Modeling GPCRs

A.C.M. Paiva, L. Oliveira, F. Horn, R.P. Bywater, G. Vriend

In memoriam Antonio Paiva

While we were working on this article, our good friend, colleague, and mentor, Antonio Paiva, died after losing the fight to cancer. We know Antonio as a stimulating force in the GPCR field. He has been one of the founding fathers of the informal GPCR club that met regularly at the EMBL in the early 1990s. The GPCRDB sprouted from these meetings. The thousands of scientists that use the GPCRDB every day owe Antonio thanks. His co-authors will miss him, and send words of condolence to his family. May they find consolation in the fact that the cancer could not stop him from finishing this article. We will miss him.

In memoriam Florence Horn

This paper is a dedication to the memory of Florence Horn who died shortly after we finished this article. Flo did so much to make genomics data come alive in a way that was meaningful to her and to thousands of researchers in bioinformatics, molecular biology, structural biology, and medicinal chemistry. Along with all these researchers, the coauthors of this paper wish to put on

record their thanks to Flo and to remember her as a colleague who bestowed her good humor and joie de vivre on all those who worked with her.

Abstract. Many GPCR models have been built over the years for many different purposes, of which drug-design undoubtedly has been the most frequent one. The release of the structure of bovine rhodopsin in August 2000 enabled us to analyze models built before that period to learn things for the models we build today. We conclude that the GPCR modeling field is riddled with common knowledge. Several characteristics of the bovine rhodopsin structure came as a big surprise, and had obviously not been predicted, which led to large errors in the models. Some of these surprises, however, could have been predicted if the modelers had more rigidly stuck to the rule that holds for all models, namely that a model should explain all experimental facts, and not just those facts that agree with the modeler's preconceptions.

1 Introduction

GPCRs are essential components in biological signaling processes in higher animals and accordingly, for humans, constitute the most important set of targets for the pharmaceutical industry, as is indicated by the fact that 52% of all medicines available today act on them (Watson and Arkinstall 1994). Approximately 16,700 GPCR sequences and are publicly available today (Bairoch et al. 2005), including 1,795 human proteins. The GPCRDB (Horn et al. 1998, 2003) is a worldwide repository for GPCR-related data. In addition to sequence data and multiple sequence alignments, the GPCRDB (www.gpcr.org/7tm/) gives access to approximately 8,000 mutations (Beukers et al. 1999; Horn et al. 2004). Binding constants are available for approximately 30,000 ligand-receptor combinations obtained from two different sources. Massive data is also available regarding chromosomal location, cDNA sequences, secondary structure, 3D models, and correlation mutation analyses. Query and navigation tools are also provided and allow users to retrieve local and remote information such as associated disease states, localizations, post-translational modifications, etc. Snake-like diagrams (Campagne et al. 2003) are used to offer a two-dimensional view of the receptors but also to combine sequence, structure, and mutation data. In the database, the data organization is based on the pharmacological classification of GPCRs. In addition to the five main classes (A–E), other putative GPCR families are also described, these are frizzled/smoothened family, ocular albinism proteins, insect odorant receptors, plant Mlo receptors, nematode chemoceptors, vomeronasal receptors, taste receptors T2R as well as numerous unclassified receptors. Bacteriorhodopsins are present for historical reasons. It is worth noting that most of the GPCRs present in the GPCRDB have not (yet) been proven to couple to G-proteins and we should rather talk about heptahelical receptors—and maybe rename the database 7TMDB.

There have been many dramatic developments in the use of modern "omics" technologies in drug design from genomics to metabonomics. Nevertheless, the chemical structure/function space is both disjoint and replete with highly redundant structures, which makes navigation difficult. Still today an element of luck is necessary, and this is reflected in the fact that the rate of discovery of new medicines has declined.

Structure-based design has occasionally been successful, but it is precisely in the GPCR area where structure-based design has not worked satisfactorily. The paucity of accurate structural data for GPCR templates and the desire to remedy this situation has spawned an entire generation of modelers intent on calculating/predicting GPCR structures. Before August 4, 2000, bacteriorhodopsin (Henderson and Schertler 1990; Pebay-Peyroula et al. 1997; Luecke et al. 1998; Takeda et al. 1998) was often used as a modeling template, but on that date the three-dimensional coordinates (Palczewski et al. 2000) of bovine rhodopsin became available, providing a much better template for GPCR modeling than bacteriorhodopsin, which is not even a GPCR. Moreover, bovine rhodopsin is not the perfect template, as we will explain in this chapter. Models produced Before the Crystal structure became available are called BC models, and those produced After these Data became available, AD models.

1.1 BC Modeling

Most BC models were based on low-resolution electron cryomicroscopic models of bacteriorhodopsin (Henderson and Schertler 1990) while the precision (but not the accuracy) was improved when X-ray crystal structures of bacteriorhodopsin became available (Pebay-Peyroula et al. 1997; Luecke et al. 1998; Palczewski et al. 2000). The situation improved with the availability of the Cα coordinates produced by J. Baldwin (Baldwin 1993) from an electron diffraction map (Unger and Schertler 1995; Schertler et al. 1993; Unger et al. 1997; Schertler and Hargrave 1995) produced by the Schertler group. A few models (Filizola et al. 1998; Prusis et al. 1997; Bramblett et al. 1995) were based on first principles, sometimes guided by low-resolution data measured from published slices of the electron density maps for the bovine or frog rhodopsin.

The BC modeling community developed a series of dogmas that are summarized in Box 1. Many of these are unfortunately still applied to this day by some modelers.

Given these dogmas, it can easily be understood why most modeling recipes followed the steps listed in Box 2.

Box 1 GPCR modeling dogmas and misconceptions from the BC era

Loops stick out into the solvent.
Isolated loops have the same structure as in a GPCR.
Polar residues point inward.
Helices stop at the membrane surface.
All helices are about equally long.
Helices must be perfect.
Helices are organized in a semicircular fashion.
Molecular dynamics software improves models.
There must be space in the apo-form for a ligand.
Activation does not require motion.
Important residues bind ligand or G-protein.
Important residues point inward.
The bacteriorhodopsin structure is a solution to the problem of how to pack seven helices in the membrane. It is therefore the only solution.
Bacteriorhodopsin is a GPCR without G-protein.
The lysines in helix VII should line up.
Proteins are simple.
Models are correct.

Box 2 Typical steps in a generic BC modeling project

First	Determine which template to use, or design your own helix-packing model.
Second	Use threading or moment calculations to determine the mapping of the GPCR sequence onto the selected template. Moment calculations can be based on hydrophobic moments (Donnelly et al. 1993), conservation moments (Pardo et al. 1992), etc., or a combination of these (Herzyk and Hubbard 1998). Threading can be based on general rules, helix bundle rules (Herzyk and Hubbard 1998; Pogozheva et al. 1997), or even bacteriorhodopsin-specific rules (Cronet et al. 1993).
Third	Find experimental data that agree with the model and add them to convince yourself or the referees that this is the only correct model.

We found very many publications that discussed poor BC models, showing that bluff will fly with referees and editors if the topic is important enough. Sadly, even the poorest models seemed to agree with all data (selected by the authors) and seemed to be perfect for designing drugs (according to the authors). As a courtesy, we will not list those articles here.

1.2 The Bovine Rhodopsin Structure

The high-resolution structure of rhodopsin (Palczewski et al. 2000; Schertler 2005) reveals a seven-helix bundle with a central cavity surrounded by helices I–III and V–VII (see Fig. 1).

The helices are in blue-purple. The β-hairpin in the N-terminal domain and the β-hairpin between helices IV and V (commonly known as the second extracellular loop) are in orange. The retinal is in yellow. Irregular parts are in blue-green. The topmost helix is helix IV.

Helix IV is not part of the cavity wall in this structure and makes contacts only with helix III. However, helix IV has been suggested to make contacts with some agonists which, if correct, is one of the many pieces of evidence that the active structure differs from the inactive one represented here. The conserved tryptophan at position 420 (we use GPCRDB residue numbering throughout this article) in helix IV is far away from any other residue known to have a functional role. This tryptophan might therefore play a role in receptor dimerization. Dimeriza-

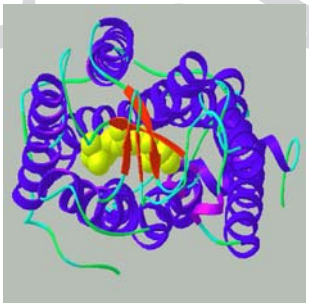


Fig. 1. The structure of bovine rhodopsin seen from above

tion by helix IV–helix IV contacts would allow for a force on the loop IV–V that might regulate the ligand entry in the other partner of the dimer. The rhodopsin crystal dimer structures do not resolve this question, as the experimental structures show antiparallel helix bundle pairs, whereas the natural dimers must be parallel bundle pairs. The β-hairpin between helices IV and V prevents access from the outside. This hairpin lies entirely between the helices, roughly parallel to the membrane surface. It has contacts with side chains of most of the helices. The most prominent contact is a disulphide bridge (Cys315–Cys480) to helix III. This calls for an explanation as to how ligands enter the binding cavity. For lipophilic ligands, like retinal itself (Schadel et al. 2003), entry/exit is expected to proceed via the membrane, as lipophilic ligands will accumulate in the membrane. For hydrophilic ligands, which include some peptides, insertion of the ligand will require some rearrangement of the loops including the hairpins. Not only will the hairpins have to make some adjustments, but the TMs will also move relative to one another. A number of clues as to what changes are likely to take place in the transition between the active and inactive structures have been published (Gouldson et al. 2004). In that work, and in a number of experimental studies cited therein (Gether and Kobilka 1998; Javitz et al. 1998), there is a movement of TM6 relative to TM3 and TM5. The crystal structure of bovine rhodopsin showed (Teller et al. 2001) that TM6 is unique in having only one hydrogen bond to another TM (TM7), while the other TMs are anchored by three or more interhelical hydrogen bonds.

2 Methods

Much GPCR-related research relies on access to all available data in a single easy-to-use data system, the GPCRDB. The principal data types contained in the GPCRDB are sequences, mutations, and structural information. Other GPCR-related information is accessible from the database's home page. Here we will describe the main steps of the GPCRDB update, its contents, and some of its functionalities.

The GPCRDB update procedure is handled by a series of python scripts, a MySQL database, and the WHAT IF (Vriend 1990) software.

Only the classification of new proteins having remote sequence similarity with already classified GPCRs, the definition of new families, and data checking in general, require some manual intervention and expertise. The other steps are fully automated.

GPCR proteins are imported from the Uniprot server (Bairoch et al. 2005). Receptors are then classified into the defined classes, families, and subfamilies using a profile-based method implemented in WHAT IF. Sequences that failed in the automatic classification step are further examined and classified manually. Fragments and short isoforms are put aside and are not used in the alignments in order to offer the highest possible alignment quality. For each class, family, and subfamily, WHAT IF is used to build multiple sequence alignments, phylogenetic trees, and other sequence-derived data in an automated manner. The profiles used for the alignments contain the location of the transmembrane domains and therefore allow us to ensure that the most conserved regions of the receptors are aligned without insertions and deletions. WHAT IF also produced the HTML pages to access the family-specific sequence data. cDNAs are imported from the EMBL databank (Cochrane et al. 2006) and align to their corresponding proteins using the genewise (Birney et al. 2004) program. Mutation data are identified and extracted from full-text articles with the MuteXt software (Horn et al. 2004). The latter automatically retrieves the corresponding UniProt entries, validates point mutations using sequence data and text mining approaches, and builds HTML pages to display mutation data as a function of receptors, articles, or residue positions. Multiple sequence alignments and snake-like diagrams (Campagne et al. 2003) are used to combine sequence, secondary structure, and mutation data. The use of the GPCRDB residue numbering system (Oliveira et al. 1993) permits this combination of many heterogeneous data types. A number is attributed to each residue in the seven transmembrane domains for all GPCR classes. This numbering system allows for fast comparisons between cognate residues in different receptors. In the mutation section of the database, the numbering system defined by Ballesteros and Weinstein (1995) is also indicated. Tables of available cross-references are provided for each GPCRDB entry to list all the different local information available and to ease navigation toward remote databases. The cross-references have been extracted from the UniProt entries and

other databases. Among other data query and retrieval tools, a Blast search against the GPCRDB can be performed via the CMBI server. The GPCRDB data content is available via anonymous FTP from ftp://www.gpcr.org/7tm/. A complete copy of the whole GPCRDB can be obtained upon request.

Copious amounts of data and speculative hypotheses can be found at the GPCRDB and they will not be reproduced here. The GPCRDB also contains a necropolis of earlier attempts at constructing GPCR models, and, more auspiciously for the future, assuming a steady accretion of good template structures, a detailed recipe for building models. Bovine rhodopsin and bacteriorhodopsin (Henderson and Schertler 1990; Pebay-Peyroula et al. 1997; Luecke et al. 1998) are sufficiently differently organized to make any detailed structural comparison meaningless (Unger et al. 1995, 1997; Schertler et al. 1993; Teller et al. 2001). However, in order to evaluate the quality of models based on the bacteriorhodopsin template, this superposition must be made. We therefore did this structure superposition by hand. Our recipe for determining the quality of bacteriorhodopsin-based BC models is given in Box 3.

Box 3 Recipe for judging BC model quality

Extract from the GPCRDB the alignment of the sequence of the GPCR model with the sequence of bovine rhodopsin.
Use the superposed structures to align the bovine rhodopsin sequence onto the bacteriorhodopsin sequence.
Extract from the modeling article how the authors aligned their GPCR with bacteriorhodopsin. If this alignment is not given, it can be extracted from a superposition of the bacteriorhodopsin-based GPCR model on the real bacteriorhodopsin structure.

This produces the alignment used for modeling. A comparison of the optimal alignment with the alignment used by the modeler is a good indication of the model's quality. This same method is used by the CASP competition judges to evaluate threading results (Venclovas et al. 2001). Our recipe for obtaining these BC-model alignment shifts differs from what is normally used because only the structure of bovine rhodopsin is known, while the beta-adrenergic receptor is the most modeled GPCR.

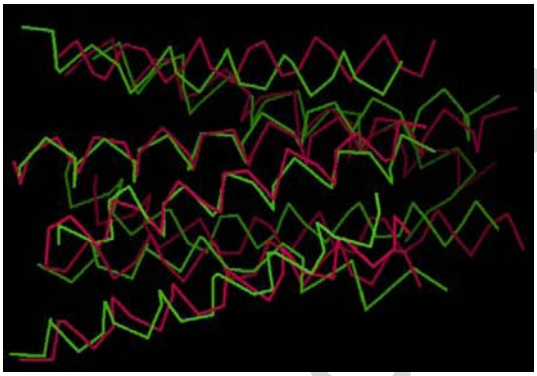


Fig. 2. Superposed bovine rhodopsin structure and BC-model

3 Results and Discussion

3.1 The Quality of BC Models

Figure 2 shows the superposition of the structure (Palczewski et al. 2000) and a very good BC model built, published (Baldwin 1993), and deposited before August 2000. It can be seen that the gross features are modeled reasonably well. The Cα- and all-atom modeling errors (i.e., displacements between the model and the X-ray structure) are 2.5 Å and 3.2 Å, respectively. Although impressive, this model is still too poor to be of any use for rational drug design purposes.

The bovine rhodopsin structure (in red) is shown superposed on the BC model built (in green) and deposited before August 2000. As only the helices were modeled, the loops in the structure are also not shown.

We selected a superposition with a large overlap of the two retinal molecules. A shift in the structure superposition leads to a shift of three or four positions in the sequence alignment. Shifting the structure superposition up or down by one entire helical turn does not improve the alignments. Therefore, the subjective nature of the superposition does not influence our conclusions. We believe that all GPCR models (including our own) that are based on the bacteriorhodopsin template are poor, and none can have made a positive contribution to rational

drug design projects, other than that common knowledge was confirmed (which is not surprising, as the models were first made to agree with that common knowledge).

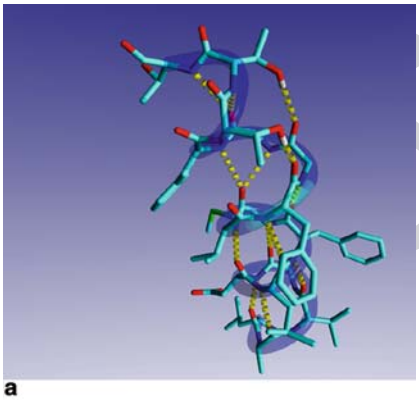
Sequence alignments extracted from deposited GPCR models revealed that we could publish models that had residues misaligned by as many as ten positions. We think this holds a warning for the future.

3.2 How Could This Happen?

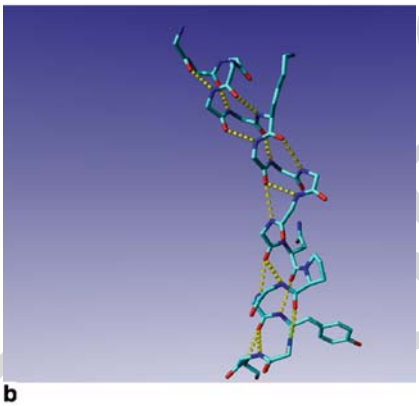
An extensive discussion of BC models can be found in the article section of the GPCRDB (Horn et al. 1998). None of the BC modelers had located the IV–V hairpin correctly between the helices. While all were aware of the Cys315–Cys480 disulphide bridge, which firmly anchors this loop to the top of TM III, all modelers knew that loop IV–V had to be external. Often bizarre reasoning was used to reconcile these two contradicting claims and to justify the position of helix III. The experimental data enabling the correct prediction of the IV–V hairpin location was all the while available to the BC modelers. It also was known that in opsins His474 and Lys477 in this hairpin form a chloride-binding site that regulates the optimal absorption wavelength of the retinal (Wang et al. 1993). A reasonable conclusion from this is that since this site modifies the wavelength, it should be located near the retinal. Unfortunately, the common knowledge that the loops stick out into the solvent overcame the experimental and *in silico* (Kuipers et al. 1996) data about the chloride site. This provides a strong lesson for the future: models must explain all available data. If certain data seem untrustworthy, think twice. Most likely you do not trust those data only because it disagrees with your model.

Another problem that seriously hampered the quality of BC models are the massive irregularities in the transmembrane helices. Figure 3 shows some individual transmembrane helices. Helix II, for example, contains an α-bulge, i.e., one residue pair has a hydrogen-bonding pattern as if they are in a so-called π-helix. It is by no means certain that this is reproduced in other GPCRs; it may well be rhodopsin-specific (Bywater 2005).

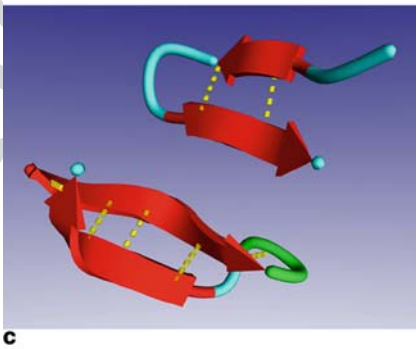
The bovine rhodopsin structures provided us with a large number of structural surprises. Common knowledge had it that π helices (if it



a



b



c

Fig. 3a–c. Irregularities in bovine rhodopsin helices. **a** The α-bulge in helix II. It can be clearly seen that the backbone C=O oxygen of the leucine located one turn after the aspartic acid of the conserved, well-known, LXXXD motif, forms hydrogen bonds with two backbone N–H protons, while all other backbone hydrogen bond donors and acceptors are satisfied. **b** The 3_{10} helix in the middle of helix VII. The lysine at the *top right* is the lysine that binds the retinal. The backbone C=O of residue one turn below this lysine forms hydrogen bonds with two backbone N–H protons, while three backbone C=O groups and one backbone N–H group are not involved in hydrogen bonds. **c** The two β-hairpins near the *top* (extracellular side) of the structure. The two β-hairpins in the N-terminal arm and the loop IV–V are just 1 Å too far away from each other to form one contiguous sheet. It is unclear how uncommon such a separation is, but it certainly came very unexpectedly to the modeling community

involves one residue only, we had better call it an α bulge) are rare. In the excellent course "Principles of Protein Structure Using the Internet" (Johansson 1999) we find an extensive explanation of why this should be a rare event (see Box 4).

Box 4 Why the π helix (α bulge) is rare (Johansson 2006)

The π helix is an extremely rare secondary structural element in proteins. Hydrogen bonds within a π helix display a repeating pattern in which the backbone C=O of residue i hydrogen bonds to the backbone HN of residue i+5. Like the 3_{10} helix, one turn of the πhelix is sometimes found at the ends of regular α helices but π helices longer than a few i, i+5 hydrogen bonds are not found. The infrequency of this particular form of secondary structure stems from the following properties:

The Φ and Ψ angles of the pure π helix (–57.1, –69.7) lie at the very edge of an allowed, minimum energy region of the Ramachandran (Φ,Ψ) map.

The π helix requires that the angle τ (N–CA–C′) be larger (114.9) than the standard tetrahedral angle of 109.5°.

The large radius of the π helix means the polypeptide backbone is no longer in van der Waals contact across the helical axis forming an axial hole too smallfor solvent water to fill.

Side chains are more staggered than the ideal 3_{10} helix but not as well as the α helix.

Obviously there are big differences between an α bulge and a π helix, but while predicting the unpredictable, such fine details are easily overseen.

3.3 The Quality of AD Models

We were surprised to find many modeling studies performed after the release of the bovine rhodopsin three-dimensional coordinates into which very little knowledge of this template was incorporated. Ballesteros et al. (2001) wrote that amine receptors can be modeled from the bovine rhodopsin template. They neglect the IV–V hairpin, crystal contacts, and the fact that many residues cannot be detected in the X-ray structure. Orry and Wallace (2000) docked endothelin in an endothelin receptor model based on a rhodopsin model by Pogozheva (Pogozheva et al. 1997). The authors write in a note added after submission that the bovine rhodopsin structure became available after the paper was submitted, and claim that their model and the bovine rhodopsin structure are similar. Their model is not deposited, but from the figures in the article, it can be seen that the endothelin molecule is docked where one would expect the IV–V β hairpin and that this hairpin is modeled as a hyperexposed loop. These are just two of the many examples of neglect of details of the bovine rhodopsin structure. A survey of recent GPCR modeling-related literature revealed a series of flaws (see Box 5).

Box 5 Flaws detected in AD models

Total neglect of loops, especially the IV–V β-hairpin (Lopez-Rodriguez et al. 2001, 2002; Shim et al. 2003).

Modeling loop-based data for individual loops obtained from NMR experiments or from sequence similarity with another PDB file (Lequin et al. 2002; Chung et al. 2002; Yang et al. 2003; Mehler et al. 2002).

Molecular dynamics (MD) compacted the IV–V β hairpin (Pellegrini et al. 2001).

Models based on a frog electron density map (Church et al. 2002). It is regrettable that an MD publication on a homology model can be accepted for publication when the author has failed to show what the same protocol does to the bovine rhodopsin structure.

3.4 AD GPCR Modeling

The availability of the bovine rhodopsin structure opens new alleys for modeling GPCRs. Box 6, however, lists some warnings for would-be AD modelers.

Box 6 Warnings for Would-Be AD Modelers

The bovine rhodopsin structure is the inactive form of the protein, while the active form is a much more appropriate modeling goal for pharmaceutical purposes.

The rhodopsin crystal structure is an antiparallel dimer, whereas GPCR dimers must be parallel.

It is far from certain that the bovine rhodopsin structure can be used as a template for all GPCRs, because sequence analysis indicates that opsins differ very much from the pharmaceutically interesting (Class A) GPCRs.

Modeling studies start with a sequence alignment between the bovine rhodopsin template and the GPCR model sequence. The percentage sequence identity between bovine rhodopsin and other (Class A) GPCRs can be as low as 20%.

Normally, when the sequence identity between the model and the template falls below 30%, the sequence alignment is the main bottleneck in the homology modeling procedure. Class A GPCRs might be an exception to this rule, because each helix contains one or two highly conserved residues that allow for an unambiguous alignment.

The observed structure of many loops seems to be determined by crystal contacts.

It is difficult to model the loops by homology, because most cytosolic loops cannot be seen in an electron density map, and most observed extracellular loop structures are probably induced by crystal packing forces. In any case, the sequence identity between most GPCRs and bovine rhodopsin is too low to derive any reliable loop alignment.

Several of the irregularities in the rhodopsin transmembrane helices seem rhodopsin-specific, whereas others seem more generic. It is not clear how to unambiguously decide which irregularities can be carried over from the rhodopsin template to, for example, an amine receptor model.

The bovine rhodopsin structure, combined with extensive sequence analyses (Horn et al. 1998), science philosophy, and all what we learned from the above, however, provides a series of hints (see Box 7).

Box 7 Hints for Would-Be AD Modelers

At three locations, however, features can be seen that give hope for modeling. These are the highly conserved:

- Trp280 and Gly295 in loop II–III.
- Loop IV–V and the Cys315–Cys480 disulphide bridge.
- Tyr734 at the bend between the helices VII and VIII and the adjacent sequence motif Phe800, Arg/Lys801 in helix VIII.
- The associated WWW pages (articles section of the GPCRDB) lists many special positions.

3.5 The Active Form

Modeling the active form of Class A GPCRs depends critically on the hypothesized mechanism of that activation process (Gouldson et al. 2004). We therefore start with a summary of possible activation mechanisms. These activation models consist of essentially the same three general steps that are shown in Box 8.

Box 8 The three steps of the activation process

Entry of the ligand into the ligand binding pocket.
The receptor moving from the inactive state into the active state, or the active state being frozen by the ligand.
The G-protein being activated, or the activated state being frozen.

The clearest lesson to be learned from the BC experience is that molecular dynamics technology has not reached a level of maturity needed to aid with the prediction of the differences between the active and the inactive state. Not only is there a need for improved force fields that *inter alia* should reflect the membrane environment, but there are only few attempts to model the solvent (lipid bilayer with water above and below it) and none of these take account of, for example, the fact that the two leaflets of the bilayer are mutually asymmetric in character.

The low-resolution structure of light-activated bovine rhodopsin suggests that an outward motion of helices V and VI might be the major difference relative to the dark state (Szundi et al. 2006). Unfortunately, Murphy took care that those helices V and VI, in Schertler's crystal

form (Schertler 2005), are involved in crystal contacts so that definitive conclusions about this motion cannot be drawn. The idea that especially the cytosolic half of helix VI moves a great deal upon GPCR activation would be in agreement with the results of our sequence analyses.

3.6 New Rules to Replace the Old Dogmas

Even if it may be deemed desirable to model loops (e.g., for studying the interactions between cytosolic loops and G-proteins, or between external loops and large peptide ligands) we would advise against this. There are no accurate template structures and precious little homology. The work by Yeagle et al. (1997, 1995, 2000) makes clear that determination of the structure of the loops independently from the rest of the molecule is not successful. Paradoxically, it may seem, it is harder to model oligopeptides than folded proteins, and this is because the former have no tertiary structure, and secondary structure is either absent or hard to predict.

Nevertheless, for most purposes (e.g., ligand design), it will be enough to model the seven transmembrane helices and the IV–V hairpin. The alignment of the helices should be based on the conserved motifs. Extrapolating from the performance of GPCR modelers over the years, we can only advise sticking to the bovine rhodopsin helix backbone coordinates. Any attempt to improve this for other GPCRs will undoubtedly make things worse rather than better. This unfortunately contradicts the earlier remarks that some of the helix irregularities might be rhodopsin-specific (Bywater 2005). The IV–V hairpin should be modeled from bovine rhodopsin. If this loop is not present in the model sequence, it seems doubtful that a reliable model structure can be built.

The bovine rhodopsin three-dimensional coordinates represent the inactive form of this receptor. To model the (pharmaceutically much more interesting) active form of GPCRs, one should not rely on molecular dynamics, but rather on the outcome of experiments that can be interpreted unambiguously (Gouldson et al. 2004). Molecular dynamics simulations might fill in the details once the relevant motions have been determined experimentally.

Modeling GPCRs based on their homology to rhodopsin is seriously hampered by the fact that we cannot predict well if the irregularities

in the rhodopsin helices are rhodopsin-specific, or general GPCR or Class A GPCR features. Modeling is further hampered by the low level of identity between the sequences of the rhodopsin template and the other (Class A) receptors. However, it is commonly observed in homology modeling projects that the areas in and around active sites and other binding motifs are better conserved than the rest of the protein, and it can safely be assumed that this will also be the case for GPCRs. In The Class A GPCRs, the so-called rhodopsin-like family, the active site (G-protein binding) and the regulatory site (ligand binding) are both located in the transmembrane helix bundle. Consequently, we expect the conserved residues and the residues that are conserved at the subfamily level to reside in these transmembrane helices. Modeling only the bundle of seven helices, loop IV–V, and perhaps helix VIII should therefore suffice to arrive at a model that, although certainly wrong in most of its details, will at least be useful for designing experiments, especially mutagenesis experiments, which provide valuable feedback in terms of structure.

Experimental data clearly show the function of several residues. For example, the arginine at position 340 is the crucial switch in G-protein coupling, the tyrosine at position 528 is involved in G-protein binding, etc. Having coordinates available for just one GPCR, we have to resort to sequence-based techniques to obtain information about the location and role of residues for which the experimental function determination is less trivial. Correlated mutation analyses [CMA (Oliveira et al. 1993)] and entropy variability analyses [EVA (Oliveira et al. 2003; Laerte Oliveira et al. 2006)] have revealed several types of residue positions that show recognizable conservation/variability patterns. Obviously, knowledge about these conservation patterns aids in the alignment of GPCR sequences to the rhodopsin structure. These conservation patterns range from very conserved in and around the active site where they have a role in maintaining the right structure and mobility required for the function, up to near maximal variability at positions where only very weak evolutionary pressures work in only a small subset of the sequences.

In practice, the most conserved residues form the active site, which for GPCRs is the G-protein-binding site. Residues known to be involved in G-protein binding must be aligned in the multiple sequence align-

ment. The residues involved in maintaining the structure of the active site are less conserved than the active site residues themselves, but are still much more conserved than all other residues. The next group of yet lesser conserved residues are the core residues that form critical contacts between secondary structure elements. At these positions, one tends to observe mainly intermediately large residues that normally are hydrophobic, but in GPCRs also can be polar, or sometimes even charged. Several of these residues are involved in signal transduction through the receptor and show a corresponding EVA pattern. Occasionally, cysteines in a bridge are so conserved that automatic methods might see them as functionally important. But in general, residues that are only involved in maintaining the structural integrity of the molecule tend to be less conserved than active site residues. We therefore speculate that the cys–cys bridge between the external end of helix III and loop IV–V has a functional role, presumably in the regulation of ligand entry. Obviously, the residues facing the lipid membrane are in majority hydrophobic. Nonhydrophobic outward-pointing residues most likely have a functional role, perhaps in dimer formation.

We can draw the general conclusion that modeling is possible for all residue positions with recognizable conservation patterns, or with clearly recognizable variability patterns. Clearly, most opsins can be modeled from the rhodopsin template over nearly the full length. It also seems likely that the cytosolic halves of most Class A GPCRs can be modeled as well. It is at this moment still a matter of modeler's religion rather than exact science whether one models the whole helices or not. It is of crucial importance for pharmaceutical purposes to model the loop IV–V. This will be difficult in all cases where the loops have a different length from this loop in rhodopsin and in all cases where the sequence is very different from rhodopsin. Obviously, the cysteine in this loop IV–V must be aligned with the bridged cysteine in rhodopsin. Several sequences (e.g., melanocortin, mas, cannabinoid, and a few more) do not have this loop, which makes it highly unlikely that their helical organization will be anywhere similar to that of the rhodopsin helices. In a few cases, loop IV–V has several cysteines, which in many cases will require experimental determination of the right cysteine to align for the cys–cys bridge. In a few classes of GPCRs, a highly conserved proline is found at position 348. In case it is desired that helix VIII be

part of the model, one can use the highly conserved motif F(Y,L)810-R(K)811 that is present in rhodopsin and in several other GPCR classes.

So, in summary, we can say that all BC GPCR models were poor, and all AD GPCR models will not be up to CASP-competition (Protein Structure Prediction Center 2006) standards. We have given the beginning of a new GPCR modeling recipe. One day, hopefully soon (!) this recipe will be proven wrong, but it is the best we can do given current data and Ockham's razor. We can, however, draw some hope from the (paraphrased) quote of Bax that "All GPCR models are wrong, but sometimes these models can be useful." And when used with care, GPCR models are often a powerful tool to aid us with the design of experiments that can shed light on the sequence-structure-function-human health relation of this intriguing class of molecules.

Acknowledgements. The GPCRDB was initiated as an EC sponsored project (PL 950224). GV acknowledges financial support from BioRange and BioSapiens. The BioSapiens project is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health," contract number LSHG-CT-2003–503265. BioRange is a programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

References

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2005) The universal protein resource (UniProt). Nucleic Acids Res 33:D154–D159

Baldwin JM (1993) The probable arrangement of the helices in G protein-coupled receptors. EMBO J 12:1693–1703

Ballesteros JA, Weinstein H (1995) Integrated methods for modeling G-protein coupled receptors. Methods Neurosci 25:366–428

Ballesteros JA, Shi L, Javitch JA (2001) Structural mimicry in G protein-coupled receptors: implications of the high-resolution structure of rhodopsin for structure–function analysis of rhodopsin-like receptors. Mol Pharmacol 60: 1–19

Beukers MW, Kristiansen I, Ijzerman AP, Edvardsen O (1999) TinyGRAP database: a bioinformatics tool to mine G protein-coupled receptor mutant data. TiPS 1999 20:475–477

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14:988–995

Bramblett RD, Panu AM, Ballesteros JA, Reggio PH (1995) Construction of a 3D model of the cannabinoid CB1 receptor: determination of helix ends and helix orientation. Life Sci 56:1971–1982

Bywater RP (2005) Location and nature of the residues important for ligand recognition in G Protein-coupled receptors. J Mol Recognit 18:60–72

Campagne F, Bettler E, Vriend G, Weinstein H (2003) Batch mode generation of residue-based diagrams of proteins. Bioinformatics 19:1854–1855

Chung DA, Zuiderweg ER, Fowler CB, Soyer OS, Mosberg HI, Neubig RR (2002) NMR structure of the second intracellular loop of the alpha 2A adrenergic receptor: evidence for a novel cytoplasmic helix. Biochemistry 41:3596–3604

Church WB, Jones KA, Kuiper DA, Shine J, Iismaa TP (2002) Molecular modelling and site-directed mutagenesis of human GALR1 galanin receptor defines determinants of receptor subtype specificity. Protein Eng 5:313–323

Cochrane G, Adelbert P, Althorpe N et al. (2006) EMBL Nucleotide Sequence Database: developments in 2005. Nucleic Acids Res 34:D10–D15

Cronet P, Sander C, Vriend G (1993) Modelling of transmembrane seven helix bundles. Protein Eng 6:59–64

Donnelly D, Overington JP, Ruffle SV, Nugent JH, Blundell TL (1993) Modelling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. Protein Sci 2:55–70

Filizola M, Perez JJ, Carteni-Farina M (1998) BUNDLE: a program for building the transmembrane domains of G protein-coupled receptors. J Comput Aided Mol Des 12:111–118

Gether U, Kobilka BK (1998) G Protein receptor activation: II. Mechanism of agonist activation. J Biol Chem 273:17979–17982

Gouldson PR, Kidley NJ, Bywater RP, Psaroudakis G, Brooks HD, Diaz C, Shire D, Reynolds CA (2004) Toward the active conformations of rhodopsin and beta-2-adrenergic receptor. Proteins 56:67–84

Henderson R, Schertler GFX (1990) The structure of bacteriorhodopsin and its relevance to the visual opsins and other seven-helix G protein-coupled receptors. Philos Trans R Soc Lond B Biol Sci 326:379–389

Herzyk P, Hubbard RE (1998) Combined biophysical and biochemical information confirms arrangement of transmembrane helices visible from the three-dimensional map of frog rhodopsin. J Mol Biol 281:741–754

Horn F, Weare J, Beukers MW, Horsch S, Bairoch A, Chen W, Edvardsen O, Campagne F, Vriend G (1998) GPCRDB: an information system for G protein-coupled receptors. Nucleic Acids Res 26:275–279

Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G (2003) GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res 31:294–297

Horn F, Lau AL, Cohen FE (2004) Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics 20:557–568

Javitch JA, Ballesteros JA, Weinstein H, Chen J (1998) A cluster of aromatic residues in the sixth membrane-spanning segment of the D2 receptor is accessible to the binding-site crevice. Biochemistry 37:998–1006

Johansson K (1999) Bioinformatics practical. http://alpha2.bmc.uu.se/~kenth/bioinfo/structure/secondary/08.html. Cited 24 November 2006

Kuipers W, Van Wijngaarden I, Ijzerman AP (1994) A model of the serotonin 5-HT1A receptor: agonist and antagonist binding sites. Drug Des Discov 11:231–249

Kuipers W, Oliveira L, Paiva ACM, Rippmann F, Sander C, Vriend G, Ijzerman AP (1996) Sequence-function correlation in G protein-coupled receptors. In: Findlay JBC (ed) Membrane protein models. BIOS Scientific, Oxford

Kuipers W, Oliveira L, Vriend G, IJzerman AP (1997) Identification of class-determining residues in G protein-coupled receptors by sequence analysis. Receptors Channels 5:159–174

Lequin O, Bolbach G, Frank F, Convert O, Girault-Lagrange S, Chassaing G, Lavielle S, Sagan S (2002) Involvement of the second extracellular loop (E2) of the neurokinin-1 receptor in the binding of substance P. Photoaffinity labeling and modeling studies. J Biol Chem 277:22386–22394

Lopez-Rodriguez ML, Murcia M, Benhamu B, Olivella M, Campillo M, Pardo L (2001) Computational model of the complex between GR113808 and the 5-HT4 receptor guided by site-directed mutagenesis and the crystal structure of rhodopsin. J Comput Aided Mol Des 15:1025–1033

Lopez-Rodriguez ML, Vicente B, Deupi X, Barrondo S, Olivella M, Morcillo MJ, Behamu B, Ballesteros JA, Salles J, Pardo L (2002) Design, synthesis and pharmacological evaluation of 5-hydroxytryptamine(1a) receptor ligands to explore the three-dimensional structure of the receptor. Mol Pharmacol 62:15–21

Luecke H, Richter HT, Lanyi JK (1998) Proton transfer pathways in bacteriorhodopsin at 2.3 angstrom resolution. Science 280:1934–1937

Mehler EL, Periole X, Hassan SA, Weinstein H (2002) Key issues in the computational simulation of GPCR function: representation of loop domains. J Comput Aided Mol Des 16:841–853

Oliveira L, Paiva ACM, Vriend G (1993) A common motif in G protein-coupled seven transmembrane helix receptors. J Comput Aided Mol Des 7:649–658

Oliveira L, Paiva PB, Paiva AC, Vriend G (2003a) Identification of functionally conserved residues with the use of entropy-variability plots. Proteins 52:544–552

Oliveira L, Paiva PB, Paiva AC, Vriend G (2003b) Sequence analysis reveals how G protein-coupled receptors transduce the signal to the G protein. Proteins 52:553–560

Orry AJW, Wallace BA (2000) Modelling and docking the endothelin G protein-coupled receptor. Biophys J 79:3083:3094

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. Science 289:739–745

Pardo L, Ballesteros JA, Osman R, Weinstein H (1992) 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. PNAS 89:4009-4012

Pebay-Peyroula E, Rummel G, Rosenbusch JP, Landau EM (1997) X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipid cubic phases. Science 277:1676–1681

Pellegrini M, Bremer AA, Ulfers AL, Boyd ND, Mierke DF (2001) Molecular characterization of the substance P*neurokinin-1 receptor complex: development of an experimentally based model. J Biol Chem 276:22862–22867

Pogozheva ID, Lomize AL, Mosberg HI (1997) The transmembrane 7-alpha-bundle of rhodopsin: distance geometry calculations with hydrogen bonding constraints. Biophys J 72:1963–1985

Protein Structure Prediction Center (2006) http://predictioncenter.gc.ucdavis.edu/. Cited 24 November 2006

Prusis P, Schiöth HB, Muceniece R, Herzyk P, Afshar M, Hubbard RE, Wikberg JES (1997) Modelling of the three-dimensional structure of the human melanocortin 1 receptor, using an automated method and docking of a rigid cyclic melanocyte stimulating hormone core peptide. J Mol Graph Model 15:307–315

Rippmann F, Bottcher E (1993) Molecular modelling of serotonin receptors. 7TM 3:1–27

Schadel SA, Heck M, Maretzki D, Filipek S, Teller DC, Palczewski K, Hofmann KP (2003) Ligand channeling within a G-protein-coupled receptor. The entry and exit of retinals in native opsin. J Biol Chem 278:24896–24903

Schertler GF (2005) Structure of rhodopsin and the metarhodopsin I photointermediate. Curr Opin Struct Biol 15:408–415

Schertler GFX, Hargrave PA (1995) Projection structure of frog rhodopsin in two crystal forms. PNAS 192:11578–11582

Schertler GF, Villa C, Henderson R (1993) Projection structure of rhodopsin. Nature 362:770–772

Shi L, Javitch JA (2002) The binding site of aminergic G protein-coupled receptors: the transmembrane segments and second extracellular loop. Annu Rev Pharmacol Toxicol 42:437–467

Shim JY, Welsh WJ, Howlett AC (2003) Homology model of the CB1 cannabinoid receptor: sites critical for nonclassical cannabinoid agonist interaction. Biopolymers 71:169–189

Szundi I, Ruprecht JJ, Epps J, Villa C, Swartz TE, Lewis JW, Schertler GF, Kliger DS (2006) Rhodopsin photointermediates in two-dimensional crystals at physiological temperatures. Biochemistry 45:4974–4982

Takeda K, Sato H, Hino T, Kono M, Fukuda K, Sakurai I, Okada T, Kouyama T (1998) A novel three-dimensional crystal of bacteriorhodopsin obtained by successive fusion of the vesicular assemblies. J Mol Biol 283:463–474

Teller DC, Okada T, Behnke CA, Palczewski K, Stenkamp RE (2001) Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of G protein-coupled receptors. Biochemistry 40: 7761–7772

Unger VM, Schertler GFX (1995) Low resolution structure of bovine rhodopsin determined by electron cryo-microscopy. Biophys J 68:1776–1786

Unger VM, Hargrave PA, Baldwin JM, Schertler GFX (1997) Arrangement of rhodopsin transmembrane alpha-helices. Nature 389:203–206

Vaidehi N, Floriano WB, Trabanino R, Hall SE, Freddolino P, Choi EJ, Zamanakos G, Goddard WA 3rd (2002) Prediction of structure and function of G protein-coupled receptors. PNAS 2002 99:12622–12627

Venclovas C, Zemla A, Fidelis K, Moult J (2001) Comparison of performance in successive CASP experiments. Proteins Suppl 5:163–170

Vriend G (1990) WHAT IF: a molecular modeling and drug design program. J Mol Graph 8:52–56

Wang Z, Asenjo AB, Oprian DD (1993) Identification of the Cl-binding site in the human red and green colour vision pigments. Biochemistry 32: 2125–2130

Watson S, Arkinstall S. The G-protein linked receptor Facts Book. 1994, Academic Press Ltd, ISBN 0-12-738440-5

Yang X, Wang Z, Dong W, Ling L, Yang H, Chen R (2003) Modeling and docking of the three-dimensional structure of the human melanocortin 4 receptor. J Protein Chem 22:335–344

Yeagle PL, Alderfer JL, Albert AD (1995) Structure of the third cytoplasmic loop of bovine rhodopsin. Biochemistry 34:14621–14625

Yeagle PL, Alderfer JL, Albert AD (1996) Structure determination of the fourth cytoplasmic loop and carboxyl terminal domain of bovine rhodopsin. Mol Vis 2:12–19

Yeagle PL, Alderfer JL, Salloum AC, Ali L, Albert AD (1997) The first and second cytoplasmic loops of the G protein-receptor, rhodopsin, independently form betaturns. Biochemistry 36:3864–3869

Yeagle PL, Salloum A, Chopra A, Bhawsar N, Ali L, Kuzmanovski G, Alderfer JL, Albert AD (2000) Structures of the intradiskal loops and amino terminus of the G-protein receptor, rhodopsin. J Pept Res 55:455–465