Homology Modeling of Class A G-Protein-Coupled Receptors in the Age of the Structure Boom in resolved structures

Asma Tiss, Rym Ben Boubaker, Daniel Henrion, Hajer Guissouma, Marie Chabbert

Homology modeling of class A G-protein-coupled receptors in the age of the boom in resolved structures

Asma Tiss[1,2#], Rym Ben Boubaker[1#], Daniel Henrion[1], Hajer Guissouma[2] and Marie Chabbert[1]*

1 UMR CNRS 6015 – INSERM 1083, Laboratoire MITOVASC, Université d'Angers, Angers, France

2 Laboratoire de Génétique, Immunologie et Pathologies Humaines, Département de Biologie, Faculté des Sciences de Tunis, Université de Tunis El Manar, Tunisie

* To whom correspondence should be addressed

# Equal contribution to the work

Running head (60 characters): Homology modeling of GPCRs

Key Words: Homology modeling, MODELLER, GPCR, membrane receptor, loop modeling, computational biology

Summary

With 700 members, G protein-coupled receptors of the rhodopsin family (class A) form the largest membrane receptor family in humans and are the target of about 30% of presently available pharmaceutical drugs. The recent boom in resolved structures of GPCRs led to the structural resolution of 57 unique receptors in different states (39 receptors in inactive state only, 2 receptors in active state only and 16 receptors in different activation states). In spite of these tremendous advances, most computational studies on GPCRs, including molecular dynamics simulations, virtual screening and drug design, rely on GPCR models obtained by homology modeling. In this article, we detail the different steps of homology modeling with the MODELLER software, from template selection to model evaluation. The present structural boom provides closely related templates for most receptors, except for the LGR (Leucine-rich repeat) and MRG (Mas-related) receptors. If, in these templates, some of the loops are not resolved, the numerous available structures give the opportunity to find loop templates with similar length for equivalent loops. However, simultaneously, the large number of putative templates leads to model ambiguities that may require additional information based on multiple sequence alignments or molecular dynamics simulations to be resolved. Using the modeling of the human bradykinin receptor B1 as a case study, we show how several templates are managed by MODELLER, and how the choice of template(s) and of template fragments can improve the quality of the models. We also give examples of how additional information and tools help the user to resolve ambiguities in GPCR modeling.

1. Introduction

Class A (rhodopsin-like) G protein-coupled receptors (GPCRs) form the largest family of transmembrane receptors in the human genome [1, 2]. They include about 300 non-olfactory receptors classified into a dozen of sub-families (Fig. 1) and 400 olfactory receptors. These receptors allow the transfer of information from an extracellular signal to the cell cytoplasm. The extracellular signal is usually a ligand that, after binding to the receptor, induces a conformational change from an inactive to an active conformation, which in turn binds to and activates effector proteins such as G proteins and arrestins. GPCRs participate in numerous physiopathological processes and are the target of about 30% of presently used drugs [3].

The pharmacological importance of GPCRs explains the considerable effort spent to resolve their molecular structure. These receptors share a common fold comprising seven transmembrane helices (TM), with highly conserved anchor positions in each helix [4]. The first structure of a GPCR, rhodopsin, was resolved in 2000 [5]. The resolution of a second receptor, the β2 adrenergic receptor, required seven years [6]. Since then, several technical locks were broken, and new GPCR structures from different families followed almost non-stop. Now in 2020, the structures of 57 unique class A receptors (all of them being non-olfactory receptors), totalizing more than 300 structures in different complexes or activation states have been resolved by X-ray crystallography, serial femtosecond crystallography, cryo-electron microscopy or solid state NMR, providing a deeper understanding of the mechanism of action of GPCRs [7-10]. Nevertheless, in spite of this avalanche of structures, about 80% of the non-olfactory GPCRs still do not have resolved structures.

Among the 57 class A receptors with at least one structure, 39 receptors have been resolved in inactive state only, 2 receptors in active state only and 16 receptors in different activation states. These structures reveal the structural diversity of the transmembrane fold (Fig. 2) and have evidenced the large conformational change occurring upon activation (e.g. the type 1 angiotensin II receptor, AT1, in Fig. 3a) with a pivotal motion of TM6 that opens an intracellular cavity allowing effector binding [11]. They have also revealed an allosteric binding site for the sodium ion (e.g. the δ opioid receptor, OPRD, in

Fig. 3b) that acts as an allosteric modulator [12] and should be taken into account in molecular docking to GPCRs [13]. Inactive states of GPCRs are adapted for drug design or virtual screening of antagonists/inverse agonists, whereas active states should be adapted for drug design of agonists [14, 15]. Design of biased agonists (agonists specific of a signaling pathway) raises an additional level of complexity [16]. Deorphanisation of orphan receptors often relies on virtual screening [17, 18]. In addition, understanding of the mechanisms of action of GPCRs requires MD simulations of receptors in different activation states. These studies still rely heavily on molecular modeling. Thus, even in the age of the GPCR structural boom, molecular modeling by homology is still necessary.

This chapter discusses basic and advanced features of molecular modeling of class A GPCRs with the homology-based MODELLER software [19, 20]. The concept of homology modeling is based on evolution. Proteins are homologous when they share a common ancestor, which result in structure and, to a lesser extent, to sequence and function similarity [21]. The unknown structure of the target protein is modeled from the known structure of (at least) one homologous protein (the template) and the sequence alignment of the target versus the template(s). With keeping in mind that homology modeling programs always give a model, the main questions concern the quality of the model(s) and the way(s) to improve them. The inputs have to be carefully prepared and the outputs carefully evaluated to take into account all available information, with critical assessment of the assumptions made.

Difficulties of homology modeling depend on the sequence similarities between the target and the template. It is generally assumed that, above a threshold of 30% of identity, homology modeling may be quite straightforward, even if caveats can occur [22]. Below this limit, modeling usually becomes increasingly difficult because of structural variations. In spite of their common fold, each GPCR structure is unique [7-10], with structural variations to adapt to the variety of ligands (Fig. 2). It is thus mandatory to carefully select template(s) to correctly translate sequence into structural similarities. Moreover, as active structures are far less frequent than inactive ones (Fig.1), modeling of active states requires special attention.

Here, we detail the procedure of molecular modeling of rhodospsin-like GPCRs and the customization of the modelling process that is possible with standalone MODELLER. We emphasize the importance of evolutionary information [23-26] for template selection and model customization. Using the human

bradykinin receptor B1 with bound sodium ion in the inactive state as a case study, we show how MODELLER manages multiple templates and how the quality of the resulting model(s) can be improved by the careful choice of multiple template(s) and of template fragment(s).

2. Materials

2.1 Hardware

A computer running Linux/Unix, Apple Mac OS X (10.6 or later), or Microsoft Windows (XP or later)

2.2 Software

The MODELLER 9.23 program [19, 27] can be downloaded and installed from http://salilab.org/modeller. It is written in Fortran 90 and uses Python for its control language. Thus, all input scripts to MODELLER are Python scripts. In addition to MODELLER, several tools are required:

- a text editor capable of outputting plain text files, such as the free and open-source software gedit, available for Linux, Windows and Mac OS X.
- a molecular viewing tool, such as the PyMOL Molecular Graphics System, Schrodinger, LLC or UCSF Chimera [28]. The structural alignment tools they include are usually sufficient.
- a multiple sequence alignment (MSA) program, such as ClustalW [29], T-COFFEE [30], or MUSCLE [31]
- a visual software for MSA editing. We recommend GeneDoc (Multiple Sequence Alignment Editor, Analyzer and Shading Utility) developed at Pittsburgh Supercomputing Center's National Resource for Biomedical Supercomputing [32].

2.3 Input Files

MODELLER needs three kinds of input files:

- The .py script file written in Python
- At least one .pdb file containing the structure of one template

- The .ali file indicating the alignment between the target and the template(s). This alignment file has a specific format for MODELLER.

2.4 Additional tools and web sites

- For template structures: The Protein Data Bank (https://www.rcsb.org/)
- For the search of homologous templates: direct mining of the Protein Data Bank with the Blastp utility in UniProt (https://www.uniprot.org/), Expasy (https://www.expasy.org/), or NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi)
- For general information on receptors: UniProt that includes Swiss-Prot with reviewed entries and UniProtKB with automatic entries. Swiss-Prot centralizes functional information on proteins with detailed annotations that are curated by experts.
- For secondary structure prediction: JPred4 [33], accessible at http://www.compbio.dundee.ac.uk/jpred4/index.html
- For quality check of the model: PROCHECK [34] accessible at https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/
- For building non-redundant homologous sequence sets: the nrdb90.pl perl script [35] that can be found at ftp://biodisk.org/Program/Perl/Bioperl/nrdb90.pl.

2.5 GPCR specific web sites

Two web sites gather invaluable, updated information on GPCRs:

- The GPCRDB database [10, 36], accessible at https://gpcrdb.org, which gathers sequences, structures, genetic variations and structure-based alignments, with a classification based on ligands
- The Zhang Lab web site at the university of Michigan (https://zhanglab.ccmb.med.umich.edu) with online services for GPCRs (GPCR-EXP: Experimental structures, GPCR-RD: Experimental restraints, GLASS: GPCR-ligand association database).

3 Methods

The general principles of homology modeling by satisfaction of spatial restraints as implemented in MODELLER are detailed in Note 4.1. Here we will detail the steps necessary for homology modeling of class A GPCRs in different activation states with MODELLER. For clarity purpose, as a case study, we will model the human bradykinin receptor B1 (UniProt access code: BKRB1_HUMAN, thereafter B1), in an inactive state with a bound sodium ion at the allosteric binding site.

3.1 Gaining information on your target receptor

Before initiating your modeling project, gather available information on your target. In addition to literature, check UniProt. In its reviewed part, each entry includes not only core data (amino acid sequence, name, citations..) but also classifications, cross-references, mutational data, natural variants, amino acid modifications (modified residues, glycosylation, lipidation, disulfide bonds ...), binary interactions, subunit structure, and possibly 3D structures or models. In addition to UniProt, two web sites provide updated, GPCR specific information. The GPCRDB [36] gathers sequences, structures, structural alignments and mutational data. The Zhang Lab web site centralizes experimental structures and distance restraints on GPCRs, and a GPCR-ligand association database [37].

3.2 Gain information on your receptor sub-family

Class A receptors are classified into a dozen of evolutionary based sub-families that make consensus in the literature [2, 25]. Within sub-families, sequence identity is usually around 25-30% but may be as low as 15% between sub-families. In addition, around twenty receptors cannot be related to any sub-family and are unclassified (UC). Evolutionary classification gives information on several sequence and structural patterns (Note 4.2). A first pattern concerns two transmembrane helices, TM2 and TM5, which

have variable proline motifs and, thus, structural variability [24, 25, 38], as exemplified in Fig. 2. A second pattern concerns the disulfide bonds stabilizing the receptor extracellular domain.

In addition, careful analysis of the multiple sequence alignment of the target sequence with receptors from its sub-family may reveal unusual patterns that should be taken into account in the modeling procedure and/or the subsequent computational studies, in particular (1) indels in transmembrane helices [24, 38], (2) mutations in highly conserved sequence motifs of GPCRs (DRY in TM3, CWXP in TM6, NPXXY in TM7) [4], and (3) mutations in the sodium binding site [12, 26].

For example, in our case study, the receptor B1 belongs to the sub-family of chemotaxic (CHEM) receptors (Fig. 1). It is closely related to a set of receptors for vasoactive peptide receptors (the bradykinin receptor B2, the angiotensin II receptors AT1 and AT2, the apelin receptor) and to the chemokine receptors. As these receptors, it possesses the P2.58/ P5.50 proline pattern characterized by a proline kink in TM2 and a proline bulge in TM5. In addition, as most chemotaxic receptors, it possesses two disulfide bonds between TM3 and extracellular loop 2 (ECL2) and between the N-terminus and ECL3. Please note that in the family classification found in UniProt, it is classified as class A GPCR (IPR000276 GPCR_Rhodpsn), bradykinin receptor B1 (IPR001186 Brdyknn_1_rcpt) and bradykinin receptor (IPR000496 Brdyknn_rcpt which includes B1 and B2 receptors). The intermediary sub-family level is not provided by UniProt (Note 4.3) but may be found in the literature [2, 24].

3.3 Searching for suitable template(s)

The primary requirement for homology modeling is the identification of at least one known structure with similarity to the target sequence to be used as template. Now, numerous structures of GPCRs are available and the choice of the "best" template(s) has to be done carefully, keeping in mind that selection of closest homolog based on sequence identity does not guaranty model accuracy [39, 40]. Nevertheless, as each sub-family presents unique structural features that have to be taken into account for modeling, it is recommended to select at least one template from the same sub-family, with similar proline and cysteine patterns. This should be possible in most cases except for the LGR (leucine-rich repeat GPCRs) and MRG (Mas-related GPCRs) sub-families for which no structure has been resolved.

Search of close homologs with resolved structures in the Protein data Bank can be carried out straightforwardly by blasting the target sequence from UniProtKB (hits with 3D structures or models), Expasy or NCBI (hits with 3D structures only). Homology modeling may be based on several templates, rather than a single one. Indeed, the use of several templates approximately equidistant from the target sequence, with a weighting based on sequence similarity, generally increases the model accuracy [41]. For our test case B1, the closest hits (in decreasing order of similarity) are AT1 (6 structures), AT2 (5), CCR5 (6), CCR9 (1), APJ (1), and then, the opioid receptors (6, 4 and 4 for delta, kappa and mu, respectively). All these receptors have the same TM2 and TM5 proline patterns as B1 and all, except the opioid receptors, have the double disulfide bonds in the extracellular domain. All, except AT2, have inactive state structures. AT1, AT2 and the opioid receptors have active state structures. Among these structures, only the δ opioid receptor (OPRD) has an inactive structure with a sodium ion bound at the allosteric binding site [42]. In these structures, N-and C-termini and loops may be missing, because of truncation, intrinsic disorder or replacement by a fusion protein for crystallization purposes.

3.4 Selecting suitable template(s)

Template selection depends on (1) the target state to be modeled (see the large outward motion of TM6 that differentiates the inactive from the active state of AT1 in Fig. 3a), (2) the resolution of the N- and C-termini and of the loops, and (3) the inclusion of bound ligands (receptor agonists or antagonists, ions, lipids and water molecules). The experimental resolution of the crystal structures is usually not the most crucial factor. By contrast, the impacts of modifications on the receptor done for crystallization (mutations, nanobodies, and insertions, with special concern for junctions) and of quaternary and crystal contacts have to be carefully evaluated. The specific sequence/structure patterns of GPCRs also need careful evaluation. It is necessary to check what is "normal" or not in the structures under scrutiny, such as curvature of helices, additional bends or kinks, unusual distortions (π or 3_{10} helices), unusual structural motifs, and helix orientations, in particular for the C-terminal helix 8 (H8).

Fig. 3 displays four putative templates for the modeling of B1: AT1 (4YAY) [43], CCR5 (4MBS) [44] and OPRD (4N6H) [42] in inactive states, and AT1 in an active state (6OS0) [45]. Among these templates, only OPRD has a resolved ICL3 loop and a bound sodium ion with coordination water, but

this receptor does not possess the second disulfide bond between the N-terminus and ECL3. Inactive AT1 has an unstructured ICL2 and an unconventional, tilted positioning of the C-terminus (also observed in CCR5, but not in active AT1). In CCR5, active AT1 and OPRD, ICL2 is structured as a helix. In addition, CCR5 has an unusual outward orientation of the C-terminus of TM6. Thus, different templates will be necessary to model the entire sequence of B1, but their selection will affect the resulting models.

3.5 Mining and analyzing receptor homologs

In many cases, ambiguities on "best" templates are observed. For example, in the case of B1, should we model ICL2 with an α-helix, as observed in the OPRD, CCR5 and active AT1 templates or as a coil as observed in inactive AT1 template? Other examples of ambiguities can be found when the target receptor does not possess one of the anchor prolines in the TM helices [38] or when there are insertions/deletions in loops of the target versus the template(s). In ambiguous cases, analysis of orthologous or paralogous sequence sets may be informative [24, 38]. To build these sets, use the InterPro identifiers (Note 4.3) and mine UniprotKB. The subsequent analysis may be facilitated by building a non-redundant set with sequence identities lower than 90%. This can be carried out easily with the nrdb90.pl perl script [35]. The multiple sequence alignment of the resulting set should be carefully checked and, if necessary, manually corrected with Genedoc [32]. This procedure will considerably reduce the number of orthologous sequences and, subsequently, allow to better visualize key evolutionary events. In some cases, secondary structure prediction using JPred4 [33] and customized MSA may be informative (see example below).

3.6 Including ligand or non-protein residues

If the template contains a ligand, water molecules, ions, or other non-protein residues (anything marked as HETATM in the PDB file), MODELLER can include them into the generated model. By default all HETATM records are ignored. They are read when the `env.io.hetatm` and `env.io.water` Booleans are set to `TRUE`. Ions and water molecules are indicated by 'i' or 'w' in the alignment and participate in the refinement step. The unrecognized residues or ligands are taken into account with the

BLK (‘.’) residue type (both in the template and the target sequences) to copy them as rigid bodies into the model. The atom coordinates are transferred but the BLK residues are static and do not participate in the refinement step.

The HETATM records are read from the templates, in the order they're written in the PDB file. Thus they must be indicated by the appropriate symbol (‘.’, ‘w’. or ‘.i’) in the *same order* as in the alignment. If the template includes extra HETATM ligands that must not be taken into account in the model, manually delete them in the PDB file or align them with a gap ('-') in the target sequence. If a chain break ('/') is added immediately before the '.' residues in the alignment, this will force the ligands to have a different chain identifier (ID). If you model a ligand peptide, you need to add the chain break (‘/’) between the receptor and the peptide ligand that will have different chain IDs.

To model the sodium binding site in an inactive receptor, use as template the sodium binding site of another receptor with similar binding mode [12]. Depending on the similarities with the other templates, you may prefer using the entire receptor or the sodium binding fragment including residues from TM2, TM3, and TM7. Most importantly, do not forget to include water molecules in the sodium binding site (Fig. 3b). They will not be present in the model if you do not indicate them. Be cautious when you select water molecules to be included. A strong overlap between the preliminary target model and a water molecule may lead to the crash of the MODELLER job. In that case, carefully check the water molecules to be included.

3.7 Building receptor chimera

In several cases, it may be interesting to use different templates for different parts of the target. In addition to the sodium binding site detailed above, two other cases are worth mentioning:

1. Modeling of active state target: Comparison of active/inactive structures reveals that largest structural changes during activation occur in TM5, TM6 and TM7 [6]. As active template are less frequent than inactive ones (Fig. 1), an active state target may be better modeled as a chimer of (parts of) closely related inactive templates and farer related active templates. This may be done by using either (i) both an active template for the overall fold (TM1 to TM7) and an inactive template for the most stable part (TM1 to TM4) or (ii) inactive templates for TM1 to TM4 and active

templates for TM5 to TM7. In this latter case, however, it will be necessary to include the active template TM3 to correctly adjust the orientation of the active and inactive templates. The precise determination of the active and inactive template regions to be used requires careful visual inspection.

2. Loop modeling: This is a difficult part of a modeling procedure. Loops are frequently missing in the template structure(s) and are difficult to reliably model *ab initio*. MODELLER proposes two functions (`loopmodel` and `dope_loopmodel`) to automatically generate/refine loops. However, the loops obtained with these methods may markedly differ from structures revealed in GPCR structures. Now that a large set of GPCR structures are available, it is highly preferable to search for resolved equivalent loops with the same length and use them as templates. Fig. 4 summarizes, for each loop, the lengths that have a structural template. In case of ambiguities, SS predictions based on judicious MSA may help. In addition, we note than insertions in loops have frequently an α-helical structure and may result in a protruding TM helix. A first example is given by the orexin receptor 1 (6TP3) in Fig. 2a, in which the long ECL3 (87 residues between the residues 5.50 and 6.50) is structured as a protruding N-terminal part of TM6. A second example is given by CXCR1, AT1 and CCR5 (Fig. 2 and 3) for which the long ECL3 is structured as a protruding N-terminal part of TM7.

3.8 Preparing template file(s)

After checking putative templates, prepare your PDB template file(s). MODELLER allows the user to select the first and last position of a *single contiguous* segment of the template to be used in the modeling procedure, but this selection is not possible in case of *discontinuous* segments. In this latter case, due for example to insertion of a fusion protein in the template, there is no alternative to manual editing of the pdb file to avoid long and unmanageable indels in the alignment file. Breaks in the structure have to be indicated with a '/' symbol in the alignment file. In any case, even when this is not strictly necessary, manual editing of the PDB file(s) to excise the regions of the template that will not be used in the modeling procedure is recommended.

3.9 Aligning model sequence with the template(s) and preparing .ali file

This is a strategic part of the modeling procedure that has to be done very carefully. Several points have to be kept in mind:

1. Identification of the best templates generally involves alignments of the target sequence with a set of available template sequences and structures. However, the "best" alignments obtained by automatic alignment programs depend on the parameters used and may not be "optimal". They do not take into account user's additional information that may improve the alignment and the resulting model. Manual corrections of alignment may be necessary for that purpose.
2. In MODELLER, the alignment file creates *spatial restraints* (Note 4.1). Thus it may be useful to "de-align" residues to remove special restraints, increase flexibility and let MODELLER deal with stereochemical restraints only. The easiest way to remove structural constraints is adding gaps in the .ali file.
3. The sequence of the template *must* match the sequence in the pdb file. Thus unresolved parts in the pdb file must be removed from the sequence.
4. The .ali file has a special format. The first line give the sequence name in the pir fomat (`>P1; name`). The third line gives the sequences. Each sequence must be terminated by the terminating character '*'. Each chain break must be indicated by a single '/'.
5. The second line gives information on the nature of the sequence and the region to be used. There must be 10 fields, separated by 9 colons (':'). The first field indicates if the sequence is a template (structure, structureX, structureN..) or the target (sequence), the second field indicates the sequence name, the fields 7 to 10 are optional but not the colon characters ':'. The fields 3 to 6 indicate first position and chain, last position and chain. If the template is the single contiguous segment, simply specify here the beginning and ending residues that will be used for modeling. Examples can be found in the MODELLER tutorial. However, when templates contain noncontiguous segments, the easiest way is to edit the templates to remove any atom or heteroatom not used in the modeling procedure and then to use the '.' character between the colons as in `'structure:name:.:.:.:.: : : '`. This indicates that all the residues of the template pdb file have to be read.

3.10 Adding or suppressing restraints

Two commands in MODELLER allow either adding restraints (`restraints.add()`) or removing restraints (`restraints.unpick()`). Different types of restraints can be added in a MODELLER script. However, for GPCR modeling, most useful ones are:

1. Secondary structure constraints. This allows extension of a TM helix, modeling of helix 8 which is often not resolved in templates, structuration of a missing loop as an α-helix or forcing the β-strand structure of ECL2.
2. Dihedral constraints to reorient side chains or favor interactions.
3. Distance constraints to maintain or create interactions

Disulfide bonds not present in the template(s) may be added by the `special_patches()` command. Suppressing restraints is also possible, in particular for structural restraints built from the alignment. This can be done with the `unpick(*atom_ids)` command. Alternatively, it can be obtained by de-aligning the alignment with the introduction of gaps. For example, before adding dihedral restraints on a sidechain, the residue can be de-aligned to avoid conflicts between restraints. It is worth to note that these commands work on the atoms or residues of the *target*, so that their correct identifiers in the target (and not in the template) have to be provided.

3.11 Model building

This is the easiest step in the procedure. Select the number of models and the refinement procedure of models, and, optionally of loops (not recommended in GPCR modeling except for short loops or loop regions). Among the five procedures proposed ('none', 'very.fast', 'fast', 'slow', 'very.slow'), the 'fast' and 'slow' options give optimal results. From 20 to 50 models usually ensure a good representability of the available conformational space. A good comprise is to initiate the modeling procedure with 20 models/fast refinement option and terminate with 50 models/slow refinement option, after optimization of restraints.

3.12 Evaluating models

Once one model or a set of models have been generated, there are different ways to further assess them. The DOPE potential [46] provided by MODELLER allows a comparison of model and template profiles and the visualization of putative problematic regions in the alignment. However, for GPCRs, when several models are generated from the same alignment, the DOPE scores are very similar and the MODELLER molpdf scores [19] that indicate the violations of the restraints are more discriminative. Several steps have to be carried out to evaluate models:

1. Check the log file from the modeling run for runtime errors and restraint violations. The global molpdf scores are indicated at the end of the log file, which allows determining the "best" models that will be used for further analysis.
2. Visually inspect the models or the "best" models to insure that no coarse mistake has been done in the alignment file or in the script. Superpose all the models to visualize the conformational space available, especially for loops.
3. Now focus on the "best" models (typically 5 out of 20 or 50 generated models). The sums of the violations for each residue are indicated at the B factor position in the PDB files of the models (*B999*.pdb). Visualize them on the model structures. For regions with higher violations, inspect the *V999* files that give the type of restraints that is violated. If necessary, modify the alignment or the restraints accordingly.
4. Additional methods can be used to assess the quality of the model, for example PROCHECK [34] that verifies the stereochemistry of the model.
5. In the selection of the "optimal" model used for further computational studies, be very careful to the orientation of key residues or motifs such as W4.50, W6.48 or DRY in TMH3. Be also very carefully to the positions of N- and C- termini and to the structures of the loops.

3.13 Return to our test case

Here we will analyze the molecular models of the human B1 receptor, with bound sodium ion, obtained from different templates. The aim is highlighting MODELLER practice to help future users in template selection. In Fig. 3, we have shown 4 representative structures of homologs: AT1 in the inactive state (closest homolog), OPRD (resolved ICL3 and closest homolog with bound sodium ion), CCR5

(homolog with resolved ECL2), and AT1 in an active state (resolved ECL2). Among these putative templates, we removed CCR5 because of the unusual outward positioning of TM6 on the extracellular side.

First, we used inactive AT1 (4YAY) and OPRD (4N6H) as templates. Since, in OPRD, there is no the disulfide bond linking the N-terminus and ECL3, we aligned the sequences of the templates with the target, except for the N-terminus and ECL3. Thus, both templates are used for ICL1, ICL2, ECL1, ECL2 and H8. Nevertheless, in the resulting models of B1 (Fig. 5a), the ICL1, ICL2 and ECL1 loops of the target match those of the AT1 template only (closest template). The ICL3 and ECL3 loops match their unique template, respectively, OPRD and AT1. The models differ strongly in the modeling of the ECL2 hairpin, since there is no template for this segment. Interestingly, concerning H8, MODELLER does not privilege one template over the other one, and the resulting models have either the tilted orientation observed in AT1 (4YAY) or the horizontal orientation observed in OPRD (4N6H), but not an average orientation.

In the second procedure, we aimed at improving the modeling of the ECL2 hairpin (Fig. 5b). As the length between TMH4 and the ECL2 cysteine is identical in B1 and AT1, we selected a fragment of active AT1 (6OS0) in which ECL2 is resolved. The fragment includes TM3 and TM4 for proper positioning, ICL2 (in helical conformation) and ECL2 up to the Cys residue. With the addition of this fragment, there are 3 templates for ICL2. Now, MODELLER privilege the helical structure of ICL2 and the tilted orientation of H8 in the B1 models.

These different models rise the issues of the orientation of H8 and of the structure of ICL2. Concerning H8, the influence of H8 modeling on subsequent MD simulations is detailed in Note 4.4, with the example of AT1, and we recommend to privilege the horizontal orientation of H8. Concerning ICL2, we note that, both in active AT1 and inactive CCR5, ICL2 has a helical structure. To help resolve the uncertainty between the helical and loop conformations of ICL2, we carried out secondary structure prediction using JPred4 [33]. Fig. 6 displays SS predictions for B1 using either automatic BLAST search or the MSA of the 52 human sequences that share two properties. First, they belong to the evolutionary related CHEM and PUR sub-families [24] and, second, they possess the double disulfide bonds in the

extracellular domain. The former approach is not informative. By contrast, the latter approach with a customized MSA predicts a helical structure for ICL2, supporting the helical conformation.

In the third procedure, the C-terminus of inactive AT1 was "de-aligned" to exclude it from structural restraints in MODELLER. In this case, the resulting models have a helical ICL2 and a horizontal H8 (Fig. 5c). Finally, in the fourth procedure, we were concerned with the modeling of the N-terminus from the cysteine (first residue in our models) to TMH1. This segment in B1 is one residue longer than AT1 or CCR5. In the first to third procedures, it was modeled with the insertion of one residue compared to the AT1 template. In this fourth procedure, since SS predictions suggest an N-terminal extension of TM1, we checked whether the length of this segment in B1 is compatible with a helical extension. Thus, we also aligned the N-terminus of OPRD in the alignment file. In this latter case, the B1 models do have an additional helical turn at the extracellular side of TM1 without violations of stereochemical restraints (Fig. 5d).

3.14 Concluding remarks

In this Chapter, we have shown how the choice of the template(s) is determinant for the resulting target models and how MODELLER deals with multiple templates. Rather than an average conformation, MODELLER estimates the probability of each conformation, and may propose several clearly different models, as exemplified by the H8 orientation in B1 modeling. We emphasize the importance of generating and comparing several models to estimate the conformational space compatible with the restraints provided by the template(s) and the alignment.

We have also shown that carefully selected receptor fragments can greatly improve modeling. Template-based combination of fragments is a powerful approach for molecular modeling of GPCRs that has been implemented in several web sites for automatic GPCR modeling, such as GPCR-I-TASSER [47] and GPCR-SFFE [48]. The automatic approaches are very efficient for the modeling of the transmembrane domain but do not provide to the user the possibility to use target specific information to guide modeling, especially for loops, now that a large variety of templates is available (Fig. 4). Understanding the details

of the homology modeling procedure by MODELLER will help the user to make rational choices that will allow improving the quality of customized GPCR models.

4 Notes

4.1 Modeling by satisfaction of spatial restraints

MODELLER belongs to a class of modeling methods that work by satisfying constraints or restraints on the structure of the target sequence using its alignment to related protein structures as a guide [19, 20]. The program is designed to use (1) spatial restraints based on the template structure(s) and the alignment file, (2) restraints based on statistical analysis of structures from homologous proteins, (3) restraints based on the CHARMM22 force field, and (3) restraints based on additional information about the target. These restraints are expressed as conditional probability density functions. For a receptor with about 350 residues and 2500 atoms, MODELLER builds and optimizes about 50000 restraints [20].

Spatial homology-derived restraints*:* In the first step of model building, distance and dihedral angle restraints on the target sequence are derived from the template structure and the alignment between the target and the template. Distance restraints are obtained by assuming that the corresponding distances between **aligned residues in the alignment file** are similar. Dihedral angle restraints are also derived from **aligned residues** in the alignment file. The forms of the restraints are based on empirical knowledge of structures of homologous proteins.

Stereochemical restraints*:* In the second step, additional restraints based on the CHARMM22 force field are added to enforce proper stereochemistry [49].

Adding restraints: Additional structural or functional information on the target cannot be derived from the template PDB file. This is the case of, for example, a suspected disulfide, cross-linking restraints, site directed-mutagenesis results, and predicted secondary structure. Most commonly added restraints

for GPCRs are distance restraints, dihedral angle restraints and secondary structure restraints (α-helix restraints for helix termini and β-strand/sheet restraints for extracellular loop 2). Information can also be obtained from careful examination of sequence alignment of the target orthologs which may provide general knowledge about receptor specificities. Adding restraints is a general way of taking into account these considerations. This can be easily done with MODELLER and may markedly improve the quality of the target structure, with the `special_restraints()` and `special_patches()` functions.

Removing restrainsts: MODELLER can unselect all the restraints on specified atoms with the restraints.unpick() command. It is also very easy to remove alignment-based spatial restraints by modifying the alignment with the addition of gaps.

Optimization/refinement*:* Finally, all the restraints are combined in an objective function that is optimized in Cartesian space. The optimization is carried out by the use of the variable target function method [50] to obtain the model. Then the model is refined by using conjugate gradients and simulated annealing (SA) [19]

Model reproducibility

Several slightly different models can be calculated by varying the initial structure (random shift), and the variability among these models can be used to estimate the errors in the corresponding regions of the fold.

Loop modeling: Loop modeling is an important aspect of comparative modeling for GPCRs, first because it may be the most variable part of sequences and structures, and second because many loops are not resolved or have been substituted by the insertion of a protein aimed at improving the crystallization. There are two main classes of loop-modeling methods: (i) database search approaches that scan a database of all known protein structures to find segments fitting the anchor core regions; (ii) conformational search approaches that rely on optimizing a scoring function. This is the case of

MODELLER with an optimization function, relying on conjugate gradients and molecular dynamics with simulated annealing [51].

Evaluating a model

Template selection and alignment accuracy usually have a large impact on the model accuracy, especially for models based on low sequence identity to the templates. If several models are calculated for the same target, this allows gaining information on regions with putative structural flexibility or variability. In that case, the "best" model can be determined by selecting the model with the lowest value of the MODELLER objective function or the DOPE [46] or SOAP [52] assessment scores, which are reported at the end of the log file. None of these scores are absolute measures, in the sense that they can only be used to rank models of the target.

4.2 GPCR classification

GPCRs are present in fungi, amoeba, animals and possibly plants [23, 53]. They have known a stupendous evolutionary success in animals where they highly diversified. Species-specific sub-families make difficult a general classification among the animal reign but, in vertebrates, the GRAFS classification into five families has been widely adopted [2, 53]. Among these 5 families, the rhodospsin-like or class A is the largest one with 700 out of the 800 total human receptors. These 700 receptors include 400 olfactory and 300 non-olfactory receptors. Finally, the 300 human non-olfactory receptors can be further classified into a dozen of evolutionary based sub-families, to which can be added the UC class of "unclassified" receptors. The nomenclature for these sub-families is as follows: PUR (purinergic receptors), CHEM (chemotaxic receptors, including vasoactive peptide receptors), SO (somatostatin and opioid receptors), PEP (peptide receptors), MLT (melatonin receptors), OPN (opsins), PTG (prostaglandin receptors), MEC (melanocortin, EDG and cannabinoid receptors), AD (Adenosine receptors), AMIN (Amine receptors), MRG (MAS-related receptors), LGR (leucine-rich repeat receptors). However, the evolutionary relationships between these sub-families are not obvious. The classification into four α, β, γ, and δ groups [2] has been widely adopted in the literature but is not consistent with the order of sub-family apparition during evolution [24, 53]. Based upon a multi-

dimensional scaling analysis, we proposed a model of radial evolution of GPCRs from ancestral peptide receptors with three main evolutionary pathways [25].

The model of radial evolution provides a framework to rationalize the variable proline patterns in TM2 and TM5 observed in class A GPCRs and is consistent with the order of sub-family apparition. Peptide receptors (PEP) are among the most ancient GPCRs Most PEP receptors are characterized by proline residues at position 2.59 and 5.50 and bulges in TM2 and TM5, as observed for the orexin receptor 1 [54] shown in Fig. 3a. The first evolutionary pathway is related to the deletion of one residue in TM2 and led the SO, CHEM and PUR receptors. This deletion occurred in an ancestor of the SO receptors that led by divergence to the CHEM and PUR receptors [24]. The SO, CHEM and PUR sub-families are characterized by the P2.58 and P5.50 proline patterns, corresponding to a kinked TM2 and a bulged TM5, respectively, as observed for CXCR1 [55] (Fig. 3b). The second pathway is related to divergence of amine (AMIN) and adenosine (AD) receptors, characterized by the same sequence and structural patterns as PEP receptors. The third pathway corresponds to independent evolution of several sub-families with mutations of the proline residues in TM2 and TM5, which are often correlated. These mutations can lead to straight TM2 and TM5, as observed in the cannabinoid receptor 1 [56]from the MEC sub-family (Fig. 3c). Prostaglandin receptors (PTG) present an example of sequence variability for TM2 (P2.59 or no proline) with a conserved bulge observed in PE2R3 (P2.59), PD2R (P2.59) and TA2R (noP) whereas the absence of proline in TM5 led to a bulged TM5 in PD2R and to straight helices in PE2R3 and TA2R [57-59]. The MRG and LGR sub-families have no proline residues in TM2 and TM5 and no structurally resolved member. Thus no *a priori* hypothesis can be done on the structure of these helices. Combination of cellular biology experiments, extensive sequence analysis and molecular dynamics simulations were necessary to propose a bulged TM2 and a straight TM5 in TSHR, a LGR receptor [38].

In addition to the proline pattern in TM2 and TM5, a second sequence pattern presents interesting features in class A GPCRs. It concerns the disulfide bond(s) stabilizing the receptor extracellular domain. A first disulfide bond links the N-terminal part of TM3 to the extracellular loop 2 (ECL2) and is present in most sub-families, except MEC and MRG. A second disulfide bond links the N-terminus to an extracellular extension of TM7 in some purinergic and chemotaxic receptors. Alternative disulfide

bonds have been observed in MEC receptors, whereas an additional disulfide bond is frequently found in ECL3 of amine receptors. Thus, each sub-family presents unique structural features which have to be taken into account for modeling.

It is worth noting that the GPCRDB [60] uses a classification based on the chemical nature of the ligands because it is focused on molecular docking and drug design. However, this may be confusing for homology modeling because similar ligands may bind receptors from different sub-families. For example, the lysophosphatidic acid receptors LPAR1-3 and LPAR4-6 are members of the MEC and of the PUR sub-families, respectively with no proline in TM2 and TM5 for LPAR1-3 and two proline residues at P2.58 and P5.50 for LPAR4-6.

4.3 Mining GPCRs in UniProt

In Uniprot, class A GPCRs can be easily mined by searching entries with the correct family identifiers. They are identified as PF00001 (7tm_1) in Pfam, IPR000276 (GPCR_Rhodpsn) or IPR017452 (GPCR_Rhodpsn_7TM) in InterPro, PS00237 (G_PROTEIN_RECEP_F1_1) or PS50262 (G_PROTEIN_RECEP_F1_2) in PROSITE. In InterPro, the IPR017452 identifier is broader than the IPR000276 identifier and includes a number of taste and vomeronasal receptors. In PROSITE, the PS50262 identifier, based on sequence profiles, is of higher quality than the PS00237 identifier, based on motifs. The three identifiers PF00001, IPR000276 and PS50262 are equivalent. Olfactory receptors can be identified as IPR000725 (Olfact_rcpt).

When the user wishes to go deeper in the classification tree, InterPro provides several useful levels of classification. For example, the angiotensin II receptors that include type 1 and type 2 have the reference IPR000248 (ATII_rcpt), the type 2 has the reference IPR000147 (ATII_AT2_rcpt), allowing hierarchical selection. However, the intermediary level of classification corresponding to the 12 sub-families is usually not taken into account either in UniProt or in GPCRDB, based on the nature of the ligand (see Note 4.2). Lists of receptors belonging to the 12 sub-families that make consensus in the literature may be found in [2, 24].

4.4 Positioning of helix 8

The user should never avoid critical assessment of the template(s). For example, in the structure of inactive AT1 shown in Fig. 3 (4YAY), the orientation of H8 is tilted. Such tilted orientation has been observed in other GPCR structures, e.g. for CCR5 (4MBS). However, it has not been observed in other structures of AT1, such as the recently resolved active structures of AT1 in complex with angiotensin II (6OS0) or angiotensin derivatives (6OD1, 6OS1, 6OS2). This suggests that the tilted orientation observed in the inactive structure of AT1 might result from artifacts, due to truncation of the C-terminus and experimental conditions.

To answer this question, we carried out molecular dynamics simulations of AT1 (Fig. 7). Starting from the 4YAY structure of AT1, we built two models of the receptor, the first one with H8 positioned as in the 4YAY structure and the second one with H8 positioned as observed in the OPRD 4N6H structure. The models were inserted into a POPC bilayer using the charmm-gui interface (www.charmm-gui.org) and then underwent a short equilibration step (1 ns heating procedure with progressive release of structural constraints) followed by 280 ns of production run with NAMD [61]. When H8 was modeled in an orientation parallel to the membrane bilayer, it remained stable in this orientation during the 280 ns of the simulations. When the starting structure had a tilted H8 orientation, this one was not stable. H8 underwent a large seesaw motion that induced strong perturbations not only of H8 but also of the intracellular sides of TM6, TM7 and TM1 (Fig. 7c). These results highlight the importance of the orientation of H8 on the stability of MD simulations and strongly suggests that the canonical horizontal orientation of H8 should be privileged in molecular modeling.

5 References

1. Bockaert, J. and J.P. Pin, *Molecular tinkering of G protein-coupled receptors: an evolutionary success.* EMBO J, 1999. **18**(7): p. 1723-9.
2. Fredriksson, R., et al., *The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints.* Mol Pharmacol, 2003. **63**(6): p. 1256-72.
3. Overington, J.P., B. Al-Lazikani, and A.L. Hopkins, *How many drug targets are there?* Nat Rev Drug Discov, 2006. **5**(12): p. 993-6.
4. Sealfon, S.C., et al., *Related contribution of specific helix 2 and 7 residues to conformational activation of the serotonin 5-HT2A receptor.* J Biol Chem, 1995. **270**(28): p. 16683-8.
5. Palczewski, K., et al., *Crystal structure of rhodopsin: A G protein-coupled receptor.* Science, 2000. **289**(5480): p. 739-45.
6. Cherezov, V., et al., *High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor.* Science, 2007. **318**(5854): p. 1258-65.
7. Xiang, J., et al., *Successful Strategies to Determine High-Resolution Structures of GPCRs.* Trends Pharmacol Sci, 2016. **37**(12): p. 1055-1069.
8. Garcia-Nafria, J. and C.G. Tate, *Cryo-Electron Microscopy: Moving Beyond X-Ray Crystal Structures for Drug Receptors and Drug Development.* Annu Rev Pharmacol Toxicol, 2020. **60**: p. 51-71.
9. Katritch, V., V. Cherezov, and R.C. Stevens, *Diversity and modularity of G protein-coupled receptor structures.* Trends Pharmacol Sci, 2012. **33**(1): p. 17-27.
10. Munk, C., et al., *An online resource for GPCR structure determination and analysis.* Nat Methods, 2019. **16**(2): p. 151-162.

11. Rasmussen, S.G., et al., *Crystal structure of the beta2 adrenergic receptor-Gs protein complex.* Nature, 2011. **477**(7366): p. 549-55.

12. Katritch, V., et al., *Allosteric sodium in class A GPCR signaling.* Trends Biochem Sci, 2014. **39**(5): p. 233-44.

13. Margiotta, E., G. Deganutti, and S. Moro, *Could the presence of sodium ion influence the accuracy and precision of the ligand-posing in the human A2A adenosine receptor orthosteric binding site using a molecular docking approach? Insights from Dockbench.* J Comput Aided Mol Des, 2018. **32**(12): p. 1337-1346.

14. Nygaard, R., et al., *Ligand binding and micro-switches in 7TM receptor structures.* Trends Pharmacol Sci, 2009. **30**(5): p. 249-59.

15. Congreve, M., J.M. Dias, and F.H. Marshall, *Structure-based drug design for G protein-coupled receptors.* Prog Med Chem, 2014. **53**: p. 1-63.

16. Shonberg, J., et al., *Biased agonism at G protein-coupled receptors: the promise and the challenges--a medicinal chemistry perspective.* Med Res Rev, 2014. **34**(6): p. 1286-330.

17. Diaz, C., P. Angelloz-Nicoud, and E. Pihan, *Modeling and Deorphanization of Orphan GPCRs.* Methods Mol Biol, 2018. **1705**: p. 413-429.

18. Stockert, J.A. and L.A. Devi, *Advancements in therapeutically targeting orphan GPCRs.* Front Pharmacol, 2015. **6**: p. 100.

19. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. **234**(3): p. 779-815.

20. Webb, B. and A. Sali, *Comparative Protein Structure Modeling Using MODELLER.* Curr Protoc Protein Sci, 2016. **86**: p. 2 9 1-2 9 37.

21. Devos, D. and A. Valencia, *Practical limits of function prediction.* Proteins, 2000. **41**(1): p. 98-107.

22. Sanchez, R. and A. Sali, *Advances in comparative protein-structure modelling.* Curr Opin Struct Biol, 1997. **7**(2): p. 206-14.

23. Chabbert, M., et al., *Evolution of class A G-protein-coupled receptors: implications for molecular modeling.* Curr Med Chem, 2012. **19**(8): p. 1110-8.

24. Deville, J., J. Rey, and M. Chabbert, *An indel in transmembrane helix 2 helps to trace the molecular evolution of class A G-protein-coupled receptors.* J Mol Evol, 2009. **68**(5): p. 475-89.

25. Pele, J., et al., *Multidimensional scaling reveals the main evolutionary pathways of class A G-protein-coupled receptors.* PLoS One, 2011. **6**(4): p. e19094.

26. Taddese, B., et al., *Evolution of chemokine receptors is driven by mutations in the sodium binding site.* PLoS Comput Biol, 2018. **14**(6): p. e1006209.

27. Webb, B. and A. Sali, *Comparative Protein Structure Modeling Using MODELLER.* Curr Protoc Bioinformatics, 2016. **54**: p. 5 6 1-5 6 37.

28. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.

29. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-8.

30. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* J Mol Biol, 2000. **302**(1): p. 205-17.

31. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.

32. Nicholas, K.B., N.H.B. Jr, and D.W.I. Deerfield, *GeneDoc: Analysis and Visualization of Genetic Variation.* EMBNEW.NEWS, 1999. **4**: p. 14.

33. Drozdetskiy, A., et al., *JPred4: a protein secondary structure prediction server.* Nucleic Acids Res, 2015. **43**(W1): p. W389-94.

34. Laskowski, R.A., et al., *AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR.* J Biomol NMR, 1996. **8**(4): p. 477-86.

35. Holm, L. and C. Sander, *Removing near-neighbour redundancy from large protein sequence collections.* Bioinformatics, 1998. **14**(5): p. 423-9.

36. Isberg, V., et al., *GPCRdb: an information system for G protein-coupled receptors.* Nucleic Acids Res, 2016. **44**(D1): p. D356-64.

37. Chan, W.K., et al., *GLASS: a comprehensive database for experimentally validated GPCR-ligand associations.* Bioinformatics, 2015. **31**(18): p. 3035-42.

38. Chantreau, V., et al., *Molecular Insights into the Transmembrane Domain of the Thyrotropin Receptor.* PLoS One, 2015. **10**(11): p. e0142250.

39. Castleman, P.N., et al., *GPCR homology model template selection benchmarking: Global versus local similarity measures.* J Mol Graph Model, 2019. **86**: p. 235-246.

40. Costanzi, S., et al., *Homology modeling of a Class A GPCR in the inactive conformation: A quantitative analysis of the correlation between model/template sequence identity and model accuracy.* J Mol Graph Model, 2016. **70**: p. 140-152.

41. Srinivasan, N. and T.L. Blundell, *An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure.* Protein Eng, 1993. **6**(5): p. 501-12.

42. Fenalti, G., et al., *Molecular control of delta-opioid receptor signalling.* Nature, 2014. **506**(7487): p. 191-6.

43. Zhang, H., et al., *Structure of the Angiotensin receptor revealed by serial femtosecond crystallography.* Cell, 2015. **161**(4): p. 833-44.

44. Tan, Q., et al., *Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex.* Science, 2013. **341**(6152): p. 1387-90.

45. Wingler, L.M., et al., *Angiotensin and biased analogs induce structurally distinct active conformations within a GPCR.* Science, 2020. **367**(6480): p. 888-892.

46. Shen, M.Y. and A. Sali, *Statistical potential for assessment and prediction of protein structures.* Protein Sci, 2006. **15**(11): p. 2507-24.

47. Zhang, J., et al., *GPCR-I-TASSER: A Hybrid Approach to G Protein-Coupled Receptor Structure Modeling and the Application to the Human Genome.* Structure, 2015. **23**(8): p. 1538-1549.

48. Worth, C.L., et al., *GPCR-SSFE 2.0-a fragment-based molecular modeling web tool for Class A G-protein coupled receptors.* Nucleic Acids Res, 2017. **45**(W1): p. W408-W415.

49. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* J Phys Chem B, 1998. **102**(18): p. 3586-616.

50. Braun, W. and N. Go, *Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm.* J Mol Biol, 1985. **186**(3): p. 611-26.

51. Fiser, A. and A. Sali, *Modeller: generation and refinement of homology-based protein structure models.* Methods Enzymol, 2003. **374**: p. 461-91.

52. Dong, G.Q., et al., *Optimized atomic statistical potentials: assessment of protein interfaces and loops.* Bioinformatics, 2013. **29**(24): p. 3158-66.

53. Fredriksson, R. and H.B. Schioth, *The repertoire of G-protein-coupled receptors in fully sequenced genomes.* Mol Pharmacol, 2005. **67**(5): p. 1414-25.

54. Rappas, M., et al., *Comparison of Orexin 1 and Orexin 2 Ligand Binding Modes Using X-ray Crystallography and Computational Analysis.* J Med Chem, 2020. **63**(4): p. 1528-1543.

55. Park, S.H., et al., *Structure of the chemokine receptor CXCR1 in phospholipid bilayers.* Nature, 2012. **491**(7426): p. 779-83.

56. Hua, T., et al., *Crystal Structure of the Human Cannabinoid Receptor CB1.* Cell, 2016. **167**(3): p. 750-762 e14.

57. Fan, H., et al., *Structural basis for ligand recognition of the human thromboxane A2 receptor.* Nat Chem Biol, 2019. **15**(1): p. 27-33.

58. Wang, L., et al., *Structures of the Human PGD2 Receptor CRTH2 Reveal Novel Mechanisms for Ligand Recognition.* Mol Cell, 2018. **72**(1): p. 48-59 e4.

59. Morimoto, K., et al., *Crystal structure of the endogenous agonist-bound prostanoid receptor EP3.* Nat Chem Biol, 2019. **15**(1): p. 8-10.

60. Isberg, V., et al., *GPCRdb: an information system for G protein-coupled receptors.* Nucleic Acids Res, 2017. **45**(5): p. 2936.

61. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* J Comput Chem, 2005. **26**(16): p. 1781-802.

Acknowledgments: This study was supported by institutional grants from INSERM, CNRS and University of Angers. This work was granted access to HPC resources of IDRIS (GENCI grant 100567 to MC). MC is supported by CNRS. AT is supported by a fellowship from the University of Carthage (Tunisia). RB is supported by a fellowship from the University of Angers (France).

Figure Captions

Fig. 1: Evolutionary tree of human class A GPCRs indicating structurally resolved receptors by sub-family. The color code of the circles depends on the GPCR sub-family (PUR: light green, CHEM: blue, SO: red, MAS: teal, PTG: violet, PEP: dark green, MLT: light grey, LGR: khaki, OPN: orange, MEC: magenta; AD: dark grey; AMIN: cyan). The circle is open when the receptor has no structure and closed when the receptor has at least one structure. Receptors with at least one active state structure are indicated by squares (black squares for both inactive and active state structures, open squares for active state structures only). Receptors with a sodium-bound structure are indicated by red stars. For the resolved structures, orthologous receptors from any organism are taken into account. The receptor names in the tree correspond to the UniProt identification name without the "HUMAN" extension. The arrows indicate the B1 target and the CCR5, AT1 and OPRD templates that are discussed in the text.

Fig. 2: Structural diversity of class A GPCRS. (a) The orexin receptor 1, OX1R, has a bulged TM2 (P2.59), a kinked TM (P5.50) and a disulfide bond linking the extracellular terminus of TM3 to ECL2; (b) The receptor CXCR1 has a kinked TM2 (P2.58), a bulged TM5 (P5.50), the TM3-ECL2 disulfide

bond, and an addition disulfide bond linking the extracellular terminus of TM7 to the N-terminus of the receptor; (c) The cannabinoid receptor 1, CBR1, has two straight TM2 and TM5 helices and an unusual disulfide bond in ECL2. In the three cases, the receptors are in an inactive state. The TM2 and TM5 proline residues are magenta. The PDB codes are 6TOD (OX1R), 2NLN (CXCR1) and 5TGZ (CBR1).

Fig. 3: Comparison of templates for B1 modeling. The inactive structure of AT1 (PDB 4YAY) is superposed on the structure of (a) active AT1 (PDB 6OS0), inactive OPRD (PDB 4N6H) and inactive CCR5 (PDB 4MBS). The structures are shown as white ribbons with differences highlighted in magenta for inactive AT1 and slate for the other structures. The sulfur atoms of the disulfide bonds are shown as magenta and slate balls for inactive AT1 and the other receptors, respectively. In (a), the arrow indicates the pivotal motion of TM6 upon activation. The tilted orientation of H8 in the inactive structure is not observed in the active structure. In (b) the sodium ion and coordination water molecules present in the structure of OPRD are shown as yellow and grey spheres, respectively. The orientation highlights the tilted orientation of H8 in AT1 and the structure of ICL3 in OPRD. In (c), the orientation highlights the difference in the structures of ICL2 and the tilted orientation of the TM6 extracellular terminus in CCR5.

Fig. 4: Statistical analysis of the loop lengths in human GPCRs. The length is measured as the number of residues between the anchor residues *n*.50 present in each helix *n*. Blue bars indicates that resolved loops of the indicated length are present in the available GPCR structures. Grey bars indicate the absence of resolved loops. For ICL3, the arrows for the long loops correspond to a length of 87 residues for the PEP receptor OX1R (6TP3) and to a length of 78 residues for the AMIN receptor ADRB2 (6MXT).

Fig. 5: Influence of templates on the resulting B1 models. In each panel, the top 5 models (out of 20) obtained with MODELER are superposed with the templates used in modeling procedure. For clarity purpose, all the structures are shown as white ribbon, except the regions of interest in the templates. In (a), B1 is modeled from OPRD (blue) and inactive AT1 (magenta). In (b), B1 is modeled from OPRD (blue), inactive AT1 (magenta) and a fragment from active AT1 encompassing TM3, ICL2, TM4 and

ECL2 up to the Cys residue (green). In (c), B1 is modeled as in (b) except that the C-terminus of AT1, in light pink, has been "de-aligned". In (d), B1 has been modeled as in (c), except for the N-terminus. In (a-c), the N-terminus has been modeled from the AT1 template only. In (d), the OPRD template has been included, resulting in an additional helical turn at the N-terminus of TM1.

Fig. 6: Secondary structure predictions for B1 using JPred4. Automatic Jpred4 prediction is based on automatic BLAST search starting from the B1 sequence. The customized prediction is based on a user-provider MSA of 52 human receptors with the two disulfide bonds in the extracellular domain as observed in B1. The SS drawing corresponds to the experimental SS of AT1 in the 4YAY structure.

Fig. 7: Consequences of the orientation of H8 on MD simulations of the AT1 receptor. MD simulations were run for 280ns with NAMD, starting from an AT1 model with horizontal H8 (a) and tilted H8 (b). The starting conformations are blue in (a) and magenta in (b). Representative snapshots from the beginning to the end of the production run indicate the conformational space sampled by H8 during the simulations. The snapshots are shown as white ribbons, with increasing greying of the C-terminus from white to dark grey with the simulation time. In (c), comparison of RMSF of AT1 when the trajectory was started with the horizontal (blue) and tilted orientation (magenta) of H8.

Homology modeling of class A G-protein-coupled receptors in the age of the boom in resolved structures

Asma Tiss[1,2#], Rym Ben Boubaker[1#], Daniel Henrion[1], Hajer Guissouma[2] and Marie Chabbert[1*]

Figure 1

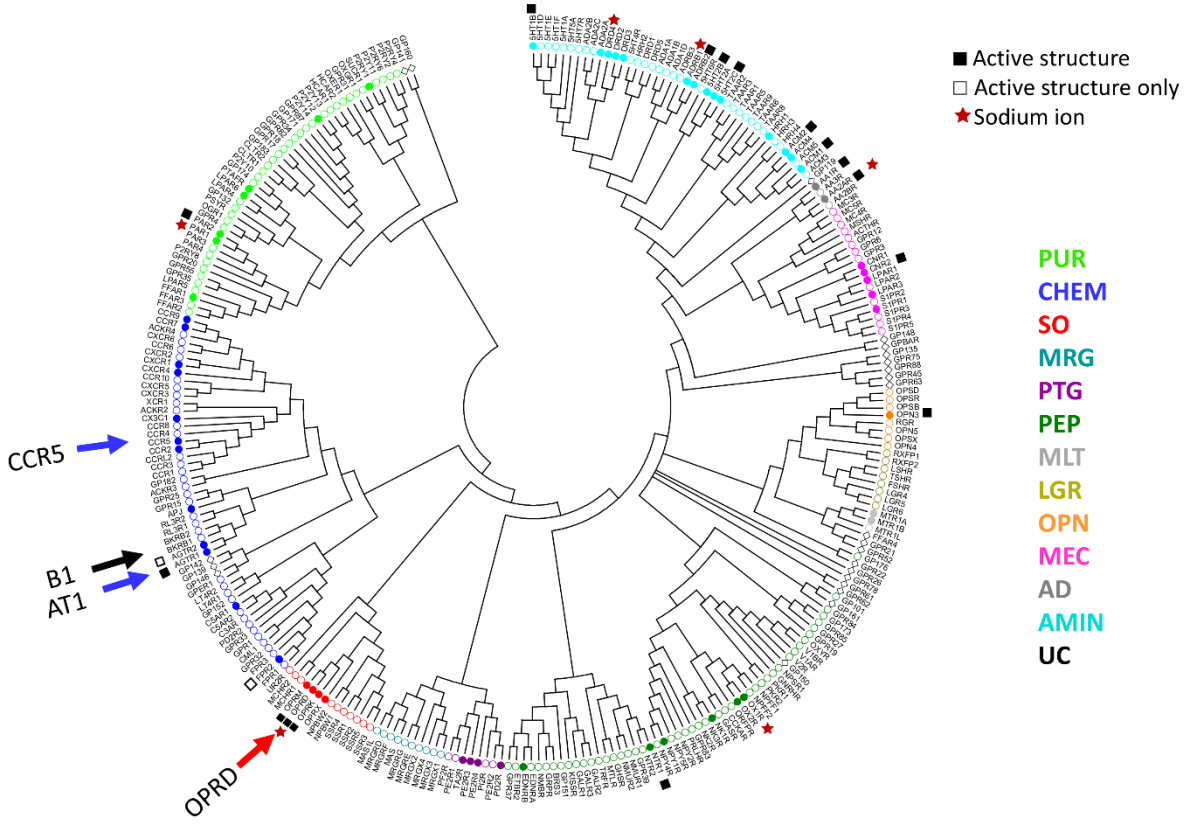


Figure 2

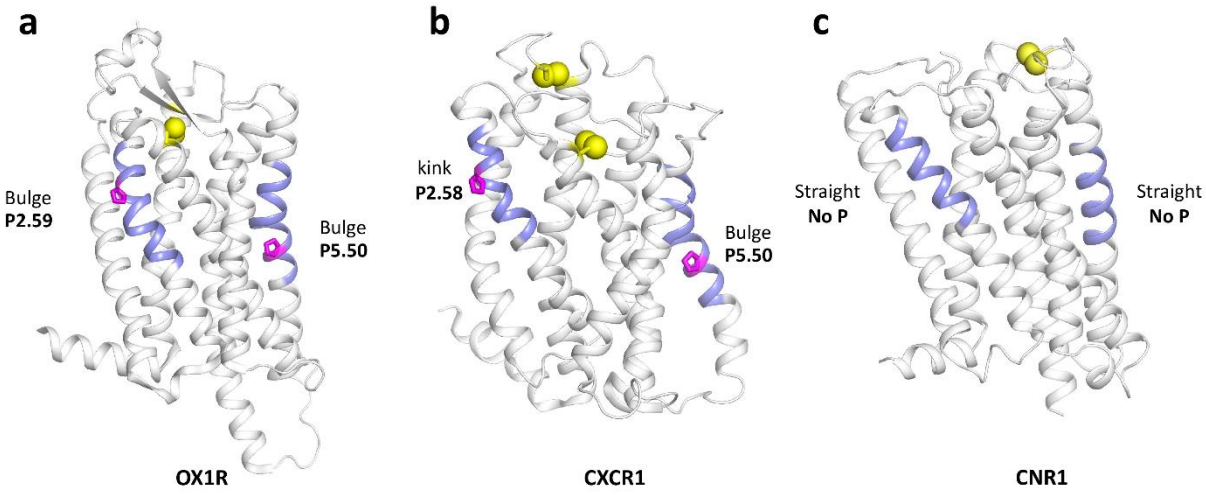


Figure 3

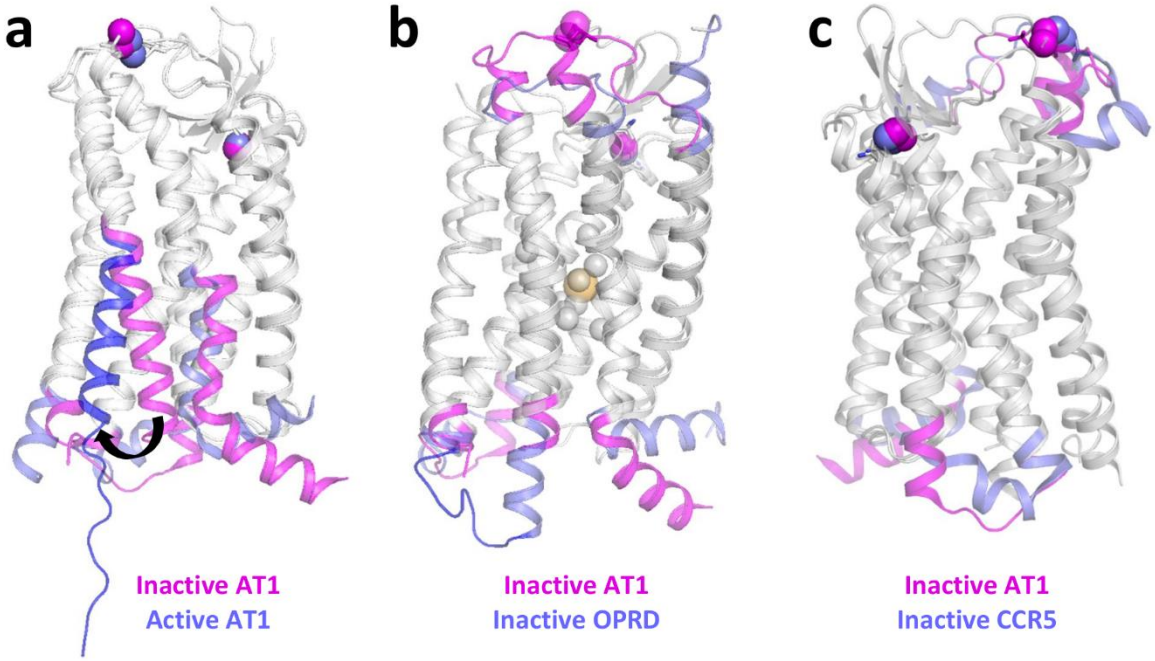


Figure 4

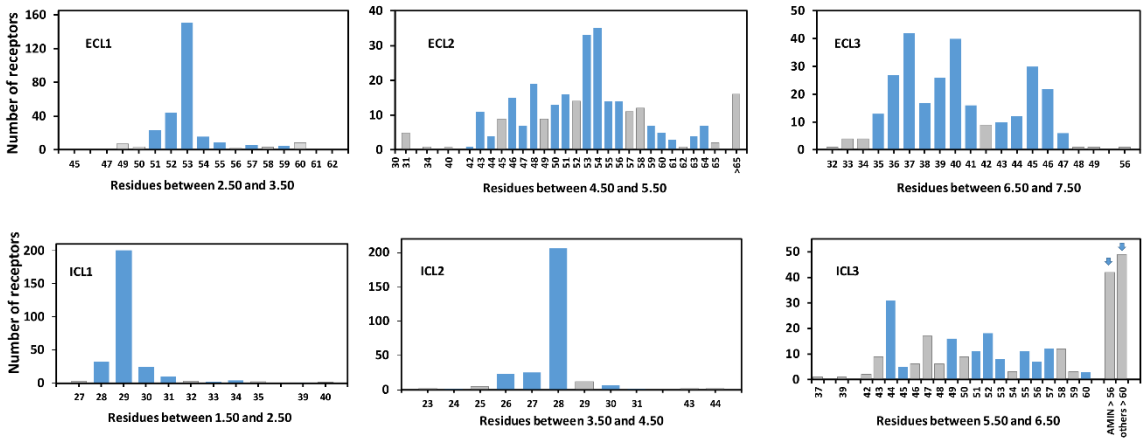

Number of receptors
ECL1
Residues between 2.50 and 3.50
ECL2
Residues between 4.50 and 5.50
ECL3
Residues between 6.50 and 7.50
ICL1
Residues between 1.50 and 2.50
ICL2
Residues between 3.50 and 4.50
ICL3
Residues between 5.50 and 6.50
AMIN > 56
others > 60

Figure 5

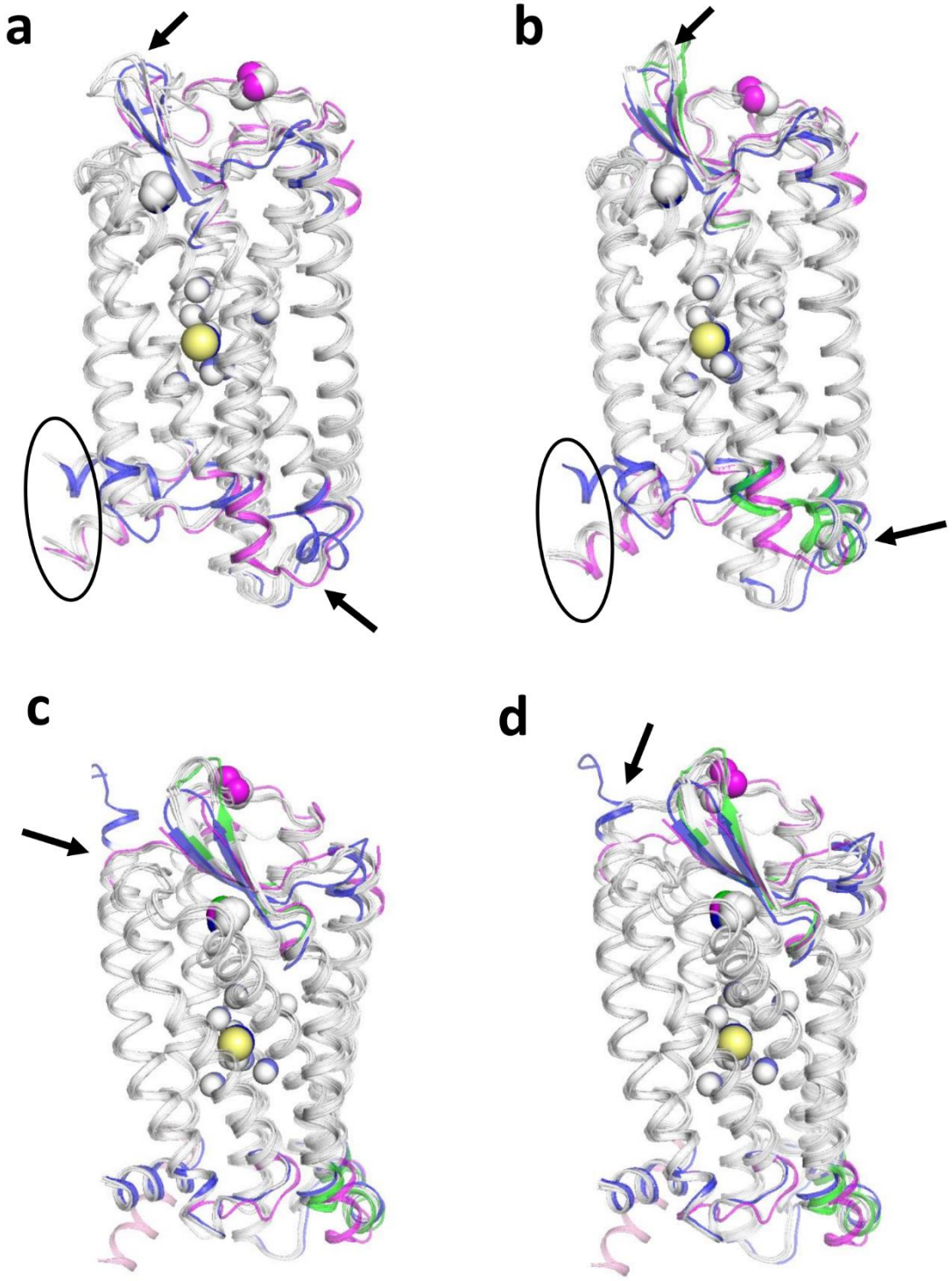


Figure 6

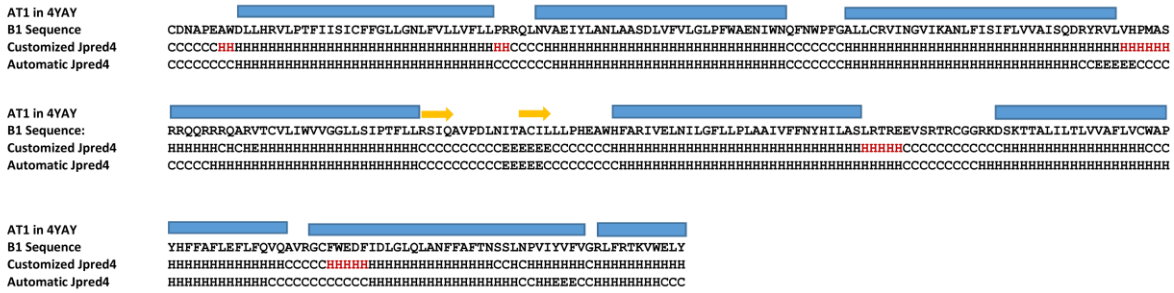
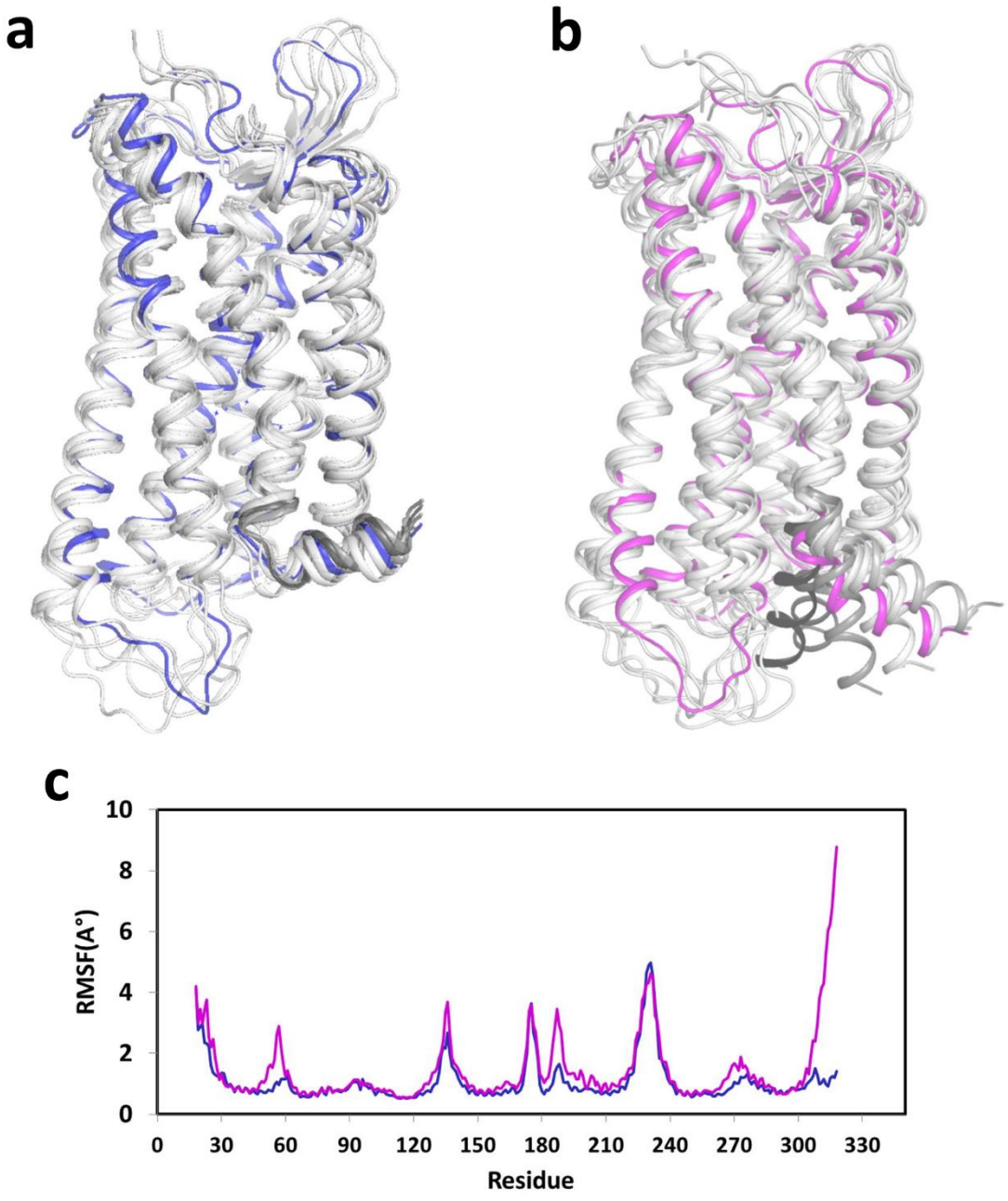

AT1 in 4YAY
B1 Sequence
CDNAPEAWDLLHRVLPTFIISICFFGLLGNLFVLLVFLLPRRQLNVAEIYLANLAASDLVFVLGLPFWAENIWNQFNWPFGALLCRVINGVIKANLFISIFLVVAISQDRYRVLVHPMAS
Customized Jpred4
Automatic Jpred4
AT1 in 4YAY
B1 Sequence:
RRQQRRRQARVTCVLIWVVGGLLSIPTFLLRSIQAVPDLNITACILLLPHEAWHFARIVELNILGFLLPLAAIVFFNYHILASLRTREEVSRTRCGGRKDSKTTALILTLVVAFLVCWAP
Customized Jpred4
Automatic Jpred4
AT1 in 4YAY
B1 Sequence
YHFFAFLEFLFQVQAVRGCFWEDFIDLGLQLANFFAFTNSSLNPVIYVFVGRLFRTKVWELY
Customized Jpred4
Automatic Jpred4

Figure 7


a
b
c
RMSF(A°)
10
8
6
4
2
0
0
30
60
90
120
150
180
210
240
270
300
330
Residue