. 2.6. Phylogeny of GPCR families, from [43]. The diagram also includes putative GPCRs that have been shown to actually not be GPCRs, such as the insect olfactory receptors (isolated tree on the right)

- α group: 89 members in total. It includes subgroups such as the prostaglandin receptors, all the amine receptors (serotonin, muscarinic, adrenergic etc.), the opsins, the melatonin receptor cluster and the MECA cluster (melanosin/endothelial differentiation factor/cannabinoid/adenosin receptors). Perhaps the best known group, with several members such as rhodopsin, A_{2a} and $\beta 2$ adrenergic receptor characterized structurally very deeply.
- β group: 36 members in total, all peptide receptors-notable members include the oxytocin receptor, the thyrotropin-releasing hormone receptor, the ghrelin receptor, neurotensin receptor, and vasopressin receptor.
- γ group: 59 members in total. It includes three subgroups of which perhaps the most notable is the opioid receptor subfamily, which has been completely structurally solved. Other clusters include the chemokine receptor subgroup and the two MCH receptors.
- θ group: 58 members in total (plus olfactory receptors). Apart from the olfactory receptors, members of this group include the MAS oncogene receptor

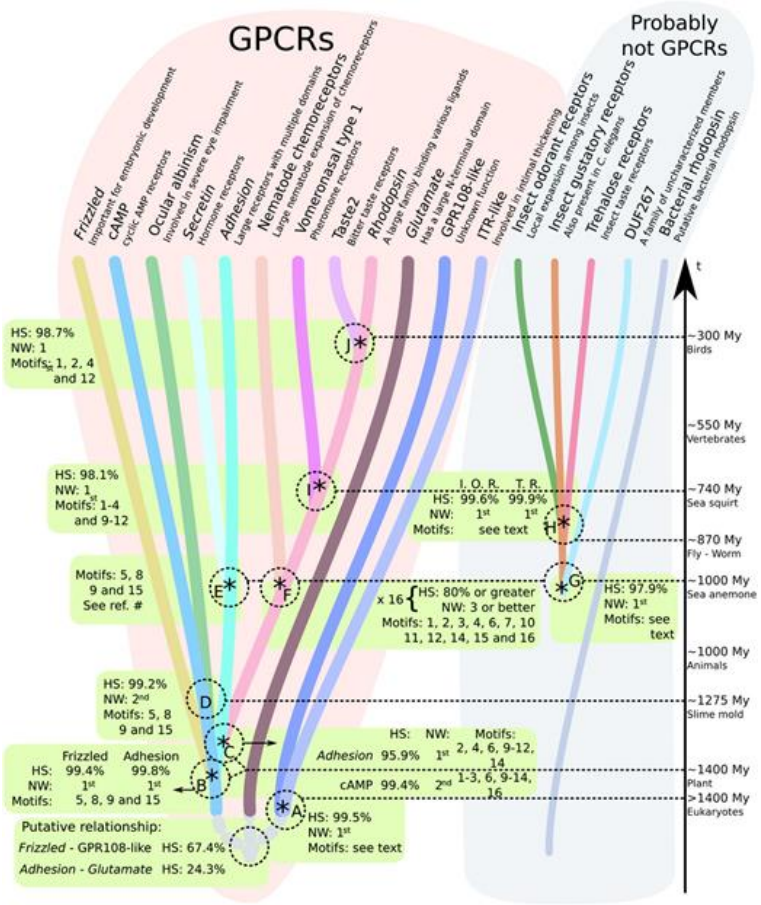

Fig. 2.7. Phylogeny of GPCR families, from [43]

cluster, the glycoprotein hormone receptors (FSHR, TSHR, LHCGR), leucine-rich repeat-containing GPCRs (LGRs), and the purin receptor cluster, which includes the nucleotide receptor (P2Y receptors), the formyl peptide receptors (FPR1-3) and others.

Class A GPCRs share several features between its receptors, such as a compact 7-TM fold, a small N-terminal domain, a variable (in shape and conservation) extracellular loop 2 (ECL2). As far as it is known today, they all bind their native ligands in the so-called orthosteric binding cavity within the 7 TM bundle (in contrast with Class B1/B2 and Class C receptors, which feature distinct extracellular ligand-binding domains), even if an allosteric binding site has been identified and studied recently [45–49].

Regarding the transmembrane helixes, GPCRs are characterized by a common share 7-TM topology, which is well conserved even among divergent GPCR classes. The N-terminal is located in the extracellular part of the cell and the C-terminal in the intracellular part. The seven helices are connected by three extracellular and three intracellular loops (ECL1-3 and ICL1-3) (see Figure 2.8). The seventh helix is often followed by an intracellular amphipatic helix (H8) that is localized parallely to the membrane. This helix, in the majority of the cases is not crystallographically solved but replaced by homology modeling. Functionally, the structure can be understood as formed by three regions (following [14]): an extracellular region which modulates ligand access; a transmembrane region that binds ligands and transmits the activation signal and an intracellular region that interfaces with cytosolic partners, most importantly Gα and β-arrestins. The C-terminal sequence can contain some palmitoylation sites [30], necessary for dimerization of GPCRs. In general the sequence and structure of the intracellular half is more conserved than the one of the extracellular half [50].

Now we are going to treat these different parts more in detail: Extracellular loops (ECLs) in general, and especially ECL2, are considered the most variable parts in a GPCR. A disulfide bridge connecting the ECL2 with the end of TM3

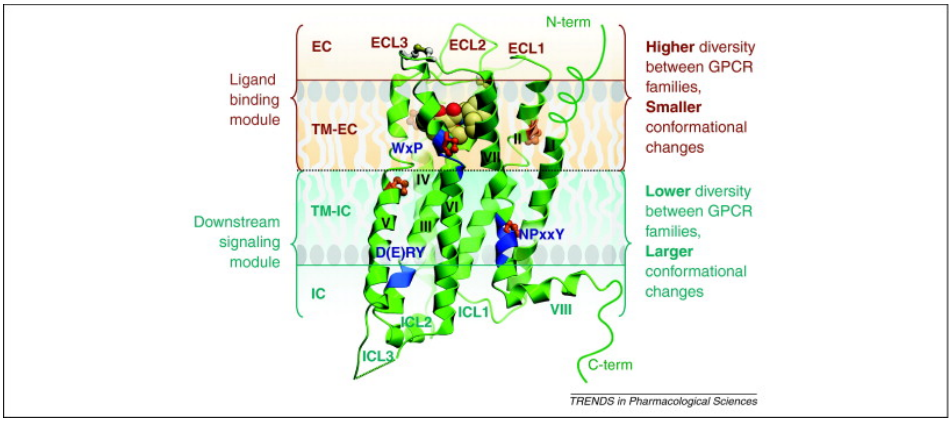

Fig. 2.8. General structure of Class A GPCR with detailed functional regions [30].

is very well conserved in class A; however sequence, secondary structures and positioning with respect to the 7-TM bundle are largely different. Peptide binding receptors for example tend to have a larger ECL2 and to shape it in a β-hairpin, while β-adrenergic receptors feature a short α-helix. ECL1 and ECL3 are shorter (5-6 and 6-8 residues long on average, respectively [30]) but still feature distinct conformations in different receptors, sometimes stabilized by more than one disulfide bond or salt bridges [30]. The ECL2 contains a glycosylation site in about 32% of class A receptors, and glycosylation could help to stabilze the ECL2 conformation [51]. Finally, ECL2 can also have an important role in activation, acting surprisingly, as a damper of the inactive/active transition and locking TM5, TM6 and TM7 extracellular boundaries in an inactive conformation [51] (see Figure 2.9).

Regarding the 7-TM bundle, the characteristic core of all GPCRs is the bundle of seven transmembrane *alpha*-helices (7-TM bundle). All seven helices completely span the plasma membrane and form a distinct binding pocket located in almost all the extracellular half of the receptor [30]. The helices are only weakly tilted with respect to the membrane plane, with the distinct exception of TM3, which is instead markedly slanted and crosses the inner part of the TM bundle [14].

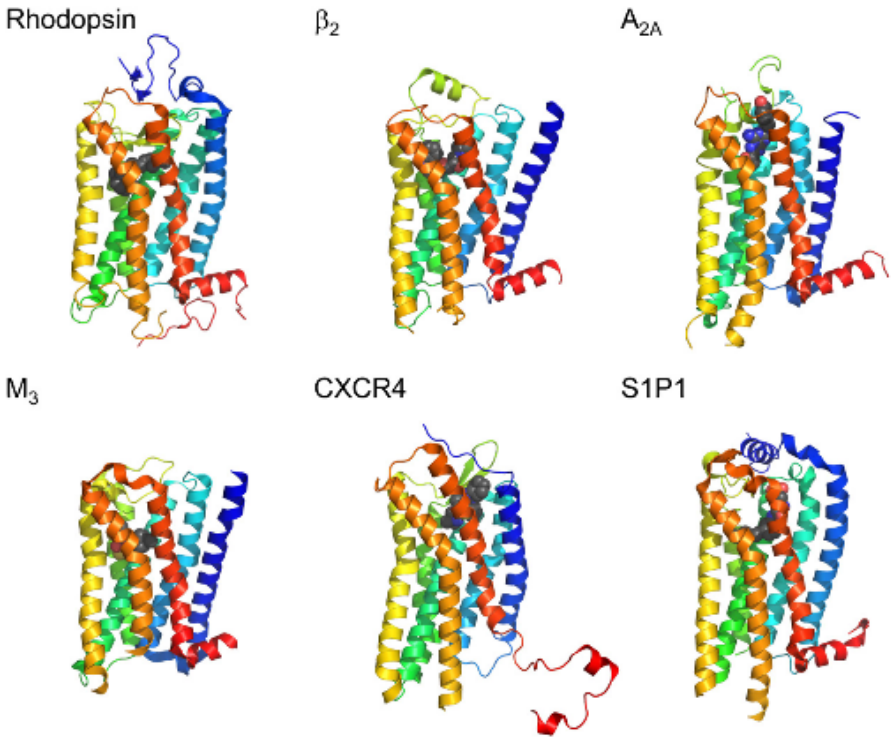

Fig. 2.9. Side-by-side comparison of the crystal structures of six representative receptors. [52].

As such, contacting most of the other helices, TM3 acts as a central structural hub in GPCRs activation [14]. Despite the sequence diversity of GPCRs, that can range often below 30%, even between members of the same subclass, the structure of the 7-TM bundle is remarkably well conserved. The average RMSD between different GPCRs rarely exceeds 3Å [30]. Local variations are however present, especially at the extracellular half, where deviations in the position of the extracellular helical tips can reach 7Å [30]. The structures of the helices is in fact not regular, most often in the extracellular half of TM2 and TM5 [14, 53, 54]. Other bulges/irregularities can be found however in all TMs, depending on the receptor [54]. The overall conservation of the TM helical structure is high enough to allow a direct residue-to-residue mapping between different GPCRs, which has been initally formalized for class A as the Ballesteros-Weinstein numbering system [55]. To take into account the occasional 1-residue insertions or deletions present across the TMs of some receptors [14], [55], [54], a structurally-based update to the Ballesteros-Weinstein numbering has been published, and it is available on the GPCRDB web-server [56]. This is so-called generic GPCRdb numbering that will be used throughout all this thesis. The TMs contain several short conserved sequence motifs, which have functional significance, especially in receptor activation. Among them the most important are the $C_{6.47}WxP_{6.50}$ motif, the $N_{7.49}PxxY_{7.53}$ motif, the $D(E)_{3.49}RY_{3.51}$ motif [57], [58]. These motifs are quite conserved beyond class A. Their importance will be discussed in details later. The intracellular region of GPCRs, contrarily to the extracellular part, is more structurally conserved. This is, in a certain sense, expected, since the role of the intracellular region is that of performing large-scale motions related with activation and interfacing with partner proteins and effectors. The short (six residue) ICL1 in particular shows a similar conformation in most crystallized GPCRs [30], ICL2 seems to either fold in a short α-helix parallel to the membrane or exist as an unstructured loop. The two conformations can be observed in crystal structures of the same receptor, and even by different receptor molecules in the same asymmetric units [30]. ICL3 is the most variable intracellular region of GPCRs, spanning a range of sizes from five to hundreds of residues considering different receptors. There is evidence from proteolysis experiments that it is a very flexible region [30]. ICL3 is often substituted with a fusion protein that helps crystallization, and hardly ever can be solved. As such our knowledge of ICL3 structure is very limited. This region is however functionally very significant, since it is, along with ICL2, the main determinant of G-protein specificity [30], [58] and activation. Helix 8 (H8) is a non-transmembrane helix that runs parallel to the membrane, conserved in almost all GPCRs whose structure has been solved so far (also beyond class A), featuring a conserved F(RK)xxFLxxxLF amphiphilic motif and palmitoylation sites [30].

The orthosteric binding cavity is thus considered the canonical binding site and it is very well studied. Unfortunately sometimes it does not offer the variability necessary to study and propose a new ligand (especially when it is super well conserved as the case of Muscarinic receptors). Consequently, in the last five years, allostery became an appreciated phenomenon in GPCRs (see Figure 2.10), for both artificial and endogenous ligands [59]. From a drug discovery point of view, many GPCRs seem to possess multiple druggable allosteric sites. The existence of allosteric sites in GPCRs has been dramatically evidenced styding the muscarinic

receptors and in the last years after the crystal structure of the P2Y1 purine receptor in complex with the antagonist BPTU [45]. An unusual binding site had been found also for the allosteric agonist TAK-875 in the FFAR1 crystal structure, the ligand inserting between TM3 and TM4 [60]. Thus, contrarily to the orthosteric binding site, the allosteric binding site involves different positions in different helixes accordingly to the receptor that we are considering. Extracellular loops have been recognized experimentally equally important in receptor specificity and ligand binding of many different GPCRs [61], [62], [63]. The binding mechanism of ligands in the GPCR allosteric cavity was investigated by large scale molecular dynamics simulations for the muscarinic [64] and the $\beta 2$ adrenergic receptor [65]. A distinct vestibular binding site, close to the extracellular space was discovered. Ligands spend a significant time in the vestibular site en route to the orthosteric binding site [66], [67].

Another particular allosteric effects on GPCRs and the on transmission of the signal is the sodium ion that at physiological concentrations (140 mM) were observed in several GPCRs since the 1970s [48]. Mutagenesis experiments evidenced that the highly conserved residue $D_{2.50}$ mediates the allosteric effects of sodium [48]. Confirm of the binding of Na+ in GPCRs however was found only in 2012, within the crystallographic A2a adenosine receptor structure [68]. A single sodium ion is bound right below the bottom of the orthosteric GPCR binding cavity, coordinated by several water molecules and conserved residues. This pocket is structurally and sequence-wise conserved in other inactive GPCR structures even if the sodium electron density was not resolved, but it collapses in active GPCRs (see Figure 2.11). Despite the fact that the coordination of the sodium shows some difference between receptors, all of them share the same positioning below position 6.48 and coordinated by some waters and $D_{2.50}$; position $N_{7.49}$ of the NPxxY

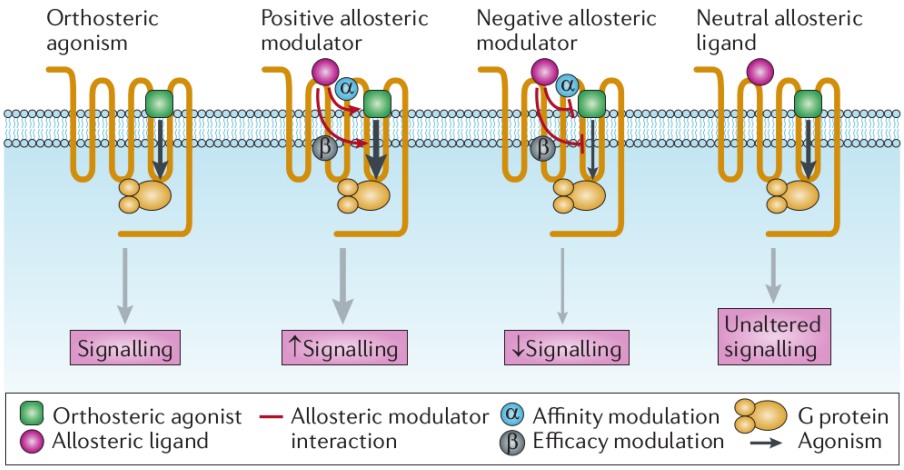

Fig. 2.10. How different type of allosteric ligands influence the transmission of the signal. [59].

motif is also close to the coordination site [48]. The collapse of the sodium binding site in the active structures and the rearrangement of the pocket polar network suggests an important role of this site in GPCRs activation. We studied this particular binding site for the GPR3 receptor. This work will be discussed in the in the Results section.

How GPCRs are crystallized? Membrane proteins are notoriously hard to crystallize, and as such our structural knowledge of the GPCRs family is relatively limited. While some experimental data on GPCRs helical arrangement were already found in 1993 [69], GPCR structural biology officially started in 2000, with the publication of the first crystal structure of rhodopsin [70]. Rhodopsin however remained the only crystallized GPCR until 2007, when the first structure of $\beta 2$ adrenergic receptor was published [71]. The structure was not only a indicator of how a ligand is located and interact with the receptor, but it paved the way, experimentally, to all the other GPCR crystal structures. In less than a decade from the $\beta 2$ adrenergic receptor structure, the Protein Data Bank accumulated structures for 50 vertebrate GPCRs, plus one mollusk (squid rhodopsin) and one viral (US38) GPCR; both a GPCR/G-protein and GPCR/β-arrestin complex have been solved so far. GPCRs present a series of difficulties when trying to solve their structure: generally low native expression, low solubility and high flexibility [72], [73]. Large-scale expression has been achieved first by optimizing expression systems in insect cells (via baculovirus vectors), but also Pichia pastoris yeast or even Escherichia coli bacteria have been used successfully [72]. Still, GPCRs crystals are grown slowly in lipidic environments, and as such they are tiny and fragile [73]. They thus require specialized data collecting strategies at X-ray crystallography facilities [72]. Structural stabilization and engineering of crystallizable surfaces has been achieved mainly in three ways [72]:

1. Co-crystallization with some antibody fragments or nanobodies in order to reduce the conformational flexibility and yielding a larger polar surface amenable

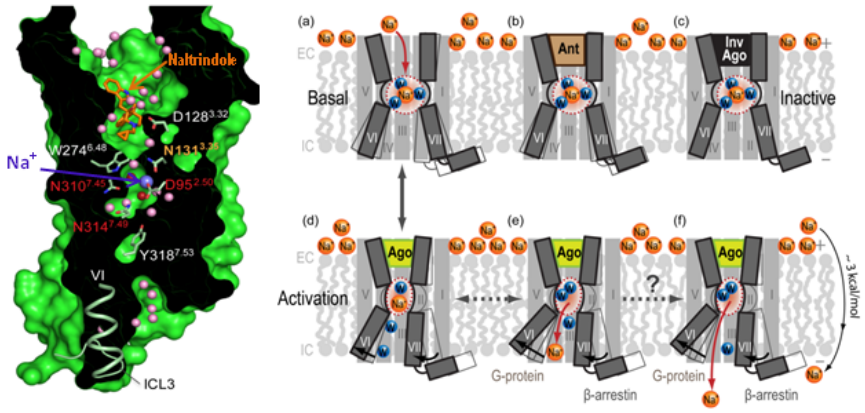

Fig. 2.11. Allosteric Binding of Na+ and its functional role in the activation mechanism of GPCRs (image from the Katritch laboratory).

to crystallization. The usage of nanobodies has been especially important to obtain active GPCRs structures [72], [73], [74].
2. Fusion proteins such as T4 lysozyme (T4L). These proteins are usually inserted in place of the highly flexible and variable ICL3. Substituing T4L with ICL3 removes the flexible loop and restricting the motions of TM5 and TM6. Fusion proteins also provide an additional polar surface for crystal formation. [72], [73]. It has been suggested however that they could lead to some artefact, e.g. preventing native ionic interactions and difficulty in building a proper topology for molecular dynamics simulations. [73].
3. Thermostabilization by mutagenesis. Mutants are selected for their ability to bind ligands at increased temperatures. Thermostabilized mutants are often used in conjunction with the strategies above. Thermostabilizing mutations do not significantly change the structure of GPCRs [72, 73].
4. Addition of high stability ligands allows the receptor to fall preferentially into a single energy minimum, making the population homogeneous and thus helping nucleation and packing. It also helps expression, by reducing receptor recycling at the cell surface [72].

However, even with the contemporary boom of GPCRs structures crystallization, structural coverage of the GPCR superfamily is very small: structurally solved receptors are less than 5% of the total number of human receptors. At the current rate of structural coverage expansion, it would take more than a century to obtain crystal structures for all the human GPCR superfamily. Therefore, structural prediction techniques are more than necessary in order to study these receptors. Currently, efforts in modeling GPCRs and GPCR/ligand complexes are assessed every few years in the GPCR Dock competition [75–77]. The homology modeling technique is particularly suitable for s strucuture prediction due to the high structural conservation of the 7-TM bundle. Although homology modeling of GPCRs shows that there is a rough correlation between sequence identity and model RMSD from the real structure [76, 77], in some cases sequence identity with the template is not necessarily the best predictor of model accuracy. Other structural descriptors (e.g. conservation of residues through helices, native ligand size and correlation with binding cavity volume, etc.) might be more important [78, 79]. Flexible docking is also helpful in producing more accurate models of the binding cavity and the ligand/receptor binding mode [80]. However, the recovery of correct ligand/receptor contacts, even with the best models approaching experimental accuracy, does not go beyond 50% for static homology models [77]. Nevertheless, virtual screening on GPCRs homology models has been moderately successful in the last years[80]. Combined techniques such as modeling and docking coupled with molecular dynamics simulations and backed by multiple experimental data can yield accurate predictions [81]. However, as for many proteins, the greatest challenge in GPCRs structure prediction is loop modeling. While the error in backbone prediction ranges between 1 and 4Å of RMSD, ECL2 error is almost always beyond 4Å and it can easily reach beyond 10Å [77]. The error is such big that docking of ligands on GPCR models can perform better when loops are removed [82].

How GPCRs transduct the signal? The transduction of the GPCR signal requires the receptor to pass from an inactive (R) to an active state (R), and

the latter to be able to bind/activate downstream effectors such as G-proteins. One of the best and most approved theoretical frameworks for GPCRs activation is the conformational selection theory [78, 83]. Following this model, GPCRs constantly explore active and inactive states, and ligands act by binding to the active state, stabilizing those conformations while destabilizing them in the inactive state. However for at least some GPCRs, such as angiotensin receptors and rhodopsin, induced fit, that is a mechanism where ligands bind directly the receptor to the active state, providing an increase in terms of energy, may be relevant as well [83]. Many GPCRs are known to possess also a significant basal activity [84, 85], thus supporting the hypothesis that ligands act by shifting an already present equilibrium. G-proteins binding acts as the final activation step, inclining definitely the energy landscape towards the activated state (Figure 2.12)[83]. Studies on the rhodopsin receptors have clarified that the active-inactive transition is not binary, but requires a multiplicity of metastable receptor states [86]. This has also been clarified by accelerated molecular dynamics simulations on the A2a adenosine receptor [87]. Agonists, inverse agonists and partial agonists modulate the relative depth of the various energy minima. Millisecond-scale massively parallel molecular dynamics simulations have shown that, in $\beta 2$ adrenergic receptor, the transition between active and inactive conformations is strongly shifted by the presence of agonists or inverse agonists, and that it can follow a multiplicity of structural pathways [88]. Also, multiple final active states could exist, each leading to binding of a different effector and thus triggering different downstream pathways. When ligands preferentially trigger alternative active conformations, this is known as biased agonism. These states can be trigged by external influences, e.g. phosphorylation of the receptor favouring pre-existing conformations that lead to arrestin binding [89] or by mutagenesis [90]. Today there are available different GPCR structures in both active/partially active and inactive states, that allow us to better understand the conformational changes related to activation. However, being them static structures, it is hard to identify the pathway by which such conformational changes propagate. Nevertheless, the comparison of inactive and active structures, coupled to mutagenesis data, allow us to gather information on the possible switches involved in GPCRs activation. Active structures are so far available only for Class A GPCRs (except for the Class B glucagon-like receptor (GLP-1), solved in an active state), so structural details of activation are basically unknown in other classes. The crystal structure of the active GPCR/Gα complex has been resolved in 2011 [35], a landmark achievement that landed Brian Kobilka his Nobel Prize a year later. Comparison of active and inactive structures has permitted to identify large-scale motions common to all active-state GPCRs. The most dramatic is the motion reagrds the TM6, which rotates and swings outward by almost 10-14Å [91–93], opening a crevice that allows for the insertion of the Gα C-terminal helix within the GPCR intracellular side [35]. At the same time there is an upward motion of TM3, inward motions of TM5,TM7 and TM1 (Figure 2.13) [93]. On a smaller scale, several conformational changes have been related to class A GPCR activation. Here we are going to list only the most relevant. A general pattern is that almost all switches are broadly but not absolutely conserved. For example, for the rotamer toggle switch (see below) there is a substantial 20% of class A receptors with a different chemical identity for position 6.48. In the

conformational selection model of GPCR activation this is not an insormountable problem, given that switches do not work necessarily as dominoes, rolling conformational changes in a sequential, inevitable manner; rather they are a series of cascades interactions controlling the equilibrium between R and R forms. If one switch weakens or modifies, others can strengthen to take its place [94] changing in this way the equilibrium. Thus the universality of switches has to be understood statistically rather than absolutely, and their contribution energetically rather than only structurally. This has consequences for the work described in this thesis.

One of the first identified switches involved in activation was the rotameric switch of the highly conserved (W in 65% of Class A receptors; aromatic in over 80% [93, 94]) residue W6.48 , belonging to the conserved Class A $C_{6.40}WxP_{6.50}$ motif. This is often called the rotamer toggle switch [96] or transmission switch. W6.48 is located at the bottom of the orthosteric binding cavity and it either contacts ligands or it is affected by the interaction with ligands contacting neighbouring residues. In active GPCR structures, often W6.48 moves from pointing towards TM7 to pointing towards the binding cavity [97] but it can take also alternative conformations,e.g. in the active neurotensin receptor [98]. W6.48 is involved in several interhelical interactions with TM3 and TM5 (notably with positions L3.40, P5.50 and with L5.51 [99]), which in turn can propagate the conformational change downstream to the intracellular side. Mutagenesis experiments confirm that W6.48 is a critical residue in GPCR activation, its mutations often leading to higher or lower constitutive activity [100]). Molecular dynamics simulations and experimental structures suggest that the rotameric state of W6.48 can be controlled by agonists or inverse agonists. There have been however doubts in the literature on the universality of the role of W6.48, noticing that in some otherwise structures its rotameric state does not change significantly [101].

The ionic lock is a fairly conserved interaction complex including a salt bridge between neighbor residues D(E)3.49 and R3.50 , plus their interaction with E6.30 and T6.34 [85]. It has been already noticed in the first solved rhodopsin structure [70]. The conservation of the (D/E)RY motif (96% for R 3.50 and 88% for D(E) 3.49), coupled to the observation that, in rhodopsin, it opens in active structures, led to the hypothesis it was a critical activation swtich. However in many solved inactive GPCR structures the ionic lock is not conserved, and as such its universality is now decreased [70]. It could have a more general validity as concentration of polar interaction between TM3 and TM6 around R3.50, for example including the hydrogen bond between Y3.60 and H 6.31 in B2AR, or the R3.50 /Q6.36 hydrogen bond in the histamine H1 receptor [70] or can be specifically important only for some residues. Similar, even if less discussed, is the 3-7 lock, the interaction between (in rhodopsin) K7.43 and E3.28.

Perhaps the closest thing to a universal activation switch in GPCRs is the unlocking of the hydrophobic hindering mechanism (HHM) (Figure 2.15). This is a thick network of hydrophobic interactions pivoted around the triple helical interaction X6.40/F6.44-L3.43 (where X=aliphatic hydrophobic residue). This interaction is locked in the inactive state and it changes n the active state by the rotation of TM3 and concurrent upward motion of TM6 [93]. The unlocking of this hydrophobic interhelical interaction is associated with the switching of several other hydrophobic contacts, such as the X3.40/L3.43-P5.50, X3.40/L3.43-L2.46 [93], and

the motions of other residues such as positions 2.43 and 3.46 [93]. This unlocking is in turn related to the motions of Y7.53 and the opening of a water channel. Extensive mutagenesis data confirm the role of these residues in activation, as mutations correlated with decreased capability for activation or increased basal activity [93]. Correlated with the series of changes in the HHM is the motion of Y7.53, belonging to the N7.49PxxY7.53 motif, a well conserved motif of class A GPCRs. In the inactive state, Y7.53 points towards TM1, TM2 or H8. In active

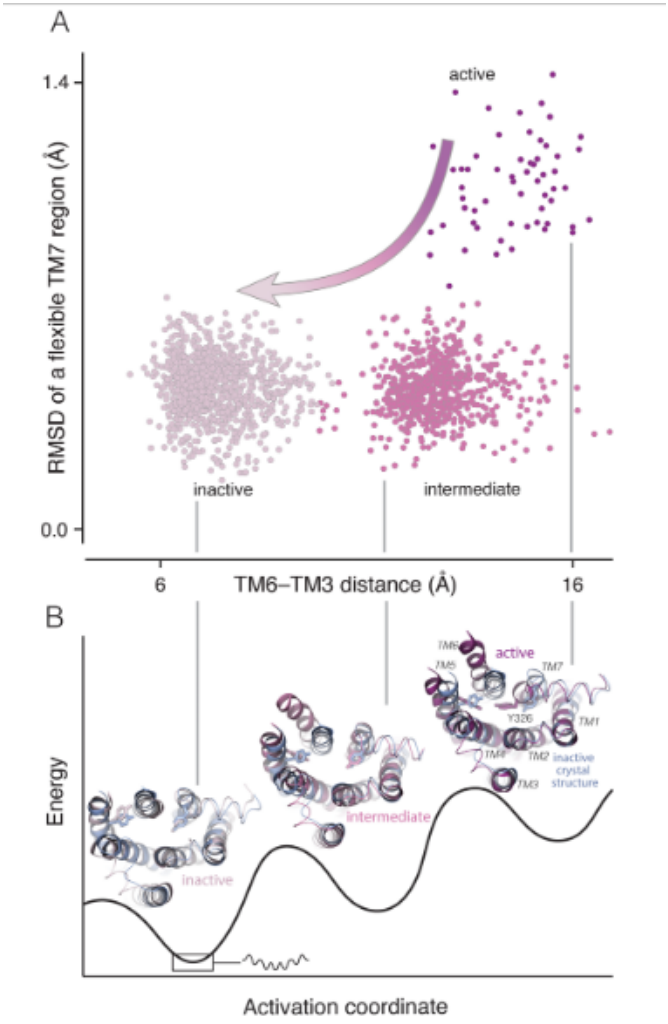

Fig. 2.12. Protein conformations cluster into distinct conformational state using molecular dynamics simulations. (A) Plotting an MD simulation trajectory along two geometric coordinates reveals three distinct conformational states during the process of $\beta 2$ adrenergic receptord eactivation (top). RMSD is the root-mean-square deviation. (B) Snapshots from simulation, representing each of the three conforma- tional states (light pink, magenta and dark purple), are overlaid with the inactive-state crystal structure (blue) [95]

GPCR structures, Y7.53 inserts into the space that would be occupied by TM6 in the inactive state, pointing towards the axis of the 7-TM bundle [101]. This motion correlates with the formation of a hydrogen bond, at least in β2AR and rhodopsin, with Y5.58 , which also moves inwards upon activation [93].

GPCR structures contain internal water molecules in the intracellular side. The importance of structural waters have been correlated with activation even before active state structures became available, on the basis of indirect experimental

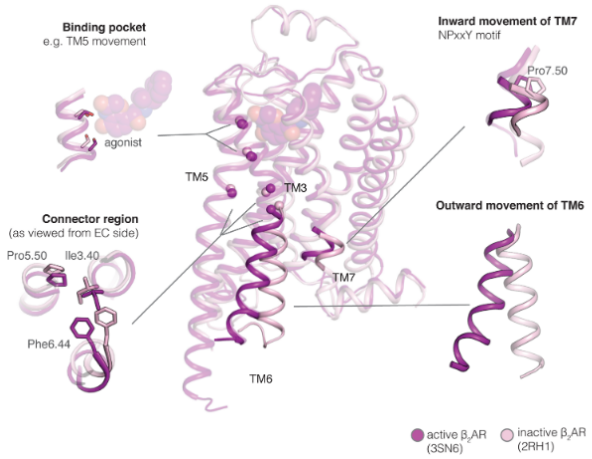

Fig. 2.13. Structural rearrangements during GPCR activation. Inactive (light pink) and active (dark purple) conformations of the β2 adrenergic receptor show differences in helix position and side-chain orientatio

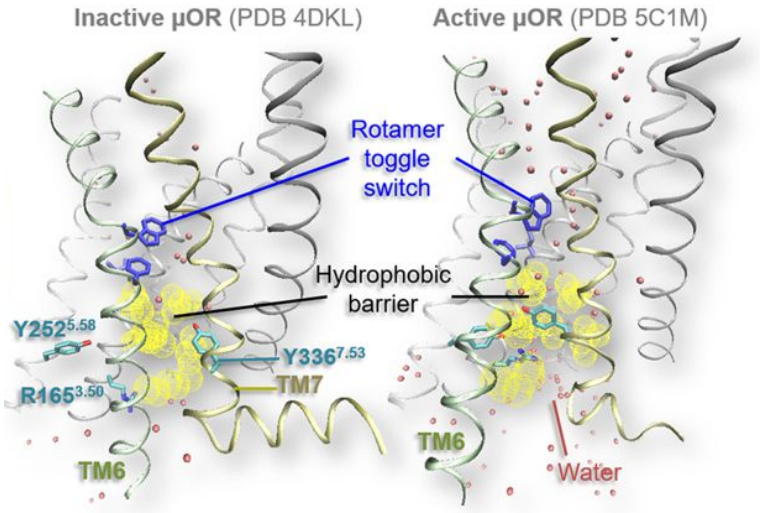

Fig. 2.14. Rotamer toggle switch in the inactive and active state of the μ opioid receptor [102]

evidence [103]. The restructuring of waters within the receptor and the opening of a continuous water channel, connecting the orthosteric binding site and the extracellular side, is made possible by the unlock of the HHM, and confirmed by crystallographic data e.g. for the A2a adenosine receptor. A stable, fully open water channel probably requires G-protein binding [93]. Molecular simulations support the opening of a water channel in the intracellular side as a critical event upon activation [104].

The intracellular hydrogen bond network connecting the orthosteric cavity with the intracellular side also rearranges upon activation. The details of the polar network has emerged when high resolution structures, such as that of the inactive δ-opioid and of the active μ-opioid allowed to resolve the position of structural waters, uncovering a network of water-mediated hydrogen bonds on the intracellular side (Figure 2.16) [105]. Residues involved in the polar network are often overlapping with the previous discussed switches: W6.48 is involved, as isY 7.53 and X6.40. The critical event is the loss of the coordinated allosteric sodium with the motions of N3.35 and S3.39 [98, 105]. On a more general scale, GPCRs activation has been found to coincide with the formation of a continuous buried ionizable network connecting the extracellular to the intracellular side, which is instead disaggregated in the inactive receptors [41]. Surprisingly, in the same study an analogous networks is found in the prokaryotic 7-TM proton pump, hinting at a conserved fundamental similarity despite the extreme sequence divergence [41].

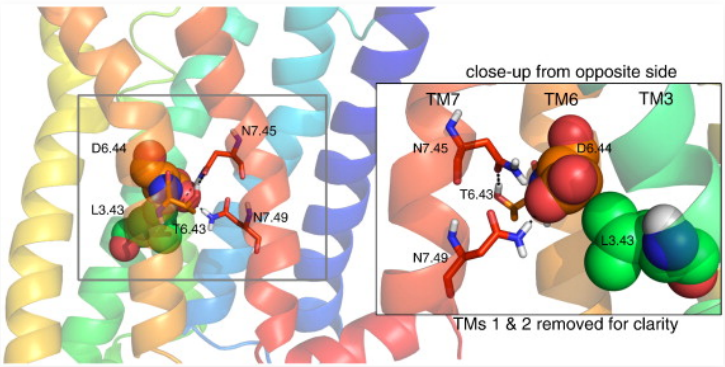

Fig. 2.15. Hydrophobic Hindering mechanism in GPCRs[93]

a

Polar network residues in inactive states

BW numbering	δOR	β_2AR	M2R	Rhodopsin
1.50	67N	51N	41N	55N
2.50	95D	79D	96D	83D
3.32	128D	113D	103D	
3.39	135S	120S	110S	
5.46	221A	207S	194A	211H
6.40	266V	278I	392I	257M
6.44	270F	282F	396F	261F
6.48	274W	286W	400W	
7.41	307G	315G	429C	295A
7.45	310N	318N	432N	
7.46	311S	319S	433S	
7.49	314N	322N	436N	302N
7.53	318Y	326Y	440Y	306Y

b

Polar network residues in active states

BW numbering	μOR	β_2AR	M2R	Rhodopsin
1.50	86N	51N	41N	55N
2.43	107I	72I	62L	76L
2.50	114D	79D	69D	83D
3.32	147D	113D	103D	117A
3.39	154S	120S	110S	
3.43	158L	124L	114L	128L
5.58	252Y	219Y	206Y	223Y
6.40	285V	278I	392I	257M
6.48	293W	286W	400W	265W
7.38	322I	312N	426Y	292A
7.41	325G	315G	429C	295A
7.45	328N	318N	432N	298S
7.46	329S	319S	433S	
7.49	332N	322N	436N	302N
7.53	336Y	326Y	440Y	306Y

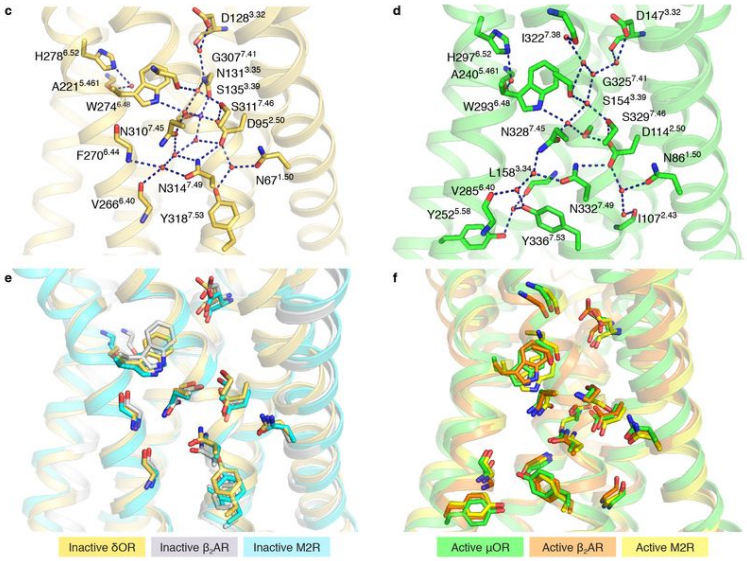

Fig. 2.16. Polar networks involved in GPCR activation [106]

3

Methods

3.1 Homology Modeling

In the last years, the genomic revolution has allowed us to have informations on the whole genomes of a big number of organisms. The JGI Genomes OnLine Databases [107] lists 15.238 complete and published genome sequences as of November 2018. The idea of sequencing biomes in full is not peregrine, and some projects are already doing so-called metagenomics for microbial communities [108]. Yet, it is clear the concept that we basically only begun to have an idea of the decoding genomes. One of the main problems to deal with between genome sequencing and full mechanistic understanding of a living organism is to translate genomic information into structural information because reading genomic sequences does not tell us the structures and functions of the entities it will produce -namely, proteins and RNA. This results in an huge discrepancy, of one order between the big number of protein-coding genes of which we know the sequence and the few ones of which we know the detailed three-dimensional structure. The number of Protein Data Bank [109] unique entries, as of November 2018, is 45538. Contrarily, the number of non-redundant protein sequences known as the same date (from the UniParc non-redundant database) [110] is beyond 111 millions, almost 1000 times larger. At a first sight, the correspondence between protein sequences and structures could be biunivocal: each sequence could produce a unique, completely different protein fold. In practice, it is totally the opposite. Today we know that the total fold space of proteins is immensely smaller than the sequence space. The number of currently known folds, depending on databases, is less than 1400, and new folds are at this point discovered very rarely so that the total is almost a few thousands at most [111]. Most importantly, we know that similar sequences tend to fold in very much the same way, but the opposite is not true. Sequences that share the same fold are not necessarily similar [112]. This means that, once we know the fold and the sequence of a given protein, we automatically know the fold of practically all variants of this protein. **How far can we go**? The correspondence between sequence similarity and structure similarity has been investigated in the 1980s, by the landmark contributions of Chothia and Lesk, M.Levitt, R.F.Doolittle etc. In general, it turns out that a sequence identity of <50% between two proteins will result in a root mean square deviation (RMSD) larger than 1Å [113] (Figure 3.1)

and that, in general, one can safely predict that two proteins with >35% sequence identity will share the same fold [114]. This is an average relation. Indeed it is known that actually a few superfolds dominate the fold space, so that even vastly distant proteins (< 10-20% identity) can share a generally analogous structure. The distribution of folds follows, in fact, a power law [115]. From these concepts emerge the idea that a reasonable strategy to predict protein structures is that of using homologous, structurally solved proteins as a template. This possibility was recognized early: a structural model of α-lactalbumin based on lysozyme was built already at the dawn of structural biology, in 1969, by Browne et al. [116]. The first systematic homology modeling approaches begin to be developed and formalized in the 80s, which also saw the introduction of the concept of multitemplate modeling [117]. MODELLER, still the most used and successful homology modeling software, was first developed in 1993 [118]. Today homology modeling is a remarkably and the most successful technique. Homology modeling is by far the most widely used computational approach to predict the 3D structures of proteins, and almost all protein structure prediction servers rely on homology modeling, as seen in the community-wide blind benchmark Critical Assessment of Techniques for Protein Structure Prediction (CASP) (Meier and Soding, [119]). The general procedure for homology modeling is depicted in Figure 3.2 (from [120]).

The first step in homology modeling is finding structurally solved structures with possibly high sequence identity with our target. To find the best templates,

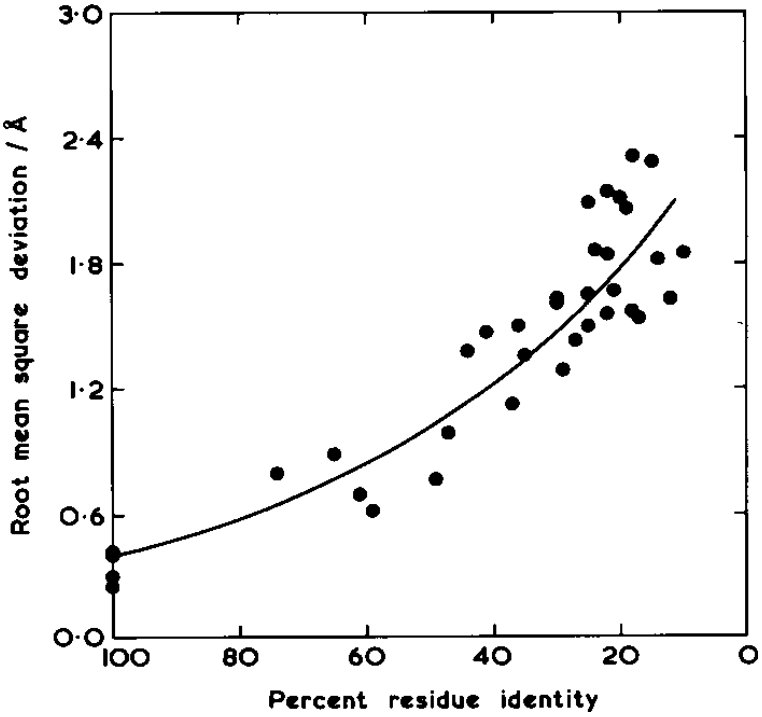

Fig. 3.1. Original diagram of the sequence identity/RMSD relationship in protein structures, from [113]

a classical sequence search software such as the common BLAST [121] can be used as a first step. BLAST is a heuristic sequence search algorithm, which align short sequence segments, on top of which it then builds a whole alignment. As such it is simple and very fast, but it is not guaranteed to successfully find all distant homologues -nor to align them correctly. In this case profile-based search algorithms are required. These algorithms use the individual query sequence to create a multiple-sequence profile, which condenses the evolutionarily relevant features of the query sequence. This profile is then used for the research of the best template. BLAST variants such as PSI-BLAST use this approach to find distant relatives of a query sequence [122]. PSI-BLAST works by finding closely related sequences to the target, then using such sequences to build a sequence profile. The profile can be then used for another query, using it to retrieve further (and less related) sequences. The process can be iterated multiple times. The downside is that an iterated PSI-BLAST search can actually run too deep, ending up retrieving sequences which are actually unrelated to the target and thus corrupting the search. Automated protocols such as HHsearch have employed empirical strategies (e.g. ignoring the N- and C-termini of the profile or loops) to mitigate these issues [123]. A more recent alternative to PSI-BLAST is HHblits, which relies on the comparison of precalculated hidden Markov models (HMMs) with the target sequence [124]. In this case, the query is bootstrapped by building a tentative HMM by calculating pseudocounts of chemically similar aminoacids depending on the context of each query residue. In all cases, the final output is a series of protein sequences represented in the Protein Data Bank, which form a list of suitable

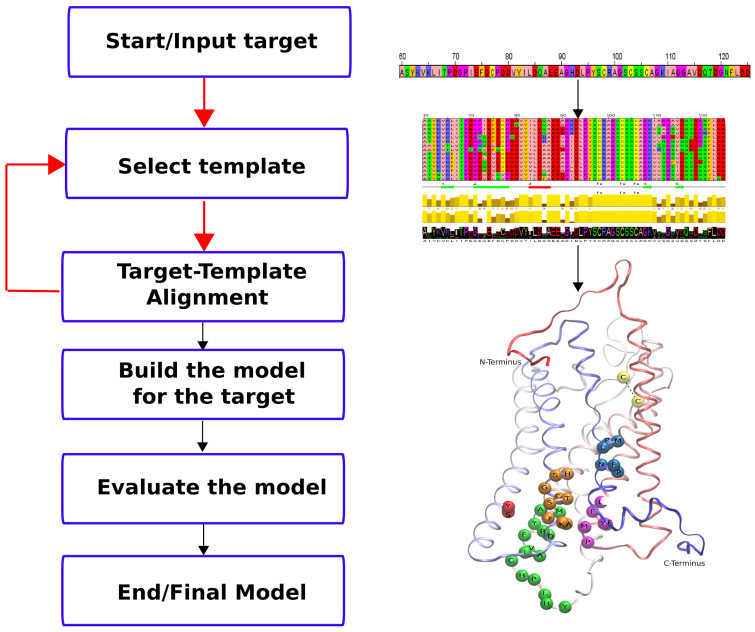

Fig. 3.2. Flux diagram of the basic homology modeling procedure. The principal steps investigated in this work are depicted with red arrows

structural templates. Of course the quality of the profile-based template search depends on the quality of the protein profile used as a query. Most importantly, if the same profile is then used for the template-target alignment, further caution must be used. The template-target alignment is the essential input, along with the template structure(s), for the modeling process. In the long tradition of garbage in, garbage out, homology modeling cannot adjust a wrong alignment [125], meaning that, no structural algorithm can compensate a fundamental error in the input alignment. Algorithms such as BLAST and HHblits are specifically designed for search speed over alignment accuracy, and as such the quality of the multiple sequence alignments they generate is suboptimal. Most important, the quality of the profiles they generate from the query is not perfect: they either derive from a set of suboptimal alignments (PSI-BLAST) or statistically based algorithms (HHblits). A completely mathematically rigorous multiple sequence alignment (MSA) is computationally prohibitive beyond a dozen of sequences [126]. Yet a few tools have been published and made available online for a robust and fast MSA, using progressive pairwise alignments. We relied on PROMALS as a choice, since it is especially performant in aligning distant sequences [127]. PROMALS uses, again, HMMs of the query sequences to perform a profile-based alignment, and also incorporates secondary structure information/prediction in its profile-based alignment. This is especially useful in protein families, such as GPCRs, where the primary sequence can be divergent but the secondary structure is highly conserved. It must be stressed that, regardless of how careful the automated strategies are, experimental information and knowledge of the target have to be taken into consideration for a correct assessment. An automated alignment of a target sequence to a distant template can however introduce artefacts, e.g. implausible gaps in highly conserved structural elements such as transmembrane helices, or off-by-one alignment shifts that move highly conserved sequence patterns. In this case alignments might be corrected and/or selected by hand, of course rigorously justifying each human decision. Once a template-target alignment is made, the coordinates of the target can be built. so far, three strategies have been developed there to generate the model [120]:

1. **Rigid body assembly.** In this strategy, the most conserved regions between target and template, such as the transmembrane core, are identified, and the coordinates of Cα atoms in such regions are used to build a framework. The backbone atoms are then superimposed on this rigid framework.
2. **Segment matching.** The template here provides a guide of atomic positions. A library of peptide structures taken from the Protein Data Bank is then searched and the best candidates fitted on the guide positions.
3. **Satisfaction of spatial restrains.** The template is used as a source of probabilistic geometric restrains and the algorithm attempts to minimize violations of these restraints on the target atoms.

In this thesis we used the third strategy, which is also one of the most successful. The spatial restraints strategy was first designed by Sali and Blundell [120], and it is implemented in MODELLER, the homology modeling software used in this work. The algorithm retrieves a set of spatial restrains (interatomic distances, ϕ and ψ backbone angles, etc.) from the template and applies them to the tar-

get sequence, each with an associated probability distribution function. These template-derived restrains are then enriched by general stereochemical restraints derived from an all-atom force field (in the case of MODELLER, the force field is CHARMM [128]). Both classes of restraints are combined in a single function and the model is then built so to minimize the violations of the geometric restrains. In MODELLER, minimization is divided in two steps. The first exploits the variable target function method [129]. Then, the model is further optimized by molecular dynamics with simulated annealing. An advantage of the spatial restrain approach is that it is straightforward to include experimental information (e.g. NMR) by simply adding or modifying restraints. For example, NMR information can be easily incorporated in the modeling process. Cross-linking experiments, FRET data, etc. can also be incorporated. However, the most difficult part of a protein to be modeled are loops. They are a critical challenge for every protein structure prediction algorithm and they are often highly functional parts of the protein, thanks to their flexibility and exposure to the solvent. Loops do not necessarily follow the rules of standard secondary structures. They are also rarely conserved even between closely related proteins, hence they are poorly predictable based on the homology technique; moreover the same loop can have different structures depending on the surrounding protein environment [130] and references therein). Even worse is the case when loops are not solved due to their flexibility, thus defeating homology modeling strategies even when a close template is otherwise available. For these reasons, loop prediction has been described as a mini-folding problem [120]. Therefore even in the context of homology modeling, loop prediction often requires *ab initio* structure prediction techniques, such as the ROSETTA algorithm [131]. Multiple techniques can be compared and even used together, such as combining coarse-grained search of conformational spaces with spatial restraints statistical techniques [132]. For loop modeling, MODELLER, the software we used in this work, employs a combination of the molecular mechanics force field CHARMM along with statistical potentials derived from structure databases for the backbone and side chain angles. Random conformations are generated and then optimized so to minimize its disagreement with the force field and the statistical spatial restrains, by conjugate gradients minimization and molecular dynamics [131]. For short and medium-sized loops (up to 10-12 residues) the error is in the range of 1-2.5Å, comparable with the average fluctuation of the loops at room temperature [131].

Homology modeling is of course not solving the folding problem from first principles: each model is the result of a local energy minimization, and it is rappresentative of a global energy minimum. For example the rotameric state of aminoacid side chains in regions of poor sequence identity is almost guaranteed to be somehow wrong. Statistical algorithms such as MODELLER therefore usually generate a number (from tens to thousands) of potential models, from which the best one has to be chosen. In theory, best would mean the one closer to the real structure, but of course this is unknowable *a priori*. In practice what best means is ambiguous, but in general we look for the model that best fits expectations about protein structural parameters and statistics. The most immediate quality control is an analysis of the basic structural parameters using for example the Ramachandran plot. There are several services for this aim, one of the most common being

PROCHECK [133]. The server VADAR [134] is a similar but slightly more modern alternative. It quickly calculates more than thirty structural parameters, including not only Ramachandran plots but also side chain packing (by measuring the fractional excluded volume), evidence of buried charges, omega angle, etc., with clear indication of problematic values. More physically detailed and compact evaluations can be quantified by a scoring function. Scoring functions can be divided into physics-, knowledge and learning-based. Quoting [135]:

For physical scoring functions, the goal is to describe the physics of the interaction between atoms as accurately as possible. These functions are often parameterized on much smaller systems than proteins, and the typical example is a molecular mechanics force field such as OPLS [136], CHARMM [128] or Amber [137]. Knowledgebased scoring functions in contrast derive probability distributions from features extracted from native structures [138, 139]. Finally, learning-based functions are trained on structural features to distinguish between correct and incorrect models and to predict the actual quality of a given model. MODELLER includes and calculates two scoring functions, DOPE and GA341. DOPE (Discrete Optimized Protein Energy) is a purely statistical scoring function, based on an atomic distance-dependent statistical potential derived from almost 1500 crystallographic structures. DOPE takes into account, for its reference state, the size of the protein, improving score accuracy with respect to size-unaware potentials that average between structures of different sizes. GA341 is a combined score, which condenses together a statistical potential that measures structural parameters, such as residue distances, accessibilities and model compactness, but also adds a scoring based on the template sequence identity to the target [140]. An interesting feature of the DOPE score implementation in MODELLER is that it returns both an absolute and a normalized score. The latter allows to compare models derived from different templates and of different sizes. It is important in practice, during the model assessment, that poorly modeled regions can weigh a lot on the final resulting DOPE score. In particular, in many models, N- or C-terminal tails can be poorly modeled since they are unresolved in the template. In this case the (poor) scoring of the tails will dominate on the score of the structured part of the model, making it hard to find the best model. For this reason it is best to cut the N and C terminal tails from the target sequence before modeling. The final test of a model is, of course, its consistency with experimental data. Mutagenesis and functional data can yield reasonable expectation on the localization of functional residues and side chains orientation (e.g. in a binding or catalytic site). Often, inconsistent models of this type derive from an incorrect template-target alignment, insufficient template/target similarity in specific regions, or poor loop modeling. While alignment problems cannot, in practice, be recovered, strategies such as molecular dynamics can in some cases mitigate the latter two problems.

3.2 Docking

As we know now very well, protein molecules do not act alone. While modeling predicts the conformation of a protein, what is often needed in order to predict the biological activity is the interaction between the protein with its molecular

partners, being them other proteins or small molecules. The computational prediction of multi-molecular complex structures is known as docking. While protein structural prediction can use sequence information and reasonable assumptions, as we have seen, to infer analogies between structures, docking algorithms are forced to rely on chemical and physical information. The task of docking is thus a daunting search of the conformational space to find the most energetically favorable conformation of the molecular complex. Just as structure prediction, the formally correct docking algorithm would be full exploration of the components energy landscape. This strategy is almost always impractical, due to the enormous computational cost. *In silico* drug screening, for example, requires the evaluation of protein binding poses of thousands of small molecules. Empirical strategies are thus needed. In practice, usually a docking protocol consists of three phases (which can take place sequentially or simultaneously):

- A rigid docking starting point, where the partners of the complex are treated as rigid object and evaluated by a simple, quick function such as shape complementarity.
- A rescoring function of a subset of previously obtained structures with a more accurate (but expensive) scoring function.
- Finally, flexibility, that is molecular degrees of freedom, is introduced and the previous docking poses refined.

Docking generates and ranks solutions by maximizing (minimizing) a given scoring function. In theory this would be free energy, but this is unaccessible. We need thus an easy to calculate function that, hopefully, approximates the behavior of a free energy. The most fundamental quantity to evaluate is geometrical complementarity. While usually insufficient by itself, complementarity scores are often a basis for the scoring functions of many contemporary docking programs, such as ZDOCK, PIPER, PatchDock, MolFit or HEX (reviewed in [141]). This can be done either by computing surfaces and then solvent-accessible/solvent-excluded surfaces, or by dividing the molecules in voxels using a grid in space and then counting the occupied/unoccupied voxels in proximity of the two molecules (see [142] and references therein). The next step to improve scoring accuracy is including a force field -often adapting one used for molecular dynamics, to evaluate electrostatic, Lennard-Jones (LJ), and hydrophobic energy terms. Unsurprisingly, scoring functions involving electrostatics work systematically better than scoring functions without it, even if, curiously, considering only LJ terms can perform equal or better than consider LJ plus electrostatics, in some cases [141]. Evaluation of such terms often rely, for performance reasons, on empirical approximate algorithms. In particular, hydrophobic interactions, while theoretically complex entropic effects, are known to be in practical situations proportional to the buried solvent accessible surface (bSAS) between two atomic surfaces and parametrized as such. A critical requirement for energy functions is softening: smoothing the energy/scoring function so that it does not diverge to infinity (e.g. LJ potentials for r=0). Softening however reduces the accuracy of the scoring function, thus increasing the amount of false positives [142]. It must be emphasized that, often, scoring function performance is not universal. A given scoring function can work well for one target and dismally for another, as exempled by a test using the same docking

software (GOLD) and two different scoring functions for both thymidine kinase and estrogen receptors. In general, scoring functions are optimized on a training set of protein-protein or protein-ligand complexes, and therefore the quality of the scoring function is often dependent on that of the set. Unfortunately, (i) datasets of protein-ligand complexes and their curation is still suboptimal, (ii) biological affinity data are too often contradictory or of poor quality.

Docking algorithms must search the possible orientations of ligand and target and evaluate them. Several techniques are used to explore the 6-dimensional orientation space (3 translational + 3 rotational variables). FFT search for example is the most common and ancient (1992) and is based on the fast Fourier transform (FFT) search by Katchalski-Katzir, Vakser et al. [143]. In their own words: "The algorithm involves an automated procedure including: (i) a digital representation of the molecules (derived from atomic coordinates) by three-dimensional discrete functions that distinguishes between the surface and the interior; (ii) the calculation, using Fourier transformation, of a correlation function that assesses the degree of molecular surface overlap and penetration upon relative shifts of the molecules in three dimensions; and (iii) a scan of the relative orientations of the molecules in three dimensions. The algorithm provides a list of correlation values indicating the extent of geometric match between the surfaces of the molecules; each of these values is associated with six numbers describing the relative position (translation and rotation) of the molecules. The procedure is thus equivalent to a six-dimensional search but much faster by design, and the computation time is only moderately dependent on molecular size. (from [143]). With the era of the computer vision a new docking method was born in the, the geometric hashting [144]. Here, the shape of both partners is encoded by collecting critical points. Relative 3D coordinate systems are then built using as basis each possible triplet of points (2 points define the X axis, the third point defines the Y axis; Z axis is normal to the XY plane, following the right hand rule). The relative location of other points is then discretized and calculated accordingly to the relative coordinate systems. Each point location/coordinate system couple is stored in a hash table, indexed by the point locations. To find the complementarity, the hash table of one partner is compared with the one of the second: if more than a threshold number of points coincide for a given basis, shape complementarity is defined. Clusters of points which correspond in the lookup are then superimposed to generate the docking pose. The algorithm is fast because, once the preprocessing step is made, it requires little more than table lookup.

Most recently, function optimization algorithms can be used to search the docking poses without having to resort to exhaustive search. The program thus explores the scoring function landscape as if any other function to minimize/maximize, reducing docking to a general class of optimization problems. Examples are genetic algorithms, quasi-Newton optimization (e.g. BFGS), Metropolis Monte Carlo, particle swarm optimization and simulated annealing [145]. Common to all these algorithms is that solutions are found iteratively, starting from one or more random guesses, generating variants of the guess, rejecting some of the results and using the others as basis for the next iteration - differences are mostly in how the variants are generated and the rejection/acceptance criteria. The local gradient of the function can be taken into account in some cases (e.g. by BFGS). Flexibility can

be directly introduced in these algorithms: rotations and motions can be included among the parameters to vary between iterations.

So far, we have considered the protein and the ligand as rigid objects. This is of course unphysical. Molecular rearrangements are however computationally costly, so they are usually considered only after rigid conformational search and scoring have selected most binding poses out. In the case of proteins, the hardest issue is the rotameric state of aminoacid side chain. The search is limited thus to known populated rotamers for each residue, but even in this case the remaining conformational space is too vast. An alternative is adding flexibility in a non-biased manner using a short molecular dynamics, with or without explicit water.

3.2.1 Docking on homology models

The structure of most proteins of pharmacological or biological interest is unknown. We have seen that homology modeling can help us in guessing the protein structure when it is not solved. Applying docking on homology models (from now on DHM) appears therefore to be the natural approach to predict a protein-ligand complex for a structurally unsolved protein. However the feasibility of DHM is not straightforward. The prediction has two sources of error: the error in protein prediction inherent to the homology modeling, and the error intrinsic to the heuristic and limited energy landscape surveying of docking algorithms. The two errors are not independent: we can expect errors in the protein structure prediction to affect the binding region/cavity, to the detriment of the docking accuracy. In a systematic performance analysis of protein-protein DHM [146], the single parameter of the protein model most correlated with docking success (as measured by RMSD with a real, crystal structure) is target-template sequence identity. However, even more important, in the same analysis, is the quality of experimental information. Even twilight-zone homology models can produce near-native docked poses with high-quality information. Conversely, low-quality experimental information degrades docking performance quickly even in the case of high homology. This highlights the necessity of using information-driven docking when dealing with homology models, compensating the lack of direct structural information with indirect structural evidence. In specific cases the above trends may be thwarted. A recent analysis of small molecule docking on GPCR homology models has shown that DHM quality and target-template sequence identity can be uncorrelated. In the case of GPCRs the 7-transmembrane helix 3D topology is in fact extremely conserved even between very distant homologues, making the sequence identity less concerning. Docking quality is thus more a function of the binding cavity volume and prediction of loops, which are flexible and poorly conserved even between close homologues. Indeed, in the case of GPCRs, DHM is known to perform better once loops are removed from the model.

So far we have considered the docking problem as something completely obscure. However, we often have partial experimental data on the complex: for example, NMR chemical shift data of the holo- and apo- forms; or mutagenesis data indicating which residues are plausibly part of the binding cavity. This information alone does not tell us the binding pose, but it can be added to guide the docking procedure, to help the algorithm in finding the bet pose. While not the only one,

possibly the most successful softwares that uses an information-driven approach is HADDOCK, written and maintained by Alexandre Bonvin and coworkers [147]. Originally conceived for protein-protein docking, HADDOCK is now applicable also at small molecule/protein docking. It performs as a more or less standard docking program, but the user can add several types of information using the AIRs (Ambiguous Interaction Restrains) file. Mutagenesis, NMR and FRET data are examples of experimental data that can be encoded as AIRs in HADDOCK. All of them are basically converted to distance constraints that the software attempts to satisfy. Other geometrical constrains such as e.g. angles can be also added by the user. An AIR is defined as follows: An AIR is defined as an ambiguous intermolecular distance with a maximum value of 3Å between any atom of an active residue of protein A and any atom of both active and passive residues of protein B (and inversely for f protein B). The effective distance for each restraint is calculated using a the equation in figure 3.3, where N atoms indicates all atoms of a given residue and N res the sum of active and passive residues for a given protein. In this way, the passive residues do not have direct AIRs to the partner protein but can satisfy the partner protein active restraints. The 3Å limit represents a compromise between hydrogen-hydrogen and heavy-atom/heavy-atom minimum van der Waals distances. There is a distinction, when defining AIRs, between active and passive residues. Active residues are the ones who are known to be involved in the interaction. Passive residues are residues that can participate as well in the interface, and are normally the neighborhood of active residues. Any active residue that does not contact an active or passive residue of the other partner carries an energy penalty. Furthermore, when generating configurations, HADDOCK by default discards at random (for each configuration) 50% of the residues. This is intended to deal with noisy data (e.g. residues supposed, by mutagenesis, to be directly involved in binding but that instead impact it indirectly), however not very helpful.

The HADDOCK protocol performs three steps, each step (after the first) refining the top-scoring subset of the previous one:

1. **Rigid body energy minimization**. The partners are distanced a minimum of 25Å, randomly rotated and oriented around their respective centers of mass, and then docked as rigid bodies by energy minimization. Thousands of structures (usually 2000) are generated at this stage, clustered and scored.
2. **Semiflexible simulated annealing refinement**. The top structures (usually a few hundreds) of the previous run are refined. Three or four simulated annealings are performed. In the first step, partners are rigid bodies and the

$$d_{iAB}^{eff} = \left(\sum_{m_{iA}=1}^{N_{atoms}} \sum_{k=1}^{N_{N_{res}B}} \sum_{n_{kB}=1}^{N_{atoms}} \frac{1}{d_{m_{iA}n_{kB}}^{6}} \right)^{-1/6}$$

Fig. 3.3. Haddock equation for distance restrains

orientation is optimized (default: 1000 steps from 2000 to 50K with 8 fs time steps). In the second, side chain motions are introduced (default: 4000 steps from 2000 to 50K with 4 fs time steps). Last, flexibility is introduced also in the backbone (1000 steps from 500 to 50K with 2 fs time steps): one rigid, one allowing sidechain motion, one allowing sidechain and backbone motion.

3. textbf Fully flexible refinement with molecular dynamics in explicit water. The output structures of the previous step are immersed in a 8Å shell of TIP3P water molecules. Again, there are three substeps. In the first, the system is heated to 300 K stepwise (100K, 200K, 300K, 500 steps each temperature). The side chains in the interface are free to move, but the rest of the structure is hold by position restrains. 5000 MD steps are then calculated at 300K, with position restrains on the heavy atoms not belonging to the interface. Finally, the structure is cooled (300K, 200K, 100K, 1000 steps each temperature) with position restrains only in the backbone atoms at the interface.

HADDOCK evaluates structures in two ways. The first is a physical scoring function, actually the non-bonded part of the OPLS force field, including full electrostatic and van der Waals energy terms. HADDOCK also clusters solutions together based on their respective RMSD. Similar docking solutions, RMSD-wise, can thus be analyzed quickly together, by looking at the clusters average statistics. RMSD clustering allows one to choose, in some instances, the most populated cluster, instead of the one with the lowest energetics. This is because the OPLS force field has no intrinsic description of entropy: the population of a cluster can be thus used as a (remarkably crude) proxy for the effective Boltzmann population of the docking pose. (see Figure 3.4)

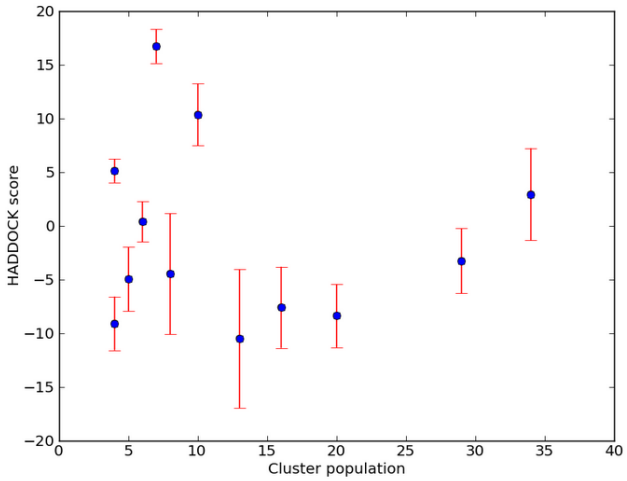

Fig. 3.4. Example of score/population diagram for a final (water refinement) HADDOCK docking. The number of structures for each RMSD-based cluster is on the X axis, the average score on the Y axis; error bars are standard deviations

3.3 Molecular dynamics simulation

So far we have seen static prediction methods: we generate a protein structure from its sequence, then we generate a protein/ligand complex from the protein structures. The solutions of these modeling algorithms are statical, 3D structures. However, we know that chemical and biological processes happen in time. Thus, we need a way to predict the dynamical behavior of biomolecules. Following Richard Feynmans famous quote, Everything that living things do can be understood in terms of the "jiggling and wiggling" of atoms, the most straightforward approach is that of integrating the actual, physical equations that govern the "jiggling and wiggling" atomic motions, to simulate the behavior of the molecules in time. This is the definition of molecular dynamics. Biological processes happen on an immense range of timescales: from the femtoseconds (10^{-15} s) of molecular bond vibrations to the billions of years of evolution. Correspondingly, length scales go from the kilometer range of ecosystems to the picometers of individual bonds. Restricting to molecular processes, the ones we are focused on here, does not mitigate the problem too much: we still have to deal with time scales from the fs to the second, and from the pm to the μm, a range of 15 and 12 orders of magnitude, respectively. In an ideal world this would be of no concern: we could simulate every conceivable system by solving the Schrödinger equations for its particles. Our real world however is constrained by available technologies. For this reason, we have to use the appropriate computational technique depending on the time and length scales corresponding to the individual processes we want to study. Every time we jump from a magnitude range to the next we are forced to sacrifice some accuracy in the description of the system, removing degrees of freedom: we obtain speed in return. We can broadly define three main levels of description:

- Quantum mechanics (QM)
- Atomistic molecular mechanics (MM)
- Coarse-grained molecular mechanics (CG)

Beyond CG, other simulation scale levels exist (e.g. continuum mechanics, systems biology descriptions) but at this point they go well beyond the realm of molecular biology. Molecular dynamics has a long story - the first implementations of the methodology date from before the 1960s [148]. Today several software suites are available for molecular dynamics: the most widespread are GROMACS [149], NAMD[150], AMBER [151]. In the work presented in this thesis, GROMACS has been used, but the discussion below applies to most softwares.

3.3.1 Molecular dynamics: an overview

Molecular mechanics (MM) techniques allow the investigation of timescales from the 10^{-12} to the 10^{-5} s, and system sizes up to 10^6 atoms, thus bringing us into the nanometer length scale. This is a system size comfortably suited to investigate individual proteins and small protein complexes along with the solvent, and to investigate several molecular processes of biological interest, from folding of small proteins to ligand binding to short-time conformational transitions. There

are several MM techniques, such as Monte Carlo (MC) simulations, energy minimization etc. These techniques however do not give direct information on the time evolution of the system -that is, the dynamics. Here we focus on molecular dynamics, or MD, where instead we get a model of the evolution of a molecular system in time. The unit of the system in MM is the individual atom. 4 Atoms are treated as points in space which obey to the laws of Newtonian mechanics, which in MD are integrated at discrete timesteps -differential equations are thus reduced to difference equations. The forces by which the atoms interact are described by a force field: a simplified description of both bonded and non bonded interactions, as depending on distances and angles between (two or more) atoms. Electrons and their interactions are thus implicit, simplifying immensely the calculations, at the cost of losing detail such as polarization. Any chemical reactivity is of course lost. MM therefore models essentially conformational changes in a system, keeping the identity of molecules constant during the simulation. Ideally a MD simulation aims at reproducing the macroscopic (thermody-namic) measurable properties of the simulated ensemble. It does so under the ergodic assumption, that is, on the limit of infinite time, the average of an observable over time equals the average of an observable over the phase space. In other words, all accessible microstates have the same probability at the limit of infinite time. In practice ergodicity is not attained, a more convenient proxy is usually simulation convergence, that is when observables do not change significantly anymore over time (Figure 3.5).

Our simulated system is not infinite. Boundaries conditions are thus require. There are basically two choices. The first is having some sort of walls enclosing the system, effectively constraining it into a box. The problem of this approach is that the behavior of atoms close to the boundary will be necessarily unphysical, and display surface effects which might affect the simulation. The most used approach is thus periodic boundary conditions, or PBC. In this case, there are no hard, impassable boundaries. On the contrary, the simulation takes place in a finite but unbounded topology, equivalent to the 3-dimensional surface of a 4-torus: particles that exit on one side re-enter on the opposite side, seamlessly. An easier way to visualize PBC is that of imagining the box as infinitely replicated in each direction.

3.3.2 Force fields in MD

In MD, a force field is a representation of the forces that govern the interactions between atoms. There are two classes of interactions that we will take into consideration: bonded interactions and non bonded interactions. Bonded interactions represent the covalent bonds between atoms, and model bond stretching, rotations, bending, dihedral angles etc. They can involve 2,3 or 4 atoms and can be represented for example by harmonic potentials. Non bonded interactions represent the electrostatic and van der Waals interactions, and are usually two-body interactions. The latter interactions often take the shape of Lennard-Jones or Morse potentials. Given that the properties of an atom depend on its chemical environment, a force field usually describes parameters for several atom types even for the same chemical element - e.g. an aminic nitrogen will be different from an amidic nitrogen [153]. There are several atomistic force fields: the most used were first developed in the 1980s-1990s and currently updated: they are CHARMM, GROMOS, AMBER, and OPLS. Among them, GROMOS is peculiar in being a united

atom force field: some atom groups such as e.g. aliphatic carbons with hydrogens are treated as a single particle. Force fields parameters are determined in various ways, usually combining both *ab initio* and experimental data. GROMOS for example is parametrized aiming at reproducing the free enthalpies of hydration and solvation [154], while AMBER relies more on *ab initio* quantum mechanical calculations [155]. The differences in the way force fields have been developed imply that force field parameters cannot be freely mixed and matched with each others to maintain self consistency.

3.3.3 Solvation

An important component to be considered in all biological systems is water. MD simulations do not model biomolecules in a vacuum (with a few exceptions) but consider some solvent effects. There are two approaches:

- **Explicit solvation**. Explicit solvation models water with atomic detail, molecules of water are thus included in the simulation. Given the amount of water requested to properly model solvation, this requires substantial computational resources (it is not unusual for water to constitute the majority of atoms in

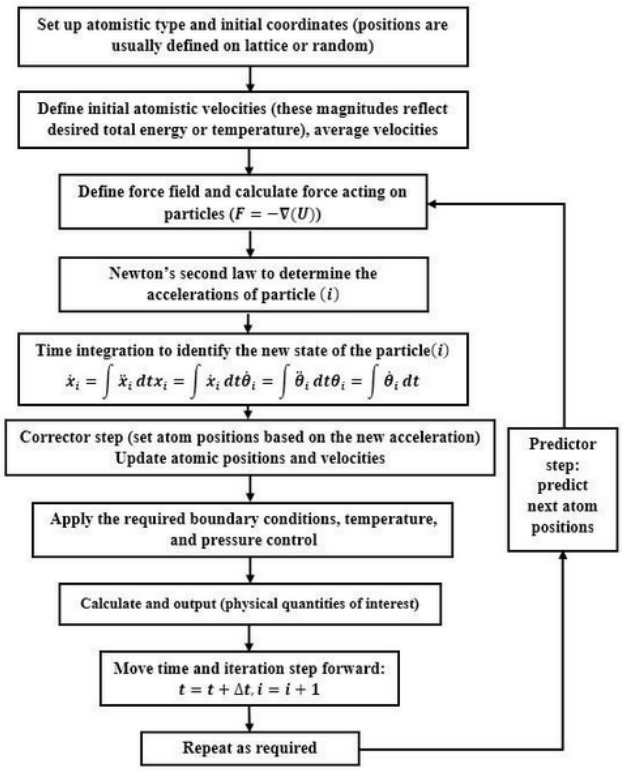

Fig. 3.5. Step in molecular dynamics simulation [152]

a simulated system), but it has the best chances of producing reliable results. Modeling water with an atomistic force field is notoriously tricky, and current water models such as TIP3P [156], SPC [157] and SPC/E [158] still do not fully reproduce the water chemico-physical parameters. In general water models show faster diffusion and lower internal structure than real water. Models that employ more than three atom points for water, attempting at modeling more closely the charge distribution of water, are sometimes used.

- **Implicit solvation**. In biomolecular simulations time spent simulating bulk water molecules is almost always mostly wasted: we are not interested as much in the microscopic behavour of water as in having a realistic environment for the biomolecule of interest. Therefore there has been interest in developing methods which approximate the solvent as a continuum potential, attempting to reduce computational costs while maintaining accuracy. These implicit solvent (IS) techniques usually model the polar component of solvent as a continuous dielectric [159]. The nonpolar component of solvation is estimated from the solvent-accessible surface [159]. While widely considered behind explicit solvent models, IS approaches have been successful in reproducing folding of small proteins [159] However, IS models cannot reproduce structural waters, a key component of many protein structures.

3.3.4 Non-bonded Interactions

Non-bonded force calculation is by far the most computationally expensive part of MD, since in principle every possible atom-atom couple should be calculated. However many types of non bonded interactions decay very quickly (usually exponentially) with distance, and as such the contribution of distant atom-atom interactions could be not considered. We should therefore define a short range cutoff: a distance after which we consider interactions to be zero, and stop calculating them. If we were to truncate abruptly the potential at the cutoff distance, however, we would have a discontinuity in the function, which would lead rapidly to artefacts. The potential is thus usually shifted so that it is actually zero at the cutoff distance (see Figure fig:ndinteractions)

The simple cost of calculating distances between all possible atom pairs becomes as well quickly prohibitive. A first approach to reduce the problem is using Verlet lists [160]: given a radius r_{cut} for our interaction cutoff, we define a $r_v > r_{cut}$. We then calculate a list of particles within r_v, and we update the list only when a particle is displaced by more than —r_{cut} r_v—.

3.3.5 Integration of the equations of motion

If we have a system of N particles, and a force field $U(\vec{r}_i)$, $i = 1, \ldots, N$ we have a 3N second-order differential equation set which represent Newtons equations of motion:

$$\vec{F}_i = m_i \ddot{\vec{r}}_i \tag{3.1}$$

In MD, we discretize differential equations as difference equations, integrated over discrete time steps. However, we want to model a system where time is in

principle continuous. The size of the timestep is thus a critical parameter, since a too coarse time step would coarse grain the integration too much and generate unphysical artefacts. The smaller the timestep, however, the higher the computational cost. The choice thus usually falls on the largest possible timestep which falls below half the characteristic time of the fastest motions of the system. In practice, the fastest motions are the vibrational motions (bond stretching and bending), which limit our time step to 1-4 fs. We want the algorithm to be numerically stable, approximating the behaviour of the system for a time step of 0 -this for example rules out the simple Euler method. To be consistent with Newtons equations, the integrator algorithm should be also in principle time-reversible and symplectic [160]. Symplecticness means that the phase space volume and total energy are both conserved in time - that is, it must obey Liouvilles theorem. The most common symplectic algorithms used in MD are [160]:

- **The Verlet algorithm**: basic integration algorithm. Given a time point t, the Verlet computation is done by the Taylor expansion of $\vec{r}(t+\Delta t)$ and $\vec{r}(t-\Delta t)$:

$$\vec{r}_i(t+\Delta t) = 2\vec{r}_i - \vec{r}_i(t-\Delta t) + \frac{1}{m}\vec{F}_i(t)(\Delta T)^2 \tag{3.2}$$

  Note two things: velocity is implicit (it can be recalculated by the difference), and it fails for $t=0$, since we need to know $(t-\Delta T)$.
- **The Velocity Verlet algorithm**: A solution to the issues above is a variant, the velocity Verlet algorithm:

$$\vec{r}_i(t+\Delta t) = \vec{r}_i(t) + \vec{v}(t)\Delta t + \frac{1}{2}\vec{F}_i(t)\Delta t^2 \tag{3.3}$$

$$\vec{v}_i(t+\Delta t) = \vec{v}_i(t) + \frac{1}{2}[\vec{F}_i(t) + \vec{F}_i(t+\Delta t)]\Delta t \tag{3.4}$$

  Here we explicitly calculate velocity: the point $(t-\Delta T)$ is not needed and we can thus start from t=0.

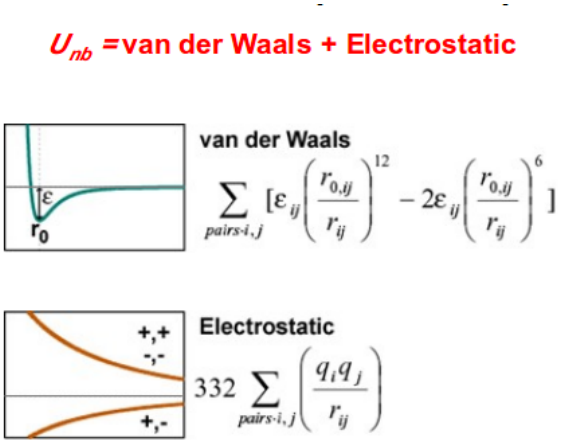

Fig. 3.6. Non-bonded components of the potential function

- **The Leapfrog algorithm**: a velocity-explicit variant of the Verlet algorithm. It is so called because velocities and positions are calculated at staggered times, each one leaping on the other.

$$r_i(t + \Delta t) = r_i(t) + v_i(t + \frac{\Delta t}{2})\Delta t \tag{3.5}$$

$$v_i(t + \frac{\Delta t}{2}) = v_i(t - \frac{\Delta t}{2}) + \frac{1}{m}\vec{F}_i(t)\Delta t \tag{3.6}$$

Let's consider now another important parameter for MD, the thermostat. Since Newtons equations conserve energy, integration of Newtons equations of motions would lead the simulation to belong to a microcanoni- cal(NVE) ensemble [160]. However we often want to mimic realistic situations, where temperature is conserved (NVT or NPT ensembles). To do this, we have to mimic the coupling of our system to a heat bath. This is done by appropriate thermostat algorithms, which correct the velocities of particles so to regenerate, with various degrees of approximation, a canonical ensemble. The simples approach would be rescaling the velocities so that they maintain, at each step, the Boltzmann distribution at the desired temperature. Velocity rescaling however would damp or erase the thermal fluctuations of the system, generating an artificial ensemble. More refined algorithms are used and available, here we treat briefly two of the most common, showing two different approaches to the problem:

- **Nohé-Hoover thermostat**. In this approach, we add to the description of the system two additional degrees of freedom: s and p_s. s can be understood broadly as the position of a virtual heat bath coupled to the system, and p_s as its momentum; a mass Q is also added to s (this is a user-chosen parameter which depends on the system). The system Hamiltonian then reads:

$$H = \frac{1}{2}\sum_i m_i|\mathbf{p}_i|^2 + U(\mathbf{r}^N) + \frac{p_s^2}{2\mathbf{Q}} + k_b(T)(3N+1)\ln s \tag{3.7}$$

where the first two terms are the standard NVE Hamiltonian, and the added last two terms are the thermostat terms. It can be then demonstrated that the microcanonical (NVE) simulation on the extended system will return a canonical (NVT) simulation on the standard system, with the correct partition function. The algorithm is also deterministic - no stochastic terms are added. The above equation describes the original thermostat. However, it is more convenient usually to reformulate p_s in terms of a friction coefficient ξ :

$$H = \frac{1}{2}\sum_i m_i|p_i|^2 + U(\mathbf{r}^N) + \frac{\xi_Q^2}{2} + k_b(T)(3N)\ln s \tag{3.8}$$

- **Langevin dynamics**. An example of a stochastic thermostat, the Langevin dynamics works directly at the level of the equations of motion, adding a friction and a noise term:

$$\vec{m}_i\ddot{\vec{r}}_i = \vec{F}_i - m_i\Gamma_i\vec{r}_i + \vec{\xi}_i(t) \tag{3.9}$$

Γ_i is the friction constant with the units of a frequency, while $\vec{\xi}_i$ is a term introducing Gaussian noise. The noise term models the Brownian collision of the atom with solvent molecules of temperature T-coupling the system, overall, with a stochastic heat bath. It has to be noticed that the Langevin thermostat does not conserve momentum (since it adds/removes randomly terms). To accurately model different molecules, they could need to be coupled to different friction terms. The friction term, damping motions, has the advantage of allowing, in principle, larger time steps to be used in molecular dynamics, without too much loss of stability for advanced methodologies using this approach.

3.3.6 Combined Molecular Dynamics Approches: MM/CG

We have seen in previous paragraph that the choice of our simulation method depends on the length and time scale of the process we want to investigate, and that each increase in accuracy is paid by a decrease in the time scale available. Unfortunately, often we want both an accurate description and a long time scale. One of the possible ways to achieve this is having, in the same simulations, two different levels of description. This is possible when the process we are interested to describe accurately is limited to a small part of the whole system, and yet we cannot isolate the subsystem from its environment. For example, QM and MM levels are routinely used together in QM/MM simulations. The part of the system that needs to be treated quantum mechanically is coupled to the rest of the system, treated by standard MM (see Figure 3.7). In this way, for example, enzymatic reactions can be studied in full QM detail with appropriate environment and constrains, provided by the MM part. The less accurate part of the simulation thus acts as a necessary scaffold for the more accurate part. In the same fashion, we can mix MM and CG levels of simulation, allowing us to achieve MM-level atomistic detail for a process of interest but greatly reducing the computational needs of our system, reducing drastically the number of particles. This allows either longer time scales and/or multiple simulations to be run at reduced computational cost. This is the approach we used in this work. The original multiscale approach was developed by Neri et al. [161]. In this, the system is divided into a MM part and a CG part, communicating through an interface region in a Langevin dynamics scheme. The potential energy of the system corresponds to the sum:

$$V = \underbrace{E_{MM} + E_{CG} + E_I}_{} + \underbrace{E_{\frac{MM}{I}} + E_{\frac{CG}{I}}}_{} + \underbrace{E_{SD}}_{} \tag{3.10}$$

E_{MM} and E_I is the atomistic GROMOS43a1 force field, all atoms are considered in the interface (I) region as well. E_{CG} is a Go model: it includes a harmonic potential for bonds between consecutive CG beads, and a pairwise Morse potential based on the pairwise aminoacid contacts, with the minimum centered on the distance between C in the native structure. The nonbonded component of $E_{\frac{CG}{I}}$ interaction potential has the same Morse potential shape and parameters of the E_{CG}, applied on both the Cα and Cβ of the residue in I. In this MM/CG model, we are not really interested in what the CG region does: we use the CG as a sensible scaffold that maintains (i)the molecule structural integrity and (ii)transmits and receives more or less correctly fluctuations to/from the MM part. The MM/CG

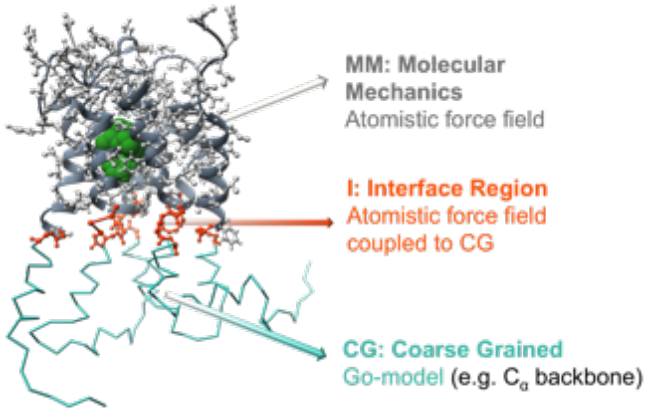

Fig. 3.7. Different components of the MMCG technique, from [161]

model has been tested on the HIV-1 protease and the human β-secretase: in both cases the enzyme active site has been modeled as MM, and the rest of the protein as CG. The MM/CG was later extended by Leguebe et al. [81] to perform simulation including also the membrane, thus an ideal system for GPCRs. The main additional ingredient is the addition of five virtual barriers, defined by five surfaces φ_i, i = 1 . . . 5

The potential of a predictive capacity of membrane MM/CG makes it suitable to applications such as binding pose prediction. We have seen that docking on homology models, while possible, is subject to several error sources due to the approximate nature of docking scoring functions, the fact that docking algorithms cannot explore, for performance reasons, the whole energy landscape of the ligand/protein complex, and the inherent error in the homology model structure. MM simulations of homology models are however fraught with peril, since simulation stability and protein structure quality are correlated. It is also known that homology model structures are hard to optimize by atomistic MD, possibly because the structure is in a frustrated local minimum with significant barriers from the global energy minimum. The same studies however show that restraining improves the results. The MM/CG approach might be used thus to simulate homology models of membrane proteins, by constraining most of the protein in a CG model, and leaving the (smaller) MM part to explore the (consequently smaller) phase space in a stable fashion. In turn, this can be used to recover correct binding poses of small ligands starting from docking configurations, with the added benefit of including dynamics in the description.

Part II

Contributions

4

Results

This chapter summarizes the main results of the thesis. Section 4.1 desribes results published in Suku and Giorgetti (2017) [162].

4.1 Common evolutionary binding mode of rhodopsin-like GPCRs

4.1.1 Abstract

G-protein Coupled Receptors (GPCRs) form the largest membrane protein superfamily in vertebrates. Advances in crystallization techniques so far resulted in the resolution of 44 unique receptors available for the GPCRs researchers community, 37 of which belong to rhodopsin-like GPCRs class (June 2017). We performed in this work the first systematic analysis of GPCRs binding cavities based on the available pool of rhodopsin-like solved structures. We pinpointed ten positions shared between all the solved receptors, namely 3.32, 3.33, 3,36, 6.48, 6.51, 6.52, 6.55, 7.35, 7.39 and 7.43, as interacting with ligands. We analyzed the conservation of amino acids present in these positions and clustered GPCRs accordingly to the physicochemical properties of binding cavities residues. Clustering supplied new interesting insights into the common binding mode of these receptors. In particular, the 3.32 position turned out to have an important role in ligand charge detection. Finally, we demonstrated that residues in these ten positions have co-evolved together, sharing a common evolutionary history. This work has been published in AIMS Biophysics in 2017.

4.1.2 Results

We used a pool of 85 complexes with known structure, belonging to the rho-GPCRs class to perform our analysis. First, we manually checked and deleted all the apo-complexes and redundant receptor-ligand complexes. Then we calculated, for each receptor, all the residues involved in ligand binding. Finally, we manually checked the type of interactions established between the ligand and the receptor, i.e. hydrogen bonds, salt bridges, -stacking or covalent interactions. In order to

refer to a unique system of coordinates, we used the generic numbering from the GPCRdb database. We extracted the binding positions by selecting only those shared in ligand binding between all the solved receptors. By using this approach, we were able to distinguish ten positions, namely 3.32, 3.33, 3.36, 6.48, 6.51, 6.52, 6.55, 7.35, 7.39, 7.43, that turned out to be functionally (positions involved in binding in all the solved GPCRs structures) conserved in 100% of the complexes, meaning that they are probably key positions for binding and function in all the structurally analyzed GPCRs. These positions include residues present in only three helixes, i.e. TM3, TM6, and TM7, in agreement also with a previous study. Afterwards, we used the GPCRdb mutant browser tool to check for the existence of experimental data involving residues in these ten positions. Indeed, significant site directed mutagenesis data, involving all the ten positions, were shown to reduce ligand binding/potency of about >5-fold further confirming the functional importance of these positions (see Figure 4.1).

The observation that these ten positions are present in all the rho-GPCRs binding cavities, without distinction of the sub-family, prompted us to suggest a role in the evolutionary history of the receptors. We thus performed a coevolution and a mutual information study on a rho-GPCRs curated alignment of 1618 eukaryotic sequences (alignment retrieved from the GPCRdb). Protein evolution is the result of natural selections of mutations that have functional advantages over other random mutations. The interactions of between a ligand and a protein from coevolution can be maintained by either direct binding or functional association. If two positions interact with each other, when one undergoes a mutation, the other position may have a compensatory mutation, otherwise, the two position cannot maintain the stability or functions of the interaction over the course of evolution. Evolutionary pressure thus creates coevolution pairs of positions that

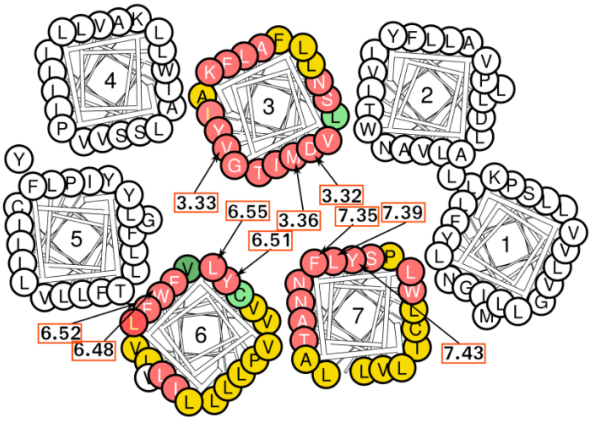

Fig. 4.1. Schematic representation of GPCRs experimental mutagenesis data, obtained by using the GPCRdb server. Red, green, and yellow positions that reduce the ligand binding/potency by >5-fold, increase the ligand binding/potency by >5-fold and have no or low effect in the binding affinity (<5-fold), respectively. Helixes are numbered from one to seven.

maintain the interactions. We observed high values of cumulative mutual information (cMI) for the ten previously calculated positions, meaning that these positions could have played an important evolutionary role. Then we calculated the amino acid conservation of the ten shared positions. From this analysis emerged a high value of conservation, of about 77%, for the tryptophan in position 6.48. This position is well known in literature since it is a hub involved either in ligand detection and receptor activation and we already discussed it in the Introduction section. All the other positions, except for 3.32, present hydrophobic amino acids, i.e. valine, methionine, leucine as the most conserved amino acids. Position 3.32 drawn our attention because it presents an aspartic acid in 22% of the rho-GPCRs. This aspartic acid is known in the literature to be responsible for interacting with amines, small positively charged molecules. We thus investigated if, using residues present in the 3.32 position as features for the clustering, could lead to a discrimination between receptors with similar physicochemical properties. Indeed, this first clustering-step showed three principal groups: (i) receptors with a negatively charged residue (amine cluster), (ii) receptors with a hydrophobic residue (hydrophobic cluster) and (iii) receptors with an aromatic residue (aromatic cluster) (see 4.2).

Then, in order to verify if these results could be correlated with the different type of ligands that activate these receptors, we cross-checked the clustering-step by grouping the endogenous ligands accordingly to their charges. We calculated the charges of 35 endogenous agonists (see Supplementary Table 2). The ligands were clustered into three main groups as well: (i) positively charged ligands that bind amine receptors as acetylcholine, histamine, adrenaline etc., that correspond to the receptors amine cluster, (ii) neutral molecules as the case of retinol, that binds opsin receptors or adenosine that binds adenosine receptors, as well as melatonin

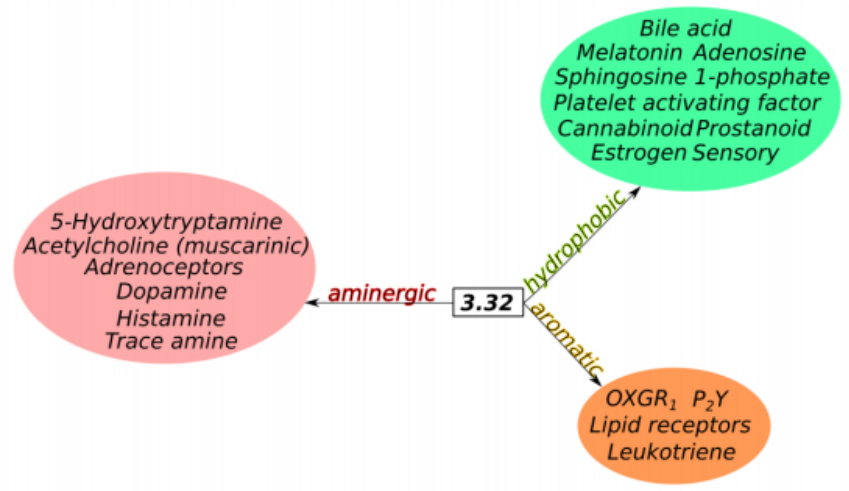

Fig. 4.2. Clustering results using the physicochemical properties of residues corresponding to position 3.32. Amine cluster is shown in red, hydrophobic cluster is shown in green and aromatic cluster in shown in orange.

that bind melatonin receptors and anandamide that binds cannabinoid receptors, all grouped in the hydrophobic cluster and (iii) negatively charged molecules as those that activate lipid or nucleotide receptors, present in the aromatic cluster. This cross-check confirmed that physicochemical characteristics of the residues in position 3.32 were enough to discriminate receptors based on the charge of their endogenous agonists. The same behavior was observed also in the case of peptide receptors that have a negatively charged residue in position 3.32. Indeed, melanin-concentrating hormone, opioids, neuropeptide W/B, somatostatin and urotensin receptors came out to be activated mostly by positively non-endogenous ligands. Taken together these results indicate a huge importance of 3.32 position in ligand charge detection during evolution and an indirect involvement of this position in receptor activation. Prompted by the encouraging result of the clustering method, we then considered all the ten previously identified positions, namely 3.32, 3.33, 3.36, 6.48, 6.51, 6.52, 6.55, 7.35, 7.39, 7.43, for a second clustering-step. As the shared positions are ten, we used ten different thresholds in our approach. At threshold equal to ten (first level of the agglomerative clustering) shared positions, receptors that have residues with similar physicochemical properties in all the positions are clustered together. On the other hand, at threshold equal to one (last level of the agglomerative clustering), receptors sharing only one similar residue are clustered together. We obtained 394 clusters for the first level and 10 clusters for the tenth level. Concerning the first level, we noticed that clustering was strongly species-dependent with mammalians sequences mostly clustered together, and other more distant species, like fishes, forming separate clusters. We focused instead our attention on the seventh level of the agglomerative cluster, with a threshold of at least three shared positions. We chose this level because it showed an optimal ratio between number of members and biological relevance of the clusters. Thus, at this point, we were interested in investigating if, sharing only three residues in the binding cavity, could be enough to explain the evolutionary history of rho-GPCRs. We compared our clusters against the GPCR network phylogenetic trees sub-branches and against data in literature containing experimental information on the GPCRs. First, all the amine receptors (Figure 4.3, cluster α) were clustered within the same cluster, similarly as in the classical GPCR phylogenetic tree. Indeed, all the amine receptors shared five positions out of ten in the binding site (Figure 4.3). In other four clusters (clusters λ, κ, μ and ν, Figure 4.3), we can distinguish Adenosine, Rhodopsin, P2Y and Chemokine receptors. While these receptors are very similar within their local branches (subfamilies), they differ throughout the rho-GPCRs class. In our clusters analysis, we observed the same trend as in the GPCRs phylogenetic tree. In cluster β positions in the binding cavity (Figure 4.3). From a biological point of view, Motilin and Ghrelin receptors are both used as drug targets for gastrointestinal disorders. Considering only their binding cavities, our clustering method was able to give informative results regarding the common pathway of these receptors. Cannabinoid receptors and Melanocortin receptors are included in the same cluster (cluster θ, Figure 4.3). These two receptors were discovered to be expressed as a chimeric protein in an isolated fragment of a leech CNS, an invertebrate species. Surprisingly, in the eukaryotic species, they share only three out of ten positions in the binding site. These positions, that could have played a key role during the evolution of these

two receptors, were captured with our method. Cluster η grouped together Lipid, Prostanoid, Cholecystokinin and Relaxin receptors (Figure 4.3). Cholecystokinin is expressed in the gastrointestinal system and is responsible for stimulating the digestion of lipids, and thus belongs to the same biological pathway of lipid receptors. In fact, Harikumar KG and collaborators have shown a high correlation between a microenvironment rich of lipids and the inactive, uncoupled state of the Cholecystokinin receptor. On the last cluster (cluster δ, Figure 4.3) we can find Opioid, Endothelin and Oxytocin receptors together. Opioid and Endothelin receptors share a common antagonist, that inhibits both receptors. Regarding the Oxytocin receptor, it seems that both Opioid and Oxytocin receptors play important roles in pain modulation. Indeed, the opioid system is involved in the oxytocin-induced antinociception in the brain of rats. Thus, our clustering method was able to capture important biological features, that are difficult to be captured using classical phylogenetic approaches.

4.1.3 Conclusion

Here we present the first global analysis on the rho-GPCRs binding cavities, based on all the solved structures of these receptors. In total, we analyzed 85 complexes and from our structural- based GPCRs analysis, we found previously uncovered properties of the binding cavities. First, we found that ten positions of the GPCRs binding cavities, namely 3.32, 3.33, 3.36, 6.48, 6.51, 6.52, 6.55, 7.35, 7.39, 7.43, are shared between all the rho-GPCRs solved structures. They are located in three helices, i.e. TM3, TM6, and TM7. These helices together with TM2 have been previously shown to play a fundamental role in GPCRs activation. This leads us

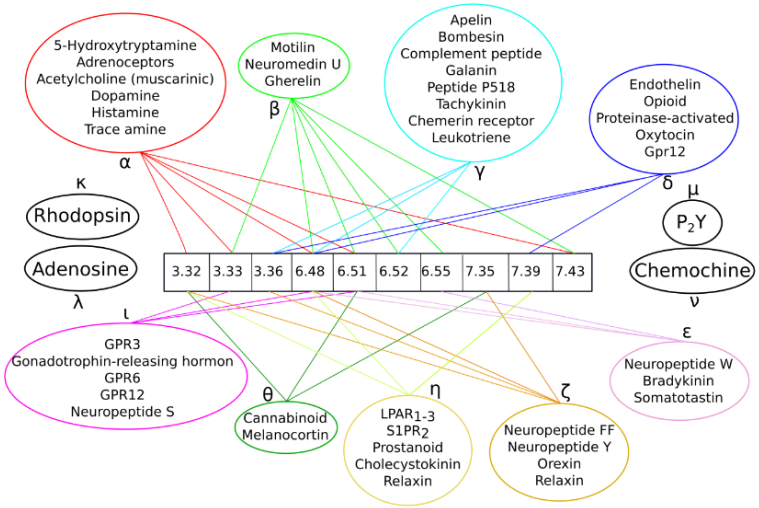

Fig. 4.3. Clustering-step using, as features, residues corresponding to the ten shared binding cavity positions. Each cluster is shown with a differently colored circle and labeled using Greek letters. The shared positions within a cluster are also illustrated with the same colored lines

to believe that our findings could be strictly connected with the activation of these receptors. In fact, the transmission of the signal in GPCRs starts with agonist binding and continue, through hinge residues interactions, towards the G-protein binding cavity. Our findings could be used as the starting point of further studies aiming at a deep learning of GPCRs activation. Using similar physicochemical properties of residues in these ten positions as features, we then were able to cluster together receptors that are very distant between each other at a sequence level, but very close in ligand recognition and binding cavities similarities. We showed that on one hand, some receptors were clustered together in a very similar way to branches of the GPCR network phylogenetic tree. On the other hand, we found clustered together receptors completely different at a sequence level but belonging to the same biological pathway. An example is the case of Opioid and Bradykinin receptors that interact with the same/very similar ligands. The method of the agglomerative clustering used here was able to capture important features of receptors binding cavities that are very difficult to be recognized using classical phylogenetic approaches. Moreover, position 3.32 that seems to have played an important role in agonists charge detection during GPCRs evolution. To the best of our knowledge, this is the first time that a similar analysis is performed on all the rho-GPCRs solved structures. We believe that our results can help in the deorphanization of GPCRs whose ligands are still unknown, as well as in suggesting novel specific drug targets.

4.2 Analysis of the Muscarinic Receptors binding site

Section 4.2 desribes results published in Radu et al. (2017) [163]

4.2.1 Abstract

Clinical and experimental studies indicate that muscarinic acetylcholine receptors are potential pharmacological targets for the treatment of neurological diseases. Although these receptors have been described in human, bovine and rat cerebral microvascular tissue, a subtype functional characterization in mouse brain endothelium is lacking. Here, we show that all muscarinic acetylcholine receptors (M1-M5) are expressed in mouse brain microvascular endothelial cells. The mRNA expression of M2, M3, and M5 correlates with their respective protein abundance, but a mismatch exists for M1 and M4 mRNA versus protein levels. Acetylcholine activates calcium transients in brain endothelium via muscarinic, but not nicotinic, receptors. Moreover bioinformatic analyses performed on eukaryotic muscarinic receptors demonstrate a high degree of conservation of the orthosteric binding site and a great variability of the allosteric site. In line with previous studies, this result indicates muscarinic acetylcholine receptors as potential pharmacological targets in future translational studies. We argue that research on drug development should especially focus on the allosteric binding sites of the M1 and M3 receptors.

4.2.2 Results: The variability of mouse muscarinic receptors lies in the allosteric binding site

The mAChRs are a sub-class of the G-protein-coupled receptors (GPCRs) family, comprising 5 subtypes (M1-M5). M1, M3, and M5 are coupled with the Gq protein and, via phospholipase C signaling pathway, generate cytosolic calcium transients; M2 and M4, on the other hand, couple with the Gi protein inhibiting the adenylyl cyclase. While obtaining mAChRs subtype-selective ligands is a primary goal in drug development, previous attempts have failed due to the highly conserved structure of the orthosteric binding site across the muscarinic receptor family members. On the other hand, the allosteric binding site seems to hold promise as a specific pharmacological target. Yet, despite considerable recent progress in crystallography and molecular modeling of mAChRs (as well as the successful crystallization of human M1, M2, M4, and rat M3 receptors), no 3D structure predictions based on homology modeling studies have been carried out for mouse muscarinic receptors. We complemented the functional studies performed by our collaborators with a thorough conservation analysis of the residues located in the allosteric and orthosteric binding cavities. To such aim, we performed a multiple sequence alignment of all eukaryotic muscarinic receptors annotated in the GPCRdb database. To define orthosteric and allosteric residues, a water level was set in the intracellular half of the receptors. We validated this subdivision against the residues known to be involved in ligand-receptor interactions of both cavities in the crystal structures. From our alignment, the orthosteric cavity reflects the full conservation of the residues (Figure 4.4). Conversely, the residues present in the allosteric binding site emerge as highly variable. To the best of our knowledge, this is the first time

that a conservation analysis has been performed considering muscarinic receptors along all annotated eukaryotic species. We then performed docking experiments on M1-M5 modeled mouse muscarinic receptors with the four antagonists: Telenzepine, J104129 fumarate, VU 0255035 and 4-DAMP. In the case of M5 we docked also its highly specific antagonist: VU 0488130 (also called ML38144). Within the limitations of the method, it can be observed that, while three of these antagonists (Telenzepine, J104129 fumarate, 4-DAMP) interact with residues located mainly in the orthosteric binding site, VU 0255035 and VU 0488130, probably due to their larger size, interact also with residues positioned in the allosteric binding sites. In particular, VU 0488130 may interact with residues located in the allosteric cavity that are exclusively present in the M5 receptor, i.e Q145 and K470. Hence, differences in affinities could be explained by the observation that the VU 0255035 and VU 0488130 networks of interactions involve less conserved regions.

A

Generic residues numbering (Class A)	Conservation residues percentage		
2.60	F=32%	Y=68%	
2.62	T=56%	L=24%	V,A=10% each
3.22	T,P=21% each	S,A=16% each	D,W,Y=1% each
3.32	D=100%		
3.33	y=100%		
3.39	S=100%		
ECL2	P=59%	L,E=16% each	M,H=1% each
ECL2	P=35%	D,A=24% each	S,L,K=1% each
ECL2	G=62%	E=24%	N=8%
ECL2	E=67%	Q=24%	T,A=1%each
5.36	E=40%	N=40%	Q=16%
5.37	P=70%	A=30%	
6.44	F=100%		
6.48	W=100%		
6.51	Y=100%		
6.52	N=98%	Y=2%	
ECL3	D=40%	A,K=18% each	Q=10%
ECL3	S=29%	L,P, D=18% each	C,N=1% each
7.31	N,K,E=23%	E,V=15%	
7.32	T=72%	F,V,R,A,E=3% each	
7.38	Y=100%		
7.42	Y=100%		

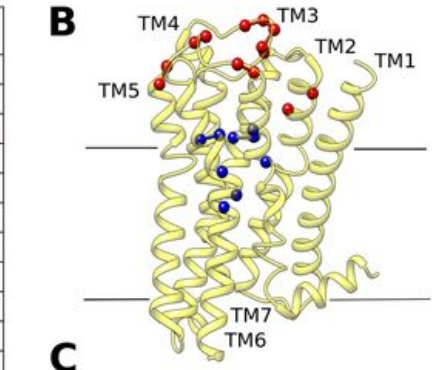

Fig. 4.4. Conservation of orthosteric and allosteric site residues among annotated eukaryotic muscarinic receptors. (A) Conservation values of residues located in allosteric (red) and orthosteric (blue) binding sites. (B) Orthosteric and allosteric binding-site residues mapped on the M3 solved structure (PDB code: 4MQS). The residues are indicated as colored balls, following the same color code as in (A). (C) Plot of residues conservation values; residues of the orthosteric binding site (blue) and residues in the allosteric binding site (red).

4.3 Analysis of the Kisspeptine Receptors binding site

Section 4.3 desribes results published in Mechaly et al. (2018) [164]

4.3.1 Abstract

Kisspeptin receptors are G-Protein-Coupled Receptors that regulate GnRH synthesis and release in vertebrates. Here, we report the gene structure of two kisspeptin receptors (kissr2 and kissr3) in pejerrey fish. Genomic analysis exposed a gene structure with 5 exons and 4 introns for kissr2 and 6 exons and 5 introns for kissr3. Two alternative variants for both genes, named kissr2-v1 and kissr2-v2, and kissr3-v1 and kissr3-v2, were revealed by gene expression analyses of several tissues. For both receptors, these variants were originated by alternative splicing retaining intron 3 and intron 4 for kissr2-v2 and kissr3-v2, respectively. In the case of kissr2, the intron retention introduced two stop codons leading to a putatively truncated protein whereas for kissr3, the intron retention produced a reading shift leading to a stop codon in exon 5. Modeling and structural analysis of Kissr2 and Kissr3 spliced variants revealed that truncation of the proteins may lead to non-functional proteins, as the structural elements missing are critical for receptor function. To understand the functional significance of splicing variants, the expression pattern for kissr2 was characterized on fish subjected to different diets. Fasting induced an up-regulation of kissr2-v1 in the hypothalamus, a brain region implicated in control of reproduction and food intake, with no expression of kissr2-v2. On the other hand, fasting did not elicit differential expression in testes and habenula. These results suggest that alternative splicing may play a role in regulating Kissr2 function in pejerrey.

4.3.2 Results

In 2001, a member of the Rhodopsin family, the kisspeptin receptor KISS1R (previously named GPR54) was shown to be activated by polypeptides kisspeptin-54,-14,-13, and-10. A few years later, kisspeptin and its receptor were regarded as essential regulators of the reproductive axis, since hypogonadotropic hypogonadism in both humans and mice was shown to be associated with mutations of KISS1R. Moreover, kisspeptin and its receptor were linked to other functions such as insulin secretion, vasoconstriction, tumor biology and the metastatic process, antioxidant function in oxidative stress, anticoagulation, and brain sex differentiation. In the present study, we report the predicted structure of two kissr genes, kissr2 and kissr3 in pejerrey fish (Odontesthes bonariensis). We also identify new alternative spliced variants for each receptor and provide preliminary evidence suggesting loss of function of variants due to intron retention. We also test the expression pattern of kissr2-v1 and kissr2-v2 in pejerrey hypothalamus after fasting, because a similar condition was reported to increase not only hypothalamic kiss2 but also kissr2 in S. senegalensis.

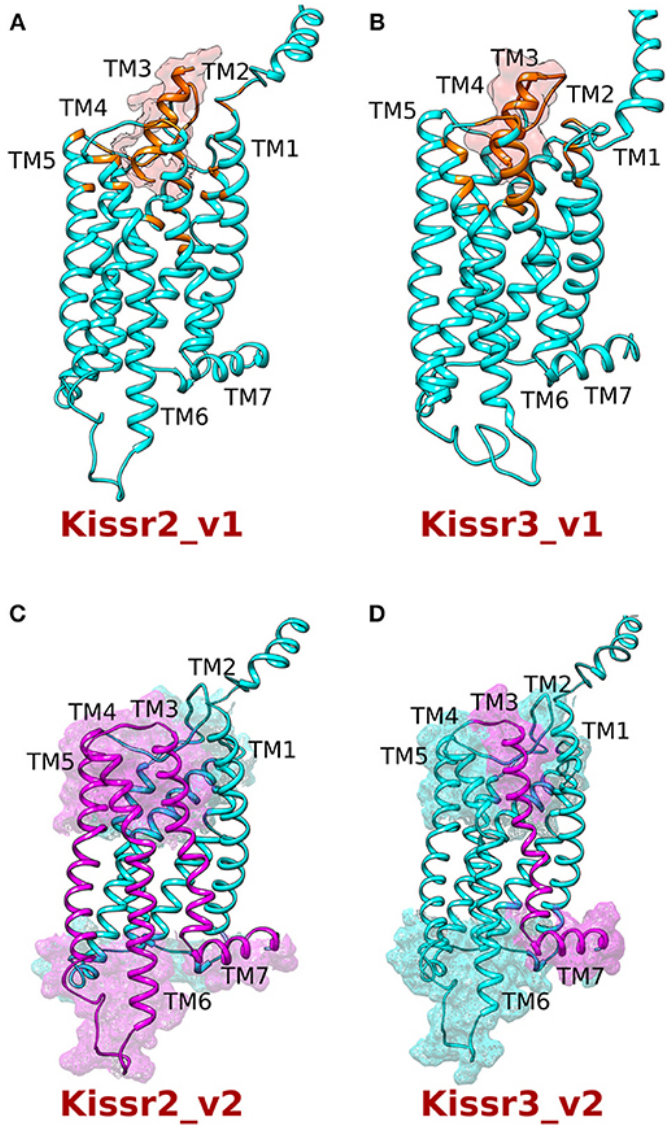

Fig. 4.5. Homology modeling of the kisspeptin receptors (Kissr2 and Kissr3) with their respective peptides kiss1 and kiss2. (A) Kissr2-v1 receptor (cyan) and (B) Kissr3-v1 receptor (cyan) with their respective peptides (orange). The residues located within 5 from the peptides are shown in orange. (C,D) The truncated region of both the receptors is shown in violet. While receptor Kissr2-v2, lacks TM5-TM7 helixes; Kissr3-v2 loses TM7 and portion of the extracellular loop 3.

4.3.3 Conclusions

Our findings suggest a novel kissr2 gene regulatory mechanism in the hypothalamus involving expression of alternatively spliced variants with intron retention that produce potentially non-functional proteins. Homology 3D models of pejer-

rey Kissr2 and Kissr3 structures were built using the on-line platform GOMoDo (Figure 4.5). Their respective peptides were then docked in the predicted binding cavities by using the Haddock program accessible also through the GOMoDo server. From the models it can be observed that: (i) in Kissr2-v1 the putative ligand binding cavity is formed by residues of TM3 (Gln125, Gln126, Val129, Gln130), ECL3 (Tyr197, Cys198, Glu200), TM5 (Gln215, Tyr220), TM6 (Leu276, Trp281, Ile284, Gln285), and TM7 (Asn311, Tyr315) (Figure 4A); and (ii) in pejerrey the Kissr3-v1 putative binding cavity is formed by residues of TM3 (Gln114, Gln115, Ala118, Gln119), ECL3 (Gln183, Thr184, Cys186), TM5 (Ser203, Tyr208), TM6 (Leu264, Trp269, Ile272, Gln273), and TM7 (His296, Tyr300). It is important to note that in both receptors the residues that are putatively crucial for ligand and G-protein binding (according to the prediction of the method used) belong to helices TM5-7, just like several other GPCRs that we analyzed before. This evidence suggests that loss of these helices in variants kissr2-v2 and kissr3-v2 could compromise receptor structure, function, or dimerization.

4.4 Allosteric sodium binding cavity in GPR3: a novel player in modulation of A production

Section 4.4 desribes results published in Suku et al. (2018) [165]

4.4.1 Abstract

The orphan G-protein coupled receptor 3 (GPR3) belongs to class A G-protein coupled receptors (GPCRs) and is highly expressed in central nervous system neurons. Among other functions, it is likely associated with neuron differentiation and maturation. Recently, GPR3 has also been linked to the production of A peptides in neurons. Unfortunately, the lack of experimental structural information for this receptor hampers a deep characterization of its function. Here, using an in-silico and in-vitro combined approach, we describe, for the first time, structural characteristics of GPR3 receptor underlying its function: the agonist binding site and the allosteric sodium binding cavity. We identified and validated by alanine-scanning mutagenesis the role of three functionally relevant residues: Cys2676.55, Phe1203.36 and Asp2.50. The latter, when mutated into alanine, completely abolished the constitutive and agonist-stimulated adenylate cyclase activity of GPR3 receptor by disrupting its sodium binding cavity. Interestingly, this is correlated with a decrease in A production in a model cell line. Taken together, these results suggest an important role of the allosteric sodium binding site for GPR3 activity and open a possible avenue for the modulation of A production in the Alzheimers Disease.

4.4.2 Results: Homology modeling and molecular dynamics simulations

GPR3 is a fascinating GPCR receptor both from a structural and functional point of view. It is known to be involved in many molecular pathways, from modulating the early phases of cocaine reinforcement to the maintenance of meiotic arrest in rodent oocytes and emotional-like responses. Very recently it was discovered to play a fundamental role in modulating the amyloid-beta peptide generation in neurons through the interaction with β arrestin 2(βarr2). The finding that certain G proteincoupled receptors (GPCRs), including also the β2-adrenergic receptor in addition to the GPR3, can regulate Aβ production has offered new avenues for Alzheimers drug discovery. In fact, whereas genetic ablation of GPR3 reduced Aβ levels, the overexpression of the latter increased A production in Alzheimers mouse model. The fact that GPR3 could be the key to find new treatments for the Alzheimers Disease, makes this receptor an ideal case of study especially from a structural point of view. Unfortunately, the GPR3 receptor does have neither a known 3D structure nor a known endogenous agonist. Only a study on its constitutive activity has been reported for this receptor as well as some data on two non-endogenous ligands: (i) an agonist, DPI34 and (ii) an inverse agonist, cannabidiol41. Although the latter has been recently associated to GPR3, it is not specific for this receptor as it interacts also with a close homolog, GPR6. Thus, forthis work we did not consider it. The model of the GPR3 receptor was built

based on the active structure of human A2a adenosine receptor (PDB code: 5G53), using the GOMoDo webserver. The GPR3 receptor shared almost 23% of sequence identity with the template and this value was within the range of the identities between the target and its best templates. However, this template turned out to be the most reasonable in terms of MODELLER score and its conformational active state. The target-template alignment was then manually checked in order to verify the presence of ALL the conserved features of the GPCRs family as the X.50 in each transmembrane helix, the DRY motif in transmembrane helix 3 and the NPxxY motif in transmembrane helix 746. All the conserved features were preserved except for the disulfide bridge between the extracellular loop 2 (ECL2) and transmembrane helix 3 (TM3). Indeed, GPR3 has no cysteines in the TM3. The generated model was then used to perform *in silico* docking experiments using the Haddock program through the GOMoDo webserver. The residues located in the top half part of the receptor were predicted as located in the putative binding cavity and used as ambiguous interaction restraints (AIR) for the docking step. Once the last docking step was completed, all the complexes (200 in total) were clustered. The best docking GPR3-DPI pose (Figure 4.6) was chosen as the one with the lowest HADDOCK score within the most populated cluster. In that conformation, the synthetic agonist DPI is positioned inside the canonical GPCRs orthostatic cavity. GPR3s putative binding cavity results mostly hydrophobic, with the phenyl rings of DPI interacting with Leu2837.39, Leu1133.32, Trp2606.48 (Figure 4.6), and Val186ECL2. Among all the interactions, two specific interactions captured our attention, (i) a halogen-bond interaction49 between the iodine atom of DPI and Cys2676.55 (almost 4$\mathring{A}$) and (ii) a sandwich-like conformation in which DPI is inserted between two phenylalanine residues, Phe1203.36 and Phe2636.51 (Figure 4.6).

Then, following the Methods section, in order to better sample the conformational space of the ligand within the putative binding cavity, the best complex was funneled to perform molecular dynamics (MD) simulations using a hybrid molecular mechanics/coarse-grained (MM/CG) approach in order to exhaustively explore the conformational space of the ligand, the binding cavity, and the hydration shell. A detailed description of the MM/CG can be found in the Methods section. The system underwent 700 nanoseconds (ns) of simulations at room temperature, reaching the stability after 300ns. We then clustered all the trajectory and analyzed the representative conformation of the most populated cluster (Figure 4.7). We noticed very few differences comparing the docking and the simulations results. The simulations relaxed and did not alter the receptor/DPI interactions compared with the initial conformation (Figure 4.7, red color and green color). Indeed, during the simulations, DPI slightly shifted and tilted from its initial position, assuming a non-planar conformation, maintaining however the interaction with Cys2676.55 which side chain moved towards to the iodine atom at distance <4$\mathring{A}$. Simulations thus confirmed the halogen-bond interaction predicted by docking experiments. Moreover, also other two residues involved in the docking predictions, Phe120$^{3.36}$ and Phe263$^{6.51}$ confirmed their contribution in the ligand binding, shifting the side chains accordingly the DPI rings position and maintaining the π-stacking interactions with the ligand.

Furthermore, we decided to study also the putative sodium (Na+) allosteric binding site, that has a fundamental importance in allosteric modulation of GPCRs, as discussed in the Introduction section. The residues that mostly contribute to sodium binding along the GPCRs family, i.e S3.39, N7.45 and D2.50 are described in literature as highly conserved. In particular, we observed that residue in position 2.50 is an aspartic acid in 90% of the eukaryotic GPCRs accordingly to the curated multiple sequence alignment of the GPCRdb. This residue can highly modulate the function of GPCRs. The role of sodium modulation is in fact well known for GPCRs. Mutagenesis studies on residues involved in Na+ coordination, and in particular Asp2.50, highlighted the different effects that allosteric sodium may have in various class A GPCRs signaling. Indeed, Asp2.50 replacement with uncharged amino acids can drastically reduce the agonist-induced G-protein activation or modulate the allosteric effect of the G-protein on ligand binding. The presence of sodium ions in the allosteric cavity can also exert different effects on the constitutive signaling of GPCRs. In many cases, the presence of bound Na+ seems to stabilize the inactive conformation of the receptor reducing the constitutive G-protein, whereas in other receptors the substitution of Na+ coordinating Asp250 abolishes the constitutive G-protein coupling and activation without affecting the agonist-stimulated activity. Exhaustive studies have also revealed that the Na+ pocket collapses due to the activation-related movements of the transmembrane helixes. In the allosteric binding site, Na+ is coordinated by a salt bridge formed

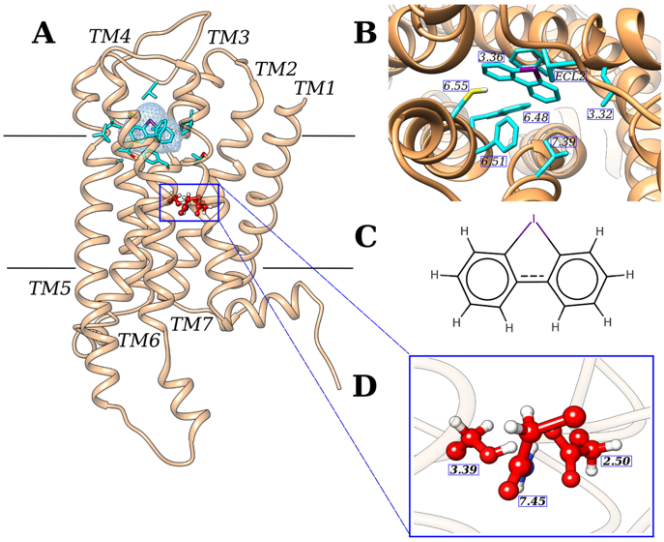

Fig. 4.6. GPR3-DPI complex docking results. In all the panels the GPR3 receptor is shown in salmon and its agonist DPI is shown in cyan. Residues side chains located in the orthosteric binding cavity are shown in cyan, while residues located in the sodium allosteric binding cavity are shown in red. The receptor is oriented with the N-terminus in the extracellular part and the C-terminus in the intracellular part (A), DPI and side chains of residues 5 distant from the agonist are shown in cyan (B), chemical structure of DPI with the iodine atom indicated in violet (C), side chains of residues involved in allosteric sodium binding are shown in red (D).

with Asp2.50 together with other additional polar interactions with Ser3.39 and Asn7.49. Most of the studies agree with the fact that the constitutive activity can be dramatically affected by mutations in Asp2.50. In the light of these results together with the results obtained with our simulations, we decided to proceed with in vitro experiments to validate/reject our hypothesis. We performed wet-lab alanine scanning mutagenesis on residues Cys267$^{6.55}$ and Phe120$^{3.36}$, putatively involved in halogen and π-stacking interactions with the ligand, respectively, and on D86$^{2.50}$, the highly conserved acidic residue present in the putative allosteric Na+ binding site. (Figure 4.8) displays dose-response curves obtained measuring the cAMP concentration in HEK293 cells transfected with WT or mutant receptors and treated with increasing concentration of DPI. The deletion of Cys267$^{6.55}$ or Phe120$^{3.36}$ side chains, predicted by *in silico* experiments to be involved in DPI binding, has the effect to increase the DPI EC50 from $\sim 2\mu$M in GPR3 WT to 5μM and 15μM, respectively. This can be explained by a decreased affinity for the agonist due to a reduction of molecular contacts in the binding cavity when mutants are introduced, and further supports the accuracy of our model. Conversely to the previous two mutants, the mutation in alanine of Asp86$^{2.50}$, putatively involved in the allosteric Na+ binding site completely abolished either the constitutive and DPI-induced stimulation of adenylyl cyclase by GPR3, suggesting that this mutation produces a totally inactive form of the receptor. These results point out that binding of allosteric Na+ is essential for GPR3 to maintain its constitutive activity or to assume an active conformation.

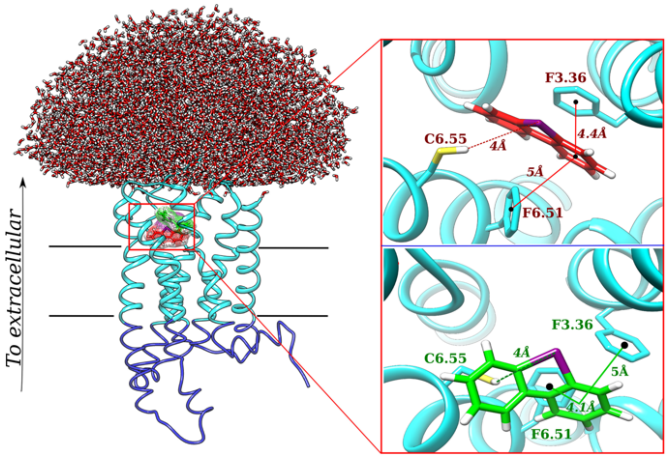

Fig. 4.7. Molecular dynamics results of DPI located in the GPR3 orthosteric binding cavity. DPI-GPR3 complex, together with the water dome that surrounds the The MM and CG regions of the receptors are shown in cyan and blue colors, respectively. The receptor is oriented with the N-terminus in the extracellular part and the C-terminus in the intracellular part. In the top right part the best docking complex is shown. While DPI located in the orthosteric binding cavity of the GPR3 receptor is shown in red, residues Cys6.55, F3.36 and F6.51 are shown in cyan. In the bottom right, MM/CG simulations of the best complex is represented. DPI is shown in green, while Cys6.55, F3.36 and F6.51 residues are shown in cyan.

Next, we investigated the effect of constitutive and agonist-induced activity of our mutants on βarr2 interaction. β-arrestin proteins are ubiquitous modulators of GPCRs signaling that recognize and bind to specific phosphorylated residues in the C-terminal tail of active GPCRs and antagonize the interaction with the G-protein. This promotes the desensitization and the internalization of the receptor. HEK293 cells transfected with the WT receptor or three mutants were treated 30minutes with DMSO (vehicle control) or 50μM DPI and the amount of intracellular cAMP was determined. Compared with the empty vector control, the constitutive activity of the WT receptor results in a 3-fold increase in cAMP level in unstimulated cells, whereas the expression of F120A and C267A mutants produces a lower, but still significant, increase of cAMP (Figure 4.9, black bars). Upon DPI stimulation, an up to 10-fold increment in cAMP concentration is observed for the WT receptor and 5 to 6-fold for F120A and C267A (Figure 4.9, red bars). Again, neither constitutive nor DPI-induced activity is detected for D86A mutant. When the same experiment is conducted in presence of co-expressed βrr2 (Figure 4.9, 5B), a sensible decrease in constitutive activity of WT, F120A and C267A mutants is observed, while no considerable effect can be detected for control and D86A mutated receptor. DPI stimulation still produces an increase in the cAMP level compared to the control (except for D86A mutant), but remarkably lower than in absence of βarr2. These findings suggest that, like WT GPR3, activated F120A and C267A mutants are negatively modulated by βarr2-mediated desensitization,

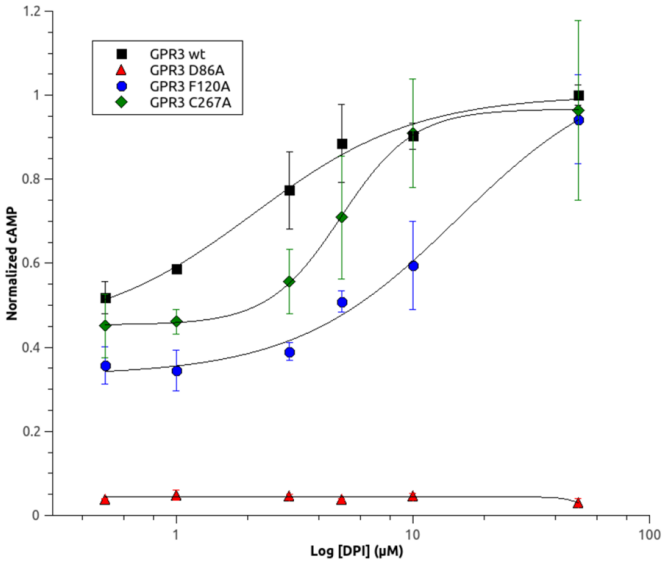

Fig. 4.8. Effect of single point mutations on DPI-induced activation of GPR3. Dose-response curves for DPI in HEK293 cells expressing WT and single point mutants of GPR3. Twenty four hours after transfection, cells were stimulated for 30minutes with increasing concentration of DPI and the intracellular cAMP level was measured. cAMP values were normalized to the maximal response. Nonlinear regression analysis was performed to generate dose-response curves and calculate concentration for 50% of the maximal effect (EC50). Data are the mean±SD of 3 independent experiments

whereas D86A, being totally inactive and therefore likely not-phosphorylated by G protein-coupled receptor kinases (GRKs), is probably unable to interact with βarr2 and this does not allow to appreciate any modulating effect. D68A mutant, although retaining proper folding and localization, appears to be completely unable to stimulate adenylate cyclase both in constitutive conditions and upon agonist stimulation. In this view, the presence of the highly conserved negatively charged Asp250, and consequently a complete and functional allosteric Na+ binding pocket, seems to be essential for this receptor to maintain its constitutive activity and to activate Gs-protein for downstream signaling. The unraveling of the functional role of the sodium ion in the activation of GPR3 certainly deserves a deeper investigation.

Moreover we investigated the correlation between the modulating effects of our mutants on GPR3 activity and the stimulated production of Aβ peptides. H4swe cells, expressing the Swedish mutation (K595N/M596L) of amyloid precursor protein (APP-swe), were transfected with WT or mutated GPR3 (alone or in presence of co-transfected βarr2) and the amount of Aβ 142 released in the culture medium was measured by ELISA 24hours after transfection (Fig. 6A). As expected, a statistically relevant (p¡0.05) increment of Aβ with respect to the control is observed for WT GPR3, F120A and C267A, while the inactive mutant D86A does not significantly increase the amount of secreted amyloid peptide. When βarr2 is co-expressed with the receptor (cyan bars), the picture is less clear. βarr2 induces a slight increment in the Aβ 142 level in the control (probably modulating other signaling pathways in the cells) but does not affect significantly the amount of peptide produced in cells transfected with WT receptor or mutants. As a result, we could not detect any statistically relevant effect of βarr2 on GPR3 in this cellular model. To further assess the correlation between the activation of GPR3 and

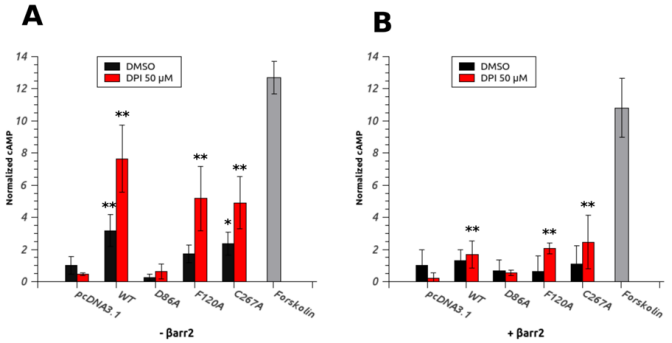

Fig. 4.9. βarr2 decreases constitutive and DPI-induced activity of GPR3. (A) HEK293 cells transfected with empty vector (pcDNA3.1), WT GPR3 and single point mutants were treated with DMSO or 50μM DPI for 30minutes and the intracellular cAMP level was measured. (B) The same experiment as in A was performed in presence of co-transfected human βarr2. 10M Forskolin, a known adenylate cyclase activator, was used as a positive control. Data are the mean±SD of four independent experiments. cAMP values were normalized to the control (empty vector untreated). Statistical significance was determined by one-way ANOVA with Bonferroni-Holmes post-hoc test comparing all samples with the relative control (empty vector) (*p<0.05, **p<0.01).

its ability to enhance the production of Aβ, we performed the same experiment in presence of an agonist. Due to the lack of knowledge regarding its physiological ligand we used DPI, the only known compound able to activate GPR3. Due to DPI poor selectivity (e.g. it is known to strongly inhibit nitric oxide synthetase from macrophages and endothelial cells and other flavoenzymes) and cell toxicity, even at micromolar concentration, DPI does not have any potential therapeutic application, but it is a useful experimental tool in studying GPR3 signaling in vitro. In this case, the transfected cells were incubated 24hours with 1μM DPI (a prolonged exposure to higher concentration of DPI resulted in higher cell toxicity) and the amount of Aβ 142 was quantified as before (Figure 4.10). Compared to the empty vector control, DPI stimulation produces a almost 50% increase in the amyloid peptide level for WT and, to a less extent, for F120A and C267A, whilst had no influence on D86A. Again, co-transfection with βarr2 produces no appreciable difference in the production of Aβ promoted by WT GPR3 or mutants and the control in these cell line. Taken together, our results prompt us to suggest that there is a correlation between the permanence of the receptor in the active state and its modulation role on γ-secretase complex, although this process has been reported to be independent of G-protein activation. Indeed, DPI stimulation proportionally increases Aβ production in WT and agonist-sensitive mutants, while D86A mutant, devoid of any cAMP stimulation activity and unable to gain access to the active state, is also ineffective in stimulating the production of amyloid peptides. Once activated, GPCRs are phosphorylated at specific positions by GPR kinases that specifically recognize the active form of the receptor and this modification considerably increase the recruitment and binding of β-arrestins. In this view, although indirectly, our findings further support the hypothesis of the involvement of β-arrestin mediated desensitization/internalization pathway in GPR3 modulation of Aβ secretion.

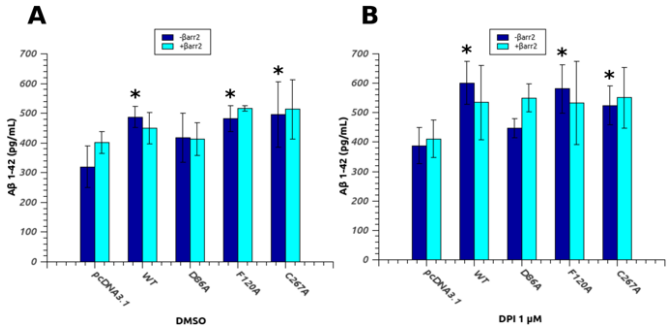

Fig. 4.10. D86A mutation reduces the GPR3-stimulated A production in H4swe cells. H4swe cells transfected with empty vector (pcDNA3.1), WT GPR3 and single point mutants, either in absence (blue bars) or in presence (cyan bars) of co-transfected human arr2, were treated with DMSO (A) or 1M DPI (B) for 24hours and the amount of A 142 in the culture medium was determined by ELISA. Data are the meanSD of three independent experiments. Statistical significance was determined by one-way ANOVA with Bonferroni-Holmes post-hoc test comparing all samples with the relative control (empty vector) (*p¡0.05).

4.5 Agonist Binding to Chemosensory Receptors: Receptor Activation Predictions

Section 4.5 desribes results published in Suku et al. (2017) [166]

4.5.1 Abstract

Human G-protein coupled receptors (hGPCRs) constitute a large and highly pharmaceutically relevant membrane receptor superfamily. About half of the hGPCRs' family members are chemosensory receptors, involved in bitter taste and olfaction, along with a variety of other physiological processes. Hence these receptors constitute promising targets for pharmaceutical intervention. Molecular modeling has been so far the most important tool to get insights on agonist binding and receptor activation. Here we investigate both aspects by bioinformatics-based predictions across all bitter taste and odorant receptors for which site-directed mutagenesis data are available. First, we observe that state-of-the-art homology modeling combined with previously used docking procedures turned out to reproduce only a limited fraction of ligand/receptor interactions inferred by experiments. This is most probably caused by the low sequence identity with available structural templates, which limits the accuracy of the protein model and in particular of the side-chains' orientations. Methods which transcend the limited sampling of the conformational space of docking may improve the predictions. As an example corroborating this, we review here multi-scale simulations from our lab and show that, for the three complexes studied so far, they significantly enhance the predictive power of the computational approach. Second, our bioinformatics analysis provides support to previous claims that several residues, including those at positions 1.50, 2.50, and 7.52, are involved in receptor activation.

4.5.2 Results: Receptor Activation Predictions

Although it cannot be excluded completely that they could also be involved in ligand binding, several residues have been previously suggested to play a role in activation for hTAS2Rs. These are residues whose mutation causes changes in receptor's response, from abolishing activation to constitutive activation. Here, we show that our bioinformatics analysis provides further support to some of these

Residue conservation percentage				
Class A hGPCRs	I=31.7%	L=30%	V=16%	M,F,Y,C=1% each
hTAS2Rs	I=76%	L=12%	V=4%	S,Q=4% each
hORs	I=88%	V=9%	L=1%	T,A,S<1% each

Fig. 4.11. Conservation of residues in position 7.52 across human class A GPCRs, hTAS2Rs and hORs

Class A GPCR	Active state	Inactive state	Species
β2-adrenergic receptor	2RH1 (2.40)	3SN6 (3.20)	human
M2 muscarinic receptor	3UON (3.00)	4MQS (3.50)	human
adenosine A2A receptor	3EML (2.60)	5G53 (3.40)	human
rhodopsin	1GZM (2.65)	3PQR (2.85)	bovine
μ-opioid receptor	4DKL (2.80)	5C1M (2.10)	murine

The corresponding PDB codes are listed, together with the crystallographic resolution (between parentheses, in Å).

Fig. 4.12. Active/inactive pairs of mammalian class A GPCR crystal structures used for the graph-based structural analysis.

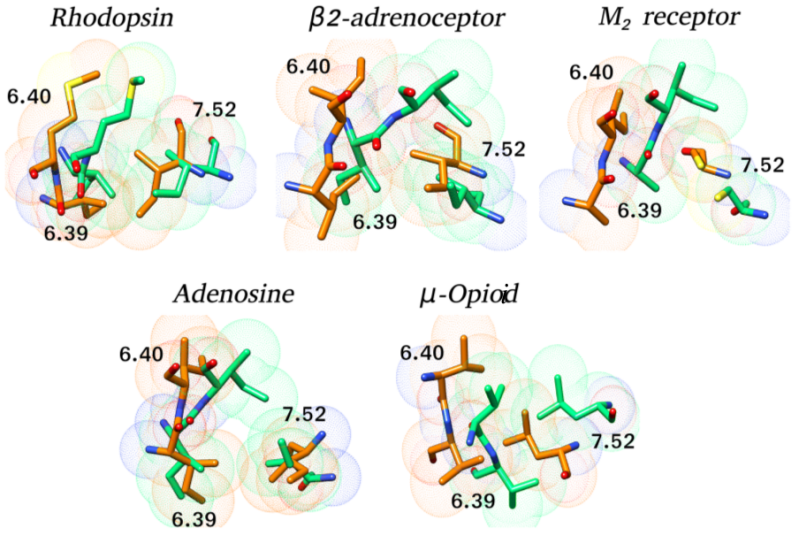

Fig. 4.13. Interactions of the residue in position 7.52 in the mammalian class A GPCR active/inactive structure pairs solved by X-ray crystallography. Inactive structures are shown in green, whereas active structures are in orange.

findings. Namely, we can distinguish three different groups of residues (hereafter indicated using the class A GPCR generic numbering).

The first group, proposed to be involved in hTAS2R1 activation, includes N24 and R55 (positions 1.50 and 2.50). These residues are highly conserved across hTAS2Rs (92 and 96% respectively). Moreover, we notice here that these two positions are also conserved in human class A GPCRs (98 and 87%) and have been shown to play a role for activation across class A hGPCRs, based on mutagenesis data and structural analyses. Therefore, this may further support the claimed role of positions 1.50 and 2.50 for hTAS2Rs activation. Nonetheless, the chemical nature of residue 2.50 changes dramatically, from a positively charged Arg in

hTAS2Rs to a negatively charged Asp in class A GPCRs. Hence, we suggest here distinct activation mechanisms on passing from bitter taste receptors to class A hGPCRs, yet converging at the same positions. Next, we consider position 7.52, which has been suggested to play a role in activation for hTAS2R38. A branched aliphatic residue (V, L or I) is present at this position in 92% hTAS2Rs (Figure 4.11). This position has never been proposed to be involved in an interaction network that changes upon activation in any class A GPCR. Therefore, to blindly investigate if this is the case, we used a pool of structures of human class A GPCRs (see Figure 4.12) for which both active and inactive structures are available and carried out a graph-based structural analysis with the aim of identifying pairs of highly conserved residues that change intramolecular interactions upon activation. This analysis not only confirms, as expected, all of the previously known residues important for class A GPCR activation, including positions 1.50 and 2.50, but also shows that (i) the hydrophobic nature of residue 7.52 is conserved across human class A GPCRs and (ii) this residue does change its interactions upon activation (Figure 4.13). This observation is in agreement with previous experimental data showing that mutations at this position modify the receptor activity in class A GPCRs. Therefore, our analysis not only confirms that position 7.52 is important for activation in hTAS2Rs, but also suggests for the first time, from a structural point of view, that this position is actively involved in a network of residues that changes upon activation in class A GPCRs. The final group of residues proposed to be involved in activation are I27 in hTAS2R1 (position 1.53), as well as S285 and H214 (positions 7.50 and 5.63) and three residues in the intracellular loop ICL3 (Q216, V234, M237) in hTAS2R4. Some of these positions (1.53, 5.63 and 7.50) are highly or fairly well conserved across hTAS2Rs (96, 96, and 68%, respectively). Interestingly, position 7.50 bears either a Ser (68%) or a Pro (28%) in hTAS2Rs, while, in human class A GPCRs, Pro is highly conserved (95%). This position belongs to the conserved TM7 motif NPxxY that is essential for class A GPCRs' activation but there are no experimental data available for this residue. In the case of ICL3 residues, they do not present high conservation values and a role in activation for these positions in human class A GPCRs has not been suggested so far (and does not emerge from our analysis). This is probably due to their intracellular location in a highly variable region and their likely participation in G-protein binding and Gα-subunit selectivity.

4.6 Extra Contributions

4.6.1 Identification of new BMP6 propeptide mutations in patients with iron overload

Section 4.6.1 desribes results published in Piubelli et al. (2017) [167]

4.6.2 Abstract

Hereditary Hemochromatosis (HH) is a genetically heterogeneous disorder caused by mutations in at least five different genes (HFE, HJV, TFR2, SLC40A1, HAMP) involved in the production or activity of the liver hormone hepcidin, a key regulator of systemic iron homeostasis. Nevertheless, patients with an HHlike phenotype that remains completely/partially unexplained despite extensive sequencing of known genes are not infrequently seen at referral centers, suggesting a role of still unknown genetic factors. A compelling candidate is Bone Morphogenetic Protein 6 (BMP6), which acts as a major activator of the BMPSMAD signaling pathway, ultimately leading to the upregulation of hepcidin gene transcription. A recent seminal study by French authors has described three heterozygous missense mutations in BMP6 associated with mild to moderate lateonset iron overload (IO). Using an updated nextgeneration sequencing (NGS)based genetic test in IO patients negative for the classical HFE p.Cys282Tyr mutation, we found three BMP6 heterozygous missense mutations in four patients from three different families. One mutation (p.Leu96Pro) has already been described and proven to be functional. The other two (p.Glu112Gln, p.Arg257His) were novel, and both were located in the propeptide domain known to be crucial for appropriate BMP6 processing and secretion. In silico modeling also showed results consistent with their pathogenetic role. The patients' clinical phenotypes were similar to that of other patients with BMP6related IO recently described. Our results independently add further evidence to the role of BMP6 mutations as likely contributing factors to lateonset moderate IO unrelated to mutations in the established five HH genes.

Results: BMP6 mutations and in silico modeling

Our next generation sequencing analysis detected 3 different heterozygous proBMP6 mutations in the probands. Concurrently, no further potentially pathogenic variant was found in the five canonical HH genes (HFE, HJV, HAMP, TFR2, and SLC40A1), with the exception of the already mentioned p.His63Asp on HFE in the patients.

Of note, one BMP6 mutation, detected in one patient, is located in exon 1 in the propeptide domain and is highly conserved (98%) across different species, and has already been demonstrated to be functional. The other two BMP6 mutations were new. The p.Glu112Gln mutation is also located in exon 1/propeptide domain, adjacent to the pathogenic variant reported by Daher et al. (p.Gln113Glu). On the other hand, the second new mutation is localized to exon 2, corresponding to a different cluster. Arg257 is fully conserved across different species (100%), while Gln112 is only 20% conserved. According to a bioinformatics prediction by

Polyphen and SIFT prediction, p.Arg257His is the most likely pathogenic mutation. Of note, the allele frequency of the two novel BMP6 mutations was quite low in the European nonFinnish population as evaluated in the Exome Aggregation Consortium (ExAC; http://exac.broadinstitute.org), which represents the most updated catalogue of human exonic variants and an invaluable tool for efficient filtering of candidate diseasecausing variants. We also checked for the presence of the three BMP6 variants in 111 Italian Caucasian subjects with normal iron parameters from another cohort at our hospital. We found two carriers of the p.Leu96Pro mutation (corresponding to an allele frequency of 0.009), while neither the p.Arg257His mutation nor the p.Glu112Gln mutation were detected. Our in silico modeling of the BMP6 dimer was built by homology, using the published solved structure of TGFβ1 as a template. The predicted conserved secondary structural elements of the prodomains across members of the TGF ligands31 allowed us to predict the putative localization and functional effects of the mutations detected,

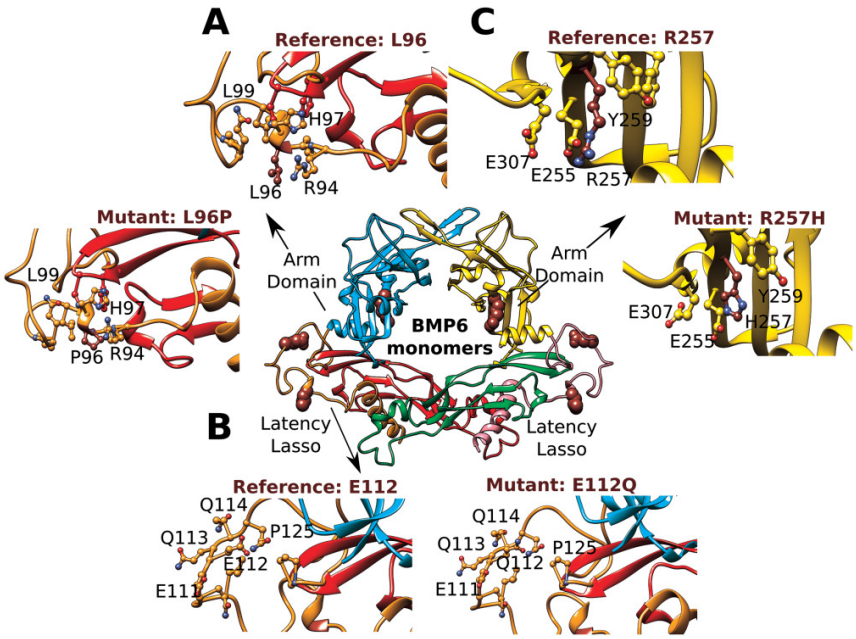

Fig. 4.14. Structural analysis of variants mapped on the 3D model of human proBMP6. Homology modeling was performed on the solved structure of proTGF1. In the central part of the figure, three parts of the two propeptide monomers are identified with different colors: on the left, the prodomain arm and latency lasso are highlighted in cyan and orange, respectively; the corresponding mature BMP6 is shown in green. On the right, the arm and latency lasso are yellow and pink, respectively, and the related mature BMP6 is represented in red. The popup panels magnify the regions with the three identified variants (panel A for p.Leu96Pro, B for p.Glu112Gln, and C for p.Arg257His). In particular, panels A and B illustrate that the two variants are localized to a region in which the mature BMP6 of one monomer is enveloped by the latency lasso of the opposing prodomain. In C, the variant is localized to a solventexposed region of the arm domain.

as illustrated in Figure 4.14. The Leu96 and Glu112 amino acids are projected to be located in the latency lasso domain, which encircles the fingers of the opposing monomer, while Arg257 is located in the arm domain. Leu96 is involved in a putative hydrophobic interaction with Leu99, and its replacement with Pro96 is predicted to cause bumps with nearby residues, creating local perturbations at the backbone level (Figure 4.14A). Glu112 is a negatively charged residue located near the dimerization interface, and its substitution with a neutral Gln112 residue may change the physicochemical (electrostatic) properties of this region, disrupting these interactions (Figure 4.14B). Our in silico model is consistent with the cytoplasmic aggregation of BMP6 mutants observed by Daher et al. through immunolocalization in transfected cells. Indeed, both p.Leu96Pro and p.Glu112Gln are located in the same amino acid cluster, in a key region clearly involved in the interaction between the two monomers (Figure 4.14) and particularly conserved across the TGF*beta* superfamily.31 The local perturbations created by these mutations likely affect this interaction, possibly leading to the formation of unfolded dimers that are subsequently degraded. ccording to our model, Arg257 may belong to a network of ionic interactions, along with two glutamic acid residues (Glu307 and Glu255). The substitution with a histidine residue could change the pH dependency of the region, altering the local stability. This region of the arm domain, enriched in charged amino acids, appears to be a common feature of TGFβ ligands and is reported to be involved in the interaction with extracellular matrix proteins. Thus, its alteration could theoretically affect either assembly/secretion or final localization of BMP6. However, further experimental data would be needed to verify this hypothesis.

5

Conclusions

We have shown in this thesis examples of how computational tools can be used on existing structural and experimental information, and provide support when it is not directly available. The studies presented here are united by scope, methodologies and generality. They complement and build experimental information by adding a dynamical dimension (time). In this sense, we generate a virtuous cycle: experiments being the basis of calculations, these leading to models, which in turn inspire further experiments.

One of the most representative work of this thesis is the study of the GPR3 receptor, in which we predicted the binding mode of the only known agonist (DPI) of this receptor, without any experimental information. This highly challenging task would have been hopeless by simple homology modeling and docking, due to the absolute lack of a close (>30% identity) template. Using the molecular dynamic technique and complementing modeling algorithms with mutagenesis information, we have produced one of the most complete structural descriptions of the receptor to date. We have also managed to produce two potentially general insights on the cascade that this receptor activates inside the cell. First we have shown that mutants on the canonical orthosteric binding cavity reduce the interaction of the βarr2 with GPR3, which was demonstrated to be heavily correlated with the formation of the plaques in the Alzheimer's Disease. This suggest that the mutants proposed by us can strongly influence the transmission of the signal inside the cell, paving the way to new drug design experiments. Second, we found a new possible allosteric binding site, which mutagenesis experiments demonstrated to kill completely the signal of the GPR3 receptor. This unexpected discovery reveals a new way for prevention and new therapeutic interventions in Alzheimers disease. Both of these biological insights will require further testing and validation, both experimental and computational: yet they would have been very difficult to extract without the aid of computational tools. The repeated success of the technique that has been developed in out laboratory, the hybrid molecular mechanics/coarse grained (MM/CG), opens the way to the use of molecular simulation as a high-throughput tool for ligand binding prediction and characterization. The approach we used can be applied naturally to any GPCR, provided enough experimental information. The powerful combination of accuracy and relatively low computational cost of the MM/CG approach suggests that in the future it could be used

for virtual screening purposes, at least as a last step. By this technique GPCRs with not a solved structure and not a high sequence identity can be finally targeted pharmacologically.

The second most relevant study instead had a general scope, giving a quantitative foundation to several previous hypotheses about ligand-receptor interactions in GPCRs. We performed the first global analysis on the Class A GPCRs binding cavities, based on all the unique solved structures of these receptors present in the PDB database. In total, we analyzed 85 complexes and from our structural-based GPCRs analysis, we found previously unexpected properties of the binding cavities. First, we found that ten positions of the GPCRs binding cavities, namely 3.32, 3.33, 3.36, 6.48, 6.51, 6.52, 6.55, 7.35, 7.39, 7.43, are shared between all the rho-GPCRs solved structures. They are located in three helices, i.e. TM3, TM6, and TM7. This leads us to believe that our findings could be strictly connected with the activation of these receptors. Moreover using similar physicochemical properties of residues in these ten positions as features, we then were able to cluster together receptors that are very distant between each other at a sequence level, but very close in ligand recognition and binding cavities similarities. To the best of our knowledge, this is the first time that a similar analysis is performed on all the Class A GPCRs solved structures. We strongly believe that our results will help in the deorphanization of most of GPCRs, as well as in suggesting novel specific drug targets. To conclude, we believe the studies presented here are convincing examples of how the convergence between experimentalists and theoreticians is becoming necessary to break new ground in molecular biology. The future of GPCR biology will depend in part on the coming of age of the already ongoing collaboration between these two approaches.

6

Acknowledgments

First I would like to thank my supervisor Prof. Alejandro Giorgetti. He always guided me in the right path of research and I'm very grateful for that and for what I am able to do today. During all these years he was my professor but also a good friend. With his help I was able to polish my skills and grow up a lot both as a student and at a personal level.

Second I want to thank all my lab, all the people I met during these years that became very good friends at the end of this PhD journey, Mirko, Mangesh, Rui, Katia, Alessandro, Ilaria and many others. They are the best and all I wanted and needed in the most difficult moments. We shared lots of beautiful moments and I will miss them a lot.

Third I want to thank my family, my mother Marie and my father Sokrat. They still don't understand what I did during my PhD but they always were ready to celebrate my best moments and share my difficulties.

Then I want to thank my husband Mauro which is the central figure of my live and of all these years as a PhD student. The words are wasted to describe all his importance and the help that he gave to me every day thus I just want to say he is simply the best of the best of all I can ever imagine and desire.

Last but not least I would like to thank all the University of Verona and all the people that helped me during these years. This experience was wonderful also thanks to all of you.

References

[1] R. Fredriksson et al. "The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints". In: *Molecular pharmacology* 63.6 (2003), pp. 1256–1272.

[2] M. C. Lagerström and H. B. Schiöth. "Structural diversity of G protein-coupled receptors and significance for drug discovery". In: *Nature reviews Drug discovery* 7.4 (2008), p. 339.

[3] I. G. Tikhonova and D. Fourmy. "The family of G protein-coupled receptors: an example of membrane proteins". In: *Membrane Protein Structure Determination*. Springer, 2010, pp. 441–454.

[4] H. Shichi et al. "Biochemistry of visual pigments I. Purification and properties of bovine rhodopsin". In: *Journal of Biological Chemistry* 244.3 (1969), pp. 529–536.

[5] J. B. Findlay, M. Brett, and D. J. Pappin. "Primary structure of C-terminal functional sites in ovine rhodopsin". In: *Nature* 293.5830 (1981), p. 314.

[6] Y. A. Ovchinnikov. "Rhodopsin and bacteriorhodopsin: structurefunction relationships". In: *FEBS letters* 148.2 (1982), pp. 179–191.

[7] P. A. Hargrave. "Rhodopsin chemistry, structure and topography". In: *Progress in Retinal Research* 1 (1982), pp. 1–51.

[8] R. J. Lefkowitz. "A brief history of G-protein coupled receptors (Nobel Lecture)". In: *Angewandte Chemie International Edition* 52.25 (2013), pp. 6366–6378.

[9] G. Pándy-Szekeres et al. "GPCRdb in 2018: adding GPCR structure models and ligands". In: *Nucleic acids research* 46.D1 (2017), pp. D440–D446.

[10] S. Schlyer and R. Horuk. "I want a new drug: G-protein-coupled receptors in drug development". In: *Drug discovery today* 11.11-12 (2006), pp. 481–493.

[11] K. Lundstrom. "An overview on GPCRs and drug discovery: structure-based drug design and structural biology on GPCRs". In: *G Protein-Coupled Receptors in Drug Discovery*. Springer, 2009, pp. 51–66.

[12] H. B. Schiöth and R. Fredriksson. "The GRAFS classification system of G-protein coupled receptors in comparative perspective". In: *General and comparative endocrinology* 142.1-2 (2005), pp. 94–101.

[13] B. K. Kobilka. “G protein coupled receptor structure and activation”. In: *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1768.4 (2007), pp. 794–807.

[14] A. Venkatakrishnan et al. “Molecular signatures of G-protein-coupled receptors”. In: *Nature* 494.7436 (2013), p. 185.

[15] M. C. Lagerström and H. B. Schiöth. “Structural diversity of G protein-coupled receptors and significance for drug discovery”. In: *Nature reviews Drug discovery* 7.4 (2008), p. 339.

[16] H. M. Stoveken et al. “Adhesion G protein-coupled receptors are activated by exposure of a cryptic tethered agonist”. In: *Proceedings of the National Academy of Sciences* (2015), p. 201421785.

[17] C. B. Felder et al. “The Venus flytrap of periplasmic binding proteins: an ancient protein module present in multiple drug receptors”. In: *AAps pharmsci* 1.2 (1999), pp. 7–26.

[18] E. Doumazane et al. “Illuminating the activation mechanisms and allosteric properties of metabotropic glutamate receptors”. In: *Proceedings of the National Academy of Sciences* (2013), p. 201215615.

[19] N. Kunishima et al. “Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor”. In: *Nature* 407.6807 (2000), p. 971.

[20] M. Behrens and W. Meyerhof. “Gustatory and extragustatory functions of mammalian taste receptors”. In: *Physiology & Behavior* 105.1 (2011), pp. 4–13.

[21] D. C. Slusarski, V. G. Corces, and R. T. Moon. “Interaction of Wnt and a Frizzled homologue triggers G-protein-linked phosphatidylinositol signalling”. In: *Nature* 390.6658 (1997), p. 410.

[22] M. Murone, A. Rosenthal, and F. J. de Sauvage. “Sonic hedgehog signaling by the patched–smoothened receptor complex”. In: *Current Biology* 9.2 (1999), pp. 76–84.

[23] R. Fredriksson et al. “The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints”. In: *Molecular pharmacology* 63.6 (2003), pp. 1256–1272.

[24] A. Krishnan et al. “The origin of GPCRs: identification of mammalian like Rhodopsin, Adhesion, Glutamate and Frizzled GPCRs in fungi”. In: *PloS one* 7.1 (2012), e29817.

[25] A. Marchiori et al. “Coarse-grained/molecular mechanics of the TAS2R38 bitter taste receptor: experimentally-validated detailed structural prediction of agonist binding”. In: *PloS one* 8.5 (2013), e64675.

[26] A. Brockhoff et al. “Structural requirements of bitter taste receptor activation”. In: *Proceedings of the National Academy of Sciences* 107.24 (2010), pp. 11110–11115.

[27] M. Sandal et al. “Evidence for a transient additional ligand binding site in the TAS2R46 bitter taste receptor”. In: *Journal of chemical theory and computation* 11.9 (2015), pp. 4439–4449.

[28] A. Brockhoff et al. “Broad tuning of the human bitter taste receptor hTAS2R46 to various sesquiterpene lactones, clerodane and labdane diterpenoids, strychnine, and denatonium”. In: *Journal of agricultural and food chemistry* 55.15 (2007), pp. 6236–6243.

[29] S. Born et al. “The human bitter taste receptor TAS2R10 is tailored to accommodate numerous diverse ligands”. In: *Journal of Neuroscience* 33.1 (2013), pp. 201–213.

[30] V. Katritch, V. Cherezov, and R. C. Stevens. “Diversity and modularity of G protein-coupled receptor structures”. In: *Trends in pharmacological sciences* 33.1 (2012), pp. 17–27.

[31] A. M. Spiegel et al. “The G protein connection: molecular basis of membrane association”. In: *Trends in biochemical sciences* 16 (1991), pp. 338–341.

[32] E. Neumann, K. Khawaja, and U. Müller-Ladner. “G protein-coupled receptors in rheumatology”. In: *Nature Reviews Rheumatology* 10.7 (2014), p. 429.

[33] N. Wettschureck and S. Offermanns. “Mammalian G proteins and their cell type specific functions”. In: *Physiological reviews* 85.4 (2005), pp. 1159–1204.

[34] N. Tuteja. “Signaling through G protein coupled receptors”. In: *Plant signaling & behavior* 4.10 (2009), pp. 942–947.

[35] S. G. Rasmussen et al. “Crystal structure of the β 2 adrenergic receptor–Gs protein complex”. In: *Nature* 477.7366 (2011), p. 549.

[36] L. M. Luttrell and D. Gesty-Palmer. “Beyond desensitization: physiological relevance of arrestin-dependent signaling”. In: *Pharmacological reviews* (2010), pr–109.

[37] D. Hilger, M. Masureel, and B. K. Kobilka. "Structure and dynamics of GPCR signaling complexes". In: *Nature structural & molecular biology* 25.1 (2018), p. 4.

[38] R. Al-Hasani and M. R. Bruchas. "Molecular mechanisms of opioid receptor-dependent signaling and behavior". In: *The Journal of the American Society of Anesthesiologists* 115.6 (2011), pp. 1363–1381.

[39] D. H. Rominger et al. "Biased ligands: pathway validation for novel GPCR therapeutics". In: *Current opinion in pharmacology* 16 (2014), pp. 108–115.

[40] R. Strotmann et al. "Evolution of GPCR: change and continuity". In: *Molecular and cellular endocrinology* 331.2 (2011), pp. 170–178.

[41] D. G. Isom and H. G. Dohlman. "Buried ionizable networks are an ancient hallmark of G protein-coupled receptor activation". In: *Proceedings of the National Academy of Sciences* (2015), p. 201417888.

[42] L. Pardo et al. "On the use of the transmembrane domain of bacteriorhodopsin as a template for modeling the three-dimensional structure of guanine nucleotide-binding regulatory protein-coupled receptors." In: *Proceedings of the National Academy of Sciences* 89.9 (1992), pp. 4009–4012.

[43] K. J. Nordström et al. "Independent HHsearch, Needleman–Wunsch-based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families". In: *Molecular biology and evolution* 28.9 (2011), pp. 2471–2480.

[44] T. K. Bjarnadóttir et al. "Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse". In: *Genomics* 88.3 (2006), pp. 263–273.

[45] D. Zhang et al. "Two disparate ligand-binding sites in the human P2Y 1 receptor". In: *Nature* 520.7547 (2015), p. 317.

[46] D. Bartuzi, A. A. Kaczor, and D. Matosiuk. "Interplay between two allosteric sites and their influence on agonist binding in human μ opioid receptor". In: *Journal of chemical information and modeling* 56.3 (2016), pp. 563–570.

[47] A. C. Kruse et al. "Activation and allosteric modulation of a muscarinic acetylcholine receptor". In: *Nature* 504.7478 (2013), p. 101.

[48] V. Katritch et al. "Allosteric sodium in class A GPCR signaling". In: *Trends in biochemical sciences* 39.5 (2014), pp. 233–244.

[49] P. R. Gentry, P. M. Sexton, and A. Christopoulos. "Novel allosteric modulators of G protein-coupled receptors". In: *Journal of Biological Chemistry* (2015), jbc–R115.

[50] K. Palczewski and T. Orban. "From atomic structures to neuronal functions of G protein–coupled receptors". In: *Annual review of neuroscience* 36 (2013), pp. 139–164.

[51] M. Wheatley et al. "Lifting the lid on GPCRs: the role of extracellular loops". In: *British journal of pharmacology* 165.6 (2012), pp. 1688–1703.

[52] K. A. Jacobson and S. Costanzi. "New insights for drug design from the X-ray crystallographic structures of G-protein-coupled receptors". In: *Molecular pharmacology* 82.3 (2012), pp. 361–371.

[53] A. Gonzalez et al. "Impact of Helix Irregularities on Sequence Alignment and Homology Modeling of G Protein-Coupled Receptors". In: *ChemBioChem* 13.10 (2012), pp. 1393–1399.

[54] R. van der Kant and G. Vriend. "Alpha-bulges in G protein-coupled receptors". In: *International journal of molecular sciences* 15.5 (2014), pp. 7841–7864.

[55] J. A. Ballesteros and H. Weinstein. "Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors". In: *Methods in neurosciences*. Vol. 25. Elsevier, 1995, pp. 366–428.

[56] F. Horn et al. "GPCRDB: an information system for G protein-coupled receptors". In: *Nucleic Acids Research* 26.1 (1998), pp. 275–279.

[57] G. E. Rovati, V. Capra, and R. R. Neubig. "The highly conserved DRY motif of class AG protein-coupled receptors: beyond the ground state". In: *Molecular pharmacology* 71.4 (2007), pp. 959–964.

[58] H. Nomiyama and O. Yoshie. "Functional roles of evolutionary conserved motifs and residues in vertebrate chemokine receptors". In: *Journal of leukocyte biology* 97.1 (2015), pp. 39–47.

[59] D. Wootten, A. Christopoulos, and P. M. Sexton. "Emerging paradigms in GPCR allostery: implications for drug discovery". In: *Nature reviews Drug discovery* 12.8 (2013), p. 630.

[60] A. Srivastava et al. "High-resolution structure of the human GPR40 receptor bound to allosteric agonist TAK-875". In: *Nature* 513.7516 (2014), p. 124.

[61] M. Michino et al. “A single glycine in extracellular loop 1 is the critical determinant for pharmacological specificity of dopamine D2 and D3 receptors”. In: *Molecular pharmacology* (2013), mol–113.

[62] B. F. Seibt et al. “The second extracellular loop of GPCRs determines subtype-selectivity and controls efficacy as evidenced by loop exchange study at A2 adenosine receptors”. In: *Biochemical pharmacology* 85.9 (2013), pp. 1317–1329.

[63] D. Wifling et al. “The extracellular loop 2 (ECL2) of the human histamine H4 receptor substantially contributes to ligand binding and constitutive activity”. In: *PLoS One* 10.1 (2015), e0117185.

[64] A. C. Kruse et al. “Structure and dynamics of the M3 muscarinic acetylcholine receptor”. In: *Nature* 482.7386 (2012), p. 552.

[65] R. O. Dror et al. “Pathway and mechanism of drug binding to G-protein-coupled receptors”. In: *Proceedings of the National Academy of Sciences* 108.32 (2011), pp. 13118–13123.

[66] R. O. Dror et al. “Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs”. In: *Nature* 503.7475 (2013), p. 295.

[67] D. Sabbadin and S. Moro. “Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR–ligand recognition pathway in a nanosecond time scale”. In: *Journal of chemical information and modeling* 54.2 (2014), pp. 372–376.

[68] W. Liu et al. “Structural basis for allosteric regulation of GPCRs by sodium ions”. In: *Science* 337.6091 (2012), pp. 232–236.

[69] G. F. Schertler, C. Villa, and R. Henderson. “Projection structure of rhodopsin”. In: *Nature* 362.6422 (1993), p. 770.

[70] K. Palczewski et al. “Crystal structure of rhodopsin: AG protein-coupled receptor”. In: *science* 289.5480 (2000), pp. 739–745.

[71] V. Cherezov et al. “High-resolution crystal structure of an engineered human β2-adrenergic G protein–coupled receptor”. In: *science* 318.5854 (2007), pp. 1258–1265.

[72] E. Ghosh et al. “Methodological advances: the unsung heroes of the GPCR structural revolution”. In: *Nature Reviews Molecular Cell Biology* 16.2 (2015), p. 69.

[73] C. L. Piscitelli et al. “A molecular pharmacologist’s guide to GPCR crystallography”. In: *Molecular pharmacology* (2015), mol–115.

[74] A. M. Ring et al. “Adrenaline-activated structure of β 2-adrenoceptor stabilized by an engineered nanobody”. In: *Nature* 502.7472 (2013), p. 575.

[75] M. Michino et al. “Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008”. In: *Nature Reviews Drug Discovery* 8.6 (2009), p. 455.

[76] I. Kufareva et al. “Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment”. In: *Structure* 19.8 (2011), pp. 1108–1126.

[77] I. Kufareva et al. “Advances in GPCR modeling evaluated by the GPCR Dock 2013 assessment: meeting new challenges”. In: *Structure* 22.8 (2014), pp. 1120–1139.

[78] C. L. Worth, G. Kleinau, and G. Krause. “Comparative sequence and structural analyses of G-protein-coupled receptor crystal structures and implications for molecular models”. In: *PloS one* 4.9 (2009), e7011.

[79] K. Rataj et al. “Impact of template choice on homology model efficiency in virtual screening”. In: *Journal of chemical information and modeling* 54.6 (2014), pp. 1661–1668.

[80] C. N. Cavasotto and D. Palomba. “Expanding the horizons of G protein-coupled receptor structure-based ligand discovery and optimization using homology models”. In: *Chemical Communications* 51.71 (2015), pp. 13576–13594.

[81] M. Leguèbe et al. “Hybrid molecular mechanics/coarse-grained simulations for structural prediction of G-protein coupled receptor/ligand complexes”. In: *PloS one* 7.10 (2012), e47332.

[82] A. Levit et al. “Homology model-assisted elucidation of binding sites in GPCRs”. In: *Membrane Protein Structure and Dynamics*. Springer, 2012, pp. 179–205.

[83] X. Deupi and B. K. Kobilka. “Energy landscapes as a tool to integrate GPCR structure, dynamics, and function”. In: *Physiology* 25.5 (2010), pp. 293–303.

[84] R. Seifert and K. Wenzel-Seifert. “Constitutive activity of G-protein-coupled receptors: cause of disease and common property of wild-type receptors”. In: *Naunyn-Schmiedeberg’s archives of pharmacology* 366.5 (2002), pp. 381–416.

[85] B Trzaskowski et al. “Action of molecular switches in GPCRs-theoretical and experimental studies”. In: *Current medicinal chemistry* 19.8 (2012), pp. 1090–1109.

[86] K. P. Hofmann et al. “AG protein-coupled receptor at work: the rhodopsin model”. In: *Trends in biochemical sciences* 34.11 (2009), pp. 540–552.

[87] Y. Miao et al. “Activation and dynamic network of the M2 muscarinic receptor”. In: *Proceedings of the National Academy of Sciences* 110.27 (2013), pp. 10982–10987.

[88] K. J. Kohlhoff et al. “Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways”. In: *Nature chemistry* 6.1 (2014), p. 15.

[89] A. K. Shukla et al. “Distinct conformational changes in β-arrestin report biased agonism at seven-transmembrane receptors”. In: *Proceedings of the National Academy of Sciences* 105.29 (2008), pp. 9988–9993.

[90] L. Valentin-Hansen et al. “Biased Gs versus Gq and β-arrestin signaling in the NK1 receptor determined by interactions in the water hydrogen-bond network”. In: *Journal of Biological Chemistry* (2015), jbc–M115.

[91] P. Scheerer et al. “Crystal structure of opsin in its G-protein-interacting conformation”. In: *Nature* 455.7212 (2008), p. 497.

[92] G. Lebon et al. “Agonist-bound adenosine A 2A receptor structures reveal common features of GPCR activation”. In: *Nature* 474.7352 (2011), p. 521.

[93] B. G. Tehan et al. “Unifying family A GPCR theories of activation”. In: *Pharmacology & therapeutics* 143.1 (2014), pp. 51–60.

[94] B. Holst et al. “A conserved aromatic lock for the tryptophan rotameric switch in TM-VI of seven-transmembrane receptors”. In: *Journal of Biological Chemistry* 285.6 (2010), pp. 3973–3985.

[95] N. R. Latorraca, A. Venkatakrishnan, and R. O. Dror. “GPCR dynamics: structures in motion”. In: *Chemical reviews* 117.1 (2016), pp. 139–155.

[96] T. W. Schwartz et al. “Molecular mechanism of 7TM receptor activationa global toggle switch model”. In: *Annu. Rev. Pharmacol. Toxicol.* 46 (2006), pp. 481–519.

[97] K. Sansuk et al. “A structural insight into the reorientation of transmembrane domains 3 and 5 during family AG protein-coupled receptor activation”. In: *Molecular pharmacology* 79.2 (2011), pp. 262–269.

[98] B. E. Krumm et al. “Structural prerequisites for G-protein activation by the neurotensin receptor”. In: *Nature communications* 6 (2015), p. 7895.

[99] X. Deupi and J. Standfuss. “Structural insights into agonist-induced activation of G-protein-coupled receptors”. In: *Current opinion in structural biology* 21.4 (2011), pp. 541–551.

[100] E. C. Hulme. “GPCR activation: a mutagenic spotlight on crystal structures”. In: *Trends in pharmacological sciences* 34.1 (2013), pp. 67–84.

[101] S. G. Rasmussen et al. “Structure of a nanobody-stabilized active state of the β 2 adrenoceptor”. In: *Nature* 469.7329 (2011), p. 175.

[102] D. M. Sena Jr et al. “Structural heterogeneity of the μ-opioid receptors conformational ensemble in the apo state”. In: *Scientific Reports* 7 (2017), p. 45761.

[103] T. E. Angel, M. R. Chance, and K. Palczewski. “Conserved waters mediate structural and functional activation of family A (rhodopsin-like) G protein-coupled receptors”. In: *Proceedings of the National Academy of Sciences* 106.21 (2009), pp. 8555–8560.

[104] S. Yuan et al. “Activation of G-protein-coupled receptors correlates with the formation of a continuous internal water pathway”. In: *Nature communications* 5 (2014), p. 4733.

[105] H. Wu et al. “Structure of the human κ-opioid receptor in complex with JDTic”. In: *Nature* 485.7398 (2012), p. 327.

[106] W. Huang et al. “Structural insights into μ-opioid receptor activation”. In: *Nature* 524.7565 (2015), p. 315.

[107] T. B. Reddy et al. “The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta) genome project classification”. In: *Nucleic acids research* 43.D1 (2014), pp. D1099–D1106.

[108] A. Escobar-Zepeda, A. Vera-Ponce de León, and A. Sanchez-Flores. “The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics”. In: *Frontiers in genetics* 6 (2015), p. 348.

[109] H. M. Berman et al. “The protein data bank”. In: *Nucleic acids research* 28.1 (2000), pp. 235–242.

[110] C. H. Wu et al. “The Universal Protein Resource (UniProt): an expanding universe of protein information”. In: *Nucleic acids research* 34.suppl_1 (2006), pp. D187–D191.

[111] D. N. Woolfson et al. “De novo protein design: how do we expand into the universe of possible protein structures?” In: *Current opinion in structural biology* 33 (2015), pp. 16–26.

[112] S. Y. Chung and S Subbiah. “A structural explanation for the twilight zone of protein sequence homology”. In: *Structure* 4.10 (1996), pp. 1123–1127.

[113] C. Chothia and A. M. Lesk. “The relation between the divergence of sequence and structure in proteins.” In: *The EMBO journal* 5.4 (1986), pp. 823–826.

[114] C. A. Orengo et al. “CATH–a hierarchic classification of protein domain structures”. In: *Structure* 5.8 (1997), pp. 1093–1109.

[115] A. Magner, W. Szpankowski, and D. Kihara. “On the origin of protein superfamilies and superfolds”. In: *Scientific reports* 5 (2015), p. 8166.

[116] W. J. Browne et al. “A possible three-dimensional structure of bovine α-lactalbumin based on that of hen’s egg-white lysozyme”. In: *Journal of molecular biology* 42.1 (1969), pp. 65–86.

[117] T. Blundell et al. “Knowledge-based prediction of protein structures and the design of novel molecules”. In: *Nature* 326.6111 (1987), p. 347.

[118] A. Šali and T. L. Blundell. “Comparative protein modelling by satisfaction of spatial restraints”. In: *Journal of molecular biology* 234.3 (1993), pp. 779–815.

[119] A. Meier and J. Söding. “Probabilistic multi-template protein homology modeling”. In: *PLoS Comput Biol* (2015).

[120] M. A. Martí-Renom et al. “Comparative protein structure modeling of genes and genomes”. In: *Annual review of biophysics and biomolecular structure* 29.1 (2000), pp. 291–325.

[121] S. F. Altschul et al. “Basic local alignment search tool”. In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[122] S. F. Altschul et al. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. In: *Nucleic acids research* 25.17 (1997), pp. 3389–3402.

[123] A. Hildebrand et al. “Fast and accurate automatic structure prediction with HHpred”. In: *Proteins: Structure, Function, and Bioinformatics* 77.S9 (2009), pp. 128–132.

[124] M. Remmert et al. “HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment”. In: *Nature methods* 9.2 (2012), p. 173.

[125] A. Fiser and A. Šali. “Modeller: generation and refinement of homology-based protein structure models”. In: *Methods in enzymology*. Vol. 374. Elsevier, 2003, pp. 461–491.

[126] D. J. Lipman, S. F. Altschul, and J. D. Kececioglu. “A tool for multiple sequence alignment”. In: *Proceedings of the National Academy of Sciences* 86.12 (1989), pp. 4412–4415.

[127] J. Pei and N. V. Grishin. “PROMALS: towards accurate multiple sequence alignments of distantly related proteins”. In: *Bioinformatics* 23.7 (2007), pp. 802–808.

[128] A. D. MacKerell Jr et al. “CHARMM: the energy function and its parameterization”. In: *Encyclopedia of computational chemistry* 1 (2002).

[129] W Braun and N Gō. “Calculation of protein conformations by proton-proton distance constraints: A new efficient algorithm”. In: *Journal of molecular biology* 186.3 (1985), pp. 611–626.

[130] A. Fiser, R. K. G. Do, et al. “Modeling of loops in protein structures”. In: *Protein science* 9.9 (2000), pp. 1753–1773.

[131] C. A. Rohl et al. “Modeling structurally variable regions in homologous proteins with rosetta”. In: *Proteins: Structure, Function, and Bioinformatics* 55.3 (2004), pp. 656–677.

[132] M. Jamroz and A. Kolinski. “Modeling of loops in proteins: a multi-method approach”. In: *BMC structural biology* 10.1 (2010), p. 5.

[133] R. A. Laskowski et al. “PROCHECK: a program to check the stereochemical quality of protein structures”. In: *Journal of applied crystallography* 26.2 (1993), pp. 283–291.

[134] L. Willard et al. “VADAR: a web server for quantitative evaluation of protein structure quality”. In: *Nucleic acids research* 31.13 (2003), pp. 3316–3319.

[135] A. Ray, E. Lindahl, and B. Wallner. “Model quality assessment for membrane proteins”. In: *Bioinformatics* 26.24 (2010), pp. 3067–3074.

[136] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. “Development and testing of the OPLS all-atom force field on conformational energetics and

properties of organic liquids". In: *Journal of the American Chemical Society* 118.45 (1996), pp. 11225–11236.

[137] S. J. Weiner et al. "A new force field for molecular mechanical simulation of nucleic acids and proteins". In: *Journal of the American Chemical Society* 106.3 (1984), pp. 765–784.

[138] M. J. Sippl. "Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins". In: *Journal of molecular biology* 213.4 (1990), pp. 859–883.

[139] R. Lüthy, J. U. Bowie, and D. Eisenberg. "Assessment of protein models with three-dimensional profiles". In: *Nature* 356.6364 (1992), p. 83.

[140] F. Melo and A. Sali. "Fold assessment for comparative protein structure modeling". In: *Protein Science* 16.11 (2007), pp. 2412–2426.

[141] S.-Y. Huang. "Exploring the potential of global protein–protein docking: an overview and critical assessment of current programs for automatic ab initio docking". In: *Drug Discovery Today* 20.8 (2015), pp. 969–977.

[142] G. R. Smith and M. J. Sternberg. "Prediction of protein–protein interactions by docking methods". In: *Current opinion in structural biology* 12.1 (2002), pp. 28–35.

[143] E. Katchalski-Katzir et al. "Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques". In: *Proceedings of the National Academy of Sciences* 89.6 (1992), pp. 2195–2199.

[144] D. Fischer et al. "A geometry-based suite of moleculardocking processes". In: *Journal of Molecular Biology* 248.2 (1995), pp. 459–477.

[145] O. Trott and A. J. Olson. "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading". In: *Journal of computational chemistry* 31.2 (2010), pp. 455–461.

[146] J. Rodrigues et al. "Defining the limits of homology modeling in information-driven protein docking". In: *Proteins: Structure, Function, and Bioinformatics* 81.12 (2013), pp. 2119–2128.

[147] C. Dominguez, R. Boelens, and A. M. Bonvin. "HADDOCK: a protein-protein docking approach based on biochemical or biophysical information". In: *Journal of the American Chemical Society* 125.7 (2003), pp. 1731–1737.

[148] B. J. Alder and T. E. Wainwright. "Studies in molecular dynamics. I. General method". In: *The Journal of Chemical Physics* 31.2 (1959), pp. 459–466.

[149] S. Pronk et al. "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit". In: *Bioinformatics* 29.7 (2013), pp. 845–854.

[150] J. C. Phillips et al. "Scalable molecular dynamics with NAMD". In: *Journal of computational chemistry* 26.16 (2005), pp. 1781–1802.

[151] R. Salomon-Ferrer, D. A. Case, and R. C. Walker. "An overview of the Amber biomolecular simulation package". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 3.2 (2013), pp. 198–210.

[152] M Mosharafi, S. Mahbaz, M. Dusseault, et al. "Molecular Dynamic Model Applications in Reservoir Geomechanics and Fracture Propagation in Pure Calcium Carbonate". In: *51st US Rock Mechanics/Geomechanics Symposium*. American Rock Mechanics Association. 2017.

[153] A. R. Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[154] C. Oostenbrink et al. "A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6". In: *Journal of computational chemistry* 25.13 (2004), pp. 1656–1676.

[155] W. D. Cornell et al. "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules". In: *Journal of the American Chemical Society* 117.19 (1995), pp. 5179–5197.

[156] W. L. Jorgensen et al. "Comparison of simple potential functions for simulating liquid water". In: *The Journal of chemical physics* 79.2 (1983), pp. 926–935.

[157] C. D. Berweger, W. F. van Gunsteren, and F. Müller-Plathe. "Force field parametrization by weak coupling. Re-engineering SPC water". In: *Chemical physics letters* 232.5-6 (1995), pp. 429–436.

[158] H. Berendsen, J. Grigera, and T. Straatsma. "The missing term in effective pair potentials". In: *Journal of Physical Chemistry* 91.24 (1987), pp. 6269–6271.

[159] J. Chen, C. L. Brooks III, and J. Khandogin. "Recent advances in implicit solvent-based methods for biomolecular simulations". In: *Current opinion in structural biology* 18.2 (2008), pp. 140–148.

[160] D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*. Vol. 1. Elsevier, 2001.

[161] M. Neri et al. "Coarse-grained model of proteins incorporating atomistic detail of the active site". In: *Physical review letters* 95.21 (2005), p. 218102.

[162] E. Suku and A. Giorgetti. "Common evolutionary binding mode of rhodopsin-like GPCRS: Insights from structural bioinformatics". In: *AIMS Press, AIMS. Biophysics* 4 (2017), pp. 543–556.

[163] B. M. Radu et al. "All muscarinic acetylcholine receptors (M 1-M 5) are expressed in murine brain microvascular endothelium". In: *Scientific reports* 7.1 (2017), p. 5083.

[164] A. S. Mechaly et al. "Evidences of alternative splicing as a regulatory mechanism for Kissr2 in pejerrey fish." In: *Frontiers in endocrinology* 9 (2018), p. 604.

[165] S. Capaldi et al. "Allosteric sodium binding cavity in GPR3: a novel player in modulation of Aβ production". In: *Scientific reports* 8.1 (2018), p. 11102.

[166] F. Fierro et al. "Agonist binding to chemosensory receptors: a systematic bioinformatics analysis". In: *Frontiers in molecular biosciences* 4 (2017), p. 63.

[167] C. Piubelli et al. "Identification of new BMP6 pro-peptide mutations in patients with iron overload". In: *American journal of hematology* 92.6 (2017), pp. 562–568.