Prediction of structure, function, and spectroscopic properties of G-protein-coupled receptors: methods and applications

Thesis by

Rene Trabanino

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2004

(Defended May 11, 2004)

Acknowledgments

First of all I would like to give the greatest thanks to my Lord and my God Jesus Christ for his inspiration and guidance throughout these years. None of this would have been possible without God, and so I would like to give glory to Him through this thesis. It is truly amazing how beautiful God's creation is and I am just grateful and humbled to have been allowed to get a glimpse of how it all works.

Next I would like to thank the kindness and mentorship of Vaidehi Nagarajan and for our discussions related not only to science but also on spirituality and its importance in everyday life. I would like to thank William Goddard for posing challenging problems and for his great insight into scientific problems. Many thanks go to the methods group of Spencer and Wely for their collaboration and discussions. I enjoyed the company and discussion, related and unrelated to science with the people in the 056C room, including Pete Huskey, Santiago Solares, John Keith, Candy Tong, Victor Kam, Jiyoung Heo, and Julius Su. Thanks to the insights, feedback, and collaborations with Joyce Peng, Shantanu Sharma, and Huazhang Shen, Yashar Kalani, and Peter Fredollino.

Thanks to my family, my parents and sister Nancy and brother Shawn for their love. To my friends from the MSTP, to my good friend and study partner Rosmery Tajiboy, my roommate Leo, and friends and brothers and sisters in Christ from the St. Joan of Arc church including Sister Virgina, Cris Rojas, Magda. And to everyone else that I am missing, you know who you are and you will always be in my heart and prayers.

Abstract

G-protein-coupled receptors are of great pharmaceutical interest, comprising the majority of targets for currently marketed drugs. The theme of my thesis is the development of the structure prediction method, MembStruk, for the superfamily of G-protein-coupled receptors. The first part of this thesis focuses on the methods and their validation. There are several steps involved in MembStruk that are detailed and tested for membrane proteins with known structures in the first few chapters (Chapters 2-6). Specifically, the first principles methods for predicting the transmembrane helical ranges and the helix hydrophobic centers are tested. The program for predicting the transmembrane helical ranges, TM2ndS, ranks in the top two when comparing performance with other top prediction methods. And because it is based on general principles, it can be applied robustly for membrane protein families for which little structural information is available. The simulation of the EC-II closing is also tested on bovine rhodopsin. The use of the MembStruk method on bovine rhodopsin as a validation case is presented in detail (Chapter 2). The large majority (71%) of the residues involved in binding in rhodopsin are predicted and the protein structure itself is 2.84 Å coordinate root mean square error in the transmembrane main chain atoms from the crystal structure.

The second part of the thesis discusses applications on various G-protein-coupled receptor systems. The application of the MembStruk method to other peptide chemokine G-protein-coupled receptors like CCR1 and CCR5 is discussed in Chapter 9. The fundamental scientific problems of G-protein-coupled receptor modulation of absorption and relaxation properties of a bound chromophore (retinal) are addressed and results are presented for the predictions of these properties.

The prediction of structure and function of G-protein-coupled receptors would allow for structure-based drug design and a rational approach to reducing drug cross-reactivity across receptor families.

Table of Contents

List of Figures

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 8

CHAPTER 9

List of Tables

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 7

Chapter 1: Introduction

The theme of my thesis is the development of the structure prediction method, MembStruk, for the superfamily of G-protein-coupled receptors (GPCRs). There are several steps involved in MembStruk that are detailed and tested for membrane proteins with known structures in the first few chapters. In particular, the use of the MembStruk method on bovine rhodopsin as a validation case is presented. This is followed by the application of the MembStruk method to other peptide chemokine GPCRs like CCR1 and CCR5.

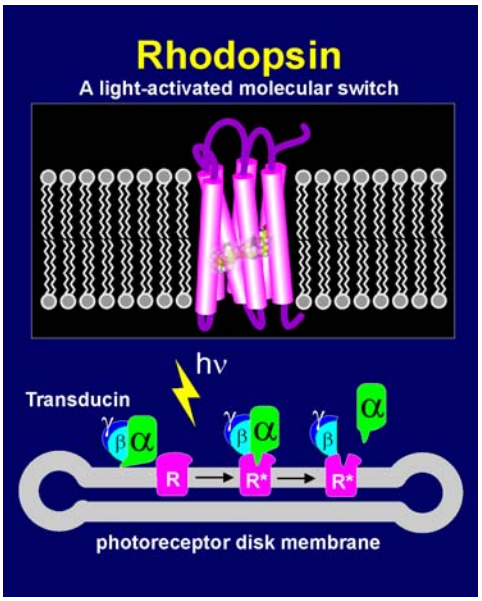


Figure 1: Schematic of rhodopsin and its interaction with the G-protein. (Figure: Christian Altenbach)

Membrane proteins make up 20-30% of genomes (Wallin et al., 1998) in various organisms and are important in various processes from ion transportation to detection of electromagnetic radiation. Within this class, the GPCR superfamily comprises about 3-4% (Schoneberg et al., 2002) of the human genome. They act by transducing an extracellular signal into an intracellular signal cascade involving G-proteins. The extracellular signal is usually chemical (peptides, lipids, neurotransmitters) but in the case of the opsin family, it is light of a specific frequency range. This is shown in Figure 1 for rhodopsin, where the particular G-protein used to propagate the signal is a heterotrimeric protein called transducin.

In the case of rhodopsin, the bound chromophore is 11cis-retinal, as shown in Figure 2. Retinal, which is derived from ingested Vitamin A, actually absorbs light in the

UV radiation range when it is free in solution. Upon absorbing this photon, the free retinal undergoes isomerization to all-trans-retinal.

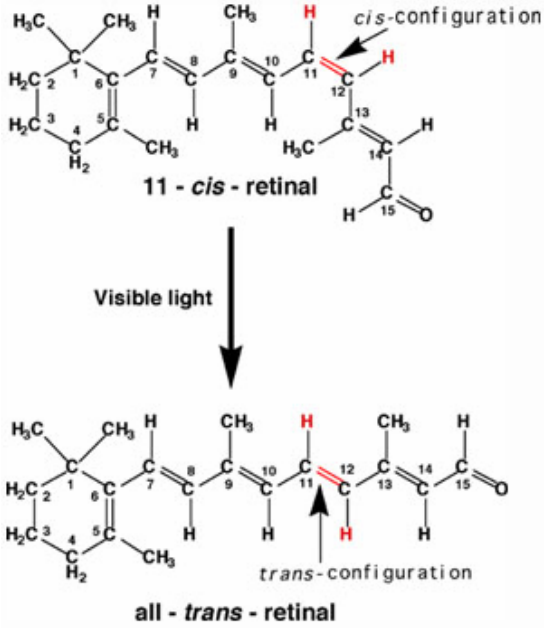


Figure 2: The photoisomerization of free retinal at the 11-cis bond. (Figure: http://wunmr.wustl.edu/EduDev/LabTutorials/Vision/Vision.html)

But this chromophore also binds to the GPCR rhodopsin and forms a Schiff base linkage with the protein, as shown in Figure 3. This bound chromophore, as opposed to its free form, absorbs light in the visible range with a maximum frequency of absorption of ~500 nm.

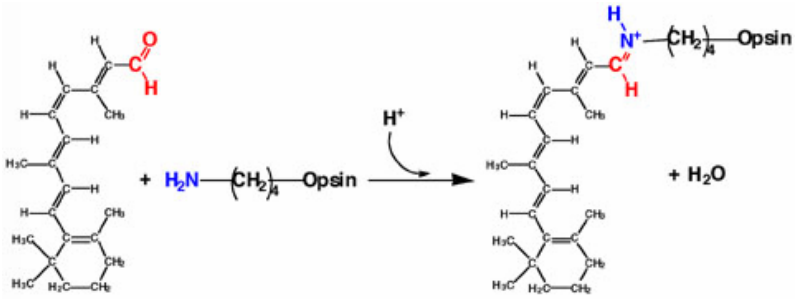


Figure 3: The Schiff base bond formation of retinal with the protein lysine. (Figure: http://wunmr.wustl.edu/EduDev/LabTutorials/Vision/Vision.html)

When bound to the protein, this chromophore undergoes isomerization (required for receptor activation) with an efficiency that is more than three times of that in its free state. In addition, within the opsin family, there are three human opsin proteins which

absorb maximally in the green, blue, and red range of visible light when bound to the same chromophore, and make possible color distinction in human vision. Thus, the roles which the GPCR plays in modulating the frequency of light absorption and isomerization efficiency are of fundamental scientific interest.

In addition to the basic scientific questions which emerge from them, GPCRs are of particular interest pharmacologically, making up more than 45% of all known drugs on the market (Horuk, 2003). For example, drugs used for the treatment of schizophrenia, high blood pressure, migraines, and ulcers target the dopamine, adrenergic, serotonin, and histamine receptors respectively.

One of the great challenges in the field of drug development for GPCRs is the reduction of cross-reactivity of drugs across GPCR families, which leads to side effects and effectively lowers the drug dosage which one may give to a patient. This is sometimes due to the similar drug libraries used to derive a lead compound for different receptor targets. It is also due to sequence similarities common to the GPCRs with which the drug is cross-reacting. An example of this is the case of BX471, which is a human CCR1 antagonist in development; it exhibits cross-reactivity with the dopamine and muscarinic receptors (Hesselgesser et al., 1998).

To aid in the drug development process, the 3D structures of the GPCRs in question would provide a rational basis for increasing the potency of a drug while reducing action on other receptors that are not of interest (known as anti-targets). For soluble proteins, the relatively large number of X-ray crystal structures available has made the use of homology modeling feasible in many cases. On the contrary, there is currently one GPCR crystal structure available for bovine rhodopsin. As such, since the

sequence identity to rhodopsin is low for most GPCRs of interest (17 % for dopamine, 14 % for serotonin), the use of homology modeling is not a feasible avenue for obtaining reliable structures for making predictions (Archer et al., 2003).

The MembStruk method (Floriano et al., 2000; Vaidehi et al., 2002; Trabanino et al., 2004) was developed by myself, Spencer Hall and Vaidehi, to predict GPCR structure using an approach very different from that of homology modeling. The MembStruk protocol determines the 3D structure of a GPCR beginning from the sequence and using mostly first principles in predicting the structure in various steps. The step which uses some crude structural information is the template building step, where the tilts of the 7Å frog rhodopsin structure (Schertler, 1998) are used to form an initial TM helical bundle. Aside from this, the TM helical extent, the translations of the helices along their axes, their rotations within the bundle, and their bends are determined from first principles. Judging from the success in predicting function (by ligand binding site and affinity determination) and direct structure (compared to crystal structure or mutagenesis studies), the use of this initial template seems to be justified (Freddolino et al., 2004; Kalani et al., 2004). Even so, currently the group is working on determining the tilts using Monte Carlo methods treating each helix as a rigid body with mesoscale forcefield interactions with adjacent helices.

This thesis is divided into two large sections, with the first section focused on the methods developed for structure prediction and bioinformatics while the second section presents applications of methods to GPCRs for the prediction of structure, function, and spectroscopic properties.

Chapter 2 of this thesis describes the MembStruk procedure in detail and provides direct structural validation for the only GPCR with a crystal structure available, bovine rhodopsin. The main chain atoms of the TM region differ by 2.87 Å CRMS (coordinate root mean square) from the crystal structure. Also, the ligand was predicted to bind in a conformation which was 2.92 Å CRMS from the crystal conformation (Figure 4). In addition, a majority (71%) of the residues interacting with the ligand in the crystal structure are predicted to interact with the ligand in the MembStruk structure. This paper was adapted from a published article (Trabanino et al., 2004).

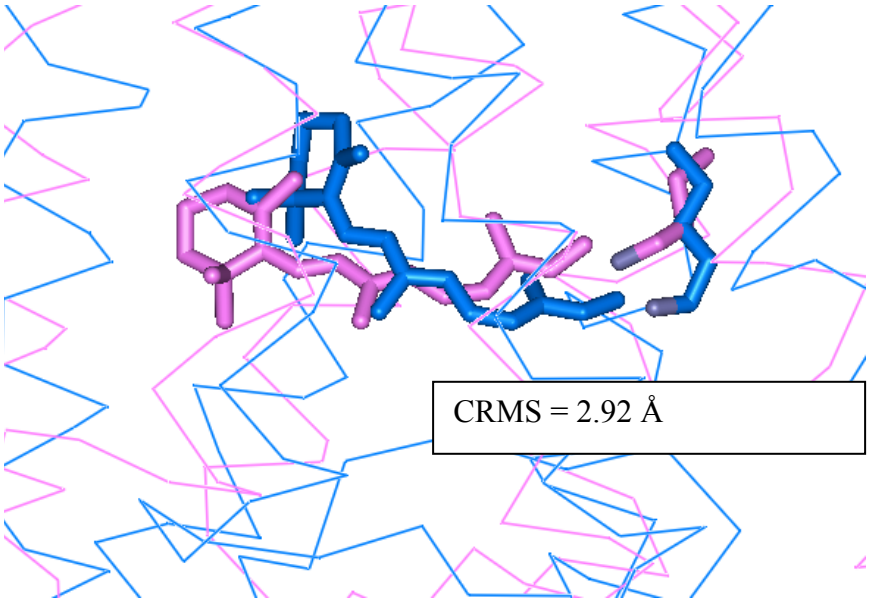


Figure 4: A comparison of the predicted (purple) and crystal (blue) conformations of bound 11cis-retinal

Chapter 3 specifically details the TM2ndS program that I developed, which predicts the TM helical regions of a membrane protein. The program is compared to all the other major prediction programs. TM2ndS emerges as one of the top two methods in performance of TM helix prediction accuracy, which is most pertinent to the GPCR focus. Of these two methods, TM2ndS is the only program which uses general principles

and does not systematically "train" its parameters for better performance on available crystal structures. Thus, it is the top hydrophobicity-based first principles methods available for high residue accuracy in predictions for membrane proteins. The details of the TM2ndS performance using redefined standards for TM helices from crystal structures are also presented.

In Chapter 4 I discuss the method for hydrophobic center prediction for membrane proteins and provides validation for its use in orienting membrane helices along their axes. The correspondence with the actual bilayer center is also presented.

Chapter 5 overviews the development of a database of TM helix and center predictions for all human GPCRs (excluding the olfactory GPCRs) in the SwissProt sequence database as well as the family relationships in order to provide an organized manner of ascertaining possible similarities between proteins which may lead to ligand cross-reactivity. In addition, the relationships between all of these GPCRs using tree diagrams of the 7 TM cores are presented. Results on the usage of TM2ndS for mining unknown GPCRs from the genome are also shown.

The extracellular loop II (EC-II) is closed over the TM barrel in the crystal structure of bovine rhodopsin. In out predicted structures, we build both the ECII loop open structure and the closed structure. Chapter 6 focuses on prediction algorithm for the closing of the ECII loop, EC_LOOP_SIM and its usage is validated in the case of bovine rhodopsin. In addition, the role of the EC-II loop in a "closed" conformation in modulating the binding of ligands is discussed in the context of a possible binding mechanism for 11cis-retinal in bovine rhodopsin.

Chapter 7 begins the section on the application of MembStruk for various GPCRs. Prediction of the structure of the three, green, red and blue colored opsins in human, and the mechanism for color distinction in humans at the molecular level are detailed in Chapter 7. The structures of the three opsin proteins responsible for color distinction are built using the crystal structure of bovine rhodopsin and the absorption spectra of the bound chromophore within the three proteins is predicted using a QM-fitted molecular dynamics together with QM/MM (quantum mechanics/ molecular mechanics) methods. The absorption maxima of the predicted spectra correspond well with the experimental maxima for the three proteins.

Chapter 8 is an extension of the previous chapter. It overviews the isomerization pathway along the potential energy surface and interprets this in terms of possible relaxation pathways for retinal after Frank-Codon excitation. In addition, the role of the protein in increasing isomerization efficiency is discussed in the context of the results.

Chapter 9 describes in detail the prediction of the structures and functions of the chemokine receptors human CCR1 (Figure 5) and CCR5 using the MembStruk protocol. This provides additional validation of the method and also provides insight into the causes of interspecies differences in binding affinity as well as manners of reducing cross selectivity of drugs to both these receptors. Specifically, the mouse and human receptors exhibit an important difference in the simulated EC-II loop (a glutamic acid in human to lysine in the mouse).

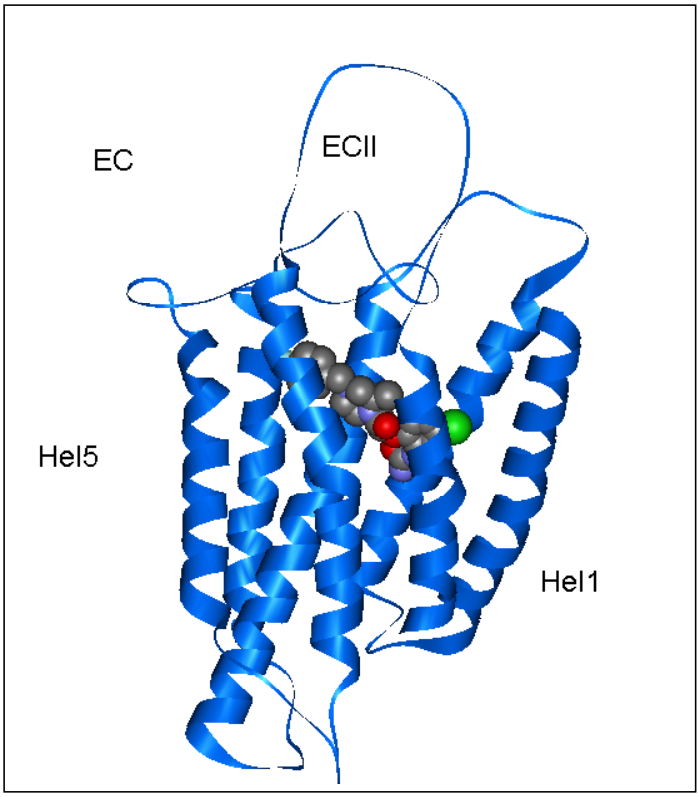


Figure 5: The predicted binding of the BX471 antagonist to human CCR1.

References

Archer, E.; Maigret, B.; Escrieut, C.; Pradayrol, L. and Fourmy, D. (2003). Rhodopsin crystal: new template yielding realistic models of G-protein-coupled receptors? *Trends Pharmacol. Sci.* 24, 36-40.

Floriano, W. B.; Vaidehi, N.; Singer, M.; Shepherd, G. and Goddard III, W. A. (2000). Molecular mechanisms underlying differential odor responses of a mouse olfactory receptor. *Proc. Natl Acad. Sci. USA.* 97, 10712-10716.

Freddolino, P. L.; Kalani, M. S. Y.; Vaidehi, N.; Floriano, W. B.; Hall, S. E.; Trabanino, R. J.; Kam, W. T. and Goddard III, W. A. (2004). Predicted 3D structure for the human beta2 adrenergic receptor and its binding site for agonists and antagonists. *Proc. Natl. Acad. Sci. USA.* 101, 2736-2741.

Hesselgesser, J.; Ng, H. P.; Liang, M.; Zheng, W.; May, K.; Bauman, J. G.; Monahan, S.; Islam, I.; Wei, G. P.; Ghannam, A.; Taub, D. D.; Rosser, M.; Snider, R. M.; Morrissey, M. M.; Perez, H. D. and Horuk, R. (1998). *J. Biol. Chem.* 273, 15687-15692.

Horuk, R. (2003). Development and evaluation of pharmacological agents targeting chemokine receptors. *Methods* 29, 369-375.

Kalani, M. Y. S.; Vaidehi, N.; Hall, S. E.; Trabanino, R. J.; Freddolino, P. L.; Kalani, M. A.; Floriano, W. B.; Kam, V. W. and Goddard III, W. A. (2004). The predicted 3D structure of the human D2 dopamine receptor and the binding site and binding affinities for agonists and antagonists. *Proc. Natl. Acad. Sci. USA.* 101, 3815-3820 (in press).

Schertler, G. F. X. (1998). Structure of rhodopsin. *Eye* 12, 504-510.

Schoneberg, T.; Schulz, A. and Gudermann, T. (2002). The structural basis of G-protein-coupled receptor function and dysfunction in human diseases. *Rev. Phys. Biochem. Pharm.* 144, 145-227.

Trabanino, R. J.; Hall, S. E.; Vaidehi, N.; Floriano, W. B.; Kam, V. W. T. and Goddard III, W. A. (2004). First principles predictions of the structure and function of G-protein-coupled receptors: validation for bovine rhodopsin. *Biophys. J.* 86, 1904-1921.

Vaidehi, N.; Floriano, W. B.; Trabanino, R.; Hall, S. E.; Freddolino, P.; Choi, E. J.; Zamanakos, G. and Goddard, W. A. (2002). Prediction of structure and function of G protein-coupled receptors. *Proc. Natl Acad. Sci. USA.* 99, 12622-12627.

Wallin, E. and von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* 7, 1029-1038.

Part 1: Methods

Chapter 2: Predictions of the structure and function of G-protein-coupled receptors: validation for bovine rhodopsin

Adapted from published article (Trabanino et al., 2004)

Abstract

G protein-coupled receptors (GPCRs) are involved in cell communication processes and with mediating such senses as vision, smell, taste, and pain. They constitute a prominent superfamily of drug targets, but atomic level structure is available for only one GPCR, bovine rhodopsin making it difficult to use structure-based methods to design receptor specific drugs. We have developed the MembStruk computational method for predicting the 3D structure of GPCRs mostly from first principles. In this paper we validate the MembStruk procedure by comparing its predictions with the high-resolution crystal structure of bovine rhodopsin. The crystal structure of bovine rhodopsin has the second extracellular (EC-II) loop closed over the transmembrane regions by making a disulfide linkage between Cys110 and Cys187, but we speculate that opening this loop may play a role in the activation process of the receptor through the cysteine linkage with helix 3. Consequently we predicted two structures for bovine rhodopsin from the primary sequence (with no input from the crystal structure), one with EC-II loop closed as in the crystal structure and the other with the EC-II loop open. The MembStruk predicted structure of bovine rhodopsin with the closed EC-II loop, deviates from the crystal by 2.84 Å CRMS (coordinate root mean square) in the transmembrane region main chain atoms.

The predicted three-dimensional structures for other GPCRs can be validated only by predicting binding sites and energies for various ligands. For such predictions we developed the HierDock first-principles computational method. We validate HierDock by predicting the binding site of 11cis-retinal in the crystal structure of bovine

rhodopsin. Scanning the whole protein without using any prior knowledge of the binding site, we find that the best scoring conformation in rhodopsin is 1.1 Å CRMS from the crystal structure for the ligand atoms. This predicted conformation has the carbonyl O just 2.82 Å from the N of Lys 296. Making this Schiff base bond and minimizing leads to a final conformation only 0.62 Å CRMS from the crystal structure.

We also used HierDock to predict the binding site of 11cis-retinal in the MembStruk predicted structure of bovine rhodopsin (closed loop). Scanning the whole protein structure, leads to a structure in which the carbonyl O is just 2.85 Å from the N of Lys 296. Making this Schiff base bond and minimizing leads to a final conformation only 2.92 Å CRMS from the crystal structure.

The good agreement of the *ab initio* predicted protein structures and ligand binding site with experiment validates the use of the MembStruk and HierDock methods. Since these methods are generic and applicable to any GPCR, they should be useful in predicting the structures of other GPCRs and the binding site of ligands to these proteins.

Introduction

Integral membrane proteins comprise 20-30% of genes (Wallin et al., 1998) in humans and other higher forms of life, playing an important role in processes as diverse as ion translocation, electron transfer, and transduction of extracellular signals. One the most important classes of transmembrane (TM) proteins is the *G-protein-coupled receptor* (GPCR) superfamily which upon activation by extracellular signals initiate an intracellular chemical signal cascade to transduce, propagate, and amplify these signals. GPCRs are involved in cell communication processes and in mediating such senses as vision, smell, taste, and pain. The extracellular signals inciting this transduction are usually chemical, but for the opsin family, it is "visible" light (electromagnetic radiation). Malfunctions in GPCRs play a role in such diseases as ulcers, allergies, migraine, anxiety, psychosis, nocturnal heartburn, hypertension, asthma, prostatic hypertrophy, congestive heart failure, Parkinson's, schizophrenia, and glaucoma (Wilson et al., 2000). Indeed although they comprise about 3-4% (Schöneberg et al., 2002) of the human genome, the GPCR superfamily represents one of the most important families of drug targets.

Within a class of GPCRs (for example, adrenergic receptors) there are often several subtypes (for example, 10 for adrenergic receptors) all responding to the same endogenous ligand (epinephrine and norepinephrine for adrenergic receptors), but having very different functions in various cells. In addition, many different types of GPCRs are similar enough that they are affected by the antagonists or agonists for other types (e.g., among adrenergic, dopamine, serotonin, and histamine receptors), leading often to undesirable side effects. This makes it difficult to develop drugs to a particular subtype

without side effects resulting from cross reactivity to other subtypes. To design such subtype specific drugs, it is essential to use structure-based methods but this has not possible because *there is no atomic level structure available for any human GPCR*. Consequently design of subtype specific drugs for GPCR targets is a very tedious empirical process, often leading to drugs with undesirable side effects. The difficulty in obtaining 3D structures for GPCRs is obtaining high quality crystals of these membrane-bound proteins sufficient to obtain high-resolution x-ray diffraction data and the difficulty of using NMR to determine structure on such membrane bound systems. Hence we conclude that to aid the structure-based drug design for GPCR targets, it is essential to develop theoretical methods adequate to predict the 3D structures of GPCRs from mostly first principles. For globular proteins there have been significant advance in predicting the 3D structures by using sequence homologies to family of known structures (Marti-Renom et al., 2000); however, this is not practical for GPCRs since a high-resolution crystal structure is available for one GPCR, bovine rhodopsin which has low homology (<35%) to most GPCR's of pharma interest.

Consequently we have been developing the MembStruk method for prediction of 3D structures for GPCRs from primary sequence without using homology modeling. MembStruk is based on the organizing principle provided by knowing that a GPCR has a single chain with seven helical TM domains threading through the membrane, which we find to provide sufficient structural information when combined with atomistic simulations (molecular dynamics and Monte Carlo) to deduce 3D structures for GPCRs adequate to predict the binding site and relative binding energy of agonists and antagonist. We have been applying MembStruk to several GPCRs, where the validation

has been based on the comparison of the predicting binding site to experimental binding and mutation data. In this paper we describe the details of the MembStruk method and validate the accuracy of the predictions by comparing with the only high-resolution crystal structure available for a GPCR, bovine rhodopsin.

Because the function of a GPCR is to signal to the interior of the cell in the presence of a particular ligand bound to the extracellular surface, it is most relevant to determine the 3D structure for the conformation of the protein involved in activating G-protein. It is widely thought that there are two distinct conformations of GPCRs, one active and one inactive, in equlibrium, even in the absence of ligands (Melia et al., 1997; Strange 1998; Schöneberg et al., 2002). This equilibrium is shifted when a ligand binds to the GPCR. Thus it would be valuable to know four structures of the protein:

- the apo-protein in both the active and inactive forms and
- the ligand-bound form in both the active and inactive forms.

so that one could study the process of GPCR activation. Even for bovine rhodopsin, there is crystal structure data for only one of these four (the ligand-bound inactive form). We postulate in this paper a model of activation involving the EC-II loop and TM3 in which the structure is assumed

- to be in the active form, when the EC-II loop is ”open” and
- to be in the inactive form when the EC-II loop is “closed.”

It is the closed conformation that is observed in the rhodopsin crystal structure (Palczewski et al., 2000; Okada, et al., 2001). In this paper we report the MembStruk predicted structures for all four structures, although comparison can be made directly to experiment only for the closed loop with ligand case.

Except for bovine rhodopsin the only experimental validation for the accuracy of predicted GPCR structures must rest on predicting the binding sites and energies for various ligands and how they are modified by various mutations. To make such predictions from first principles, we developed the HierDock method, which we validate here by predicting the binding site of retinal in bovine rhodopsin both for the experimental 3D structure and for the predicted structures (open and closed loop).

The first report on MembStruk and HierDock (Floriano et al., 2000; Vaidehi et al., 2002) focused on olfactory receptors (OR), where ligand-binding data was available for 24 simple organic molecules to 14 different ORs (Malnic et al., 1999). More recently these methods have been applied to predict the structures and function for GPCRs of such diverse subfamilies as: β1- and β2-adrenergic receptor, dopamine D2 receptor, endothelial differentiation gene (EDG) 6, sweet gustatory, and olfactory receptors (Vaidehi et al., 2002; Freddolino et al., 2003 in review; Kalani et al., 2003 communicated; Floriano et al., 2003 communicated). The HierDock technique has also been validated for globular proteins where the crystal structures are available (Wang et al., 2002; Datta et al., 2002; Datta et al., 2003; Kekenes-Huskey et al., 2003 accepted; Floriano et al., 2003 accepted). We find that the predicted structures of the adrenergic and dopamine receptors lead to binding sites for the endogenous ligands in excellent agreement with the plentiful mutation and binding experiments. Similarly the predicted binding sites and affinities for EDG 6, mouse I7 and rat I7 olfactory receptors, and human sweet receptor are consistent with the available experimental binding data.

However a quantitative assessment of the accuracy of these structure and function prediction methods can be made only for bovine rhodopsin, for which there is a high-

resolution experimental crystal structure available with ligand attached to the protein. Thus this paper provides a detailed study of rhodopsin to validate the various steps involved in our procedures for prediction of the 3D structures of GPCRs (MembStruk) and for the prediction of the binding site and the binding energy of the retinal ligand to bovine rhodopsin (HierDock).

Section 2 gives the details of the MembStruk and HierDock protocols, while Section 3 describes the results of structure and function prediction for bovine rhodopsin. These results are discussed in Section 4 followed by conclusions in section 5.

Computational methods

Force fields (FF)

All calculations for the protein used the DREIDING FF (Mayo et al., 1990) with charges from CHARMM22 (MacKerell et al., 1998) unless specified otherwise. The non-bond interactions were calculated using Cell Multipole Method (Ding et al., 1992) in MPSim (Lim et al., 1997).

The ligands were described with the DREIDING FF (Mayo et al., 1990) using charges from quantum mechanics (QM) calculations on the isolated ligand [ESP charges calculated using Jaguar (Jaguar,v4.0)]. For the lipids we used the DREIDING FF with QEq charges (Rappé et al., 1991). Some calculations were done in the vacuum (e.g., final optimization of receptor structure to approximate the low dielectric membrane environment). For structural optimization in the solvent (water) we used the Analytical Volume Generalized Born (AVGB) (Zamanakos, 2001 Caltech Chemistry Thesis) approximation to Poisson-Boltzmann (PB) continuum solvation (PB).

We use the Dreiding FF due to its generic applicability to all molecules constructed from main group elements (particularly all organics) since we will use our methods to predict the binding site and energy for a diverse set of ligands of interest to pharma. Indeed we find below that the minimized structure for bovine rhodopsin deviates from the crystal structure by only 0.29 Å, CRMS. The Dreiding FF with CHARMM22 charges has been validated for molecular dynamics simulations and binding energy calculations for many proteins (Brameld et al., 1999; Datta et al., 2003; Wintrode et al., 2003; Datta et al., 2002; Wang et al., 2002; Kekenes-Huskey et al., 2003; Floriano et al., 2003) with similar accuracy.

Validation of the force fields: The crystal structure of bovine rhodopsin (resolution 2.80 Å) was downloaded from protein database (pdb entry 1F88). The Hg ions, sugars, and waters were deleted from this structure. This crystal structure is missing 10 complete residues in loop regions and the side chain atoms for 15 additional residues. We added the missing residues and side chains using WhatIf (Vriend et al., 1990). We then fixed the TM helices and minimized (using conjugate gradients) the structure of the loop region to an RMS force of 0.1 kcal/mol/Å. Then we added hydrogens to all the residues using the PolyGraf software. The potential energy of the entire structure of rhodopsin was then minimized (using conjugate gradients) to an RMS force of 0.1 kcal/mol/Å . This minimized structure deviates from the x-ray crystal structure by 0.29 Å coordinate root mean square (CRMS) error over all atoms in the crystal structure. This is within the resolution of the crystal structure, validating the accuracy of the FF and the charges. This FF minimized crystal structure is denoted as ***ret(x)/closed(xray).***

The MembStruk protocol for predicting structure of GPCRs

MembStruk uses the hydrophobic profile of multisequence alignment of GPCRs to assign the helical TM regions. This is combined with a series of steps of Monte Carlo like systematic search algorithm to optimize the rotation and translational orientation of the TM helices. This search algorithm allows the structure to get over barriers and make the conformational search more comprehensive. This is followed by molecular dynamics (MD) calculations at a variety of coarse-grain to fine-grain levels in explicit lipid bilayer.

MembStruk was first described in Floriano et al. 2000. This method (now labeled as MembStruk 1.0) was improved to include energy optimization to determine the rotation of helices in the seven helical TM bundle in Vaidehi et al. 2002, now referred to as MembStruk2.0. In the current paper we have modified MembStruk (now denoted as MembStruk3.5) to also include optimization of the helix translations along their axes and rotational optimization using hydrophobic moment of the helices. The MembStruk3.5 procedure for predicting structures of GPCRs consists of the following steps:

1. Prediction of TM regions from analysis of the primary sequence
2. Assembly and coarse-grain optimization of the seven helix TM bundle
3. Optimization of individual helices
4. Rigid body dynamics of the helical bundle in a lipid bilayer
5. Addition of inter-helical loops and optimization of the full structure.

Henceforth in this paper any reference to MembStruk always refers to MembStruk3.5 unless mentioned otherwise. We will next discuss some of the details of these steps in MembStruk. We should emphasize here that these steps are all automated into a single MembStruk procedure. Thus the sequence is fed to MembStruk and the result at the end

is a final 3D structure for the protein in the lipid bilayer. A screenshot of the graphical user interface for part of this program is shown in Figure S5. Of course we also examine the various intermediate results generated in this procedure to allow us to detect problems, to gain insight into the validity of the various criteria, and to provide hints on improvements to make in the methods.

Step 1: Prediction of TM regions (TM2ndS): Prediction of the TM helical regions for GPCRs from the sequence rests on the assumption that the outer regions of the TM helices (in contact with the hydrophobic tails of the lipids) should be hydrophobic and that this character should be largest near the center of the membrane (Donnelly et al., 1993; Eisenberg et al., 1984). The TM2ndS method uses this concept to generate a hydrophobic profile:

Step1a: Sequence alignment. The first part of step 1 for TM2ndS uses the SeqHyd hydrophobic profile algorithm, which is based on peak signal analysis of the hydrophobic profile for each amino acid. We first tested the use of the Prift scale (Cornette et al., 1987) but we found that the hydrophobicity index value for Arg was opposite that expected for a charged residue, leading to obviously incorrect assignments. We then switched to the use of the Eisenberg hydrophobicity scale (Eisenberg et al., 1982) which is based on sound thermodynamic arguments. This scale has a range from –1.76 to 0.73 and works well for Arg and other residues to give consistent TM predictions for the many systems we have investigated. The Eisenberg scale has been used in all published MembStruk results (1.0 onward). SeqHyd requires a multiple sequence alignment using sequences related to bovine rhodopsin. This is constructed by:

- Using an NCBI Blast search (Altschul et al., 1990,1997) on bovine rhodopsin (primary accession number P02699) to obtain protein sequences with bit scores above 200 but not identical (to avoid numerical bias in later calculations) to bovine rhodopsin (E value less than e^{-100}). We prefer an ensemble of sequences providing a uniform distribution of sequence identities from 35 to 100%. For bovine rhodopsin, this leads to the 43 sequences in Table S1 of the supplemental material.
- These 43 sequences plus Bovine rhodopsin were used in ClustalW (Thompson et al., 1994) to generate a pair-wise multiple sequence alignment. This sequence alignment included sequences with identities to the bovine rhodopsin sequence as low as 40%. In general we might include sequences with higher non-zero E-values, but including too low a homology might lead to additional alignment problems.

Step 1b: Average consensus hydrophobicity and initial TM assignment. The second part of step1 of TM2ndS is to calculate the consensus hydrophobicity for every residue position in the alignment. This consensus hydrophobicity is the average hydrophobicity [using the Eisenberg hydrophobicity scale] of all the amino acids in that position over all the sequences in the multiple sequence alignment. Then, we calculate the average hydrophobicity over a *window size (WS)* of residues about every residue position, using WS ranging from 12 to 20. This average value of hydrophobicity at each sequence position is plotted to yield the hydrophobic profile, as shown in Figure 1 for WS=14. The baseline for this profile serves as the threshold value for determining the TM regions and is calculated as follows:

1) First, we obtain the global average hydrophobicity value over all residues in the protein but excluding the amino terminus region (34 residues) and the carboxyl terminus region (42 residues). This global average is 0.041 for bovine rhodopsin.

2) If the baseline in 1) does not resolve the expected seven peaks, then TM2ndS automatically changes the baselines over a range of 0.05 from the global average (thus -0.009 to 0.091 for bovine rhodopsin). The baseline closest to the average that yields the 7 peaks is used for TM region prediction. This modified baseline (*base_mod*) is shown as the pink line in Figure 1. It provides the basis for the accurate determination of the TM regions in the sequence. This final baseline may be interpreted physically as a $\Delta G=0$ value above which residues are thermodynamically stable in the transmembrane and below which they are not. This baseline is unique to the particular protein to which it is being applied, with its individual environmental factors (water clusters, ions, hydrophobic or hydrophilic ligand or interhelical interactions, membrane composition) that may change the relative stability of any particular residue.

Below WS = 12 the fluctuations in hydrophobicity ("noise") are too large to be useful. The lowest WS that yields seven peaks (with peak length greater than 10 and less than 50, and peak area greater than 0.8) is denoted as WS_{min}. The peaks ranges for WS_{min} are used as input for the helix capping module discussed in the next section.

Figure 1 shows that assigning the TM region to helix 7 is a problem because it has a shorter length and a lower intensity peak hydrophobicity compared with all the other helices. This has been observed for other GPCRs (Vaidehi et al., 2002). The low intensity of helix 7 arises because it has fewer highly hydrophobic residues (ile, phe, val, leu) and

because it has a consecutive stretch of hydrophilic residues (e.g., KTSAVYN). These short stretches of hydrophilic residues (including Lys296) are involved in the recognition of the aldehyde group of 11-cis-retinal in rhodopsin. For such cases, we use as the baseline for assigning the TM predictions the local average of the hydrophobicity (from minimum to minimum about this peak). TM2ndS automatically applies this additional criteria when the peak length is less than 23, the peak area is less than 0.8, and the local average more than 0.5 less than the base_mod. For bovine rhodopsin only TM7 satisfies this criterion and the local average (0.011) is shown by the red line in Figure1. Thus, this local average is automatically applied for proteins where the residues are relatively hydrophilic but in which the helix might still be stable because of local environmental factors (mentioned above) that stabilize these residues.

Step 1c: Helix Capping in TM2ndS: It is possible that the actual length of the helix would extend past the membrane surface. Thus, we carry out a step aimed at capping each helix at the top and bottom of the TM domain. This capping step is based on properties of known helix breaker residues, but we restrict the procedure to not extend the predicted TM helical region more than six residues. We consider the potential helix breakers (Donnelly et al., 1994) as

- P, G,
- positively charges residues (i.e., R, H, K), and
- negatively charged residues (i.e, E, D).

TM2ndS first searches up to four residues from the edge going inwards from the initial TM prediction obtained from the previous section for a helix breaker. If it finds one, then the TM helix edges are kept at the initial values. However, if no helix breaker is found,

then the TM helical region is extended until a breaker is found, but with the restriction that the helix not be extended more than 6 residues on either side. The shortest helical assignment allowed is 21, corresponding to the shortest known helical TM region. This lower size limit prevents incorporation of narrow "noise" peaks into TM helical predictions.

We have used this TM2ndS algorithm for predicting the structure for ~ 10 very different GPCR classes (Vaidehi et al., 2002). In each case the predicted binding site and binding energy agrees well with available experimental data, providing some validation of the TM helical region prediction. However only for bovine rhodopsin can we make precise comparisons to an experimental structure. Figure 2 compares the predictions of TM helical regions for bovine rhodopsin to the TM helical regions as assigned in the crystal structure (Palczewski et al., 2000). To determine which residues have an alpha helical conformation, we analyzed the phi-psi angles using PROCHECK (Laskowski et al., 1993) and considered the experimental structure to be in an α-helix if $-37 < \phi < -77$ and $-27 < \psi < -67$. This led to slightly shorter helices than quoted in the crystal structure paper. Thus the lowercase letters in Figure 2 indicate residues which are outside the above range but quoted as helices in the experimental paper. The results are:

- For TM1 our prediction adds P at the start and H at the end. In our final structure the (φ,ψ) for this P is (Not applicable [N-terminus],-43.6) and for this H is (-54.3,-32.4) whereas the values obtained in the crystal structure are (-44.3,-24.9) and (-72.5, 69.5) respectively. Since P and possibly H might be expected to break the helix, we are considering modifying our procedure to not keep such terminal P or H in the helix.

- For TM2 our prediction adds HG at the end. In our final structure the (φ,ψ) for this H and G are (-73.6,-80.9) and (-55.0,148.8) whereas the values obtained in the crystal structure are (-74.2,0.5) and (66.1,9.0), respectively. The crystal structure paper considered the H as part of the helix. Since HG could be expected to break the helix, we are considering modifying our procedure to not keep the terminal HG in the helix. In fact, the HG angles in our final structure fall outside our criteria for alpha helicity as a result of the MembStruk optimization of the structure.
- For TM3 our predictions miss the RYVVV assigned in the crystal structure to the helix. Since the first and second V do not have (φ,ψ) in the usual range for alpha helices, we consider that the VVV should be excluded. However, the polar character of RY leads TM2ndS to miss assigning them as part of the helix. The crystallographic (φ,ψ) for R and Y residues are (-55.5,-63.8), (-44.6,-56.3) whereas the values obtained in our final structure are (76.7,-51.4), (-62.9,119.2). It should be pointed out that the B-factors on the cytoplasmic end of the rhodopsin crystal structure are high in this region of the helix (pdb entry 1F88). This indicates that the helix is probably fluxional even when the receptor is not activated. Consequently caution should be used when comparing our predictions with the crystal structure at this end. Also, because the helices are translated to align hydrophobic centers in a later step of the procedure, this uncertainty in TM helical prediction may only lead to local errors in atomic structure.
- For TM4 our prediction adds G at the end and misses N at the start. The crystallographic (φ,ψ) for these N and G residues are (-43.5,-59.6) and (169.8,5.4) whereas the value obtained in our final structure are (-93.9,119.6) and (112.5,-118.4).

Thus the predictions are fine even though the G and N were misassigned. We are considering modifying our procedure to exclude a terminal G.

- Compared to the crystal structure assignment, our prediction for TM5 adds LVF at the end and misses N at the start. In addition the GQ at the end terminus in the crystal structure assignment have (φ,ψ) outside the range for alpha helices. Thus we consider that the terminal GQLVF in the TM2ndS predictions are in error, the largest error of any of the predictions. The crystallographic (φ,ψ) for these N and LVF residues are (-69.3,-51.1), (-48.2,-36.7), (-39.6,-27.1), and (-58.0,-26.5) whereas the values obtained in our final structure are (-109.9,-162.4), (-55.1,-47.8), (-63.4,-59.0), and (-81.5,59.3). The rhodopsin crystal structure has high B-factors for the intracellular end of TM5 (just as for helix 3) suggesting caution in making comparisons.
- For TM6 our prediction adds H at the end and misses EVT at the start. The crystallographic (φ,ψ) for these EVT and H residues are (-57.6,-53.0), (-54.1,-55.7), (-56.3,-52.3), and (-81.3,48.8) whereas the value obtained in our final structure are (-74.4,72.3), (-73.1,130.8), (-16.9,-53.0), and (7.1,87.7). Thus the predictions are fine despite the misassignments. We are considering modifying our procedure to exclude a terminal H. In fact, the H angles in our final structure fall outside our criteria for alpha helicity as a result of the MembStruk optimization of the structure.
- For TM7 our prediction adds P at the start and misses Y at the end. The crystallographic (φ,ψ) for the P and Y residues are (-30.2,-48.1) and (-46.0,-55.0) whereas the value for P obtained in our final structure is (-43.6,-23.2). Since the current MembStruk protocol does not model the structures of the C and N termini, we did include the Y in our structure. Thus the predictions are fine despite the

misassignments. We are considering modifying our procedure to exclude a terminal P, but it is not obvious that a modified method would automatically include the Y. In fact, the P angles in our final structure fall outside our criteria for alpha helicity as a result of the MembStruk optimization of the structure.

Overall, we consider that the predictions agree sufficiently well with the crystal structure to be useful in building them into the assembly. In addition, we can see several improvements in the capping procedure of TM2ndS that could have decreased the errors in predicting which residues near the ends are considered to be helix breakers for capping the TM helices. However, this paper is meant to validate the procedure we have been applying to many systems and we did not want to change the procedure on the basis of our only independent validation.

Step 2: Assembly and optimization of the seven helical TM bundle:

Having predicted the seven TM helix domains using TM2ndS, we next build them into the seven helical TM bundle. This involves two steps: assembly and optimization of the relative translation and rotation of the helices.

Step 2a: Assembly of the Seven TM helices into a bundle: Canonical right-handed α-helices are built for each helix using extended side chain conformations. Then the helical axes are oriented in space according to the 7.5Å electron density map of frog rhodopsin (Schertler et al., 1998). This 7.5Å electron density map gives only the rough relative orientations of the helical axes, with no data on atomic positions. This serves as the starting point for optimization of the helices in the helical bundle. It should be emphasized here that **no** information as to helical translations or rotations was used.

Since this electron density map showed no retinal present, it is not clear whether this form of rhodopsin is active or inactive. This same information has been used to build structures of ~10 other GPCR classes (Vaidehi et al., 2002). In each case the predictions of binding site and binding energy agrees well with available experimental data, providing some validation for this general approach of constructing the TM bundle of GPCRs. However for bovine rhodopsin we can make much more precise comparisons to the experimental structures, as reported below.

Step2b: Optimization of the Relative Translation of the helices in the bundle: The translational and rotational orientation of each helix in the TM bundle is critical to the nature and conformation of the binding site in the GPCR. We do *not* use homology methods to predict these quantities because many GPCRs have very remote sequence homology to rhodopsin (ranging down to 10%) making it quite risky to base a three-dimensional structure on homology modeling using the rhodopsin crystal structure as template. Also we do not use atomistic molecular dynamics (MD) and molecular mechanics (MM) methods to optimize the structure, because the large barriers between various favorable positions can trap the conformation in local minima making such approaches ineffective in repositioning the helices. *Instead we developed methods to optimize the initial packing by translating and rotating the helices over a grid of positions and by using various properties of the amino acids in the sequence to suggest initial starting points.* This Monte Carlo like systematic conformational search algorithm for rotational and translational orientation of the helices allows the system to surmount barriers in the conformational space.

Our general principle in repositioning the helices is that the outer surface of the TM bundle (at least the middle regions) should be hydrophobic in order to have stabilizing interactions with the hydrophobic chains of the lipid. We imagine a midpoint plane through the lipid bilayer corresponding to the contact of the hydrophobic chains, which we denote as the *lipid midpoint plane* (LMP). We then assume that the hydrophobic regions of the TM bundle will position themselves such that the middle of their maximum hydrophobicity lies in this plane. We tested this concept for the crystal structure of bovine rhodopsin as follows. We determined the *hydrophobic center* (HPC) for each helix as the maximum of the peak of hydrophobicity from the profiles generated with various window sizes (since we go an integer number of residues in each direction WS is always even). Our criterion for the best fit to experiment is that these 7 positions when applied to the crystal structure would all lie in a single plane that could be taken as the LMP.

As shown in Figure 3, the deviation of the calculated hydrophobic centers from lying in a single plane in the rhodopsin crystal structure is a minimum for WS 20 and 22. Thus Get_Centers calculates the overall hydrophobic center of each TM helix based on the average of centers obtained for a range of window sizes near 20. Get_Centers determines this range of window sizes as follows. First, each *hydrophobic center (HC)* is calculated for WS=20. Then, the HC are calculated for WS 12-30 (excluding WS=20). For each helix Get_Centers determines the window sizes that yield HC less then 5 residues from the HC calculated at WS=20. For example, consider helix 1 in Table 1. Here HC = 18 for WS = 20. For windows sizes 12, 14, 16, 18, 22, 24, 26, 28, 30 we find HC = 15, 13, 20, 18, 17, 18, 15, 16, 13. For WS 16, 18, 22, 24 the HC are less than 5

from the value at WS=20. Thus we consider that the hydrophobic center calculation is stable within this regime of window sizes. The HC calculated for WS 16, 18, 22, 24 for the helices 2-7 are also less than 5 residues from the centers at WS=20. Thus, Get_Centers averages the HC for window sizes 16, 18, 22, 24 and then it averages these values with the HC at WS=20 for each TM helix. Get_Centers takes these values (last column of Table 1) as the final TM helix centers. We find that for bovine rhodopsin, these seven HPC deviate by a root mean square of 1.04 Å from a common plane.

Step 2c: Optimization of the rotational orientation: Once the helices are aligned along their helical axes according to the calculated hydrophobic centers, the rotational orientation of the helices is optimized using either or both of the following steps:

i) Orienting the net hydrophobic moment of each helix to point toward the membrane (Phobic Orientation). In this procedure (denoted as CoarseRot-H), the helical face with the maximum hydrophobic moment is calculated for the middle section of each helix, denoted as the *hydrophobic mid-region (HMR).* The face is the sector angle obtained as follows: 1) the central point of the sector angle is the intersection point of the helical axis (the active helix that is being rotated) with the common helical plane (LMP). 2) The other two points forming the arc are the nearest projections (on the LMP) of the C_α vectors of the two adjacent helices. The calculation of the hydrophobic moment vector is restricted to this face angle. This allows the predicted hydrophobic moment to be insensitive to cases in which the interior of the helix is uncharacteristically hydrophilic (because of ligand or water interactions within the bundle). Currently we choose HMR to be the 1/3 of each helix straddling the predicted hydrophobic center and exhibiting large hydrophobicity. This hydrophobic moment is projected onto the common helical plane

(LMP) and oriented exactly opposite to the direction toward the *geometric center of the TM barrel (GCB).* This criterion is most appropriate for the six helices (excluding TM3) having significant contacts with the lipid membrane. The LMP is the plane that most closely intersects the hydrophobic centers as described in step 2.2.2b. The GCB is calculated as the center of mass of the positions of the alpha-carbons for each residue in the HMR for each helix summed over all seven. This procedure is called *Phobic Orientation*.

ii) Optimization of the rotational orientation using energy minimization techniques (RotMin): In this procedure, each of the seven TMs is optimized through a range of rotations and translations one at a time (the active TM) while the other six helices are re-optimized in response. After each rotation of the main chain (kept rigid) of each helix, the side chain positions of all residues for all seven helices in the TMR are optimized [currently using SCWRL (Bower et al., 1997)]. The potential energy of the active helix is then minimized (for up to 80 steps of conjugate gradients minimization until an RMS force of 0.5 kcal/mol/Å is achieved) in the field of all other helices (whose atoms are kept fixed). This procedure is carried out for a grid of rotation angles (typically every 5 ^{o} for a range of +/- 50 ^{o}) for the active helix to determine the optimum rotation for the active helix. Then we keep the active helix fixed in its optimum rotated conformation and allow each of the other six helices to be rotated and optimized. Here the procedure for each of the six helices one by one is: (1) rotate the main chain, (2) SCWRL the side chains, (3) minimize the potential energy of all atoms in the helix. The optimization of these six helices is done iteratively until the entire grid of rotation angles is searched. This method is most important for TM3, which is near the center of the GPCR TM barrel and not

particularly amphipathic (it has several charged residues leading to a small hydrophobic moment). This procedure is called *RotMin*.

For bovine rhodopsin, we used Phobic Orientation for placing the hydrophobic moments away from the GCB for all seven helices. Subsequently rotations were optimized using RotMin for helices 3 and 5 using small rotation angles of $\pm 2.5^{o}$, $\pm 5.0^{o}$, and $\pm 8.0^{o}$. This optimizes the only salt bridge in the TM region (between residues His211 and Glu122). Coarse-grain rotation optimization combining both the energy optimization and hydrophobic moments is expected to provide better optimized TM helices than either one alone.

Step3: Optimizing the individual helices:

The optimization of the rotational and translational orientation of the helices described in the above steps is performed initially on canonical helices (we also apply them again to the helices after their optimizations described in step3). To obtain a valid description of the backbone conformation for each residue in the helix, including the opportunity of G, P, and charged residues to cause a break in a helix, the helices built from the step 2 were optimized separately. In this procedure we

- first use SCWRL for side chain placement
- then carry molecular dynamics (MD) (either Cartesian or torsional MD called NEIMO [Jain et al., 1993 ; Mathiowetz et al., 1994; Vaidehi et al., 1996]) at 300 K for 500 ps,
- then choose the structure with the lowest total potential energy in the last 250 ps and minimize it using conjugate gradients.

This optimization step is important to correctly predict the bends and distortions that occur in the helix due to helix breakers such as proline and two glycines. The MD also carries out an initial optimization of the the side chain conformations, which is later further optimized within bundle using Monte Carlo side chain replacement methods. This procedure allows each helix to optimize in the field due to the other helices in the optimized TM bundle from Step 2.

Step 4: Addition of lipid bilayer and fine-grain re-optimization of the TM bundle: To the final structure from Step 3 MembStruk adds two layers of explicit lipid bilayers. This consists of 52 molecules of dilauroylphosphatidyl choline lipid around the TM bundle of seven helices. This was done by inserting the TM bundle into a layer of optimized bilayer molecules in which a hole was built for the helix assembly and eliminating lipids with bad contacts (atoms closer than 10 Å). Then we used the quaternion-based rigid body molecular dynamics (RB-MD) in MPSim(Lim et al., 1997) to carry out RB-MD for 50 ps (or until the potential and kinetic energies of the system stabilized). In this RB-MD step the helices and the lipid bilayer molecules were treated as rigid bodies and we used 1 fs time steps at 300 K. This RB-MD step is important to optimize the positions of the lipid molecules with respect the TM bundle and to optimize the vertical helical translations, relative helical angles, and rotations of the individual helices in explicit lipid bilayers.

Step 5: Loop building: Following the RB-MD, we added loops to the helices using the WHATIF software (Vriend et al., 1990). After the addition of loops, we used SCWRL (Bower et al., 1997) to add the side chains for all the residues. The loop conformations were optimized by conjugate gradient minimization of the loop conformations while

keeping the TM helices fixed. This step also allows the general option of forming selected disulfide linkages [e.g., between the cysteines in extracellular 2 (EC-II) loop (which are conserved across many GPCRs) and the N-terminal edge of TM3 or EC3]. In the case of bovine rhodopsin, the alignment of the 44 sequences from Section 2.21a indicates only one pair of fully conserved cysteines on the same side of the membrane (extracellular side). The disulfide bond was formed and optimized with equilibrium distances lowered in decrements of 2 Å until the bond distance was 2 Å. Then the loop was optimized with the default equilibrium disulfide bond distance of 2.07 Å. Annealing MD was then used to optimize the EC-II loop at this stage. This involved 71 cycles in each of which the loop atoms were heated from 50 K to 600 K and back to 50 K over a period of 4.6 ps. From each cycle the minimum potential energy structure was selected and minimized. We finally used the overall minimum energy structure for subsequent steps. During this process the rest of the atoms were kept fixed for the first 330 ps and then the side chains within the cavity of the protein in the vicinity of the EC-II loop were allowed to move for 100 ps to allow accommodation of the loop. Subsequently a full atom conjugate gradient minimization of the protein was performed in vacuum using MPSim (Lim et al., 1997). This leads to the final MembStruk predicted structure for bovine rhodopsin.

The crystal structure for the retinal/rhodopsin complex has a well-defined β-sheet structure for EC-II, which we speculate to be involved as a mobile gate for entry of 11-cis-retinal on the extracellular side of rhodopsin. Such a gating mechanism is illustrated in Figure 4 in which the helix 3 coupled to this loop by a cysteine bond is the "gatekeeper" which responds to signaling structural sub-states of rhodopsin as follows:

1) When rhodopsin binds 11-cis-retinal, the ground state conformation of the receptor is stabilized, thus shifting helix 3 towards the intracellular side (forming the D(E)RY-associated salt bridges at that end) and "closing" the EC-II loop. In fact, 11-cis-retinal has been shown to be an inverse agonist for G-protein signaling (Okada et al., 2001).

2) In response to absorption of a photon, the 11-cis retinal isomerizes to the all-trans conformation, inducing helix 3 to shift towards the extracellular side. This induction of helix 3 movement may be direct or indirect. It may be due to a direct clash of helix 3 with all-trans-retinal. This is consistent with the result of a cross-linking experiment in which the ionone ring of retinal interacts with Ala269 when the receptor is activated (Borhan et al., 2000). This may occur because the trans-retinal clashes with helix 3 of the ground state rhodopsin crystal structure (Bourne at al., 2000). The induction of helix 3 movement may also occur indirectly in the following way: 11cis-retinal as observed in the crystal structure interacts with aromatic side chains Trp265 and Tyr268 on helix 6. But all-trans-retinal does not have these stabilizing interaction with helix 6, which should decrease the energy barrier for helix 6 rotation [this has been observed in preliminary MD calculation we carried out and in reports in the literature (Saam et al., 2002)].

3) This motion (of helix 3 or helix 6) breaks the DRY-associated salt bridges (Greasley et al., 2002) at the intracellular side. Helix 3 may have fewer constraints to movement, but since it is coupled by a disulfide linkage to the EC-II loop, movement on helix 3 would likely cause an "opening" of the EC-II loop to allow Schiff base reversion and exit of the free all-trans-retinal ligand. The breaking of this DRY salt bridge would also allow hinge motion (Altenbach et al., 2001: 1 and 2) of helix 6 to

expand the molecular surface at the cytoplasmic end for G-protein binding. This model is consistent with the experimental mutations studies in which the disulfide has been shown to be important for ligand binding and receptor activation (Schöneberg et al., 2002).

Building the loops without the constraint of coupling these cysteines leads to an open EC-II loop very different from the crystal structure of bovine rhodopsin. It is likely that both the open loop and closed loop structures play an important role in GPCRs, and indeed general observations of GPCRs suggests two distinct forms one of which leads to activation of G protein and one of which does not. We consider that one of these is likely the closed form and the other the open form. It seems likely that the ligand might not be able to diffuse into the active site when the loop is closed and hence for most GPCRs (other than bovine rhodopsin) we visualize the process of activation as:

- The GPCR with the open form of EC2 loop can bind selectively to the appropriate ligand;
- Binding of the ligand favors closing of the EC2 loop;
- After closure of the loop, G protein activation may begin.

Thus we have built two structures for bovine rhodopsin (Here the MS denotes that the structure was predicted using MembStruk):

- **Apo/closed(MS)** has the cysteine coupling observed in the crystal and is the structure we compare to experiment after binding the retinal
- **Apo/open(MS)** is built without a constraint, forming what we believe would be the configuration which binds initially to the ligand.

Function prediction for GPCRs

Since there are no experimental structures available for any human GPCR, the only validation available for the accuracy of predicted structures for human GPCRs is to predict the ligand binding sites and the ligand binding energies. The accuracy in the predicted binding site can then be judged from site-directed mutagenesis experiments on the residues predicted to control selectivity. An even tougher test is to compare binding affinity of ligands to each other and to mutated proteins. For many GPCRs of pharmaceutical interest there is ample experimental data on ligand binding constants as well as agonist and antagonist inhibition constants for many GPCRs (for a compilation of this literature see http://www.gpcr.org).

To carry out such function validations for the predicted structures, it is essential to have reliable and efficient procedures for predicting binding site and binding affinities. Since the ligand-binding site is completely unknown for most GPCRs, we must scan the entire protein to identify likely binding sites and conformation of each ligand, and then we must reliably rank the relative binding energies of the various ligands in these sites. To do this we employ the HierDock procedure, which has been tested and validated for predicting ligand binding sites and ligand binding energies for many globular and membrane bound proteins (Vaidehi et al., 2002; Kekenes-Huskey et al., 2003 accepted; Floriano et al., 2003 accepted; Datta et al., 2003; Datta et al., 2002). These studies show that the multi-step hierarchical procedure in HierDock ranging from coarse-grain docking to fine-grain MD optimization leads to efficient and accurate predictions for ligand binding in proteins.

The HierDock method was first described in Floriano et al. 2000, which we label as HierDock1.0. The method was improved in Vaidehi et al. 2002, which we label as HierDock2.0. In this paper we present an improved version that we label as HierDock2.5. The various steps involved in this current procedure are as follows:

1. *Sphere generation*. We assume no knowledge of the ligand binding site in GPCRs and hence the entire molecular surface of the receptor is scanned to predict the energetically preferred ligand binding sites. The negative of the molecular surface of the protein was used to define potential binding regions within the receptor over which the various ligand conformations are to be sampled. The void regions are mapped with spheres generated over the whole receptor using the Sphgen program in DOCK 4.0. No assumptions were made on the nature or the location of the binding site in these receptors. For bovine rhodopsin this led to total of 7474 spheres, which was partitioned into 13 overlapping docking regions each with a volume of $(10Å)^3$ as shown in Figure 5. We excluded from docking regions in contact with the membrane or near the intracellular region likely to be involved in binding to the G-protein. No assumptions were made on the nature or the location of the binding site in these regions.
2. *Coarse-grain sampling*. To locate the most favorable ligand binding site(s), we used DOCK 4.0 (Ewing et al., 1997) to generate a set of conformations for binding 11-cis-retinal (a ligand known to bind to bovine rhodopsin) to each of the 13 regions. For this docking step we used a bump filter of 10, a non-distance dependent dielectric constant of 1.0, and a cutoff of 10Å for energy evaluation. The ligands were docked as non-flexible molecules to generate and score 100 conformations of the ligand in

each of the 13 regions. We then rejected any ligand conformation with less than 90% of the surface area buried into the protein and ranked the remainder by the ligand-protein interaction energy using Dreiding FF. The best binding energy conformation among the 13 regions was chosen as the putative binding region. Other conformations with binding energies within 10 kcal/mol of the best conformation were also chosen as possible binding regions.

3. *Construction of putative binding region using a more refined sampling of ligand-protein interactions.* A set of overlapping boxes were used to enclose the volume corresponding to the putative "binding region" (or regions) determined in step 2, which is now to be used for a new sampling of ligand-protein conformations similar to 2.

4. *Coarse-grain sampling of putative binding regions*. To locate the most favorable ligand binding site(s), we again used DOCK 4.0 to generate a set of conformations for binding 11-cis-retinal (a ligand known to activate bovine rhodopsin) to the putative binding region. We again used a bump filter of 10, a non-distance dependent dielectric constant of 1.0, and a cutoff of 10 Å for energy evaluation. The ligands were docked as non-flexible molecules to generate and score 1000 conformations. After eliminating ligands with less than 90% surface burial, we selected the 10% (100) with best DOCK4.0 score for further analysis.

5. *Ligand-only minimization*. The 100 best conformations selected from step 4 were (conjugate gradient) minimized keeping the protein fixed but all atoms of the ligand movable. Minimized ligand conformations that satisfied the buried surface area cutoff criterion of 75% were kept for the next step.

6. *Ligand-protein full minimization*. The ligand/protein conformations from step 5 were further energy minimized with all atoms (protein, lipid, and ligand) movable using conjugate gradients. The structure with the binding energy calculated by equation (1) was selected:

   BE_1 = Energy (ligand in protein complex) - Energy (free ligand in solvent) (1)

   Here the energy of the ligand in water is calculated using DREIDING FF and AVGB continuum solvation method. Since a substantial part of the complex is in contact with the membrane environment, we did not solvate the complex.

7. Using the best binding conformation from step 6, the side chain conformations for all the residues within 5Å of the bound 11cis-retinal conformation were optimized using the SCREAM side chain optimization program (Kam et al., unpublished). The resulting ligand-protein structure was finally optimized by conjugate gradient minimization allowing all atoms to relax.

8. Iterative HierDock (optional). The protein from step 7 (optimized with ligand bound) was saved. The steps 4-6 were repeated again in order to obtain the best possible conformation for the ligand within the protein (with side chains optimized in the presence of the ligand). This step was performed for bovine rhodopsin.

Results and discussion

We first present the results for the validation of the HierDock protocol on the crystal structure of bovine rhodopsin, followed by results on structure and function prediction for bovine rhodopsin. In order to clarify our notation we summarize it here.

- *Ret(xtal)/closed(xtal)* is obtained from the crystal structure by minimizing using the Dreiding FF. It deviates from the crystal structure by 0.29Å CRMS. It has retinal bound as in the crystal structure and has the closed form of the EC2 loop. The retinal conformations differ by 0.22 Å CRMS. This further validates the FF. Since they differ so little, the retinal in the non-minimized crystal structure, ret(xtal-noFF), is used as the reference structure for the HierDock validation step.
- *Apo/closed(xtal)* is obtained from Ret(xtal)/closed(xtal) by removing the retinal and adding the proton to Lys296. It was minimized without ligand. It deviates from the crystal structure by 0.74Å CRMS. It is likely that this is a lower bound on the change in structure upon removal of the retinal. For a more complete optimization, we would use MD.
- *Ret(HD)/closed(xtal)* is the predicted structure for 11-cis retinal obtained by applying HierDock to Apo/closed(xtal) and then forming the Schiff base linkage to Lys296 and minimizing. The ret(HD) deviates from Ret(xtal) by 0.62Å CRMS. To distinguish the error in ligand conformation due to the HierDock procedure from that due to MembStruk, the structure Ret(HD)/closed(xtal) will serve as the reference structure to compare to the predicted ligand conformations in the MembStruk structures.
- *Apo/closed (MS)* is the MembStruk predicted structure of the closed form, without the retinal. The TM bundle for this structure deviates by 2.84 Å CRMS main chain atoms from Apo/closed(xtal) (4.04 Å CRMS for all TM atoms, excluding H)
- *Ret(HD)/closed(MS)* is the predicted structure for 11-cis retinal in the Apo/closed(MS) rhodopsin structure, obtained by applying HierDock to

Apo/closed(MS) and then forming the Schiff base linkage to Lys296 and minimizing the energy The ret(HD) deviates from Ret(HD)/closed(xtal) by 2.92 Å CRMS.

- *Apo/open(MS)* is the MembStruk predicted structure of bovine rhodopsin without the retinal. There are no experiments with which to compare. This structure differs in the TM region from Apo/closed(MS) by 0.11 Å.
- *Ret(HD)/open (MS)* is the predicted structure for 11-cis retinal in rhodopsin obtained by applying HierDock to Apo/open(MS) and then forming the Schiff base linkage to Lys296 and minimizing. There are no experiments with which to compare. The retinal differs from that in Ret(HD)/closed (MS) by 1.74 Å.

Validation for function prediction: HierDock protocol for 11cis-retinal on bovine rhodopsin:

Bovine rhodopsin (a member of the opsin family) is the only GPCR to be crystallized in its entirety at a high resolution (2.8Å). Thus we used this system as a test to validate the HierDock protocol for predicting of binding sites of GPCRs.

To test HierDock, we used the apo/closed(xtal) structure with the retinal removed and minimized. First we did a complete HierDock scan as outlined above to predict the binding of 11cis-retinal to bovine rhodopsin. The crystal structure of rhodopsin has the 11cis-retinal covalently bound to Lys 296 (between the aldehyde of 11cis-retinal and the N of the Lys), but for docking we cannot have a covalent bond to the crystal. Thus we docked the full 11cis-retinal ligand (containing a full aldehyde group) and considered the Lys 296 to be protonated.

We applied steps 1-2 of the HierDock described above for all 13 overlapping regions for step 2 shown in Figure 5. The initial scan of the entire rhodopsin (step 1 to 2

in section 2.3) gave two good binding regions shown as the red boxes in Figure 5. The data for this scanning step are shown in Table 2. The final optimized best binding structure for the retinal/rhodopsin complex from step 6 of HierDock deviates by 1.11 Å CRMS from the ligand in the crystal structure as seen in Figures 6ab. The binding site (defined as the 7 residues that contribute at least 1 kcal/mol to the bonding) of this ligand is shown in Figure 9b. Lys296 has hydrophilic interactions while the other side chains have van der Waals interactions. This docked structure has the retinal O 2.72 Å from the N of Lys 296. In addition, the retinal O and the closest H of the protonated Lys296 N are just 2.35 Å apart, close enough to form an H-bond (likely an intermediate step before Schiff base formation). We then coupled these two units to form the covalent CN bond to Lys 296 while eliminating the H_2O. After minimizing the full ligand-protein structure, we find that the predicted structure for 11cis-retinal bonded to the protein deviates from the crystal structure by only 0.62Å CRMS as shown in Figures 6cd. Most of this discrepancy results because the FF minimized structure of the retinal has the ionone ring in a chair conformation which was retained in our docking procedure, whereas the crystal structure has the ionone ring in a half-chair conformation (which we calculate to be 2 kcal/mol higher in energy than the chair conformation within the minimized complex). This retinal/protein complex minimized with the Dreiding FF [denoted ret(HD)/closed(xtal)] serves as the reference structure for comparing the predicted structures in later sections. *We consider that these results validate the HierDock protocol for a GPCR.*

In addition, we used HierDock to determine the binding site and best scoring ligand conformation for all-trans-retinal, with the binding energy calculated using equation (1) above. The binding energy for 11cis-retinal was -1 kcal/mol while that for

all-trans-retinal was ~31 kcal/mol, a difference of 32 kcal/mol. This compares well with the experimental result that the retinal ligand/protein complex stores 34.7 +/-2.2 kcal/mol upon isomerization in the protein (Okada et al., 2001). This stored energy might be used to induce rigid body helical motions needed for receptor activation and G-protein binding. This excellent agreement is probably fortuitous since we have not carried out full optimizations of the all trans configuration , but it may be partly because cis and trans retinal are neutral isomers of each other with similar solvation energies.

Structure prediction of rhodopsin using MembStruk

We used MembStruk3.5 as detailed in section 2.2 to predict the structure of bovine rhodopsin using only the protein sequence. For the apo-rhodopsin we predicted two structures, one with the open EC-II loop and one with closed EC-II loop. These represent two different states of rhodopsin likely to play a role in activation of G-Protein. The crystal structure of rhodopsin has a closed EC-II loop with the 11-cis-retinal bound to it. To validate this predicted structure, we should compare to the crystal structure for apo-rhodopsin (without a bound 11cis-retinal). However, this crystal structure for the apo protein is not available. Thus instead we will compare the predicted structure to the minimized crystal structure of bovine rhodopsin after removing the 11cis-retinal. In making these comparisons, we predicted two structures for apo-rhodopsin:

- the open form where no restrictions were made on the structure of EC-II loop [*apo/open(MS)]* and
- the closed form where we assumed that EC2 makes the same cysteine linkage as observed in the crystal structure [*apo/closed(MS)]).*

The predicted TM domains are compared to the rhodopsin crystal structure in Figure 2 and discussed in section 2.2.1c

After optimization of the helices using MD (300 K for 500 ps), most helices yield the same bends as in the crystal. Thus helices 2 and 6 undergo significant bending (due to Pro267 in helix 6 and due to GLY89 and GLY90 in helix 2), which is consistent with spin-labeling EPR experiments (Farrens et al., 1996). In addition, we find that helix 7 bends near the two prolines, which has also been shown by spin-labeling experiments (Altenbach et al., 2001: 1 and 2). We find that helix 1 undergoes significant bending due to a GLY/PRO combination, but this has not yet been studied experimentally. Snapshots with even more similar bends to the crystal structure (throughout the dynamics) are also shown in Figure S2a. The overlays of all snapshots at 2.5 ps intervals in shown in Figure S2b. The dynamical nature of the bending can be seen. Such bending at "hinge" sites may be important for expanding the molecular surface needed at the cytoplasmic side to allow G-protein binding. We find similar "hinge-bending" with MD when the trans isomer is bound to the helix assembly.

The optimized helices were again inserted into a bundle. As a double-check to the rotation before the Rotmin step, a 360 degree energy scan was performed more recently (Figure S3). The alternate rotation of helix 6 at ~+90 degrees is important, since the helix with the most rotation difference than in the crystal structure is helix 6 (~100 degrees). This correlates with spin labeling experiments (Farrens et al., 1996) which point to a helix 6 rotation on receptor activation. In fact, when analyzing the hydrophobic moment orientation for a different number of middle residues from the hydrophobic center (in the optimized crystal structure), there are two conformations (Figure S4) of the helix which

emerge, corresponding to the two rotations of the active and inactive receptor. For the final structure in this current paper, this method was not used on helix 6.

After the helices were inserted into a bundle again, we carried out RotMin on helices 3 and 5, the only helix pair with a potential salt bridge The resulting 7 helix bundle was then inserted into a lipid bilayer, and optimized using rigid body molecular dynamics as described in step 4 of section 2.2. This step leads to optimization of the vertical helical positions, relative helical angles, and rotations of the individual helices within a lipid environment. The CRMS difference before and after this rigid body MD is 1.10Å for all atoms and 0.98Å for main chain atoms. This is consistent with in the changes during this optimization step for other GPCRs (Vaidehi et al., 2002).

After adding the intracellular and extracellular loops, optimizing the side chains, and then optimizing the structure in vacuum with the TM helical region fixed (to eliminate bad contacts in the loop region), we then optimized the entire structure allowing all bonds and angles to change. These *ab initio* predictions of the structure were carried out for both the open and closed forms of the EC-II loop in apo-rhodopsin leading to the Apo/open (MS) and Apo/closed(MS) structures where MS denotes a MembStruk derived structure and open or closed denotes the open or closed form of the EC-II loop. Although the crystal structure has the 11cis-retinal bound, we will compare the predicted apo-rhodopsin structures to the minimized apo-protein of the crystal structure, apo/closed(xray).

Comparing Apo/closed(MS) to Apo/closed(xray) we find a CRMS difference of 2.85 Å in the main chain atoms and 4.04 Å for all the atoms in the TM helical region. These structures are compared graphically in Figure 7a. Comparing all residues

including loops (but ignoring the residues not present or complete in the x-ray structure), the predicted structure differs from the crystal structure by 6.80 Å in the main chain and 7.80 Å CRMS for all atoms. The major contribution to this CRMS is the low-resolution loop region, which is likely to be quite fluxional and may be very different between crystal and solution. Specifically, the predicted topology and phi-psi angles of the EC-II loop are consistent with that of a beta-sheet. However, the specific twist of this beta-sheet in the xray structure was not predicted well. Although this may be partly due to packing effects in the crystal structure, we consider that our prediction of the general topology of the EC-II loop to act as a "plug" to restrict retinal binding is adequate but that specific interactions with retinal may not be predicted well. In the function prediction results discussed below in section 3.3.1, we find that there are no specific favorable interactions between the ligand and the EC-II loop before Schiff base bond formation in the crystal structure (Figure 9b). Thus the EC-II may function initially primarily as an unspecific "plug" to disfavor certain ligand conformations. After Schiff base bond formation, the ligand is then stabilized by Glu181 in the EC-II loop (Figure 10A). Thus accurate prediction of the atomic structure of the EC-II loop remains an important challenge.

We find that Apo/open (MS) deviates from Apo/closed(MS) by a CRMS difference of 0.11 Å in the main chain atoms and 0.68 Å for all the atoms in the TM helical region. These structures are compared graphically in Figure 7c. This small difference in CRMS in the transmembrane region suggests that we need to carry out long time scale molecular dynamics in order for the helices to accommodate the EC-II loop conformational change. Comparing all residues, the predicted structure differs from the

crystal structure by 4.74 Å in the main chain and 5.00 Å CRMS for all atoms. There is no experimental structure apo/open(xray) with which to compare Apo/open (MS).

HierDock function prediction for Apo_rhod (MS) structures

Except for bovine rhodopsin, essentially all applications of HierDock to GPCRs must use *predicted* structures rather than experimental structures. The question here is: given the errors in predicting the GPCR structure (2.8Å CRMS in the TM helical region) can we hope to get accurate predictions in the binding site and binding energy? We will now test how well HierDock determines the binding site of 11cis-retinal to the *predicted* rhodopsin structures Apo/open (MS) and Apo/closed(MS).

Here we repeated the full process described in section 2.3. The void space for both the Apo/open (MS) and Apo/closed(MS) structures were partitioned into fourteen 7Å x 7Å x 7 Å boxes and scanned for the putative binding site of 11-cis-retinal (using the same ab initio FF optimized ligand structure as in section 3.1). Again the molecule includes the aldehyde group (*no* assumed formation of the Schiff's base).

Apo/closed(MS)

Scanning the entire Apo/closed(MS) receptor to find the binding site and binding energy for 11-cis-retinal used the steps described in section 2. The best scoring conformation for 11cis-retinal and its associated binding site, denoted as NoSB-ret(HD)/closed(MS) are shown in Figure 9c. Here NoSB indicates the structure without the Schiff's base covalent bond between the aldehyde group of 11cis-retinal and Lys296. This conformation (no covalent attachment) differs from Ret(HD)/closed(xtal) by 3.2 Å CRMS. We should emphasize that the Apo/closed(MS) structure was constructed purely from *ab initio* predictions with MembStruk, with no input from the x-ray crystal

structure. Thus *nowhere* did we assume a lysine covalent bond with retinal in any of the docking procedures. Yet, the predicted structure identifies which Lys can bond to the retinal, with a 2.85Å between the predicted position of the retinal oxygen and the predicted position of the Lys 296 nitrogen.

Then starting with NoSB-ret(HD)/closed(MS), we formed this Schiff's base bond (eliminating H_2O), and optimized the full ligand-protein complex with conjugate gradient minimization to obtain the ret(HD)/closed(MS) structure. This differs from Ret(HD)/closed(xtal) by 2.92 Å CRMS. These structures are compared in Figures 8ab.

A second criterion for validity of the predicted binding site is to identify the residues interacting most strongly with the ligand, which can be used to predict mutational studies for validation and to design antagonists or agonists. Considering the binding site as all residues within 5.0Å of the ligand leads to 30 residues for Ret(xtal)/closed(xtal). For ret(HD)/closed(MS) we find 26 residues [26 in common with Ret(x)/closed(xtal)] and for ret(HD)/closed(xtal) we find 23 residues [15 in common with Ret(x)/closed(xtal)] in the binding site. More important is to establish which of these residues is responsible for ligand stabilization. Thus we calculated the interactions of all amino acid residues within 5Å of the ligand and kept those which interact more favorable than –1 kcal/mol interaction energy with the ligand. For ret(xtal)/closed (xtal) this leads to the 15 residues shown in Figure 10a. For ret(HD)/closed(MS) we find 10 residues (8 in common with Ret(x)/closed(xtal)) shown in Figure 10b and for ret(HD)/closed(xtal) we find 14 residues (12 of which in common with Ret(x)/closed(xtal)) shown in Figure 10c. The interactions energies of the residues are shown in Table S2. The side chains

identified to be important include Trp265 and Tyr 268, which have been implicated (Lin et al., 1996) to modulate the absorption frequency of 11-cis-retinal.

To provide an idea how the retinal binds prior to Schiff base bond formation, we also considered the binding site as all residues within 5.0Å of the ligand *before* bond formation that interact more favorable than –1 kcal/mol interaction energy with the ligand. For NoSB-ret(HD)/closed(xtal) this leads to the 7 residues shown in Figure 9b. For NoSB-ret(HD)/closed(MS) we find 8 (6 in common with NoSB-ret(HD)/closed(xtal)) shown in Figure 9c. The interactions energies of the residues are shown in Table S1. Of the top interacting residues (3 residues) in NoSB-ret(HD)/closed(xtal), 2 (Tyr268, Lys296) are also shown to rank among the top three in NoSB-ret(HD)/closed(MS). The residue which was missed (Thr118) ranked lower in NoSB-ret(HD)/closed(MS) because it is actually closer to the retinal (in comparison with the NoSB-ret(HD)/closed(xtal) structure), with distances as low as 2.8 Å (whereas an optimal van der Waals distance is ~3.4 Å) to the polyene chain of retinal.

In conclusion, we conclude that the MembStruk predicted structure is useful for predicting binding sites sufficiently well to direct mutation studies to elucidate the precise site.

Apo/open(MS).

We scanned the entire Apo/open(MS) receptor to find the binding site and binding energy for 11cis-retinal using the steps described in section 2. The best scoring conformation for 11cis-retinal and its associated binding site, denoted as NoSB-ret(HD)/open(MS) are shown in Figure 9d. The predicted structure identifies which Lys

can bond to the retinal, with 2.87 Å between the predicted position of the retinal oxygen and the predicted position of the Lys 296 nitrogen.

Then starting with NoSB-ret(HD)/open(MS), we formed this Schiff's base bond (eliminating H_2O), and optimized the full ligand-protein complex with conjugate gradient minimization to obtain the ret(HD)/open(MS) structure, This is no experimental structure with which to compare, but this structure differs from ret(HD)/closed(MS) by 1.7 Å CRMS. These structures are compared in Figures 11ab.

A second criterion for validity of the predicted binding site is in identifying the residues close to the ligand to consider for mutational studies and drug design. Considering the binding site of NoSB-ret(HD)/open(MS) as all residues within 5.0Å of the ligand, the amino acid residues which interact with less than –1 kcal/mol interaction energy with the ligand (10 residues) are shown in Figure 9d. Of these, 6 residues are also shown to interact with the ligand in the NoSB-ret(HD)/closed(MS) structure discussed in Sec. 3.3.1. We also find 4 additional residues (Phe276, Phe208, Val271, Ala272) that do not bind with 1 kcal/mol in the NoSB-ret(HD)/closed(MS) structure. This difference results from the shift in the retinal binding site upon opening of the EC-II loop. Thus, we consider that the retinal bound to the open-loop structure is partially stabilized by van der Waals interactions.

Exploring the signaling mechanism

Using MembStruk we predicted the two structures apo/open(MS) with the extracellular loop 2 (EC-II) in an "open" conformation and apo/closed(MS) with it closed. The crystal structure of rhodopsin has the closed configuration in which EC-II has a well-defined β-sheet structure with the 11-cis-retinal bound. We speculate that

changes in the structure of this loop are involved in activation of G-protein and in the entry of 11-cis-retinal on the extracellular side of rhodopsin. The idea is illustrated in Figure 4 in which the helix 3 coupled to this loop by a cysteine bond is the "gatekeeper" which responds to signaling structural sub-states of rhodopsin as follows:

1) Starting with the inactive form ret/closed with 11cis retinal covalently linked to the rhodopsin, the response to visible light causes the 11cis retinal to isomerize to all-trans-retinal, which in turn causes changes in the conformation (Altenbach et al., 2001: 1 and 2; Altenbach et al., 1996; Farrens et al., 1996) near the retinal that eventually leads a structure in which the all-trans retinal is covalently linked to the open form with a structure resembling the trans-ret(HD)/open(MS) structure from our calculations.
2) The transformation from closed to open in step 1 is caused by conformational changes responsible for activation (perhaps by the direct interaction of the trans isomer with helix 3, to induce helix 3 to shift towards the extracellular side, breaking the DRY-associated salt bridges at the intracellular side).
3) Other processes hydrolyze off the trans retinal to form a structure similar to apo/open(MS) and then other processes reattach 11cis-retinal to form a structure similar to ret(HD)/open(MS)
4) The ret(HD)/open(MS) relaxes eventually to form ret(HD)/closed(MS), the inactive form. In this process the EC-II loop closes, perhaps caused by the helix 3 shifting towards the intracellular side, reforming the DRY-associated salt bridges at that end with the final result that closes to form a structure similar to the inactive.

Thus by using MembStruk and HierDock we have generated a total of six structures (summarized later in section 5.0) for ligand/protein complexes that can now be used to explore all the processes involving ligand binding and GPCR activation. The experiment provided just one of these six structures, but the validation with experiment allows us to have greater confidence in the five structures for which experimental is not available.

Comparison to other methods

There have been attempts to model the structure of GPCRs using homology modeling methods with either the bacteriorhodopsin or bovine rhodopsin crystal structure as template (Strader et al., 1994). Since there is only one known structure, these homology applications lead to transmembrane regions very similar to the bovine rhodopsin template structure. Moreover, many important GPCR targets have only low homology to bovine rhodopsin making the models particularly unreliable (Archer et al., 2003). Thus the sequence identity of bovine rhodopsin to dopamine D2 receptor is 17%, of serotonin H1A is 14%, and of G2A is 13%, whereas good structures from homology models generally require over 45% sequence identity.

GPCR structures have also been modeled using the properties of conserved residues in multiple sequence alignments followed by optimization of the structure using distance restraint to maximize the hydrogen bonds (Lomize et al., 1990).Also distance restraints from various experiments were used to predict the structure of bacteriorhodopsin (Herzyk et al., 1995). Comparing the TM helical region of their predicted structure to a bundle of ideal helices (not bent) superimposed on the

bacteriorhodopsin electron cryo-microscopy structure, they reported a CRMS of 1.87 Å in the C-alphas.

Shacham et al. claim to have predicted the structure of bovine rhodopsin using an approach based on specificity of protein-protein interaction and protein-membrane interaction and the amphipathic nature of the helices. However they have not yet provided any details of their method or of predictions on other GPCRs.

Summary

Using MembStruk we predicted the three-dimensional structure of bovine rhodopsin protein interacting with 11cis-retinal using only primary sequence information. This led to the flowing structures:

- Apo/closed (MS) is the MembStruk predicted structure of the closed form, without the retinal. The transmembrane assembly for this structure deviates from Apo/closed(xtal) by 2.84 Å CRMS for the mainchain atoms (4.04 Å CRMS for all transmembrane atoms, excluding H). Starting with the crystal structure and minimizing using the Dreiding FF leads to a structure that deviates from the crystal by 0.29Å CRMS, indicating that the FF leads to a good description. Thus most of the 2.8 Å CRMS error is due to the MembStruk process.
- Ret(HD)/closed(MS) is the predicted structure for 11-cis retinal obtained by applying HierDock to Apo/closed(MS). This leads to close contact (2.8 Å) between the carbonyl of the retinal and the N of Lys296. Forming the Schiff base linkage and minimizing leads to the ret(HD) structure that deviates from Ret(HD)/closed(xtal) by

2.92 Å CRMS. Carrying out the same HierDock process for the minimized crystal structure leads to a predicted structure for 11-cis retinal that deviates from Ret(xtal) by 0.62Å CRMS. This indicates that it is errors in the predicted protein structure that are responsible for the errors in ligand prediction.

- Trans-Ret(HD)/closed(MS) is the predicted structure for all-trans retinal obtained by converting 11-cis retinal to all trans and allowing the protein to respond. There is no experimental structure with which to compare.
- Apo/open(MS) is the MembStruk predicted structure of without the retinal. There are no experiments with which to compare. This structure differs in the TM region from Apo/closed(MS) by 0.11 Å.
- NoSB-Ret(HD)/open (MS) is the predicted structure for 11-cis retinal obtained by applying HierDock to Apo/open(MS). There is no experimental structure with which to compare.
- Ret(HD)/open (MS) is formed from NoSB-Ret(HD)/open(MS) by forming the Schiff base linkage to Lys296 and minimizing. There are no experiments with which to compare. The retinal differs from that in Ret(HD)/closed(MS) by 1.74 Å.

The validation with experiment is sufficiently good that we can now start to explore the mechanisms by carrying out long time scale molecular dynamics and Monte Carlo calculations on these various forms to learn more about the mechanism of activation. Comparisons of the structures and energetics for these systems provide information that might be useful for understanding the mechanisms of binding and activation in rhodopsin in particular and GPCRs in general.

We have noted above several steps for which we anticipate substantial improvements and we are continuing to improve the methods. For example, the individual optimization of the helices can be performed under a more constrained environment by performing torsional dynamics of each helix in the presence of other helices or by performing torsional dynamics of all helices simultaneously. Also for improved accuracy in predicting the structures and for predicting the ligand binding energy, we intent to take into account the differential solvent dielectric environment between membrane, hydrophilic and the interfacial dielectric constants (Spassov et al., 2002).

Conclusions

These applications of TM2ndS, MembStruk, and HierDock to bovine rhodopsin validate these techniques for predicting both the structure of membrane-bound proteins and the binding site of ligands to these proteins. The predictions from such studies can be used to design experiments to test details of the structures, which might lead to improved structures. This could lead to structures, more accurate than any of these techniques individually. The HierDock and MembStruk techniques validated here should also be useful for applications to other GPCRs, particularly for targeting agonists and antagonists against specific subtypes.

In addition, these studies open the door to examination of the mechanism for activation (structural and energy changes) of signaling. Obtaining independent structures for each of the major steps involved in binding and activation (e.g., the six structures discussed for retinal-rhodopsin) provides the basis for computational studies and for experiments that should provide a basis for detailed examination of each step.

References

Altenbach, C.; Yang, K.; Farrens, D.L.; Farahbakhsh, Z.T.; Khorana, H.G.; and Hubbell, W.L. (1996). Structural features and light-dependent changes in the cytoplasmic interhelical E-F loop region of rhodopsin: A site-directed spin-labeling study. *Biochemistry* 35, 12470-12478.

Altenbach, C.; Cai, K.; Klein-Seetharaman, J.; Khorana, H.G.; Hubbell, W.L. (2001). Structure and function in rhodopsin: Mapping light-dependent changes in distance between residue 65 in helix TM1 and residues in the sequence 306-319 at the cytoplasmic end of helix TM7 and in helix H8. *Biochemistry* 40, 15483-15492.

Altenbach, C.; Klein-Seetharaman, J.; Cai, K.; Khorana, H.G.; Hubbell, W.L. (2001). Structure and function in rhodopsin: Mapping light-dependent changes in distance between residue 316 in helix 8 and residues in the sequence 60-75, covering the cytoplasmic end of helices TM1 and TM2 and their connection loop CL1. *Biochemistry* 40, 15493-15500.

Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W. and Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.

Altschul, S.F.; Gish. W.; Miller, W.; Myers,E.W.; and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

Archer, E.; Maigret, B.; Escrieut, C.; Pradayrol, L.; and Fourmy, D. (2003). Rhodopsin crystal: new template yielding realistic models of G-protein-coupled receptors? *Trends Pharmacol. Sci.* 24, 36-40.

Borhan, B.; Souto, M.L.; Imai, H.; Schichida, Y.; and Nakanashi, K. (2000). Movement of retinal along the visual transduction path. *Science.* 288, 2209-2212.

Bourne, H.R.; Meng, E.C. (2000). Structure - Rhodopsin sees the light. *Science*. 289,733-734.

Bower, M.; Cohen, F.E.; and Dunbrack, Jr. R.L. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J. Mol. Biol.* 267, 1268-1282.

Brameld, K.A.; Goddard, W.A. (1999). Ab initio quantum mechanical study of the structures and energies for the pseudorotation of 5 '-dehydroxy analogues of 2 '-deoxyribose and ribose sugars. *J Am. Chem. Soc.* 121, 985-993.

Cornette, J.L.; Cease, K.B.; Margalit, H.; Spouge, J.L.; Berzofsky, J.A.; Delisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* 195, 659-685.

Datta, D.; Vaidehi, N. Xu, X.; and Goddard III, W.A. (2002). Mechanism for antibody catalysis of the oxidation of water by singlet dioxygen. *Proc. Natl. Acad. Sci. USA*. 99, 2636-2641.

Datta, D.; Vaidehi, N.; Floriano, W.B.; Kim, K.S.; Prasadarao, N.V.; Goddard W.A. (2003). Interaction of E-coli outer-membrane protein A with sugars on the receptors of the brain microvascular endothelial cells. *Proteins*. 50, 213-221.

Ding, H.Q.; Karasawa, N.; Goddard, W.A. (1992). Atomic level simulations on a million particles - the cell multipole method for coulomb and London nonbond interactions. *Chem. Phys. Lett.* 97, 4309-4315.

Donnelly, D. (1993). Modeling alpha-helical transmembrane domains. *Biochem. Soc. T.* 21, 36-39.

Donnelly, D.; Overington, J.P. and Blundell, T.L. (1994). The prediction and orientation of alpha-helices from sequence alignments - the combined use of environment-dependent substitution tables, fourier-transform methods and helix capping rules. *Protein Eng.* 7, 645-653.

Eisenberg D.; Weiss R.M.; Terwilliger, T.C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA*. 8, 140-144.

Eisenberg, D.; Weiss, R.M.; Terwilliger, T.C.; Wilcox, W. (1982). Hydrophobic moments and protein-structure. *Faraday Symp. Chem. Soc.* 17, 109-120.

Ewing, T.A. & Kuntz, I.D. (1997). Critical evaluation of search algorithms for automated molecular docking and database screening. *J Comput. Chem.* 18, 1175-1189.

Farrens, D.L.; Altenbach, C.; Yang, K.; Hubbell, W.L. and Khorana, H.G. (1996) Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin. *Science.* 274, 768-770.

Floriano, W.B.; Vaidehi, N.; Zamanakos, G. and Goddard III, W.A. (2003). Virtual Ligand Screening of Large Molecule Databases using hierarchical docking protocol (HierVLS). *J. Med. Chem.* (accepted).

Floriano, W.B.; Vaidehi, N.; Singer, M.; Shepherd, G. and Goddard III, W.A. (2000). Molecular mechanisms underlying differential odor responses of a mouse olfactory receptor. *Proc. Natl Acad. Sci. USA.* 97, 10712-10716.

Greasley. P.J.; Fanelli, F.; Rossier, O.; Abuin, L.; Cotecchia, S. (2002). Mutagenesis and modelling of the alpha(1b)-adrenergic receptor highlight the role of the helix 3/helix 6 interface in receptor activation. *Molecular Pharmacology*. 61, 1025-1032.

Herzyk, P.; Hubbard, R.E. (1995). Automated method for modeling seven-helix transmembrane receptors from experimental data. *Biophys J.* 69, 2419-2442.

Jaguar v4.0, Schrodinger Inc. Portland, Oregon.

Jain, A.; Vaidehi, N. & Rodriguez, G. (1993). A fast recursive algorithm for molecular-dynamics simulation. *J. Comp. Phys*. 106, 258-268.

Kekenes-Huskey, P.M.; Vaidehi, N.; Floriano, W.B. and Goddard, W.A. (2003). Fidelity of phenyl alanyl tRNA synthetase, *J. Am. Chem. Soc*. (accepted).

Laskowski, R.A.; MacArthur, M.W.; Moss, D.S.; Thornton, J.M. (1993). Procheck - a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* 26, 283-291.

Lim, K-T.; Brunett, S.; Iotov, M.; McClurg, R.B.; Vaidehi, N.; Dasgupta, S.; Taylor, S. & Goddard III, W.A. (1997). Molecular dynamics for very large systems on massively parallel computers: The MPSim program. *J. Comput. Chem.* 18, 501-521.

Lin, S.W.; Sakmar, T.P. (1996). Specific tryptophan UV-absorbance changes are probes of the transition of rhodopsin to its active state. *Biochemistry.* 35, 11149-11159.

Lomize, A.L.; Poghozeva, I.D. and Mosberg, H.I. (1999). Structural Organization of G-protein-coupled receptors. *J. Comp. Aided. Mol. Design*. 13, 325-353.

MacKerell, A.D.; Bashford, D.; Bellott, M.; Dunbrack, R.L.; Evanseck, J.D.; Field, M.J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F.T.K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D.T.; Prodhom, B.; Reiher,

W.E.; Roux, B.; Schlenkrich, M.; Smith, J.C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D. & Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102, 3586-3616.

Malnic, B.; Hirono, J.; Sato, T. & Buck, L.B. (1999). Combinatorial receptor codes for odors. Cell, 96, 713-723.

Marti-Renom, M.A.; Stuart, A.C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Bioph. Biom.* 29, 291-325.

Mathiowetz, A.M.; Jain, A.; Karasawa, N.; & Goddard III, W.A. (1994). Protein simulations using techniques suitable for very large systems - the cell multipole method for nonbond interactions and the Newton-Euler inverse mass operator method for internal coordinate dynamics. *Proteins.* 20, 227-247.

Mayo, S. L.; Olafson, B.D. & Goddard III, W.A. (1990). Dreiding - A generic force-field for molecular simulations. *J. Phys. Chem.* 94, 8897-8909.

Melia, T.J.; Cowan, C.W.; Angleson, J.K.; Wensel, T.G. (1997). A comparison of the efficiency of G protein activation by ligand-free and light-activated forms of rhodopsin. *Biophys. J.* 73, 3182-3191.

Okada T, Ernst O.P.; Palczewski K.; Hofmann K.P. (2001). Activation of rhodopsin: new insights from structural and biochemical studies. *Trends Biochem. Sci*. 26, 318-324.

Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C.; Motoshima, H.; Fox, B.; Trong, I.; Teller, D.; Okada, T.; Stenkamp, R.; Yamamoto, M.; Miyano, M. (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. *Science.* 289, 739-745.

Rappé, A.K. & Goddard III, W.A. (1991). Charge equilibration for molecular-dynamics simulations. *J. Phys. Chem.* 95, 3358-3363.

Saam, J.; Tajkhorshid, E.; Hayashi, S.; Schulten, K. (2002). Molecular dynamics investigation of primary photoinduced events in the activation of rhodopsin. *Biophys. J.* 83, 3097-3112.

Schertler, G.F.X. (1998). Structure of rhodopsin. *Eye*. 12, 504-510.

Schoneberg, T.; Schulz, A and Gudermann T. (2002). The structural basis of G-protein-coupled receptor function and dysfunction in human diseases. *Rev. Phys. Biochem. Pharm.* 144, 145-227.

Shacham, S.; Topf, M.; Avisar, N.; Glaser, F.; Marantz, Y.; Bar-Haim, S.; Noiman, S.; Naor, Z., Becker, O.M. (2001). Modeling the 3D structure of GPCRs from sequence. *Medicinal Research Reviews.* 21, 472-483.

Spassov, V.Z.; Yan, L.; Szalma, S. (2002). Introducing an implicit membrane in generalized Born solvent accessibility continuum solvent models. *J. Phys. Chem. B.* 106, 8726-8738.

Strader, C.D.; Fong, T.M.; Tota, M.R.; Underwood, D. and Dixon, R.A. (1994). Structure and function of G-protein-coupled receptors. *Annu. Rev. Biochem.* 63, 101-132.

Strange, P.G. (1998). Three-state and two-state models. *Trends Pharm. Sci.* 19, 85-86 .

Teller, D.; Okada, T.; Behnke, C.,Palczewski, K.; Stenkamp, R. (2001). Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of G-protein-coupled receptors (GPCRs). *Biochemistry.* 40, 7761-7772.

Thompson, J.D.; Higgins, D.G.; Gibson, T.J. (1994). Clustal-W - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting,

position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673-4680.

Vaidehi, N.; Floriano, W.B.; Trabanino, R.; Hall, S.E.; Freddolino, P.; Choi, E.J.; Zamanakos, G. and Goddard, W.A. (2002). Prediction of structure and function of G protein-coupled receptors**.** *Proc. Natl Acad. Sci. USA*.

Vaidehi, N.; Jain, A. & Goddard III, W.A. (1996). Constant temperature constrained molecular dynamics: The Newton-Euler inverse mass operator method. *J. Phys. Chem.* 100, 10508-10517.

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

Wallin, E.; von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* 7, 1029-1038.

Wang, P.; Vaidehi, N.; Tirrell, D.A.; Goddard III, W.A. (2002). Virtual screening for binding of phenylalanine analogues to phenylalanyl-tRNA synthetase. J. *Am. Chem. Soc.* 124, 14442-14449.

Wilson, S.; Bergsma, D. (2000). Orphan G-protein-coupled receptors: novel drug targets for the pharmaceutical industry. *Drug Des Discov.* 17, 105-14.

.

Table 1: The last column shows the positions of the hydrophobic center (HC) predicted for each TM by TM2ndS for various window sizes. The first row (shaded in gray) has the window sizes chosen to calculate this hydrophobic center. Here position 1 corresponds to the first residue in the capped sequence in Figure 2.

Helix number \ Window size	12	14	**16**	**18**	**20**	**22**	**24**	26	28	30	HC
1	15	13	20	18	18	17	18	15	16	13	18.2
2	20	12	12	14	15	15	14	22	19	20	14.0
3	19	20	17	18	15	16	15	12	11	12	16.2
4	9	9	10	15	12	13	13	12	11	17	12.6
5	15	19	13	12	14	16	16	17	16	15	14.2
6	8	9	11	11	13	14	14	15	16	17	12.6
7	19	4	17	15	14	14	13	12	11	10	14.6

Table 2: Results from the coarse-grain docking step of HierDock to predict the binding site (s) in apo/closed(xray). The energies of the top 5% after ranking (level 2 of HierDock) are shown for each box. Among all boxes, the best coarse-grain score is underlined. The scores within 100 kcal/mol of the top score are shown in bold.

Box	Top 5 % after coarse-grain ranking
1	2596, 2941, 2991, 3011, 4281
2	4440, 4621, 4625, 5509, 5513
3	2338, 2375, 2409, 2566, 2571
4	5844, 5961, 6006, 6244, 6278
5	None passed buried surface criteria
6	**102, 118, 131, 136,** 208
7	1366, 1370, 1374, 1374, 1379
8	No conformations generated from DOCK
9	12026
10	**82, 139, 153,** 377, 380
11	2348, 2348, 2566, 2843, 2843
12	No conformations generated from DOCK
13	551, 734, 931, 1110, 1226

Figure Legends

Figure 1: Hydrophobicity profile from TM2ndS for bovine rhodopsin at window size WS=14. The pink line (at 0.07) is the base_mod (described in 2.2.1b) used as the baseline in identifying hydrophobic regions. The predicted TM domains are indicated by the yellow lines (after capping). The cyan lines show the predictions prior to helix capping. Each tick mark indicates the sequence number for the alignment based on bovine rhodopsin (100 residues per panel). The residues at every 5th position are indicated below each panel. The partition of helix 7 into two parts results from the hydrophilic residues near its center.

Figure 2: The transmembrane helical predictions (labeled as after capping) from TM2ndS compared with helix ranges from the bovine rhodopsin xray crystal structure. In addition the predictions before TM2ndS capping are shown. Those residues in the crystal structure which fall outside the range of alpha helicity (using analysis described in Sec. 2.2.1c) are indicated in lowercase letters.

Figure 3: The RMS deviation for various window sizes (WS) of the central residues predicted from TM2ndS for Bovine rhodopsin compared to the best fit plane to the crystal structure minimized without ligand, *Apo/closed(xtal).* This suggests that the best WS is 16 to 22.

Figure 4: Schematic for a possible signaling mechanism in rhodopsin. Note that the movement of helix 3 (caused by interaction with the trans isomer of retinal) exposes the

DRY sequence to G-protein activation and as a result closes the EC2 loop to maintain the ligand inside the bundle sequence.

Figure 5: The thirteen regions shown as boxes used in scanning the entire protein for the 11cis-retinal putative binding site. The 2 boxes chosen as binding sites by HierDock are shown in red. (a) front view with N-terminus at the bottom (b) top view obtained by rotating by 90 degrees about the horizontal-axis in a) so that the N terminus is out of the page. These two orientations are used for all structures shown in this paper.

Figure 6abcd: Validation of HierDock: (a) Front view of the 11cis-retinal conformation determined by HierDock for ret(HD)/closed(xray) (CPK color) compared to the published crystal structure (green). The CRMS difference in the ligand structures is 1.1 Å. (b) Top view of Figure a. This shows that predicted position of the retinal aldehyde oxygen is 2.8 Å from the N of LYS296, which is short enough for an H-bond. (c) Top view showing the result of making the Schiff base bond of 11cis-retinal to LYS296 in Figure a and minimizing the resulting structure (blue), compared with the crystallographic ligand structure (red). The CRMS difference between these ligand structures is 0.62 Å. (d) Top view of Figure c.

Figure 7: (a) Comparison of the predicted structure (orange) Apo/closed(MS) with the experimental structure (blue) apo/closed(xray). They differ in the TM helical region by CRMS=2.84 Å.

(b) Comparison of the predicted Apo/closed(MS) structure (orange) with the predicted Apo/open(MS) structure (cyan). They differ in the TM helical region by 0.11 Å.

Figure 8: MembStruk validation using the closed EC-II loop a) The HierDock predicted conformation of 11cis-retinal (CPK color) in the MembStruk predicted apo/closed(MS) structure, denoted NoSB-ret(HD)/closed(MS). Note that the aldehyde oxygen is 2.85 Å from the N of LYS296. b) In violet: The retinal structure after forming this Schiff base bond of 11cis-retinal to Lys296 and optimizing to form ret(HD)/closed(MS). In blue: retHD/closed(xtal). These ligand structures were found to differ by 2.9 Å CRMS.

Figure 9abcd: Comparison of predicted binding sites for retinal [those residues within 5 Å of retinal which interact strongly with the ligand (contributions to binding greater than 1 kcal/mol)] *before* Schiff base bond formation in the three rhodopsin structures.

a) All three structures and ligand conformations are shown. The colors blue, grey, orange correspond respectively to those structures analyzed in b-d.
b) NoSB-Ret(HD)/closed(xtal) structure. Here we see that seven residues bind more strongly than 1 kcal/mol.
c) NoSB-Ret(HD)/ closed (MS). Here we see that five of the seven residues in Figure (b) are predicted (only Phe208 and Hsp211, both rather weakly bound). We also find three additional residues (Phe212, Ile275, Ala117) that do not bind with 1 kcal/mol in b.
d) NoSB-Ret(HD)/open(MS). Here we see that six of the seven residues in Figure (c) bind more strongly than 1 kcal/mol. We also find four additional residues that

do not bind with 1 kcal/mol in b. This difference results from the shift in the retinal binding site upon closure of the EC-II loop.

The side chains in common with the NoSB-Ret(HD)/closed(xtal) structure (in Figure b) or with NoSB-Ret(HD)/closed(MS) (in Figure c) within the binding site around the 11cis-retinal are labeled with larger font.

Figure 10abc: Comparison of predicted binding sites of retinal with Schiff base bond formed. We considered residues within a 5 Å shell of the ligand (excluding the Lys296 to which the retinal is bound) and determined those which contribute at least 1 kcal/mol of stabilization energy for the three rhodopsin structures.

a) Ret(xtal)/closed(xtal) structure. Here we see that 15 residues bind more strongly than 1 kcal/mol.
b) Ret(HD)/closed(xtal). Here we see that 12 of the 15 residues in Figure (a) are predicted to bind strongly (Ala117 and His211 still contribute positively to bonding but are now rather weakly bound [<1 kcal/mol]). We find 2 additional residues (Cys187, Ala269) that did not bind with 1 kcal/mol in a.
c) Ret(HD)/ closed (MS). Here we find 8 of the 15 residues in Figure (a) still bind strongly. We also find 2 additional residues (Ile275, Ala269) that did not bind with 1 kcal/mol in a.

A larger font is used to label the side chains in common with the Ret(xtal)/closed(xtal) structure within the binding site around the 11cis-retinal.

Figure 11: MembStruk validation using the open EC-II loop:

a) The HierDock predicted conformation (CPK color) of 11cis-retinal in the MembStruk predicted structure to form the NoSB-Ret(HD)/open(MS) structure. Note that the aldehyde oxygen is 2.87 Å from the N of LYS296; which is short enough to form a hydrogen bond.

b) The Ret(HD)/open(MS) structure after forming the Schiff base bond (green), compared with the structure (violet) of 11cis-retinal in ret(HD)/closed(MS).These ligand structures differ by 1.7 Å CRMS. The EC-II loop may function to position the retinal ligand into its final conformation as found in the rhodopsin crystal structure.

Figure 1:

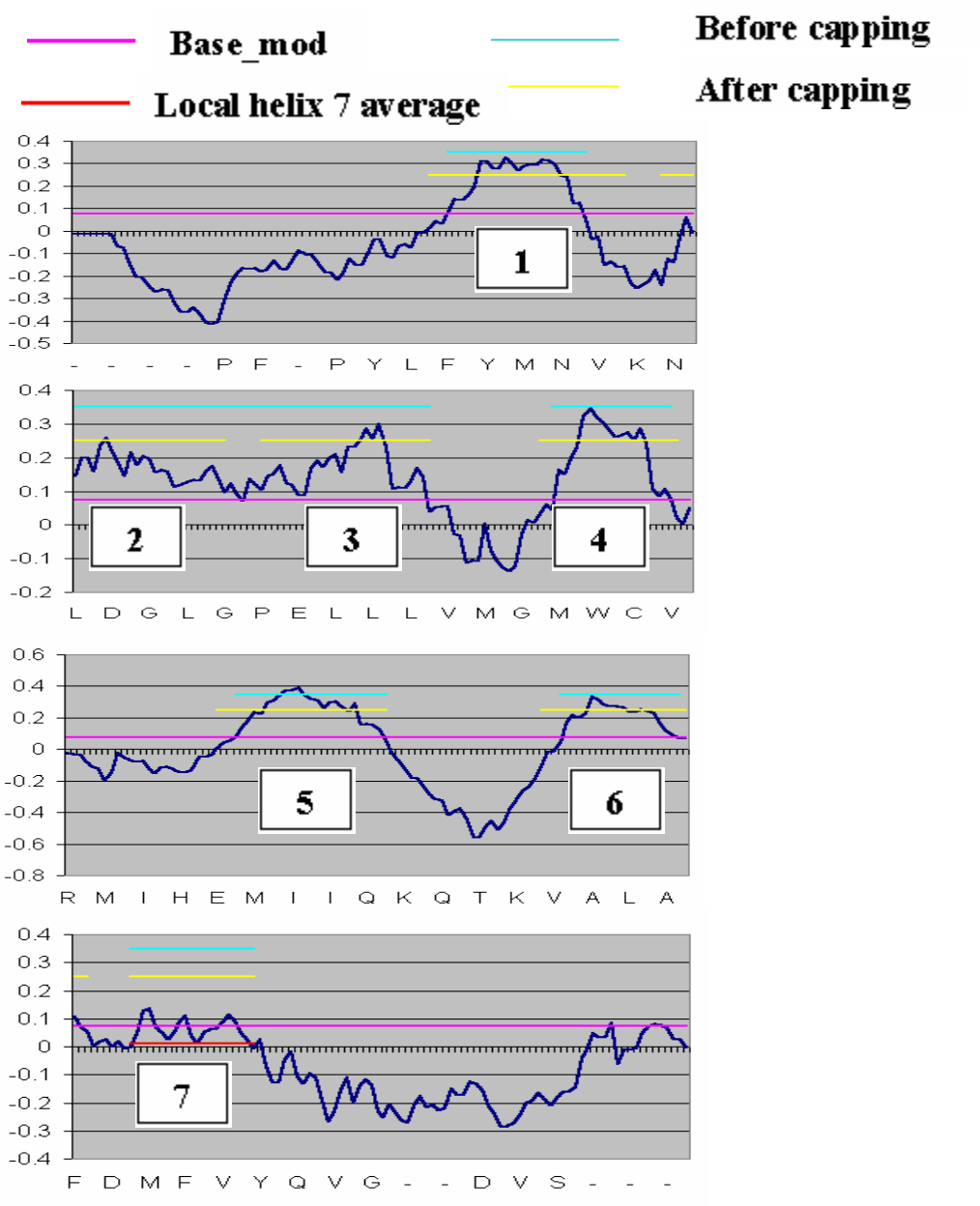

Base_mod
Local helix 7 average
Before capping
After capping
1
2
3
4
5
6
7

Figure 2:

		Size	Range
TM1:			
FSMLAAYMFLLIMLGFPINFLTL	Before capping	23	37-59
PWQFSMLAAYMFLLIMLGFPINFLTLYVTVQH	After capping	32	34-65
WQFSMLAAYMFLLIMLGFPINFLTLYVTVQ	Crystal	30	35-64
TM2:			
LNLAVADLFMVFGGFTTTLYTSLHGYFV	Before capping	28	77-104
PLNYILLNLAVADLFMVFGGFTTTLYTSLHG	After capping	31	71-101
PLNYILLNLAVADLFMVFgGFTtTLYTSLh	Crystal	30	71-100
TM3:			
VFGPTGCNLEGFFATLGGEIALWSLVVLAIE	Before capping	31	104-134
PTGCNLEGFFATLGGEIALWSLVVLAIE	After capping	28	107-134
PTGCNLEGFFATLGGEIALWSLvVLAIERYvvV	Crystal	33	107-139
TM4:			
IMGVAFTWVMALACAAPPLV	Before capping	20	154-173
HAIMGVAFTWVMALACAAPPLVG	After capping	23	152-174
NHAIMGVAFTWVmaLacAAPPLV	Crystal	23	151-173
TM5:			
VIYMFVVHFIIPLIVIFFCYGQLVF	Before capping	25	204-228
ESFVIYMFVVHFIIPLIVIFFCYGQLVF	After capping	28	201-228
NESFVIYMFVvhFIIPLIVIFFCYgq	Crystal	26	200-225
TM6:			
IIMVIAFLICWLPYAGVAFY	Before capping	20	255-274
RMVIIMVIAFLICWLPYAGVAFYIFTH	After capping	27	252-278
ekEVTRMVIIMVIAFLICwLPYAGVAFYIFT	Crystal	31	247-277
TM7:			
PIFMTIPAFFAKTSAVYNPVI	Before capping	21	285-305
PIFMTIPAFFAKTSAVYNPVI	After capping	21	285-305
IFmTIPAFFAKTSavYNPVIY	Crystal	21	286-306

Figure 3:

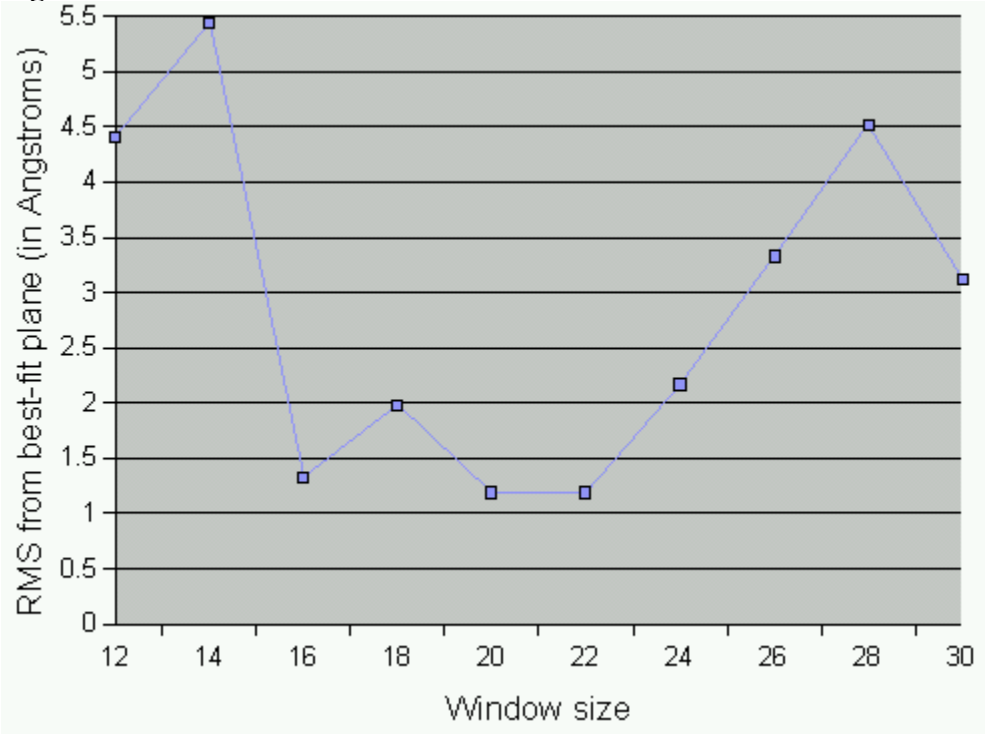

RMS from best-fit plane (in Angstroms)
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
12
14
16
18
20
22
24
26
28
30
Window size

Figure 4:

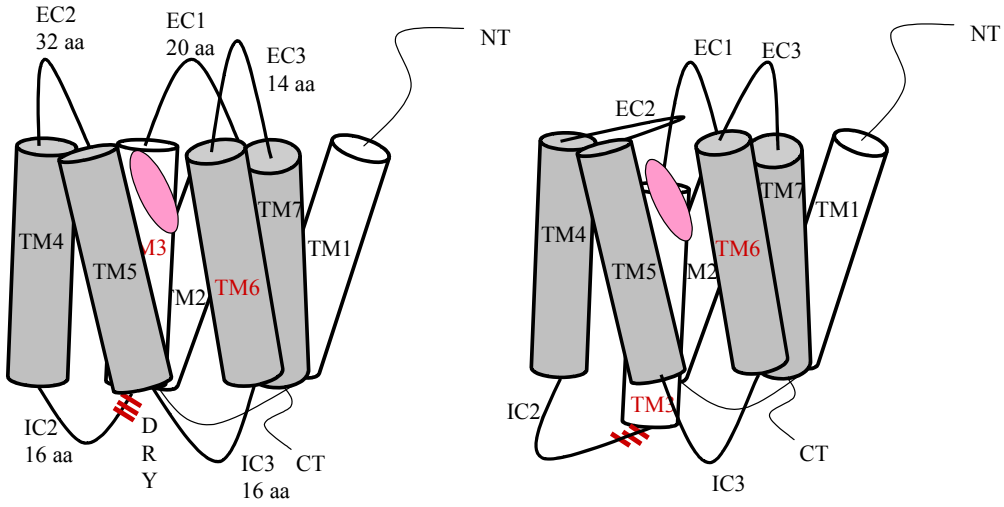

EC2
32 aa
EC1
20 aa
EC3
14 aa
NT
TM4
TM5
TM3
TM2
TM6
TM7
TM1
IC2
16 aa
D
R
Y
IC3
16 aa
CT
EC1
EC3
NT
EC2
TM7
TM1
TM4
TM6
TM5
M2
IC2
TM3
CT
IC3

Figure 5:

Figure 5a)

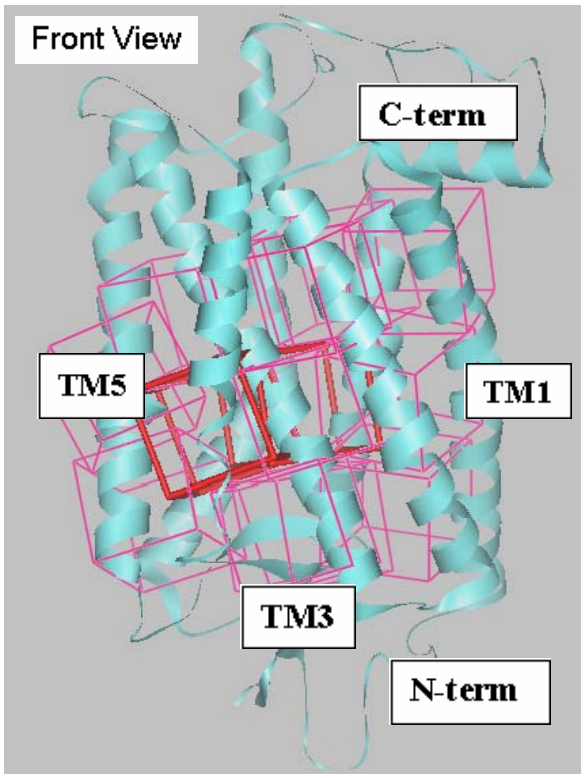


Figure 5b)

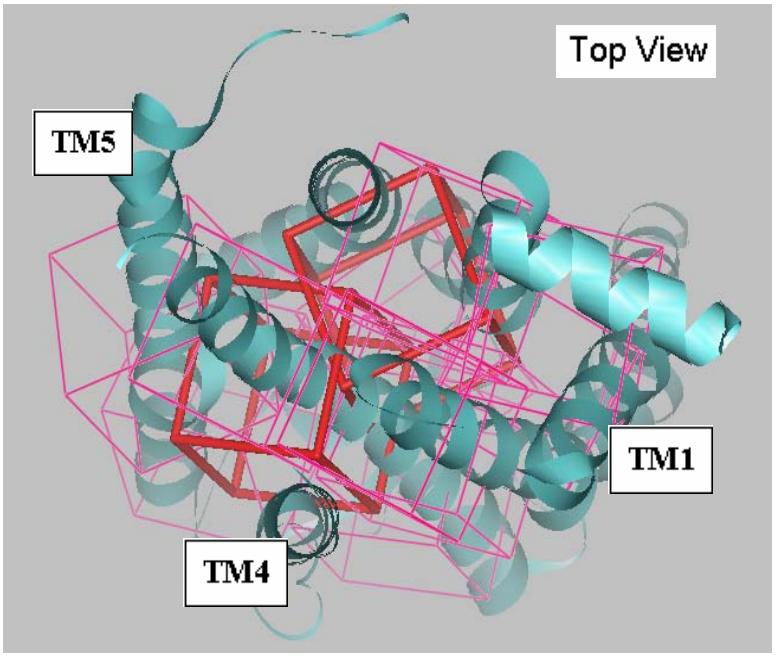


Figure 6abcd:

Figure 6a)

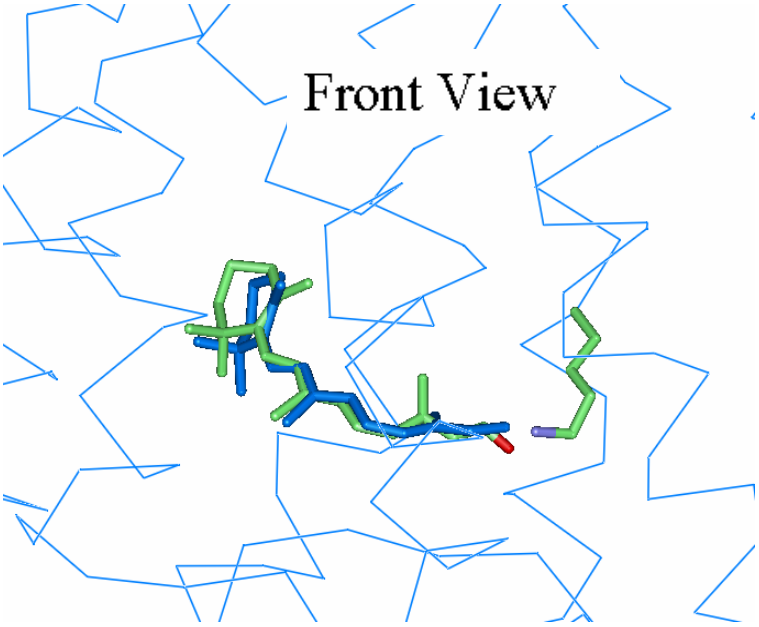


Figure 6b)

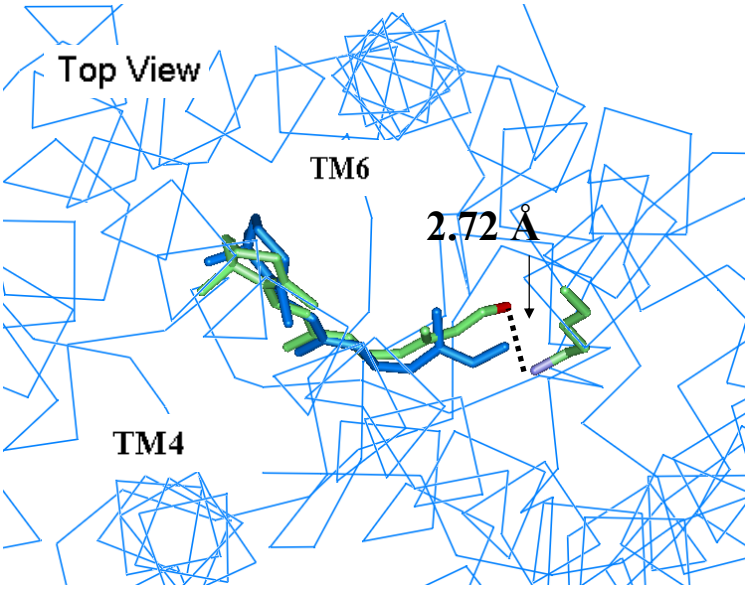


Figure 6c)

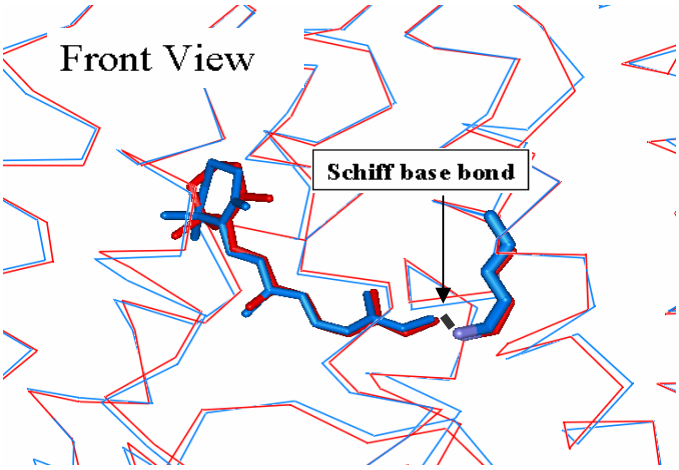


Figure 6d)

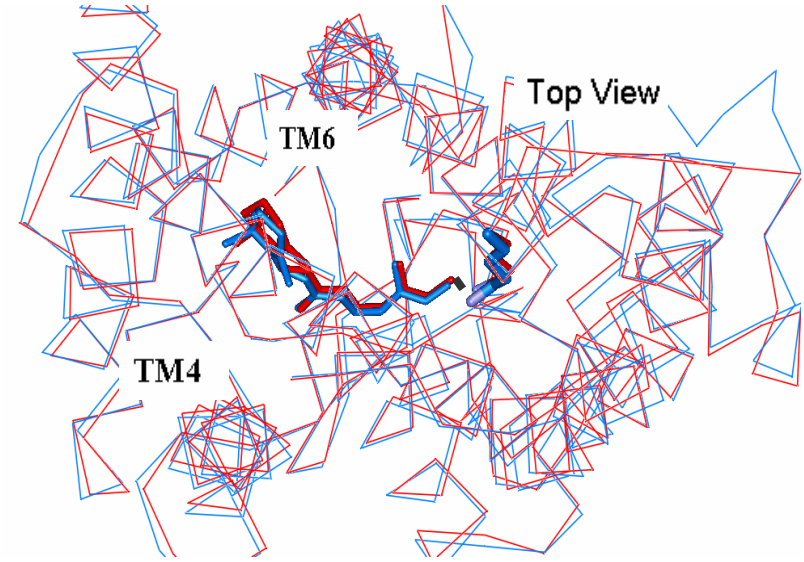


Figure 7:

Figure 7a)

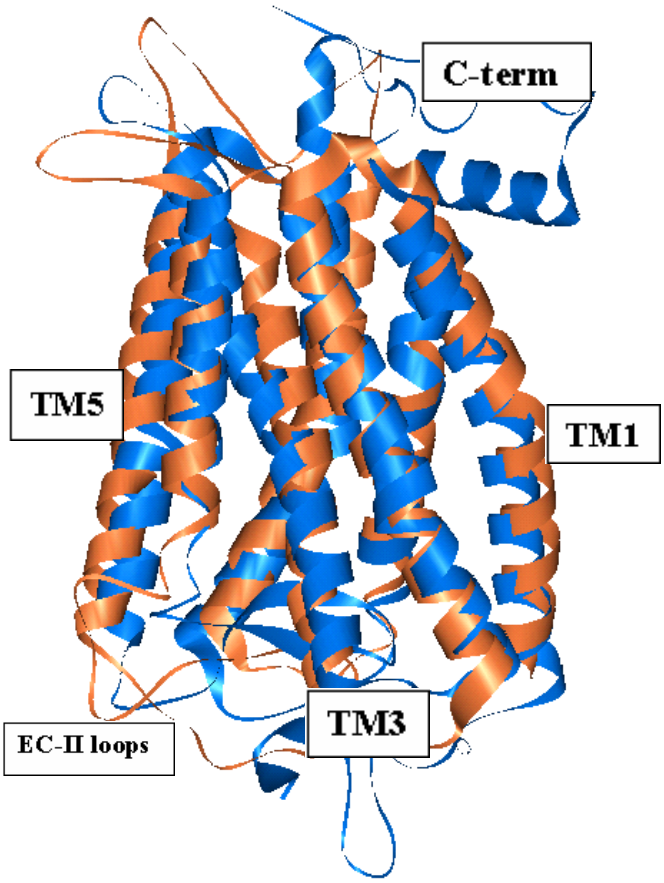


Figure 7b)

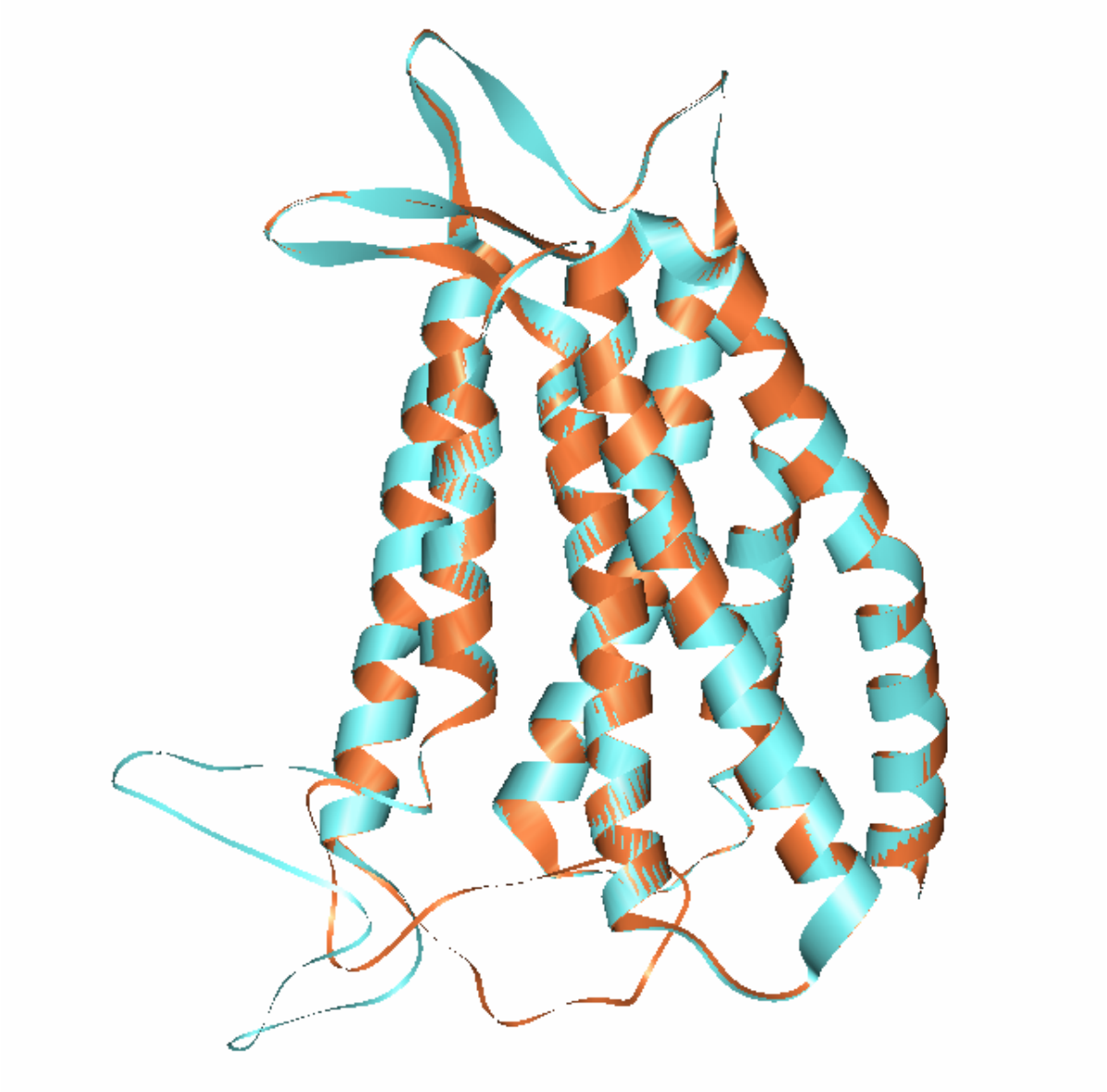


Figure 8:

Figure 8a)

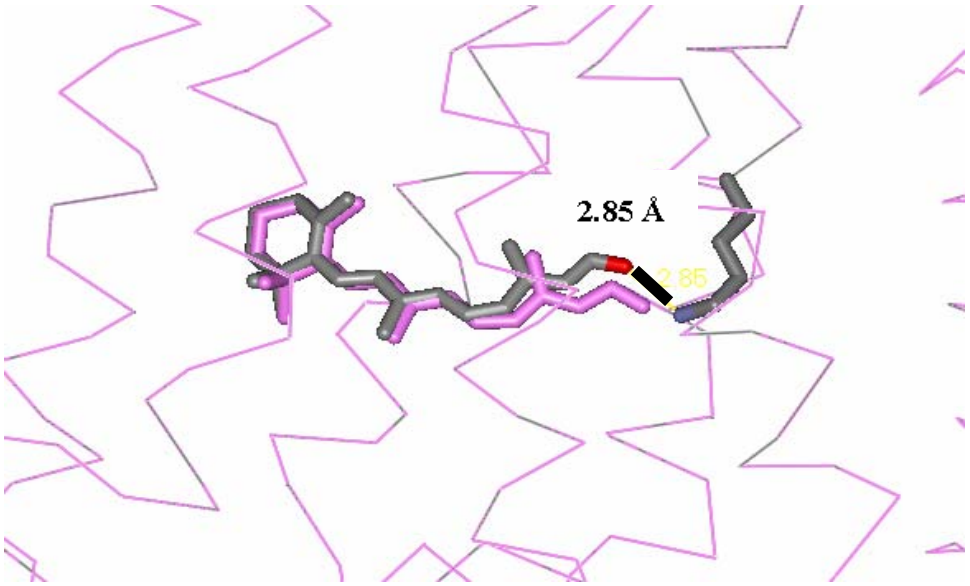


Figure 8b)

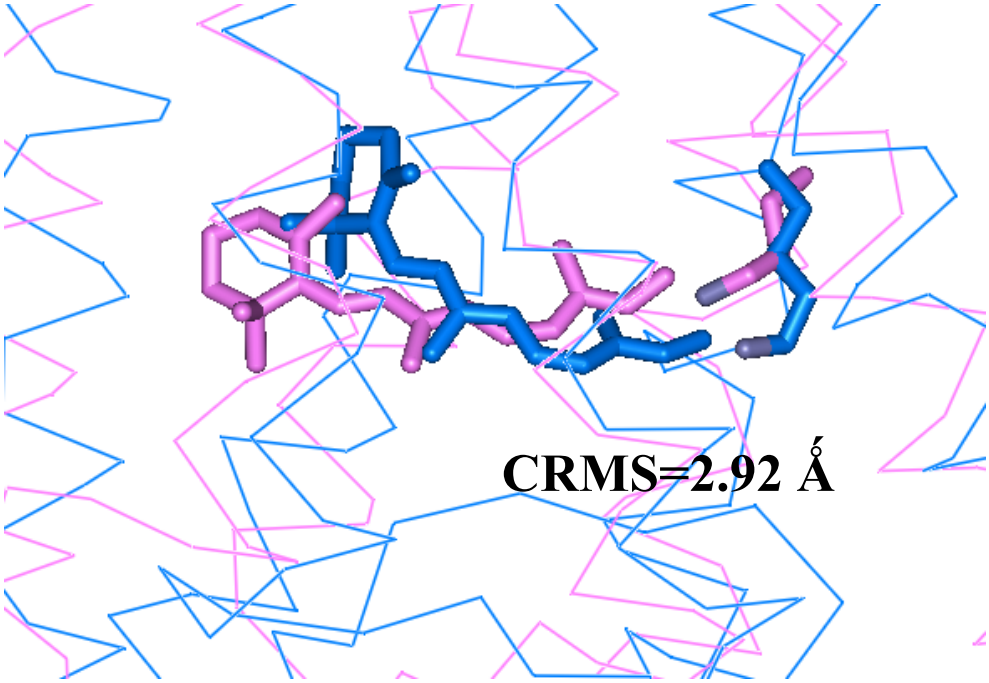


Figure 9abcd:

Figure 9a)

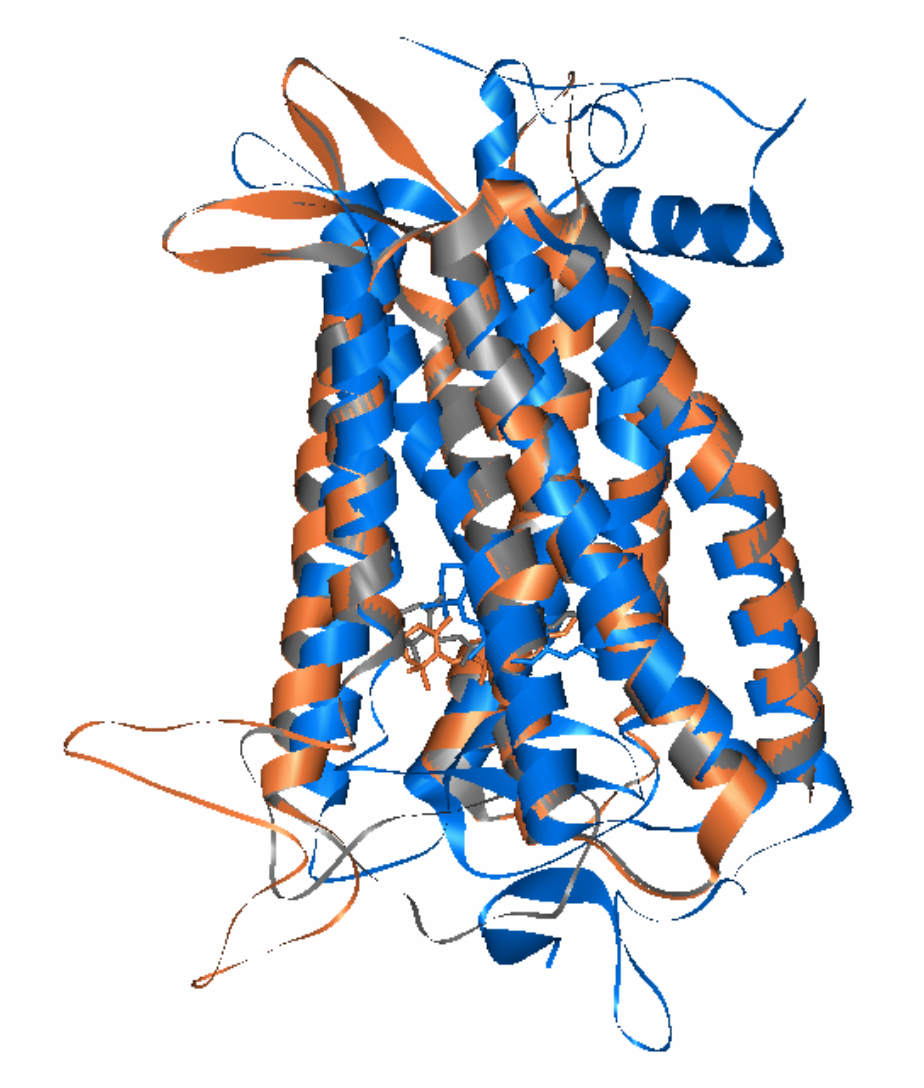


Figure 9b)

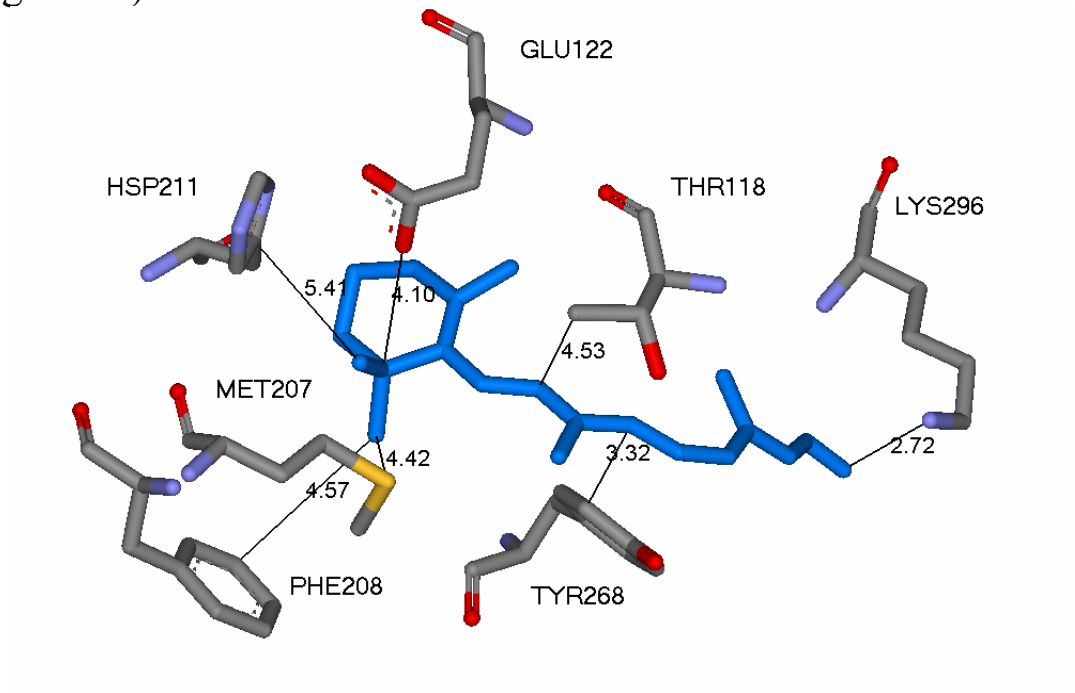


Figure 9c)

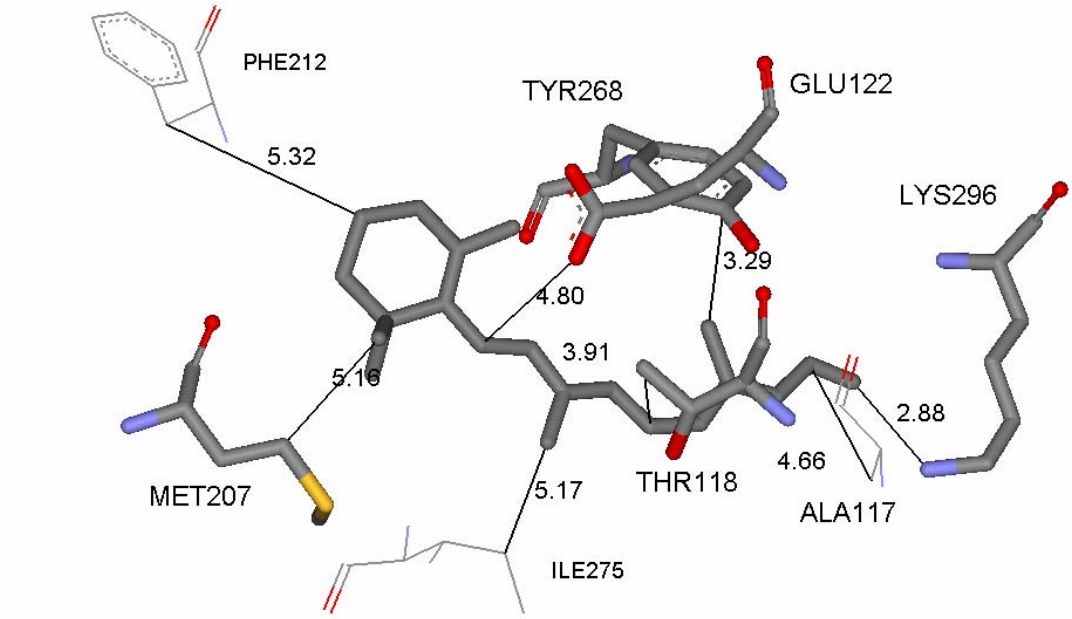


Figure 9d)

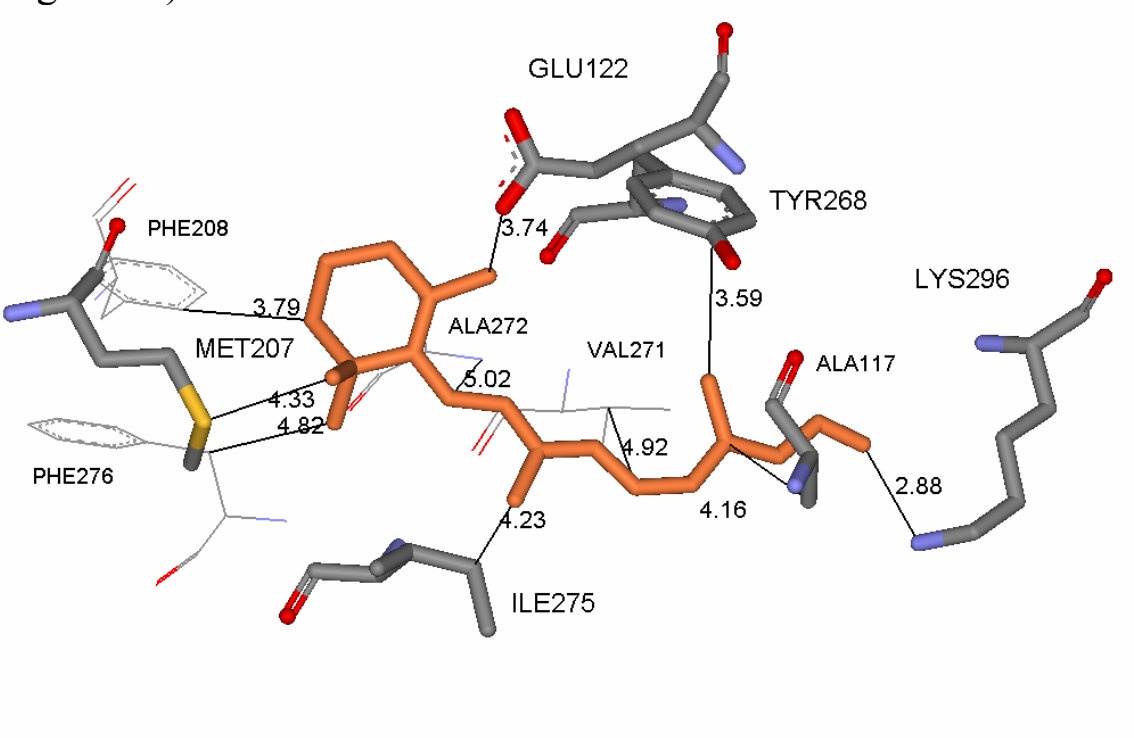

GLU122
TYR268
PHE208
3.74
3.59
LYS296
3.79
ALA272
MET207
VAL271
ALA117
5.02
4.33
4.82
4.92
PHE276
2.88
4.23
4.16
ILE275

Figure 10abc:

Figure 10a)

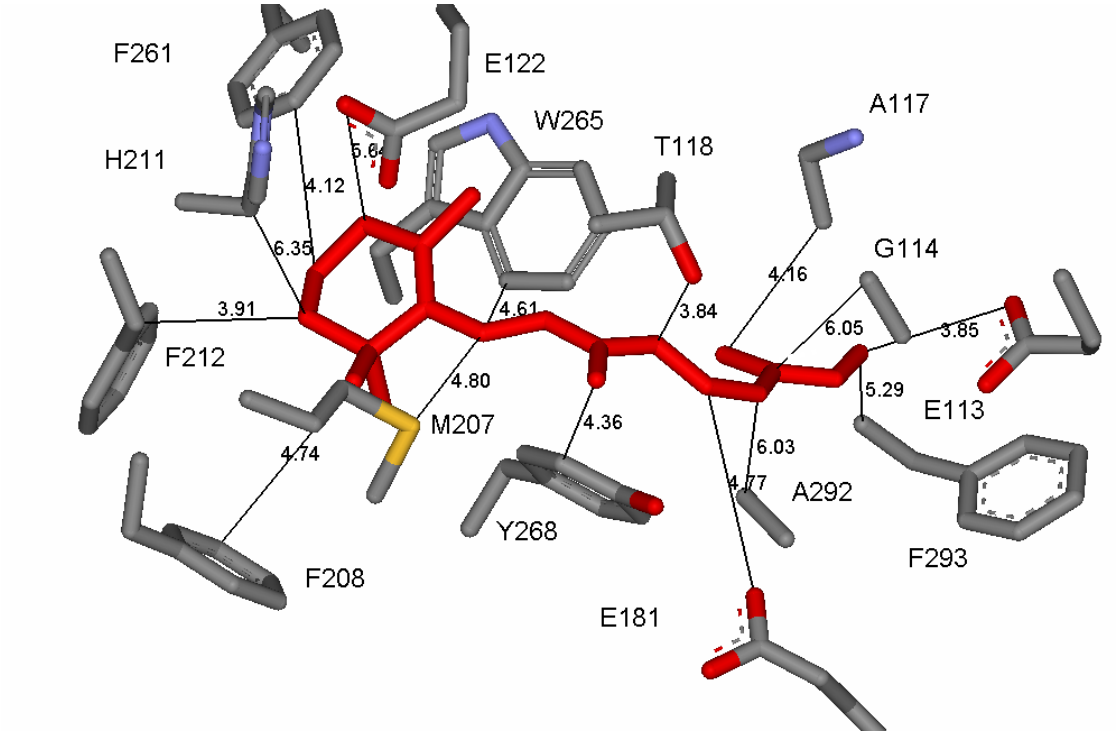


Figure 10b)

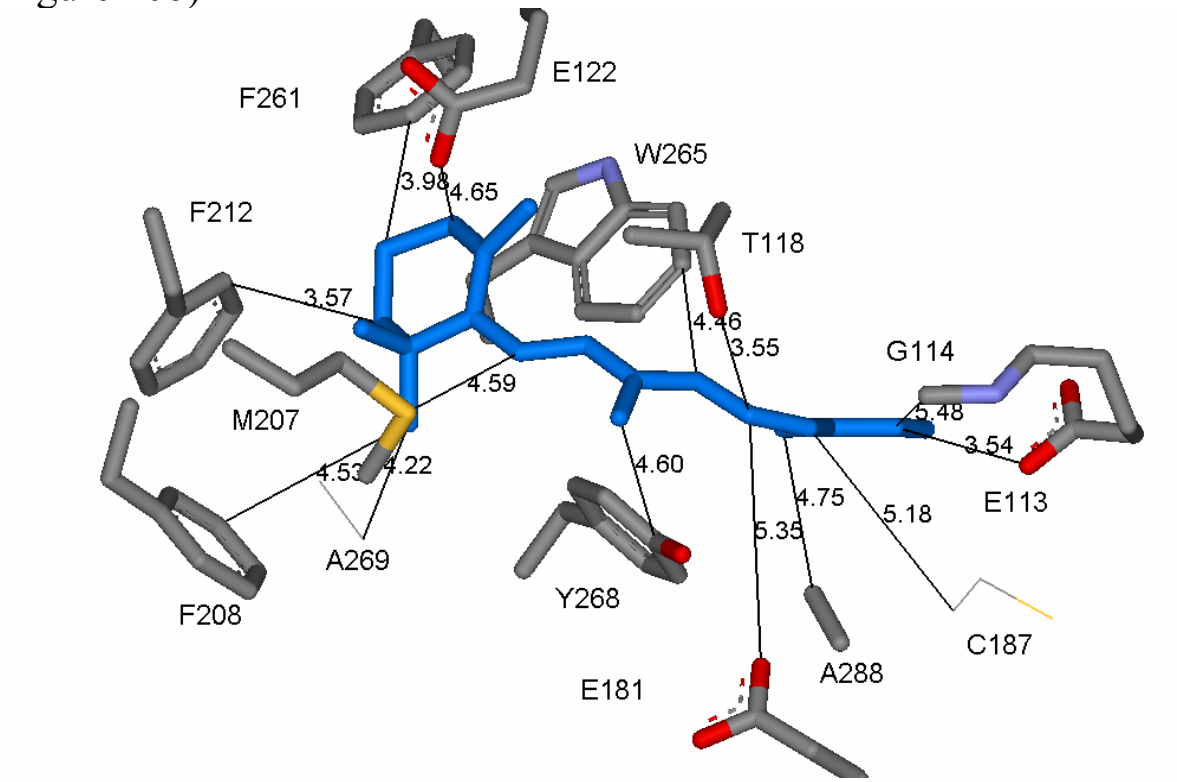

F261
E122
W265
3.98
4.65
F212
T118
3.57
4.46
3.55
G114
4.59
M207
5.48
3.54
4.53
4.22
4.60
4.75
5.18
E113
5.35
A269
Y268
C187
F208
A288
E181

Figure 10c)

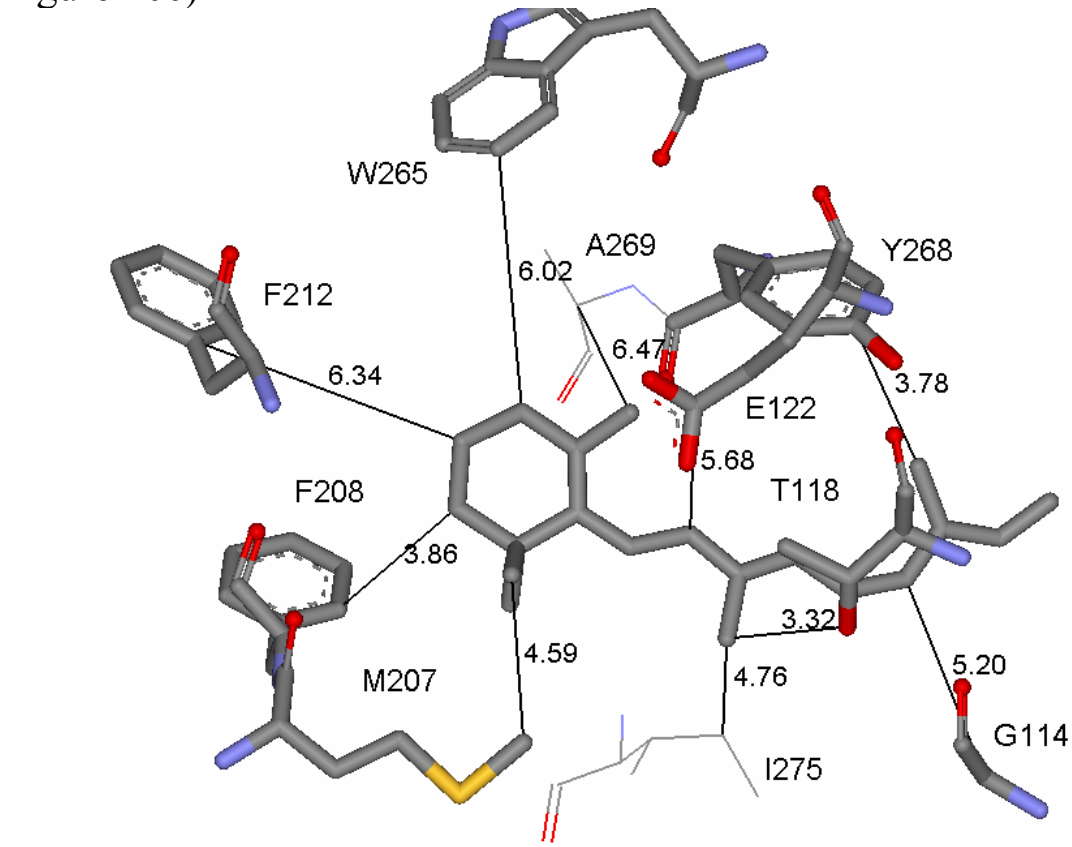

W265
A269
Y268
F212
6.02
6.47
6.34
3.78
E122
5.68
F208
T118
3.86
3.32
4.59
M207
4.76
5.20
G114
I275

Figure 11:

Figure 11a)

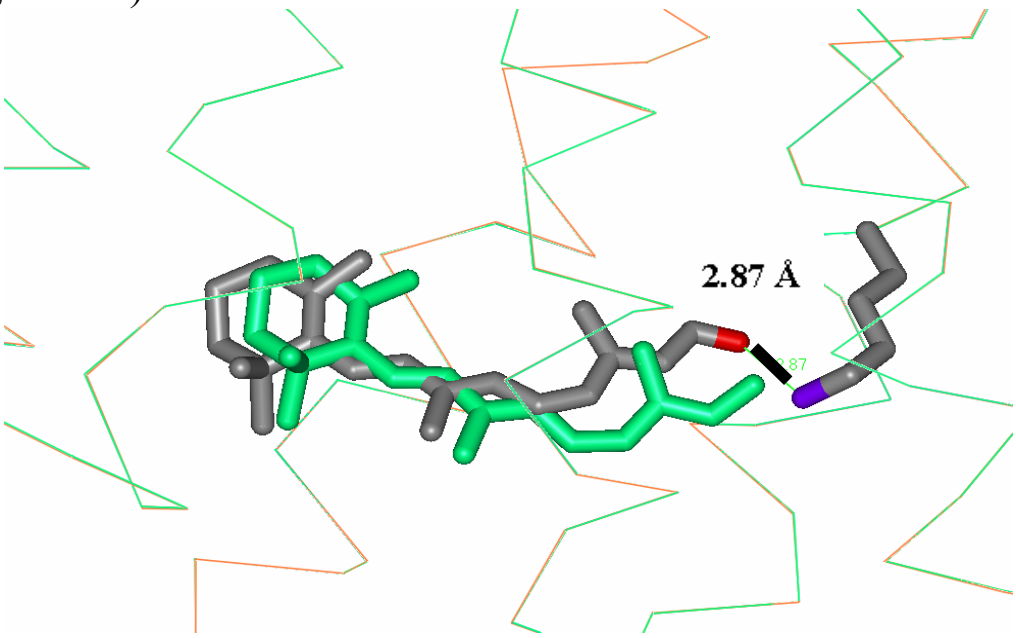


Figure 11b)

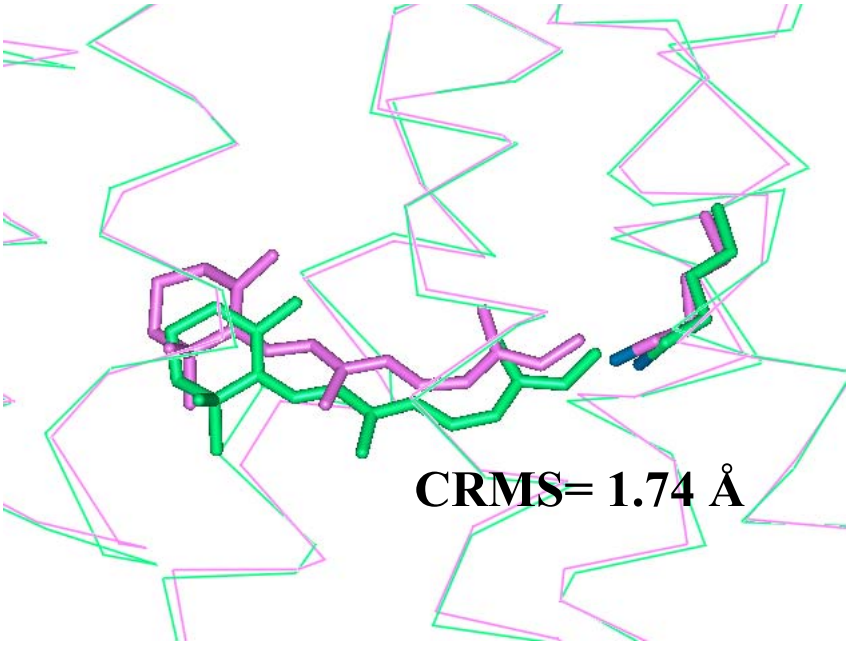


Figure S1: The sequences (43 Blast entries shown) used to generate the multiple sequence alignment with bovine rhodopsin.

```
sp|P28682|OPSB_CHICK Blue-sensitive opsin (Blue cone photorecept...
357   2e-98
sp|P51472|OPSB_ASTFA BLUE-SENSITIVE OPSIN (BLUE CONE PHOTORECEPT...
347   2e-95
sp|P32310|OPSB_CARAU Blue-sensitive opsin (Blue cone photorecept...
337   2e-92
sp|P87365|OPSB_ORYLA BLUE-SENSITIVE OPSIN (BLUE CONE PHOTORECEPT...
330   2e-90
sp|Q9W6A8|OPSB_BRARE Blue-sensitive opsin (Blue cone photorecept...
324   2e-88
sp|P51473|OPSV_XENLA Violet-sensitive opsin (Violet cone photore...
307   2e-83
sp|P51475|OPSP_CHICK Pinopsin (Pineal opsin) (P-opsin) (Pineal g...
302   6e-82
sp|P28684|OPSV_CHICK Violet-sensitive opsin (Violet cone photore...
300   2e-81
sp|P51476|OPSP_COLLI PINEAL OPSIN (P-OPSIN) (PINEAL GLAND-SPECIF...
295   1e-79
sp|O13092|OPSB_SAIBB Blue-sensitive opsin (Blue cone photorecept...
290   3e-78
sp|P51491|OPSB_MOUSE Blue-sensitive opsin (Blue cone photorecept...
290   4e-78
sp|Q90309|OPSU_CARAU ULTRAVIOLET-SENSITIVE OPSIN (ULTRAVIOLET CO...
287   2e-77
sp|P03999|OPSB_HUMAN Blue-sensitive opsin (Blue cone photorecept...
285   1e-76
sp|Q9W6A9|OPSU_BRARE Ultraviolet-sensitive opsin (Ultraviolet co...
283   3e-76
sp|Q63652|OPSB_RAT Blue-sensitive opsin (Blue cone photoreceptor...
283   5e-76
sp|P51490|OPSB_BOVIN Blue-sensitive opsin (Blue cone photorecept...
277   3e-74
sp|P87368|OPSV_ORYLA PUTATIVE VIOLET-SENSITIVE OPSIN (VIOLET CON...
276   4e-74
sp|O42490|OPSP_PETMA PINEAL OPSIN (P-OPSIN) (PINEAL GLAND-SPECIF...
273   3e-73
sp|P04001|OPSG_HUMAN Green-sensitive opsin (Green cone photorece...
269   5e-72
sp|Q95170|OPSR_CAPHI Red-sensitive opsin (Red cone photoreceptor...
268   2e-71
sp|O18913|OPSR_FELCA Red-sensitive opsin (Red cone photoreceptor...
267   3e-71
sp|P35358|OPSG_GECGE Green-sensitive opsin P521 (Green photorece...
266   4e-71
sp|P87367|OPSR_ORYLA RED-SENSITIVE OPSIN (RED CONE PHOTORECEPTOR...
266   6e-71
sp|O18910|OPSG_RABIT Green-sensitive opsin (Green cone photorece...
265   8e-71
sp|P04000|OPSR_HUMAN Red-sensitive opsin (Red cone photoreceptor...
265   1e-70
sp|O35476|OPSG_RAT Green-sensitive opsin (Green cone photorecept...
264   2e-70
```

```
sp|P34989|OPSL_CALJA Opsin, longwave 563 nm
263   3e-70
sp|Q9W6A7|OPSR_BRARE Red-sensitive opsin (Red cone photoreceptor...
263   4e-70
sp|O35478|OPSG_SCICA Green-sensitive opsin (Green cone photorece...
262   8e-70
sp|Q9R024|OPSG_CAVPO Green-sensitive opsin (Green cone photorece...
262   8e-70
sp|O35599|OPSG_MOUSE Green-sensitive opsin (Green cone photorece...
261   1e-69
sp|P32313|OPSR_CARAU Red-sensitive opsin (Red cone photoreceptor...
259   7e-69
sp|P22332|OPSR_ASTFA RED-SENSITIVE OPSIN (RED CONE PHOTORECEPTOR...
258   9e-69
sp|P41592|OPSR_ANOCA Red-sensitive opsin (Red cone photoreceptor...
258   2e-68
sp|O12948|OPSR_XENLA RED-SENSITIVE OPSIN (RED CONE PHOTORECEPTOR...
257   2e-68
sp|P22329|OPSR_CHICK Red-sensitive opsin (Red cone photoreceptor...
253   3e-67
sp|P22330|OPSG_ASTFA Green-sensitive opsin 1 (Green cone photore...
246   6e-65
sp|P22331|OPSH_ASTFA Green-sensitive opsin 2 (Green cone photore...
244   1e-64
sp|O42266|OPSP_ICTPU PARAPINOPSIN
229   6e-60
sp|O18912|OPSR_HORSE Red-sensitive opsin (Red cone photoreceptor...
222   7e-58
sp|O18911|OPSG_ODOVI Green-sensitive opsin (Green cone photorece...
222   7e-58
sp|O18914|OPSR_CANFA Red-sensitive opsin (Red cone photoreceptor...
218   1e-56
sp|O13018|OPSO_SALSA Vertebrate ancient opsin
218   2e-56
```

Table S1: Residues within a 5 Å shell (of retinal without Schiff base bond formed) which interact more favorably than an interaction energy of –1 kcal/mol with the ligand are shown in order of decreasing interaction. The interaction energy values are shown in parentheses.

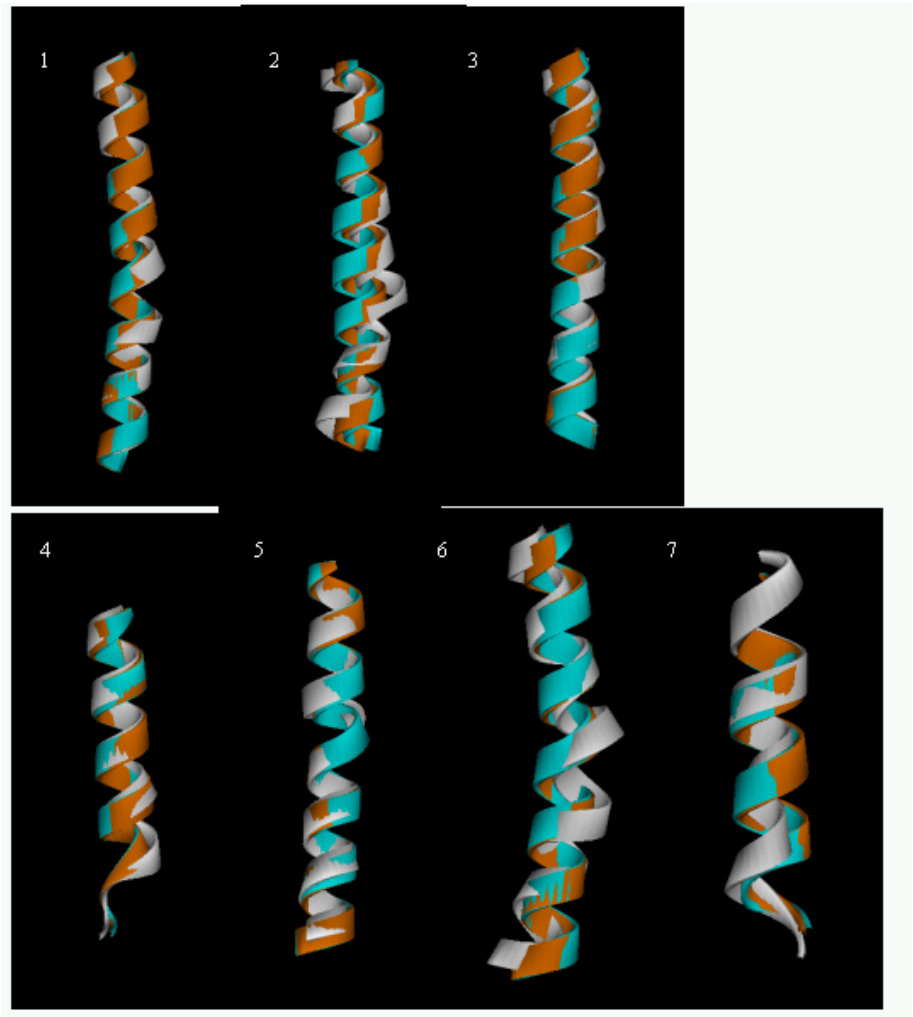
NoSB-Ret(HD)/closed(xtal)	NoSB-Ret(HD)/ closed (MS)	NoSB-Ret(HD)/open(MS)
Thr118 (-4.6)	Lys296 (-9.6)	Lys296 (-7.3)
Tyr268 (-3.9)	Glu122 (-3.0)	Tyr268 (-3.4)
Lys296 (-2.6)	Tyr268 (-2.6)	Glu122 (-2.9)
Hsp211 (-1.8)	Ala117 (-1.9)	Ala272 (-2.7)
Glu122 (-1.4)	Thr118 (-1.9)	Ile275 (-2.5)
Phe208 (-1.4)	Ile275 (-1.6)	Ala117 (-2.2)
Met207 (-1.2)	Met207 (-1.1)	Val271 (-1.9)
	Phe212 (-1.0)	Met207 (-1.4)
		Phe208 (-1.3)
		Phe276 (-1.1)

Table S2: Residues within a 5 Å shell (of retinal with Schiff base bond formed) which interact more favorably than an interaction energy of –1 kcal/mol with the ligand are shown in order of decreasing interaction. The interaction energy values are shown in parentheses.

Ret(xtal)/closed(xtal)	Ret(HD)/closed(xtal)	Ret(HD)/closed(MS)
Glu122 (-21.9)	Glu122 (-19.8)	Glu122 (-18.2)
Glu113 (-20.3)	Glu113 (-18.0)	Thr118 (-4.0)
Glu181 (-18.0)	Glu181 (-14.9)	Tyr268 (-4.0)
Trp265 (-6.5)	Trp265 (-6.8)	Phe208 (-3.2)
Tyr268 (-4.8)	Thr118 (-5.2)	Met207 (-1.2)
Thr118 (-3.9)	Tyr268 (-5.2)	Trp265 (-1.1)
Ala292 (-2.9)	Met207 (-2.7)	Gly114 (-1.1)
Ala117 (-2.3)	Ala292 (-2.3)	Ala269 (-1.1)
Phe208 (-2.0)	Phe212 (-1.9)	Phe212 (-1.0)
His211 (-1.8)	Phe208 (-1.8)	Ile275 (-1.0)
Phe212 (-1.7)	Gly114 (-1.8)	
Gly114 (-1.6)	Cys187 (-1.3)	
Met207 (-1.6)	Phe261 (-1.2)	
Phe261 (-1.4)	Ala269 (-1.2)	
Phe293 (-1.1)		

Figure S2: A graphical comparison of the individual helices from the optimized experimental structure apo/closed(xray) (white) and that created by the step 3 helical optimization of the MembStruk protocol (cyan at 100ps or after temperature stabilization, orange is best match to the crystal in all the simulation). Each helix is oriented with the N-terminus at the top. Note the large bends in the crystal helices 2 and 6 due to proline residues. B) The entire dynamics simulation after temperature stabilization or 100 ps (whichever was later) was analyzed for all helices. The helices at 2.5 ps intervals are overlayed. Helices 2 and 6 undergo the most bending at specific "hinge" residues, as has been confirmed by EPR studies. This points to the inherently dynamical nature of these helices needed for the receptor activation process and G-protein binding.

A)



B)

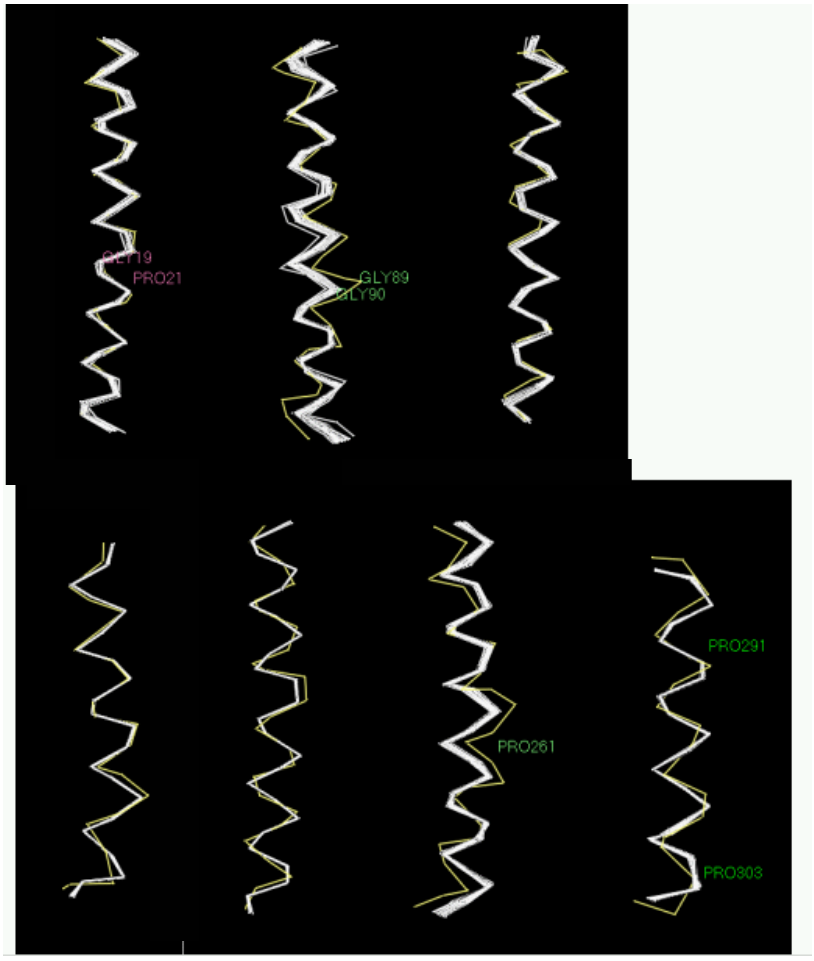
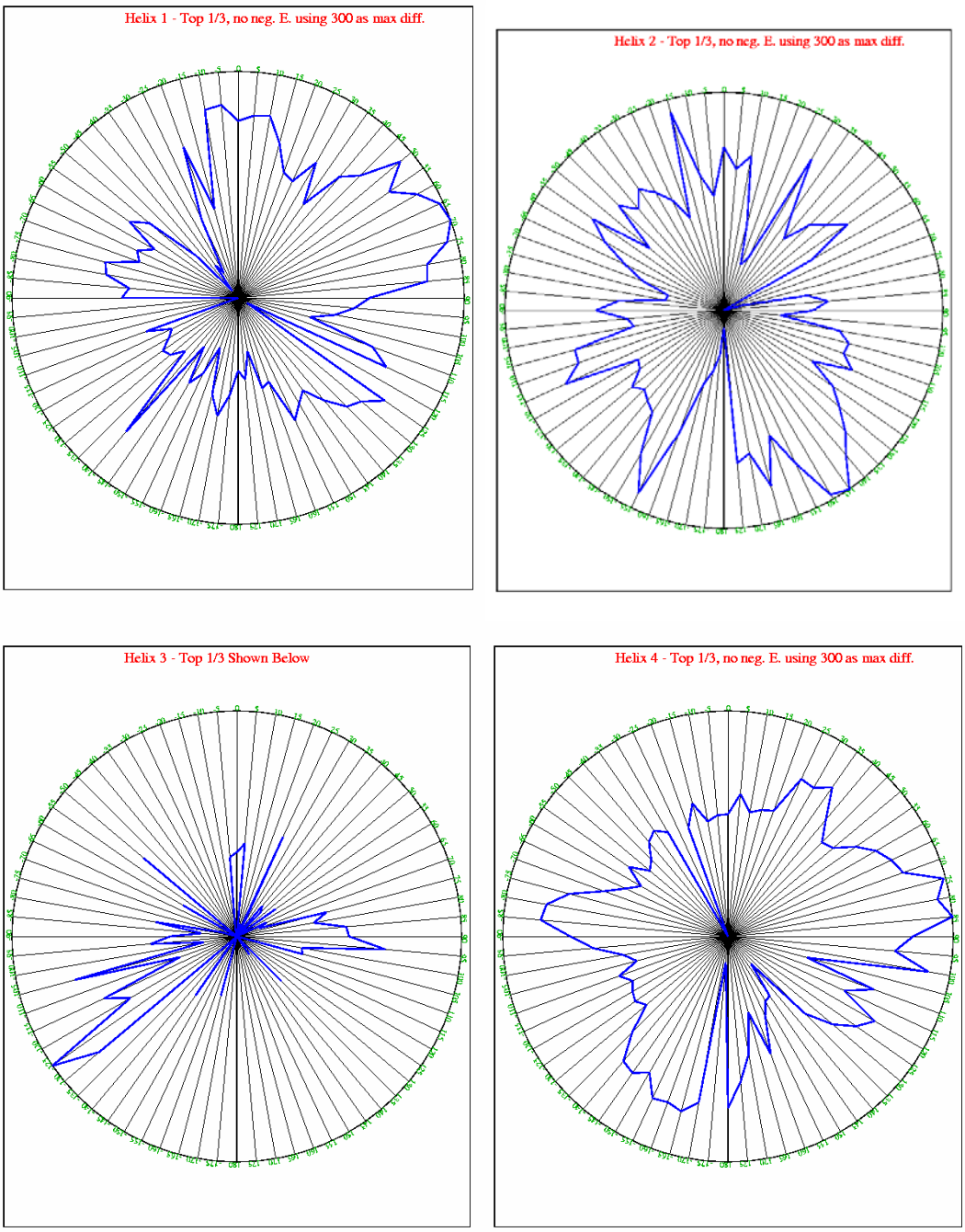
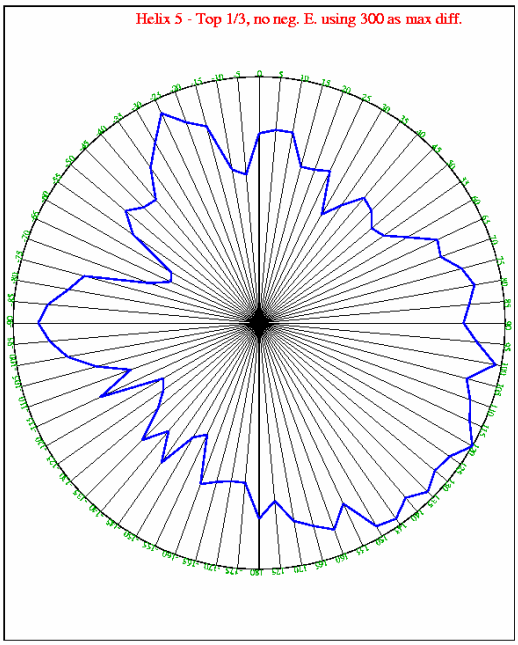
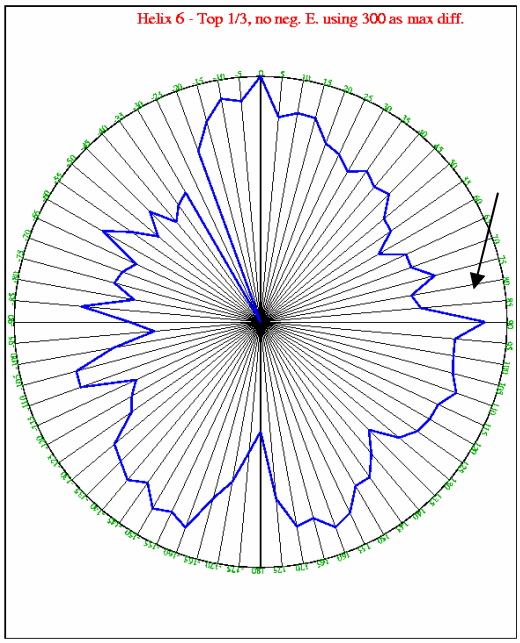


Figure S3: The 360 degree energy scans for the 7 helices of the predicted rhodopsin structure after step 3 of the MembStruk procedure. The arrow for helix 6 indicates the alternate rotation which corresponds more closely to the ground state conformation of the crystal structure.




Helix 5 - Top 1/3, no neg. E. using 300 as max diff.


Helix 6 - Top 1/3, no neg. E. using 300 as max diff.

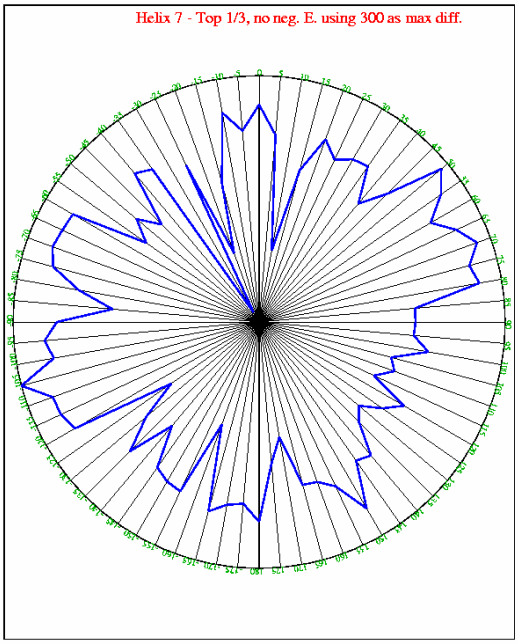

Helix 7 - Top 1/3, no neg. E. using 300 as max diff.

Figure S4: The two alternate rotations of helix 6 found by analysis of the hydrophobic moment using the "phobic face" method of MembStruk. In green is indicated this moment using 19 middle residues, and in red is indicated the moment using 15 middle residues. The alternate rotations correspond to the active and inactive forms of the receptor.

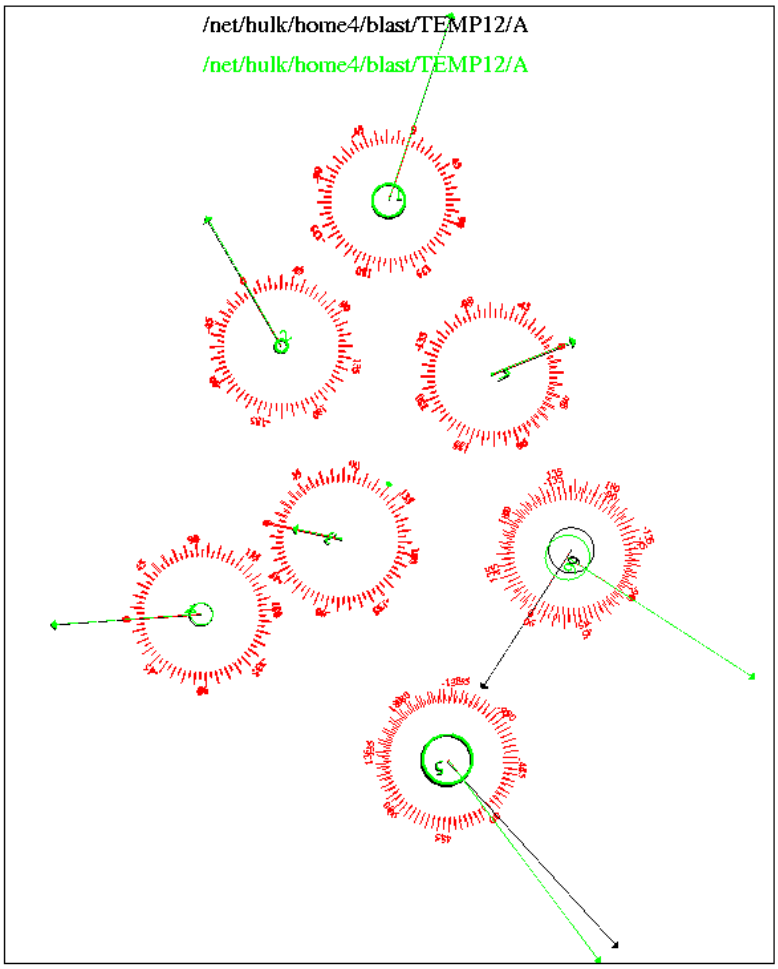
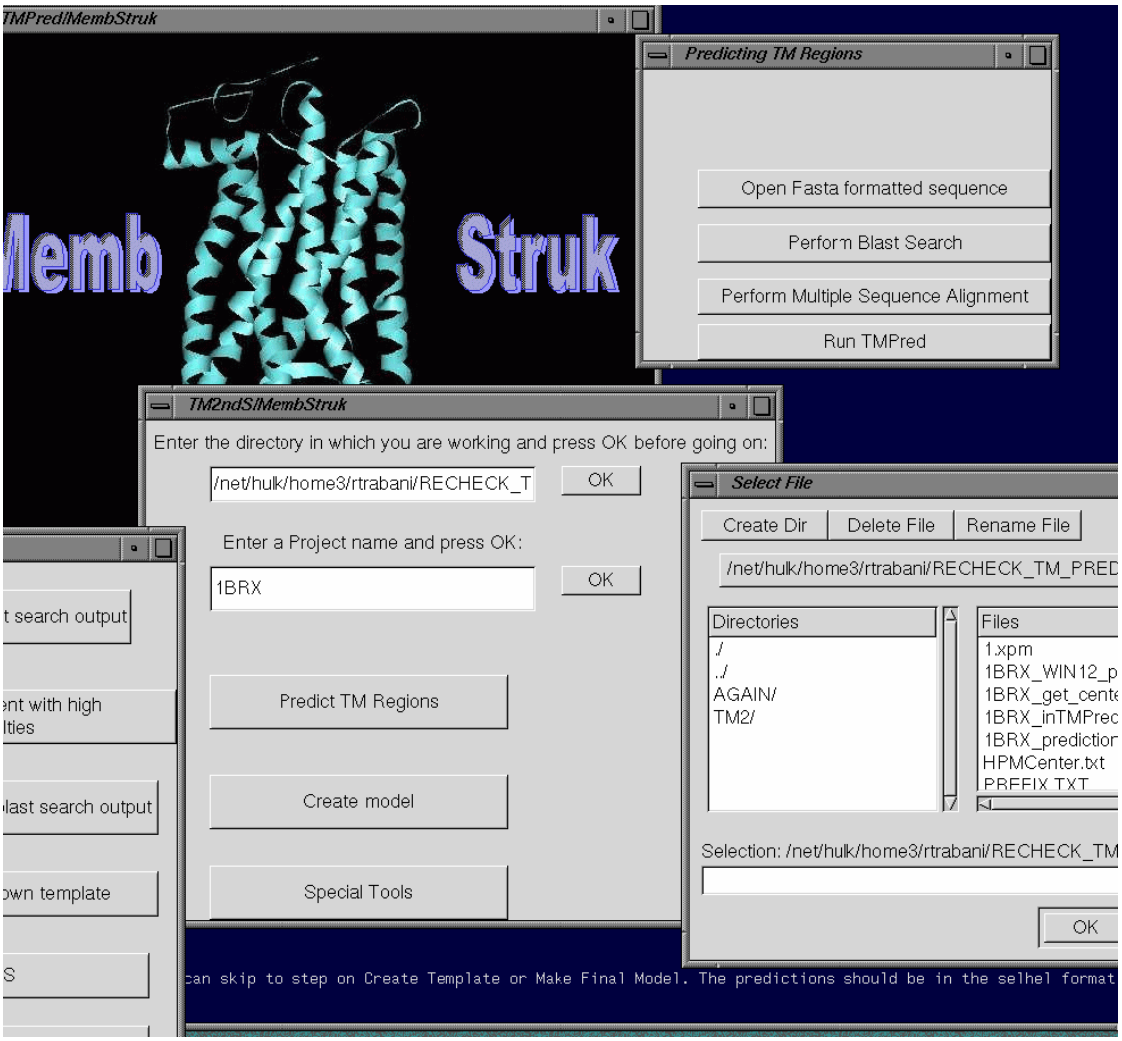


Figure S5: A screenshot of part of the graphical user interface for TM2ndS.



Chapter 3: High accuracy transmembrane helix predictions using TM2ndS

Abstract

The MembStruk was designed to predict the 3-D structures of GPCRs based on sequence alone from first principles. The first step of this method is the accurate determination of transmembrane helices from sequence using the TM2ndS protocol. This study validates TM2ndS for automatic high residue accuracy secondary structure predictions on the set of high-resolution membrane protein crystal structures (MPtopo database). In addition, the method is compared to top TM helical prediction methods. It emerges as one of the two top methods in terms of residue accuracy, and the top "untrained" method. This general method thus demonstrates great generalizability to membrane proteins of unknown structure and of various specific 3-D topologies.

Introduction

Membrane proteins are involved in many important biological processes, from ion transport to the immune response. In particular, G-protein-coupled receptors (GPCRs) are a superfamily of 7-helical transmembrane (TM) proteins which are a target for ~50 % of pharmaceutical drugs (Flower 1999). Despite this, there is a high-resolution structure for only 1 GPCR, and in general <1% of proteins with solved structures are membrane proteins (Chen 2002). In order to design receptor-specific drugs with minimal cross reactivity to other receptor subtypes, it is probably best to apply a structure-based design procedure. Because of the scarcity of atomic level structures, however, theoretical methods need to be developed to meet this need. We have thus developed a first principles method to obtain structures for GPCRs named MembStruk which has been previously described (Floriano 2000; Vaidehi 2002; Trabanino 2003).

The first step of the MembStruk procedure is the prediction of TM helical regions from sequence. There are a variety of other programs available for TM helical predictions. DAS (Cserzo 1997) and TopPred2 (Sipos 1993) use hydrophobicity plots. SOSUI (Hirokawa 1998) and TMHMM (Sonnhammer 1998) use preference functions derived from known protein transmembrane helical locations. HMMTOP (Tusnady 1998) is similar to TMHMM in that it uses a hidden Markov model method. PHDhtm (Rost 1996) uses a neural network scheme with information of sequence relationships and differences. The performance of these methods have been compared (Moller 2001), which found TMHMM to score best. The criteria did not include per-residue accuracy

however, and a recent study (Chen 2002) found there to be no clearly better method for all criteria.

As described in detail in Trabanino, 2003, and validated for the only GPCR with a reference crystal structure available, bovine rhodopsin (Palczewski 2000), MembStruk is a first principles method which applies general rules and uses no structural information from databases to aid in obtaining structures. The first step in this procedure reflects this. Within MembStruk, the determination of transmembrane helices from sequence alone is accomplished by the TM2ndS procedure. It is a hydrophobicity based method which uses multiple-sequence alignments, window scanning, and a robust baseline definition method to aid in obtaining TM regions, which are later capped by general rules of helix termination. As such, this method was developed without a training set of membrane proteins, with the intention of retaining its general applicability to proteins of unknown structure. In this paper, we present the results for TM helical predictions on a database of membrane proteins known as MPtopo (Jayasinghe 2001). As this method is applied to GPCR's, where the number of helices is known beforehand, the focus of the method is the accurate determination of TM helical regions (in terms of residues correctly predicted), although the accuracy in terms of number of TM helical regions is also analyzed (for possible future use in data mining of membrane proteins). In addition, since the primary goal is accurate TM helical residue predictions, the proteins in the MPtopo database were analyzed in terms of their phi-psi angles and crystal structure B-factors in order to provide the most accurate reference for the comparison to predictions.

Materials and methods

Preparing the reference experimental helix data

For analysis of the performance of TM helical prediction methods, there is a website (Kernytsky 2003) which automatically analyzes methods and compares with others using a common criteria. This resource was used to compare TM2ndS to to other top methods, and the results will be discussed later. Also, since TM2ndS is primarily aimed at accurate TM helical residue determination, we chose to analyze the high-resolution experimental database of MPtopo to precisely determine the true helical regions and use this as a reference for our analysis.

The MPtopo (Jayasinghe 2001) database includes accurate TM helical ranges identified from PDB structure assignments as well as visual observation to assign TM regions. However, for the purposes of this study, the helix definitions of this database needed to be refined further. We used a criteria for alpha helicity as residues with phi-psi angles in the following ranges: $-37 < \phi < -77$ and $-27 < \psi < -67$. Residues which fell outside these ranges were considered non-helical. A helix was considered broken if a stretch of two of these non-helical residues was found. In addition, the crystallographic temperature B-factors were considered in assigning TM helical regions. Any residue in which the average of backbone B-factors was over 70 was considered uncertain, and any prediction (or missed prediction) in this region was still considered valid.

The TM2ndS method

The method for TM helical prediction known as TM2ndS was described in detail in Trabanino, 2003. The procedure could be divided into three parts:

a) A multiple sequence alignment using Clustalw (Thompson 1994) of a set of sequences obtained from a Blast (Altschul 1990; Altschul 1997) search of a query sequence is obtained. Sequences above bit score 200 but non-identical (E value less than e^{-100}) were used in the alignment. An ideal set of sequences would be one in which there is uniform distribution across the bit score bins.

b) The hydrophobicity assigned by the Eisenberg scale (Eisenberg 1982) was first averaged across sequences in the alignment, and then across a window size of residues. The hydrophobicity scale was derived from combining values from various partitioning studies of residues between aqueous and apolar environments. The window size used for analysis was 14. TM regions are detected as peaks which satisfy criteria of area >=0.4 and length >6 but <=50. The baseline for these peak determinations is obtained as follows: 1) the average over the region excluding the N and C terminus was obtained 2) then the baseline is increased by 0.12 in increments of 0.01 and kept as the new baseline *base_mod* if more peaks were determined at this modified baseline.

This baseline determination method is different than that described in Trabanino, 2003 since that method focused on

GPCRs and thus looked for 7 helices. The method of modifying the original baseline was found to better resolve small interhelical loops, which has been a significant problem for prediction methods (Chen 2002).

It should be noted that TM2ndS regularly applies a local average for peaks of low intensity in the case of GPCRs. These peaks would correspond to relatively polar helices which are stabilized in the membrane by interaction with other helices (Popot 1990). However, for the current study, this local average was not applied since it is better applied when the number of helices is known beforehand, as in GPCRs.

c) Once the coarse TM regions were predicted, the regions were capped using general rules of helix breaking residues and retaining a loop size of at least 6 for all cases.

An important feature of this method is the baseline determination. Proteins are essentially thermodynamically equilibrated structures (White 2001). Thus, this final baseline may be interpreted physically as a $\Delta G=0$ value above which residues are thermodynamically stable in the transmembrane and below which they are not. This baseline is unique to the particular protein to which it is being applied, with its individual environmental factors (water clusters, ions, hydrophobic or hydrophilic ligand or interhelical interactions, membrane composition) that may change the relative stability of any particular residue.

Results and discussion

TM helical region identification

Since this method is to be currently applied to GPCRs, we focused attention on helices of certain length, namely, those in which the high-resolution experimental lengths were >=20 and <=34. Overall, the MPtopo database was made up of 24 protein complexes with 208 helices. Of these, 164 helices are of length >=20 and <=34. TM2ndS correctly identified 158 of these helices (96.3 %) with overlap of at least 15 residues. In addition, 182 helices were predicted with 158 (86.8% true positive) of these correctly overlapping at least 15 residues.

There are some problems with the structures of the proteins in this database. Thus, of these 24 protein complexes, 5 chains were excluded from the residue accuracy analysis. They were 1JSQ (6 helices) since it has only C-alpha coordinates, preventing phi-psi analysis, 1MSR (1 helix) since it is a theoretical model, and 1BGY chain E (1 helix) since it is missing residues in assigned TM helical region. Also, the proteins 1OCC chain B and 1PRC chain H were excluded due the fact that their first or last peaks were too close (less than half a peak length) away from the beginning or end of the sequence. This would not allow a good quality baseline to be determined. These cases were identified automatically by TM2ndS.

After these exclusions, 154 helices remained. TM2ndS correctly identified 148 of these helices (96.1 %) with overlap of at least 15 residues (Table 1). In addition, 162 helices were predicted with 148 (91.3% true positive) of these correctly overlapping at least 15 residues. Interestingly, if one excludes those helices predicted to be too long or

short (>34 or <20), then 148 helices are predicted with 139 of these overlapping in at least 15 residues (93.9%). Thus, there may be more false positives amongst helices which are predicted to be too short or too long. It should be noted that many of these false positives were due to signal peptides and non-TM hydrophobic regions, which is generally a problem with hydrophobicity-based methods (Lao 2002).

It should be noted that predictions were also obtained without an alignment of multiple sequences. In this case, TM2ndS correctly identified 150 of the 156 helices (96.1 %) with overlap of at least 15 residues (Table 2). But 155 helices were predicted with 148 (95.5% true positive) of these correctly overlapping at least 15 residues. Thus, for the purposes of data mining of membrane proteins from the genome, TM2ndS may perform better with one sequence instead of an alignment. However, as expected, the residue error was larger for analysis with one sequence (Johnson 1999). Thus the multiple sequence alignment method was used for evaluation of residue accuracy for this method.

Assuming that the TM helices are identified correctly (which would be the case in predictions for GPCRs, where 7 TM helices are known to exist), what is the error in determining the precise TM helical ranges?

TM helical region residue accuracy

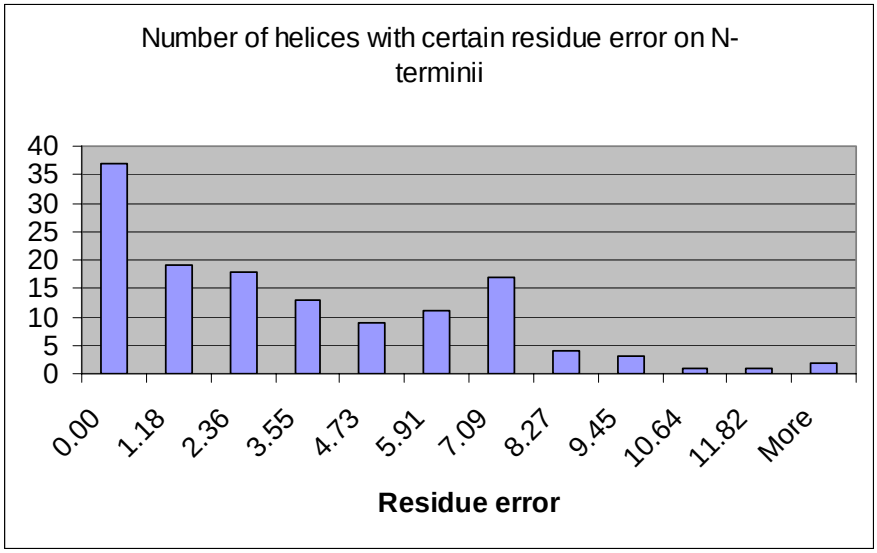
Some of the TM2ndS predicted sequences (for those proteins in which all the lengths of the crystal and predicted helices were >=20 and <=34) before and after capping are shown in Figure 1 (placed after references because of length). The predictions for all helices (which have lengths of the crystal reference helix of >=20 and <=34) in the database are analyzed are shown in Figure S1 (supplemental material after

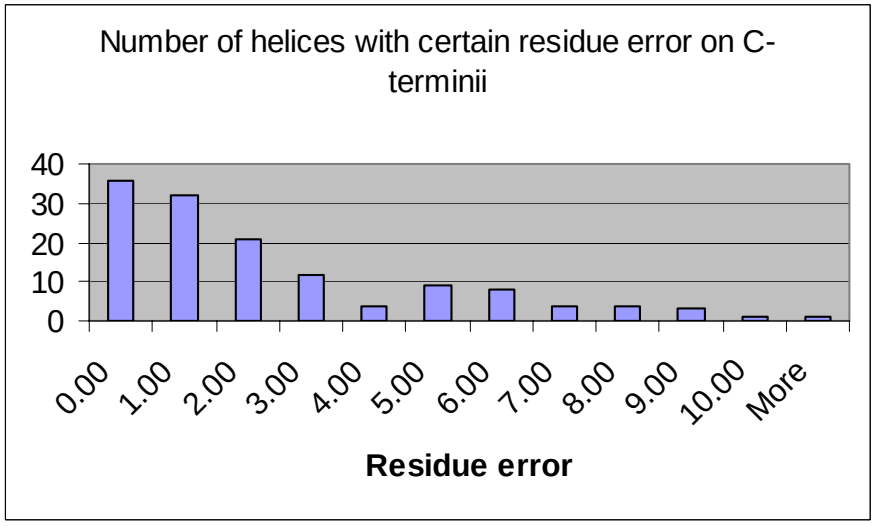
references). In addition, the experimental TM helix sequences are shown. Those residues which did not fit the criteria for alpha helicity (in phi-psi angles) are indicated in italics, while those which have B-factors over 70 are indicated in bold.

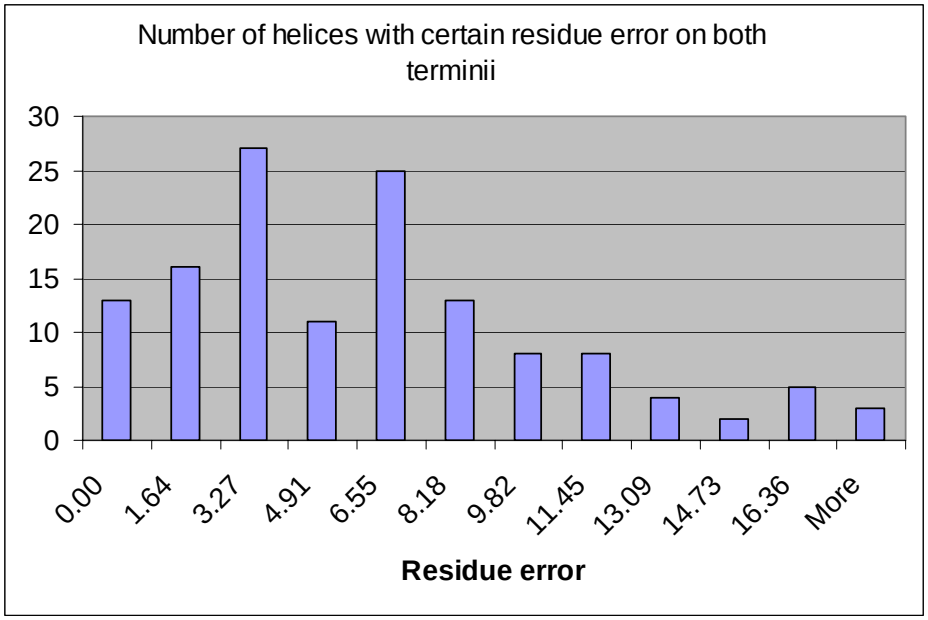
After analyzing the phi-psi and B-factors of the crystal structures, the residue accuracy errors were as follows. For all 148 helices (96.1 % of total helices as discussed in Sec. 3.1) which were predicted with overlap of >=15, the average error in residues was 5.39. This would be equivalent to an average error of 2.70 on each side of the TM helices. Of these helices, 127 (85.8%) had errors of <= 7. The average residue error for these was 4.18, corresponding to an error of 2.09 on either side.

As seen in Figure 2, the distribution of these errors is largely clustered around smaller values. The distributions of errors in the N-terminal, C-terminal, and both sides of the helices are shown in this figure.

Figure 2: The N-terminal, C-terminal , and combined errors in TM helix prediction for TM2ndS.







The purpose of such high precision is to maintain the correct secondary structure for segments of the sequence which may be involved in ligand binding or other critical function. As such, these small errors in the TM helical predictions would lead to only local errors in the protein structure prediction of MembStruk. This is because MembStruk has a translation protocol implemented which aligned helices by their hydrophobic centers, thus positioning the binding site residues of any given ligand correctly.

Gap check

There are a few certain outlier cases where errors may exceed 10. Such cases may require a better list of sequences for the alignment (more uniform distribution of homologies). In addition, a careful analysis of the profile changes with window size may lead to more accurate predictions. These outlier cases occur mostly in bacterial proteins, which have vastly different membrane compositions from the GPCRs in humans for example. The much greater membrane protein to lipid ratio in bacteria (3:1 as opposed to

1:1 in humans) and thus the higher association with other proteins contribute to this difference. So, it may not be wise to make the method more specific to do well for this database with a great number of bacterial proteins. Nevertheless, the analysis of gaps in the alignment may provide another check for the helical predictions. This is demonstrated for 4 cases, two which have larger than 10 error and two with smaller than 10 error (to evaluate the generality of this second-pass approach). In this gap check, an alignment of sequences down to low homologies (25-30%) is performed. This alignment is then checked for gaps about the TM2ndS TM helical prediction. These gaps would define an upper limit to the capping of the helix when they do not form a loop which is to small and when they do not enter the region defined by the raw (before capping) TM prediction. The cases of helices 2 and 5 of 1BRX and 3, 6 for 1OCC were analyzed by this approach. The alignments in these regions are shown in Figure 3.

Figure 3: The alignments of sequences down to lower homologies for a) 1BRX and b) 1OCCR chain C, demonstrating the gap check method. The gap check TM helix is in bold while the crystal structure TM helix in underlined.

a)

```
HELIX 2
1BRX                          -QAQITGRPEWIWLALG----------TALMGLGTLYFLVKGMGVSDPDA
gi|15790468|ref|NP_280292.1|  -QAQITGRPEWIWLALG----------TALMGLGTLYFLVKGMGVSDPDA
gi|3023375|sp|P96787|BAC3_HALS -DLLGDGRPETLWLGIG----------TLLMLIGTFYFLVRGWGVTDKDA
gi|114807|sp|P19585|BAC1_HALS1 -DLLGDGRPETLWLGIG----------TLLMLIGTFYFIVKGWGVTDKEA
gi|231626|sp|P29563|BAC2_HALS2 -DLLNDGRPETLWLGIG----------TLLMLIGTFYFIARGWGVTDKEA
gi|2829812|sp|P94854|BACR_HALV ---MPAPEGEAIWLWLG----------TAGMFLGMLYFIARGWGETDSRR
gi|2499386|sp|Q57101|BACR_HALA ---MPEPGSEAIWLWLG----------TAGMFLGMLYFIARGWGETDSRR
gi|2499387|sp|Q53496|BACR_HALS -MSTYVPGGESIFLWVG----------TAGMFLGMLYFIARGWSVSDQRR
gi|14194473|sp|O93740|BACR_HAL -PMAATVGPESIWLWIG----------TIGMTLGTLYFVGRGRGVRDRKM
gi|461611|sp|P33971|BACR_HALHP ----------IWLWLG----------TAGMFLGMLYFIARGWGETDSRR
gi|461612|sp|P33972|BACR_HALHS ------------LWLG----------TAGMFLGMLYFIARGWGETDGRR
gi|461610|sp|P33969|BACR_HALHM --------------GIG----------TLLMLIGTFYFIARGWGVTDKKA
gi|1168614|sp|P42197|BACT_HALV ------MATITTWFTLG----------LLGELLGTAVLAY-GYTLVPEET
gi|461609|sp|P33742|BACH_HALSS -EAVQSDTLLASSLWIN----------IALAGLSILLFVYMGRNVEDPRA
gi|2499383|sp|Q48315|BACH_HALH -QEIQSNFLLNSSIWVN----------IALAGVVILLFVAMGRDIESPRA
gi|2499384|sp|Q48314|BACH_HALH -QEIQSNFLLNSSIWVN----------IALAGVVILLFVAMGRDLESPRA
gi|14194475|sp|O93742|BACH_HAL -EAIQGDTLLASSLWIN----------IALAGLSILLFVYMGRNLEDPRA
gi|114809|sp|P15647|BACH_NATPH -EFVLNDPLLASSLYIN----------IALAGLSILLFVFMTRGLDDPRA
gi|2829811|sp|P94853|BACH_HALV -GEIQSNFLLNSSLWVN----------IALAGVVILLFVAMGRELESSRA
gi|15790684|ref|NP_280508.1|  -------MALTTWFWVG-----------AVGMLAGTVLPIRDCIRHPSH
gi|15789491|ref|NP_279315.1|  -AAVRENALLSSSLWVN----------VALAGIAILVFVYMGRTIRPGRP
gi|461608|sp|P33970|BACH_HALHM ---------------------------IALAGLSILLFVYMGRNVEDPRA
gi|14194474|sp|O93741|BACH_HAL -TQIRTDTLLHSSLWVN----------IALAGLSILVFLYMARTVRANRA
gi|1168615|sp|P42196|BACT_NATP ------MVGLTTLFWLG----------AIGMLVGTLAFAWAGRDAGSGE-
gi|15790610|ref|NP_280434.1|  ------MDAVATAYLGG----------AVALIVGVAFVWLLYRSLDGSPH
gi|2499388|sp|Q48334|BAC3_HALV ------MDAVAVVYGIT----------AAGFAVGVAIVGYLYASLEGSEE
gi|461613|sp|P33743|BACS_HALSS -----MTGAVSAAYWIA----------AVAFLVGLGITAALYAKLGESED
gi|14194476|sp|O93743|BACS_HAL -----MTGAVTSAYWLA----------AVAFLIGVGITAALYAKLEGSRA
gi|6685436|sp|O74631|F123_CORV -ATFHLSTHGSDWLWAA----------FSVFGVSLLTVVAWTFTRPRGA
gi|6319528|ref|NP_009610.1|   GADFHITSRGSDWLFTV---------FCVNLLFGVILVPL--MFRKPVKD
gi|1729881|sp|P51564|TCR8_PASM -NSLATHYGVLLALYAT----------MQVIFAPILGRLSDKYGRKPILL
gi|11467412|ref|NP_043269.1|  CSKLATLLRWLTHFWLFFG-------LMVLISASGFTSYEEHRDVLYYFK
gi|16763593|ref|NP_459208.1|  -FMAAVVGTLTGLVGVAFEK--A-VSWVQNMRIGALVQVADHAFLLWPLA
gi|1708329|sp|P53706|HST6_CANA IKEDTDNEKLMGVLAILRYC--S-STINGKSLLGFGILLAIFQGVSSPVF
gi|16759194|ref|NP_454811.1|  -FMAAVVGTLTGLVGVAFEK--T-VSWVQNMRIGALVQVADHAFLLWPLA
gi|15594598|ref|NP_212387.1|  FFLKEQLKAIKAELGIGDKKSSD-LEKLKTKLKALELKGEPLEVVEKELE
gi|2501369|sp|Q47085|CBRB_ERWC GALWHPDPLNVSHILVTS---------TRLSRTLIAIVVGAGLAVAGALM
gi|17568741|ref|NP_509364.1|  SIMYPKPQEKEALKDVMVFILVLGFIALIGFIYTVIEMVSRGESLKHIII
gi|15604244|ref|NP_220760.1|  SIKDSFVVTLISSEVLS-----F-IKLWGEMPMGVLFVILYSKLCNIMTT
```

Gap chosen for cap

(continued)

```
1BRX                          KKFYAITTLVPAIAFTMYLSMLLG--------------------------
gi|15790468|ref|NP_280292.1|  KKFYAITTLVPAIAFTMYLSMLLG--------------------------
gi|3023375|sp|P96787|BAC3_HALS REYYAVTILVPGIASAAYLSMFFG--------------------------
gi|114807|sp|P19585|BAC1_HALS1 REYYSITILVPGIASAAYLSMFFG--------------------------
gi|231626|sp|P29563|BAC2_HALS2 REYYAITILVPGIASAAYLAMFFG--------------------------
gi|2829812|sp|P94854|BACR_HALV QKFYIATILITAIAFVNYLAMALG--------------------------
gi|2499386|sp|Q57101|BACR_HALA QKFYIATILITAIAFVNYLAMALG--------------------------
gi|2499387|sp|Q53496|BACR_HALS QKFYIATIMIAAIAFVNYLSMALG--------------------------
gi|14194473|sp|O93740|BACR_HAL QEFYIITIFITTIAAAMYFAMATG--------------------------
gi|461611|sp|P33971|BACR_HALHP QKFYIATILITAIAFVNYLAMALG--------------------------
gi|461612|sp|P33972|BACR_HALHS QKFYIATILITAIAFVNYLAMALG--------------------------
gi|461610|sp|P33969|BACR_HALHM REYYAITILVPGIASAAYLSMFFG--------------------------
gi|1168614|sp|P42197|BACT_HALV RKRYLLLIAIPGIAIVAYALMALG--------------------------
gi|461609|sp|P33742|BACH_HALSS QLIFVATLMVPLVSISSYTGLVSG--------------------------
gi|2499383|sp|Q48315|BACH_HALH KLIWVATMLVPLVSISSYAGLASG--------------------------
```

```
gi|2499384|sp|Q48314|BACH_HALH      KLIWVATMLVPLVSISSYAGLASG-------------------------
gi|14194475|sp|O93742|BACH_HAL      QLIFVATLMVPLVSISSYTGLVSG-------------------------
gi|114809|sp|P15647|BACH_NATPH      KLIAVSTILVPVVSIASYTGLASG-------------------------
gi|2829811|sp|P94853|BACH_HALV      KLIWVATMLVPLVSISSYAGLASG-------------------------
gi|15790684|ref|NP_280508.1|       RRYDLVLAGITGLAAIAYTTMGLG-------------------------
gi|15789491|ref|NP_279315.1|       RLIWGATLMIPLVSISSYLGLLSG-------------------------
gi|461608|sp|P33970|BACH_HALHM     QLIFVATLMVPLVSISSYTGLVSG-------------------------
gi|14194474|sp|O93741|BACH_HAL     RLIVGATLMIPLVSLSSYLGLVTG-------------------------
gi|1168615|sp|P42196|BACT_NATP     RRYYVTLVGISGIAAVAYVVMALG-------------------------
gi|15790610|ref|NP_280434.1|       QSALAPLAIIPVFAGLSYVGMAYD-------------------------
gi|2499388|sp|Q48334|BAC3_HALV     RSILAALALIPGFAGISYVAMAFG-------------------------
gi|461613|sp|P33743|BACS_HALSS     RGRLAALAVIPGFAGLAYAGMALG-------------------------
gi|14194476|sp|O93743|BACS_HAL     RTRLAALAVIPGFAGLSYVGMALG-------------------------
gi|6685436|sp|O74631|F123_CORV     RLFHQIAIVVLTTGSLAYFSMASD-------------------------
gi|6319528|ref|NP_009610.1|       RFVYYTAIAPNLFMSIAYFTMASN-------------------------
gi|1729881|sp|P51564|TCR8_PASM     FSLLGAALDYLLMAFSTTLWMLYIG-------------------------
gi|11467412|ref|NP_043269.1|      RQFVFCLIGIVISNILMHFPLTLL-------------------------
gi|16763593|ref|NP_459208.1|      FILSALLAMVGYFLVRKFAPEAGG-------------------------
gi|1708329|sp|P53706|HST6_CANA     SYCFSKLLSTSLDSSIGLNSTQKI-------------------------
gi|16759194|ref|NP_454811.1|      FILSALLAMVGYFLVRKFAPEAGG-------------------------
gi|15594598|ref|NP_212387.1|      KFSLLETSSAEYIVVRNYLELITEL-------------------------
gi|2501369|sp|Q47085|CBRB_ERWC     QVLTRNPLASPGLFGINAGAMFFLI-------------------------
gi|17568741|ref|NP_509364.1|      RSLDIITIVVPPALPAAMSVGIINANSRLKKKKIFCTSPTTVNVCGLINV
gi|15604244|ref|NP_220760.1|      EQVFRIITSTFLFFFAIFGFILFP-------------------------
```

Helix too short and no alternative, so not considered for gap check

```
Helix 5
1BRX                                ------------------Y**RFVWWAISTAAMLYILY**--------------
gi|15790468|ref|NP_280292.1|        ------------------YRFVWWAISTAAMLYILY--------------
gi|3023375|sp|P96787|BAC3_HALS      ------------------ARYSWWLFSTICMIVVLY--------------
gi|114807|sp|P19585|BAC1_HALS1      ------------------ARYTWWLFSTICMIVVLY--------------
gi|231626|sp|P29563|BAC2_HALS2      ------------------ARYTWWLFSTIAFLFVLY--------------
gi|2829812|sp|P94854|BACR_HALV      ----------------GAERLVWWGISTAFLLVLLY--------------
gi|2499386|sp|Q57101|BACR_HALA      ----------------GAERLVWWGISTAFLLVLLY--------------
gi|2499387|sp|Q53496|BACR_HALS      ----------------GAERLVWWGISTGFLLVLLY--------------
gi|14194473|sp|O93740|BACR_HAL      ------------------TRIAWWAISTGALLALLY--------------
gi|461611|sp|P33971|BACR_HALHP      ----------------GAERLVWWGISTAFLLVLLY--------------
gi|461612|sp|P33972|BACR_HALHS      ----------------GAERLVWWGISTAFLLVLLY--------------
gi|461610|sp|P33969|BACR_HALHM      ------------------ARYTWWLFSTIAFLFVLY--------------
gi|1168614|sp|P42197|BACT_HALV      ------------------SYALFAVGGALFGGVIY--------------
gi|461609|sp|P33742|BACH_HALSS      ----------------LLRWVWYAISCAFFVVVLY--------------
gi|2499383|sp|Q48315|BACH_HALH      ----------------LLRWVFYGISCAFFVAVLY--------------
gi|2499384|sp|Q48314|BACH_HALH      ----------------LLRWVFYGISCAFFVAVLY--------------
gi|14194475|sp|O93742|BACH_HAL      ----------------LLRWVWYVISCAFFVVVLY--------------
gi|114809|sp|P15647|BACH_NATPH      ----------------LMRWFWYAISCACFLVVLY--------------
gi|2829811|sp|P94853|BACH_HALV      ----------------LLRWVFYGISCAFFIAVLY--------------
gi|15790684|ref|NP_280508.1|       ------------------RWLFFAVGAAGYAALLY--------------
gi|15789491|ref|NP_279315.1|       ----------------LFRWAFYAISCAFFVVVLS--------------
gi|461608|sp|P33970|BACH_HALHM     ----------------LLRWVWYGISCAFFVVVLY--------------
gi|14194474|sp|O93741|BACH_HAL     ----------------AFRWAFYLVSTAFFVVVLY--------------
gi|1168615|sp|P42196|BACT_NATP     ------------------RYALFGMGAVAFLGLVY--------------
gi|15790610|ref|NP_280434.1|       ------------------KWALFGVSSIFHLSLFA--------------
gi|2499388|sp|Q48334|BAC3_HALV     ------------------KWVLFGVSTVFHVSLFA--------------
gi|461613|sp|P33743|BACS_HALSS     ------------------KWVLFGVSSIFHVTLFA--------------
gi|14194476|sp|O93743|BACS_HAL     ------------------KWALFGVSALFHVSLFA--------------
gi|6685436|sp|O74631|F123_CORV     ------------------KWGYYTFGVSALFYIWY--------------
gi|6319528|ref|NP_009610.1|       ------------------KWGYYTIGIGAAIVVCI--------------
gi|1729881|sp|P51564|TCR8_PASM     ----------------NTVTVFFKKSLYFWLATYFI-------------
gi|11467412|ref|NP_043269.1|      ----------------ANFPWKYLLGTVFVGLSMAIT-------------
gi|16763593|ref|NP_459208.1|      ----------------PQFRYNLISIKAVFTGVIMSS-------------
gi|1708329|sp|P53706|HST6_CANA     ----------------IVSGWKLALVGISFVPLVLLV-------------
gi|16759194|ref|NP_454811.1|      ----------------PQFRYNLISIKAVFTGVIMSS-------------
gi|15594598|ref|NP_212387.1|      EV------LDPEQNVRFRDHYLDLPFDISNVFFILTANS-VETIPRPLLN
gi|2501369|sp|Q47085|CBRB_ERWC     ----------------ADRELADGAAADGLLLAALVG-------------
gi|17568741|ref|NP_509364.1|      HLNFAKALKTPRDIMESELEFLGLIVMENRLKDVTLSVINELSVANIRCV
gi|15604244|ref|NP_220760.1|      ----------------DTNEILLKSFITVILISGLIC-------------
```

b)

```
Helix 3
1OCC3                              NPSPWPLTGALSALLMTSGLTMWFH--------FNSMTLLMIGLTTNMLT
gi|28381353|sp|P00415|COX3_BOV     NPSPWPLTGALSALLMTSGLTMWFH--------FNSMTLLMIGLTTNMLT
gi|5921862|sp|O47701|COX3_ANTM     NPSPWPLTGALSALLMTSGLTMWFH--------FNSMTLLTLGLTTNMLT
gi|5921876|sp|O47695|COX3_OURO     NPSPWPLTGALSALLMTSGLIMWFH--------FNSTTLLMLGLTTNMLT
gi|5921877|sp|O47692|COX3_PELC     NPSPWPLTGALSALLMTSGLIMWFH--------FNSTTLLMLGLTTNMLT
gi|5921888|sp|O47687|COX3_TRAS     NPSPWPLTGALSALLMTSGLTMWFH--------FNSMILLMLGLTTNMLT
gi|5921861|sp|O47702|COX3_ANTC     NPSPWPLTGALSALLMTSGLIMWFH--------FNSTTLLMLGLTTNMLT
gi|5921866|sp|O47708|COX3_GAZC     NPSPWPLTGALSALLMTSGLIMWFH--------FNSTTLLMLGLTTNMLT
gi|5921865|sp|O47694|COX3_DAML     NPSPWPLTGALSALLMTSGLIMWFH--------FNSMTLLMLGLTTNMLT
gi|5921885|sp|O47685|COX3_TRAI     NPSPWPLTGALSALLMTSGLTMWFH--------YNSTILLMLGLTTNMLT
gi|6166023|sp|O48346|COX3_GAZG     NPSPWPLTGALSALLMTSGLIMWFH--------FNSTTLLMLGLTTNMLT
gi|5921864|sp|O47693|COX3_CEPN     NPSPWPLTGALSALLMTSGLIMWFH--------FNSTALLMLGLTTNMLT
gi|5921873|sp|O47700|COX3_LITW     NPSPWPLTGALSALLMTSGLIMWFH--------FNSTTLLMLGLTTNMLT
gi|6166022|sp|O48308|COX3_GAZD     NPSPWPLTGALSALLMTSGLIMWFH--------FNSTTLLMLGLTTNMLT

.

.

gi|5834878|ref|NP_006947.1|COX     SLSSYPILIFCSSLGFTSSLVVFFKNG-IFG-------GLLFCLFSIFLV


Helix 6
1OCC3                              MYQWWRDVIRESTFQGHHTPAVQKGLRYGMILFIISEVLFFTGFFWAFYH
gi|28381353|sp|P00415|COX3_BOV     MYQWWRDVIRESTFQGHHTPAVQKGLRYGMILFIISEVLFFTGFFWAFYH
gi|5921862|sp|O47701|COX3_ANTM     MYQWWRDIIRESTFQGHHTPNVQKGLRYGMILFIISEVLFFTGFFWAFYH
gi|5921876|sp|O47695|COX3_OURO     MYQWWRDVIRESTFQGHHTPTVQKGLRYGMILFIISEVLFFTGFFWAFYH
gi|5921877|sp|O47692|COX3_PELC     MYQWWRDIIRESTFQGHHTPSVQKGLRYGMILFIISEVLFFTGFFWAFYH
gi|5921888|sp|O47687|COX3_TRAS     MYQWWRDIIRESTFQGHHTPVVQKGLRYGMILFIISEVLFFTGFFWAFYH
gi|5921861|sp|O47702|COX3_ANTC     MYQWWRDVIRESTFQGHHTPNVQKGLRYGMILFIISEVLFFTGFFWAFYH
gi|5921866|sp|O47708|COX3_GAZC     MYQWWRDVVRESTFQGHHTPNVQKGLRYGMILFIISEVLFFTGFFWAFYH
gi|5921865|sp|O47694|COX3_DAML     MYQWWRDIIRESTFQGHHTSAVQKGLRYGMILFIISEVLFFTGFFWAFYH
gi|5921885|sp|O47685|COX3_TRAI     MYQWWRDIIRESTFQGHHTPTVQKGLRYGMILFIISEVLFFTGFFWAFYH
gi|6166023|sp|O48346|COX3_GAZG     MYQWWRDVIRESTFQGHHTPNVQKGLRYGMILFIISEVLFFTGFFWAFYH

.

.

gi|5834931|ref|NP_008087.1|COX     -MLWFRDIIRESTFQGMHSMFITNFLKFSMILFILSELMFFISFFWTFFH

.

gi|1169062|sp|P41775|COX3_MYTE     TFSWWRDLIREGDIG-

FHTRFVIKSFRDCVALFILSEVMFFFTFFWTFFH
```

Gap not chosen because would make helix too short

Gap chosen

Interestingly, if these same alignments with low homology sequences are used as input for TM2ndS prediction, some of the larger errors are reduced. For 1BRX, TM helices 2 and 5:

Helix 2

KFYAITTLVPAIAFTMYLSMLLG (previous error 15, now error is 5)

Helix 5

RFVWWAISTAAMLYILY too short so would use default prediction

And for 1OCCR:

Helix 3

GMILFIISEVLFFTGFFWAFYHSSLAPTPELG same error

Helix 6

DGVYGSTFFVATGFHGLHVIIGSTFLIVCFFR same error

So these additional approaches may provide a second check for the TM2ndS default predictions as well as provide alternate predictions which may be improvements in some cases.

Comparison to other methods

As mentioned previously, TM2ndS was compared to other top TM helix prediction methods by "static benchmarking" of the Rost group (developers of the PHD prediction method). This comparison is made using the same set of proteins and the same scoring method. The results for the TM helix residue accuracy are shown in Figure 4. Of all these methods, only TM2ndS and PHD performed over 75% (horizontal line) for both

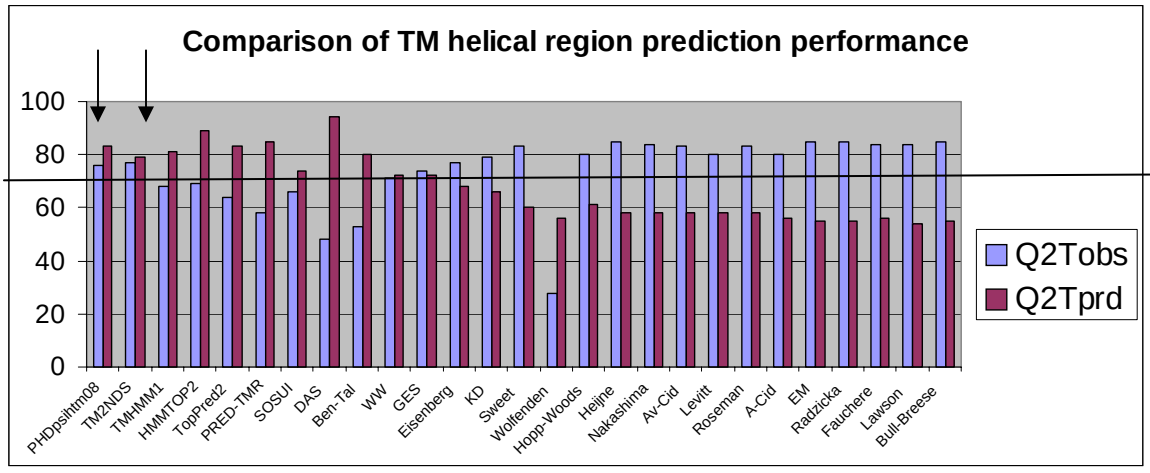
the Q2Tobs (low value indicates underprediction of TM helix lengths) and Q2Tprd (low value indicates overprediction of lengths) whose definitions are given below:

*Q2T(obs)=residues correctly predicted in TM helices/residues observed in TM helices

*Q2T(prd)=residues correctly predicted in TM helices/residues predicted in TM helices

The high values for both scores indicates a stable and accurate prediction. And of these two methods, only TM2ndS is based on hydrophobicity and general rules. It was not developed using the training set of membrane proteins, which is the case for other methods such as PHD. This indicates the general applicability of TM2ndS, especially in the case of GPCRs where there is a scarcity of information available.

Figure 4: Histogram comparing the residue accuracy performance of the top TM helical prediction methods.



Thus for most cases, the TM2ndS residue accuracy is good enough to give confidence in predictions for unknown cases. It should be reiterated that these helices were not used in "training" the method. TM2ndS is based on hydrophobicity and general rules, which contribute to its generalizability.

Conclusion

TM2ndS is based on a first principles hydrophobicity analysis, with a robust baseline calculation, and general rules for the termination of the helix. As such, it is a general method for TM helix identification which is extendable to all membrane proteins of known structure. The residue accuracy of the predictions increases when the number of helices is known beforehand. But nevertheless, this study reports the high TM helix identification accuracy of 96.1% with a minimum required overlap of 15 residues, a stringent criteria in comparison to previous such analyses. Using this same criteria, the false positives were 8.7%. When using one sequence in the analysis (instead of a multiple sequence alignment), the number for true positives and false positives were 96.1% and 95.5%, respectively. For the purposes of genome mining, this modification of TM2ndS may be used.

This method is currently incorporated into MembStruk and as such is targeted towards the predictions for GPCRs, for which the number of helices is known beforehand, and in which case the focus is on high residue accuracy to permit correct function prediction. The average error in the predictions for the TM helical proteins was 5.39 (corresponding to a 2.70 average error on any particular helix terminus accuracy). For 85.8% of the sequences, the average errors are 4.18 (2.09 on any side). The outlier

cases would very likely be improved by providing alignments with better sequence enrichment and by observing trends at different window sizes as used for GPCRs.

Because of the translation protocol of MembStruk, the small errors in the current TM helical predictions would lead to small local structure errors in the 3-D structure predictions. Thus based on this automatic method, the predictions are of sufficiently good accuracy to allow correct determination of correct ligand binding.

References

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.

Chen, C. C. and Rost, B. (2002). Long membrane helices and short loops predicted less accurately. *Protein Sci.* 11, 2766-2773.

Chen, C. P.; Kernytsky, A. and Rost, B. (2002). Transmembrane helix predictions revisited. *Protein Science* 11, 2774-2791.

Cserzo, M.; Wallin, E.; Simon, I.; von Heijne, G. and Elofsson, A. (1997). Prediction of transmembrane alpha helices in prokaryotic membrane proteins: The dense alignment surface method. *Prot. Engin.* 10, 673-676.

Eisenberg, D.; Weiss, R. M.; Terwilliger, T. C. and Wilcox, W. (1982). Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc.* 17, 109-120.

Floriano, W. B.; Vaidehi, N.; Singer, M.; Shepherd, G. and Goddard III, W. A. (2000). Molecular mechanisms underlying differential odor responses of a mouse olfactory receptor. *Proc. Natl. Acad. Sci. USA.* 97, 10712-10716.

Flower, D. R. (1999). Modelling G-protein-coupled receptors for drug design. *Biochim. Biophys. Acta* 1422, 207-234.

Hirokawa, T.; Boon-Chieng, S. and Mitaku, S. (1998). SOSUI: Classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14, 378-379.

Jayasinghe, S.; Hristova, K. and White, S. H. (2001). MPtopo: A database of membrane protein topology. *Protein Science* 10, 455-458.

Johnson, J. M. and Church, G. M. (1999). Alignment and structure prediction of divergent protein families: Periplasmic and outer membrane proteins of bacterial efflux pumps. *J. Mol. Biol.* 287, 695-715.

Kernytsky, A. and Rost, B. (2003). Static benchmarking of membrane helix predictions. *Nucleic Acids Research* 31, 3642-3644.

Lao, D. M.; Arai, M.; Ikeda, M. and Shimizu, T. (2002). The presence of signal peptides significantly affects transmembrane topology prediction. *Bioinformatics* 18, 1562-1566.

Moller, S.; Croning, M. D. R. and Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17, 646-653.

Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C.; Motoshima, H.; Fox, B.; Trong, I.; Teller, D.; Okada, T.; Stenkamp, R.; Yamamoto, M. and Miyano, M. (2000). Crystal structure of rhodopsin: A G-protein coupled receptor. *Science* 289, 739-745.

Popot, J.-L. and Engelman, D. M. (1990). Membrane Protein Folding and Oligermerization: The Two-Stage Model. *Biochemistry* 29, 4031-4037.

Rost, B.; Casadio, R. and Fariselli, P. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* 5, 1704-1718.

Sipos, L. and von Heijne, G. (1993). Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.* 213, 1333-1340.

Sonnhammer, E. L. L.; von Heijne, G. and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. Sixth International Conference on Intelligent Systems for Molecular Biology. Montreal, Canada, AAAI Press**:** 175-182.

Thompson, J. D.; Higgins, D. G. and Gibson, T. J. (1994). Clustalw - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting,

position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673-4680.

Trabanino, R. J.; Hall, S. E.; Vaidehi, N.; Floriano, W. B.; Kam, W. T. and Goddard, W. A. (2003). First principles predictions of the structure and function of G-Protein Coupled Receptors: Validation for bovine rhodopsin. *Biophys. J.*

Tusnady, G. E. and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: Application to topology prediction. *J. Mol. Biol.* 283, 489-506.

Vaidehi, N.; Floriano, W. B.; Trabanino, R.; Hall, S. E.; Freddolino, P.; Choi, E. J.; Zamanakos, G. and Goddard, W. A. (2002). Predicition of structure and function of G-protein coupled receptors. *Proc. Natl. Acad. Sci. USA.* 99, 12622-12627.

White, S. H.; Ladokhin, A. S.; Jayasinghe, S. and Hristova, K. (2001). How Membranes Shape Protein Structure. *J. Biol. Chem.* 276, 32395-32398.

Table 1: Data on true positives as well as false positives for the TM2ndS analysis using a multiple sequence alignment

	Reference crystal structure helices	Predictions
Number total	154	162
Number predicted to overlap with >=15 residues	148	148
Percents	96.1 (true +)	91.3

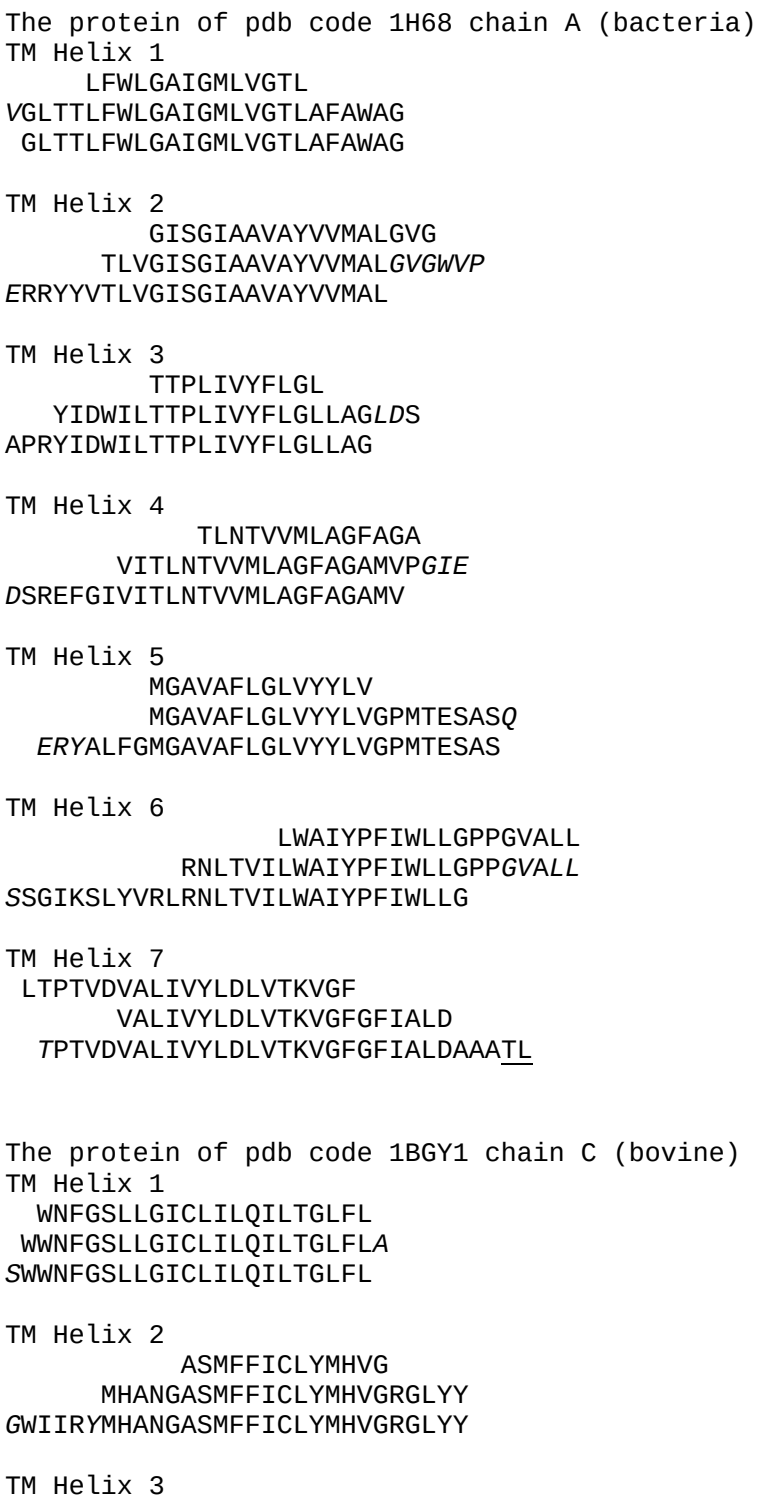
Table 2: Data on true positives as well as false positives for the TM2ndS analysis without using a multiple sequence alignment (one sequence used in analysis)

	Reference crystal structure helices	Predictions
Number total	156	155
Number predicted to overlap with >=15 residues	150	148
Percents	96.1 (true +)	95.5

Figure 1: A set of TM helix predictions before and after capping compared with the ranges from the MPtopo crystal structure database (for those proteins in which all the lengths of the crystal and predicted helices were >=20 and <=34). Those residues which did not fit the criteria for alpha helicity (in phi-psi angles) are indicated in italics, while those which have B-factors over 70 are indicated in bold.

Figure S1: A set of TM helix predictions before and after capping compared with the ranges from the MPtopo crystal structure database helices (which have lengths of the crystal reference helix of >=20 and <=34, "GPCR ranges"). Those residues which did not fit the criteria for alpha helicity (in phi-psi angles) are indicated in italics, while those which have B-factors over 70 are indicated in bold. Those helices which were predicted to be larger than 34 are indicated with a *. Those helices predicted to be smaller than 20 are indicated with **. Underlined residues have missing backbone atoms so that phi-psi angles cannot be calculated.

Figure 1:


The protein of pdb code 1H68 chain A (bacteria)
TM Helix 1
LFWLGAIGMLVGTL
VGLTTLFWLGAIGMLVGTLAFAWAG
GLTTLFWLGAIGMLVGTLAFAWAG
TM Helix 2
GISGIAAVAYVVMALGVG
TLVGISGIAAVAYVVMALGVGWVP
ERRYYVTLVGISGIAAVAYVVMAL
TM Helix 3
TTPLIVYFLGL
YIDWILTTPLIVYFLGLLAGLDS
APRYIDWILTTPLIVYFLGLLAG
TM Helix 4
TLNTVVMLAGFAGA
VITLNTVVMLAGFAGAMVPGIE
DSREFGIVITLNTVVMLAGFAGAMV
TM Helix 5
MGAVAFLGLVYYLV
MGAVAFLGLVYYLVGPMTESASQ
ERYALFGMGAVAFLGLVYYLVGPMTESAS
TM Helix 6
LWAIYPFIWLLGPPGVALL
RNLTVILWAIYPFIWLLGPPGVALL
SSGIKSLYVRLRNLTVILWAIYPFIWLLG
TM Helix 7
LTPTVDVALIVYLDLVTKVGF
VALIVYLDLVTKVGFGFIALD
TPTVDVALIVYLDLVTKVGFGFIALDAAATL
The protein of pdb code 1BGY1 chain C (bovine)
TM Helix 1
WNFGSLLGICLILQILTGLFL
WWNFGSLLGICLILQILTGLFLA
SWWNFGSLLGICLILQILTGLFL
TM Helix 2
ASMFFICLYMHVG
MHANGASMFFICLYMHVGRGLYY
GWIIRYMHANGASMFFICLYMHVGRGLYY
TM Helix 3

VILLLTVMATAFMGYVLPWGQMSFW
GVILLLTVMATAFMGYVL*PWGQ*M
*L*ETWNIGVILLLTVMATAFMGYVL

TM Helix 4
AFHFILPFIIMAIAMVHLLF
FFAFHFILPFIIMAIAMVHLLFLH
*K*ATL*T*RFFAFHFILPFIIMAIAMVHLLFLHE

TM Helix 5
ILGALLLILALMLLVLFA
KDILGALLLILALMLLVLF*AP**D*
*Y*TIKDILGALLLILALMLLVLF

TM Helix 6
GVLALAFSILILALIP
*N*KLGGVLALAFSILILALIPLL*HT*
KLGGVLALAFSILILALIPLL

TM Helix 7
FWALVADLLTLTW
PLSQCLFWALVADLLTLTWIGG*QP*
PLSQCLFWALVADLLTLTWIGG

TM Helix 8
TIGQLASVLYFLLILVLMPTAG
TIGQLASVLYFLLILVLMPTAG
HPY*I*TIGQLASVLYFLLILVLMPTAGT*I*ENKL

The protein of pdb code 1BGY5 chain J (bovine)
TM Helix 1
TSTFALTIVVGALLFER
*RR*TSTFALTIVVGALLFERAF
TSTFALTIVVGALLFERAF*DQ*G*A*DAIYEH*IN*

The protein of pdb code 1BRX chain A (bacteria)
TM Helix 1
WLALGTALMGLGTLYFL
EWIWLALGTALMGLGTLYFLVKG
EWIWLALGTALMGLGTLYFLVKG

TM Helix 2
TTLVPAIAFTMYLSMLLGYGLTMV
TTLVPAIAFTMYLSMLLG*YGLTMV*
*D*PDAKKFYAITTLVPAIAFTMYLSMLLG

TM Helix 3
TPLLLLDL
RYADWLFTTPLLLLDLALLVD*AD*Q
*A*RYADWLFTTPLLLLDLALLV

TM Helix 4
VGADGIMIGTGLVGALTKVYS
LVGADGIMIGTGLVGAL*TKVY*S

*D*QGTILALVGADGIMIGTGLVGAL

TM Helix 5
     WAISTAAMLYILYVLFF
 RFVWWAISTAAMLYILYVLFFG*FT*
YRFVWWAISTAAMLYILYVLFFG

TM Helix 6
               VVLWSAYPVVWLIGSEGAGI
       FKVLRNVTVVLWSAYPVVWLIG*S*
*R*PE*VAST*FKVLRNVTVVLWSAYPVVWLIG

TM Helix 7
 VPLNIETLLFMVLDVSAKVG
 *V*PLNIETLLFMVLDVSAKVGF
  PLNIETLLFMVLDVSAKVGFGLIL*LR*

The protein of pdb code 1OCC4 chain D (bovine)
TM Helix 1
   TVVGAAMFFIGFTALLLIWEKHYVYG
   TVVGAAMFFIGFTALLLIWEKHYV*YG*
EWKTVVGAAMFFIGFTALLLIWEKHYV

The protein of pdb code 1OCC5 chain G (bovine)
TM Helix 1
  TWRFLTFGLALPSVALCTLNS
  TWRFLTFGLALPSVALCTLNSWL***H***
ARTWRFLTFGLALPSVALCTLNSWL

The protein of pdb code 1OCC7 chain J (bovine)
TM Helix 1
             CLGGTLYSLYCLG
       RVTMTLCLGGTLYSLYCLGWAS*F*
*A*TDNILYRVTMTLCLGGTLYSLYCLGWAS

The protein of pdb code 1OCC8 chain K (bovine)
TM Helix 1
         LASGATFCVAVWVYMATQIGIE
         LASGATFCVAVWVYMATQ*IGIE*
FHD*KYG*NAVLASGATFCVAVWVYMATQ

The protein of pdb code 1OCC9 chain L (bovine)
TM Helix 1
       MTLFFGSGFAAPFF
  RLLAMMTLFFGSGFAAPFFIVRH
K*W*RLLAMMTLFFGSGFAAPFFIVRHQ*L*

The protein of pdb code 1OCC10 chain M (bovine)
TM Helix 1
LSVTFLSFLLPAGWVL
AIGLSVTFLSFLLPAGWVLYHL
PKEQAIGLSVTFLSFLLPAGWVLYH
The protein of pdb code 1AR12 chain B (bacteria)
TM Helix 1
QWLDHFVLYIITAVTIFVCLLLLIC
QWLDHFVLYIITAVTIFVCLLLLICIVR
LAHDQQWLDHFVLYIITAVTIFVCLLLLICIVR

TM Helix 2
IEVIWTLVPVLILVAIGAFSLPIL
IEVIWTLVPVLILVAIGAFSLPIL
NTPIEVIWTLVPVLILVAIGAFSLPILFR*SQE*

The protein of pdb code 1EHK2 chain B (bacteria)
TM Helix 1
EKGWLAFSLAMLFVFIALIAYTLATHTAGV
EKGWLAFSLAMLFVFIALIAYTLA***THT***AG*V*
EHKAHK*A*I*L*A***Y***EKGWLAFSLAMLFVFIALIAYTLA

The protein of pdb code 1FUM1 chain D (bacteria)
TM Helix 1
LFGAGGMWSAIIAPVMILLVGILLPLGLFP
LFGAGGMWSAIIAPVMILLV*GI*L**L***PLGLFP*
*DEP*VFWGLFGAGGMWSAIIAPVMILLV

TM Helix 2
FIGRVFLFLMIVLPL
*RVLAFAQSF*IGRVFLFLMIVLPLWCGL
FIGRVFLFLMIVLPLWCGLHRMHHAMHD*L*

The protein of pdb code 1FUM2 chain C (bacteria)
TM Helix 1
AVWFSIELIFGLF
GTAVPAVWFSIELIFGLFAL*K****NG***
YRFYMLR*E*GTAVPAVWFSIELIFGLFAL

TM Helix 2
DFLQNPVIVIINLITLAA
*GFVDFLQN*PVIVIINLITLAAALLH
PVIVIINLITLAAALLHTKTWFELA

TM Helix 3
KSLWAVTVVATIVIL
PIIKSLWAVTVVATIVILFVAL*Y*
*P*EPIIKSLWAVTVVATIVILFVAL

The protein of pdb code 1E12 chain A (bacteria)
TM Helix 1
SLWVNVALAGIAILVFV
*E*NALLSSSLWVNVALAGIAILVFVYMG
NALLSSSLWVNVALAGIAILVFVYMG

TM Helix 2
ATLMIPLVSISSYLGLLSGLTVGM
GATLMIPLVSISSYLGLL*SG*L*TVGM*
RPRLIWGATLMIPLVSISSYLGLL

TM Helix 3
STPMILLALGLLADVDLGS
RYLTWALSTPMILLALGLLAD*VDL*GS
*W*GRYLTWALSTPMILLALGLLA

TM Helix 4
VIAADIGMCVTGLAAAM
VIAADIGMCVTGLAAAMT*TSA*
*L*GSLFTVIAADIGMCVTGLAAAMT

TM Helix 5
FYAISCAFFVVVLSALVTD
FYAISCAFFVVVLSALV*T*DWAAS
LLFRWAFYAISCAFFVVVLSALV

TM Helix 6
LRVLTVVLWLGYPIVWAVGVEGLALVQSVGVT
LRVLTVVLWLGYPIVWAV*G*VE*GL*ALVQSVGVT
AEIFDTLRVLTVVLWLGYPIVWAV

TM Helix 7
DVFAKYVFAFI
GVTSWAYSVLDVFAKYVFAFILLRWV
VGVTSWAYSVLDVFAKYVFAFILLRWVA*N*

The protein of pdb code 1BL8 chain A (bacteria)
TM Helix 1
AGAATVLLVIVLLAGSYL
*WR*AAGAATVLLVIVLLAGSYLAVLAE
AAGAATVLLVIVLLAGSYLAVLAE***R***

TM Helix 2
WGRCVAVVVMVAGITSFGLVTAALA
WGRCVAVVVMVAGITSFGLVTAALATW***FVG***
WGRCVAVVVMVAGITSFGLVTAALATW

The protein of pdb code 1KZU1 chain A (bacteria)
TM Helix 1
ALLGSVTVIAILVHLAIL
IPALLGSVTVIAILVHLAILS*H*
*NPAI*GIPALLGSVTVIAILVHLAILS

The protein of pdb code 1KZU2 chain B (bacteria)
TM Helix 1
RVFLGLALVAHFLAFS
DGTRVFLGLALVAHFLAFSA*TP*
L*H*KYVIDGTRVFLGLALVAHFLAFSA

The protein of pdb code 1LGH1 chain A (?)
TM Helix 1
PSTWLPVIWIVATVVAIAVHAAV
PSTWLPVIWIVATVVAIAVHAAV
*N*PSTWLPVIWIVATVVAIAVHAAV*LAA*

The protein of pdb code 1LGH2 chain B (?)
TM Helix 1
IILAAVAHVLV
KTTFSAFIILAAVAHVLVWVWK*P*
*T*EEEAIAVHDQFKTTFSAFIILAAVAHVLVWVWK

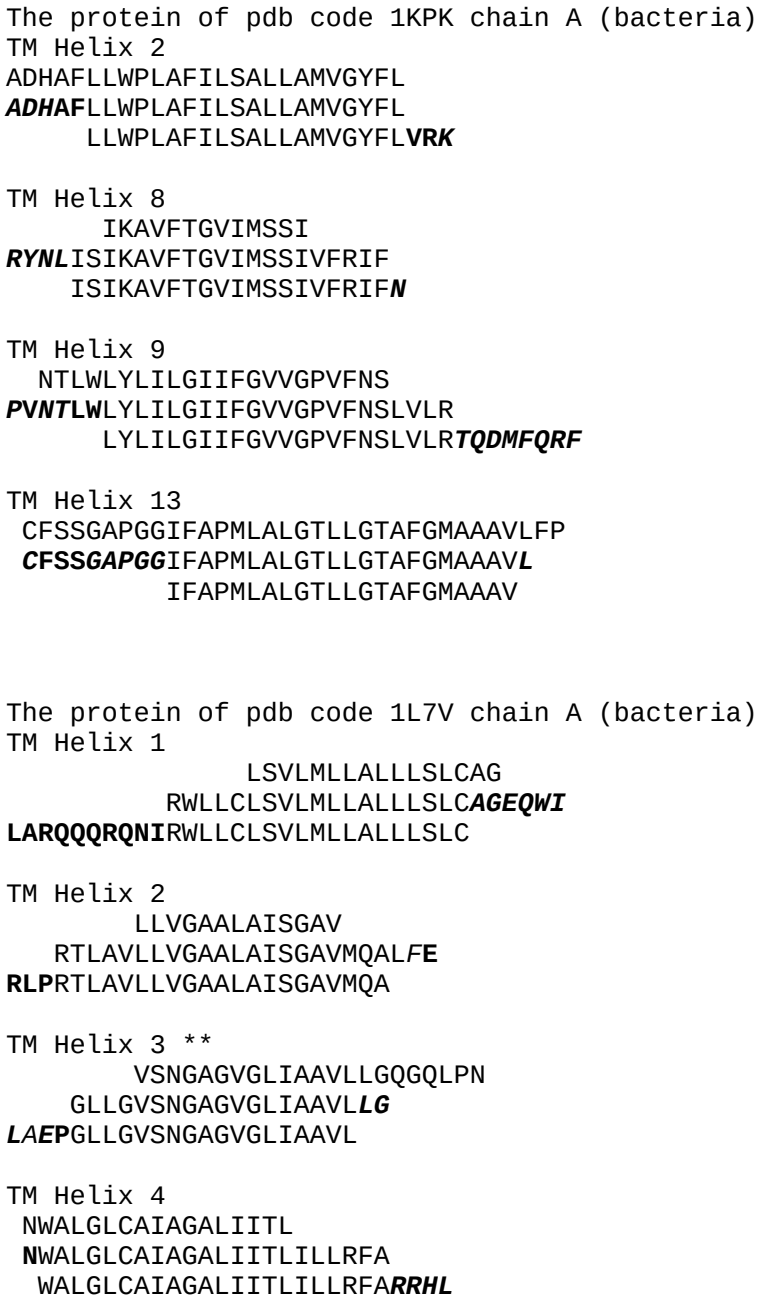
The protein of pdb code 1MSL chain B (bacteria TB)
TM Helix 1
IVDLAVAVVIGTAFTAL
*RGNI*VDLAVAVVIGTAFTALVTK
VDLAVAVVIGTAFTALVTKFT*DSI*IT*PLI*

TM Helix 2
QTIDLNVLLSAAINFFLIAFAVYFLVVL
*QTID*LNVLLSAAINFFLIAFAVYFLVVLP
LNVLLSAAINFFLIAFAVYFL

The protein of pdb code 2RCR1 chain H (bacteria)
TM Helix 1
FGNFDLASLAIYSFWIFLAGLIYYL
*FGNFDL*ASLAIYSFWIFLAGLIYYLQTE
ASLAIYSFWIFLAGLIYYLQTE*NMRE*G*Y*

The protein of pdb code 1EHK chain C (bacteria)
TM Helix 1
LVLTLTILVFWLGV
GALAVILVLTLTILVFWLGVYAVFFAR
KPKGALAVILVLTLTILVFWLGVYAVFFAR

Figure S1:


The protein of pdb code 1KPK chain A (bacteria)
TM Helix 2
ADHAFLLWPLAFILSALLAMVGYFL
ADHAFLLWPLAFILSALLAMVGYFL
LLWPLAFILSALLAMVGYFLVRK
TM Helix 8
IKAVFTGVIMSSI
RYNLISIKAVFTGVIMSSIVFRIF
ISIKAVFTGVIMSSIVFRIFN
TM Helix 9
NTLWLYLILGIIFGVVGPVFNS
PVNTLWLYLILGIIFGVVGPVFNSLVLR
LYLILGIIFGVVGPVFNSLVLRTQDMFQRF
TM Helix 13
CFSSGAPGGIFAPMLALGTLLGTAFGMAAAVLFP
CFSSGAPGGIFAPMLALGTLLGTAFGMAAAVL
IFAPMLALGTLLGTAFGMAAAV
The protein of pdb code 1L7V chain A (bacteria)
TM Helix 1
LSVLMLLALLLSLCAG
RWLLCLSVLMLLALLLSLCAGEQWI
LARQQQRQNIRWLLCLSVLMLLALLLSLC
TM Helix 2
LLVGAALAISGAV
RTLAVLLVGAALAISGAVMQALFE
RLPRTLAVLLVGAALAISGAVMQA
TM Helix 3 **
VSNGAGVGLIAAVLLGQGQLPN
GLLGVSNGAGVGLIAAVLLG
LAEPGLLGVSNGAGVGLIAAVL
TM Helix 4
NWALGLCAIAGALIITL
NWALGLCAIAGALIITLILLRFA
WALGLCAIAGALIITLILLRFARRHL

TM Helix 9
ALAGASALLLADIVA
SALLPGCALAGASALLLADIVAR
LLPGCALAGASALLLADIVAR***LA***

The protein of pdb code 1H68 chain A (bacteria)
TM Helix 1
LFWLGAIGMLVGTL
*V*GLTTLFWLGAIGMLVGTLAFAWAG
GLTTLFWLGAIGMLVGTLAFAWAG

TM Helix 2
GISGIAAVAYVVMALGVG
TLVGISGIAAVAYVVMAL*GVGWVP*
*E*RRYYVTLVGISGIAAVAYVVMAL

TM Helix 3
TTPLIVYFLGL
YIDWILTTPLIVYFLGLLAG*LD*S
APRYIDWILTTPLIVYFLGLLAG

TM Helix 4
TLNTVVMLAGFAGA
VITLNTVVMLAGFAGAMVP*GIE*
*D*SREFGIVITLNTVVMLAGFAGAMV

TM Helix 5
MGAVAFLGLVYYLV
MGAVAFLGLVYYLVGPMTESAS*Q*
*ERY*ALFGMGAVAFLGLVYYLVGPMTESAS

TM Helix 6
LWAIYPFIWLLGPPGVALL
RNLTVILWAIYPFIWLLGPP*GVALL*
*S*SGIKSLYVRLRNLTVILWAIYPFIWLLG

TM Helix 7
LTPTVDVALIVYLDLVTKVGF
VALIVYLDLVTKVGFGFIALD
*T*PTVDVALIVYLDLVTKVGFGFIALDAAATL

The protein of pdb code 1FQY chain A (human)
TM Helix 1
WRAVVAEFLATTLFVFISIGSALG
WRAVVAEFLATTLFVFISIGSALG
KL*F*WRAVVAEFLATTLFVFISIGSALG**F***K*

TM Helix 4
LMYIIAQCVGAIVATAILSGIT
LMYIIAQCVGAIVATAILSGI*TS*
LMYIIAQCVGAIVATAILSGI

TM Helix 8

  WIFWVGPFIGGALAVLIYDFIL
 HWIFWVGPFIGGALAVLIYD*F*IL
*N*HWIFWVGPFIGGALAVLIYD

The protein of pdb code 1A91 chain A (bacteria)
TM Helix 2
          GLVDAIPMIAVGLGLY
        VMGLVDAIPMIAVGLGLYVMFAV
*LL*RT*Q*FF*I*VMGLVDAIPMIAVGLGLYVMFA

The protein of pdb code 1BGY1 chain C (bovine)
TM Helix 1
  WNFGSLLGICLILQILTGLFL
 WWNFGSLLGICLILQILTGLFL*A*
*S*WWNFGSLLGICLILQILTGLFL

TM Helix 2
           ASMFFICLYMHVG
      MHANGASMFFICLYMHVGRGLYY
*G*WIIR*Y*MHANGASMFFICLYMHVGRGLYY

TM Helix 3
       VILLLTVMATAFMGYVLPWGQMSFW
      GVILLLTVMATAFMGYVL*PWGQ*M
*L*ETWNIGVILLLTVMATAFMGYVL

TM Helix 4
        AFHFILPFIIMAIAMVHLLF
      FFAFHFILPFIIMAIAMVHLLFLH
*K*ATL*T*RFFAFHFILPFIIMAIAMVHLLFLHE

TM Helix 5
     ILGALLLILALMLLVLFA
   KDILGALLLILALMLLVLF*A*P*D*
*Y*TIKDILGALLLILALMLLVLF

TM Helix 6
    GVLALAFSILILALIP
*N*KLGGVLALAFSILILALIPLL*HT*
 KLGGVLALAFSILILALIPLL

TM Helix 7
      FWALVADLLTLTW
PLSQCLFWALVADLLTLTWIGG*QP*
PLSQCLFWALVADLLTLTWIGG

TM Helix 8
    TIGQLASVLYFLLILVLMPTAG
    TIGQLASVLYFLLILVLMPTAG
HPY*I*TIGQLASVLYFLLILVLMPTAGT*I*ENKL

The protein of pdb code 1BGY2

The protein of pdb code 1BGY4

The protein of pdb code 1BGY5 chain J (bovine)
TM Helix 1
  TSTFALTIVVGALLFER
*RR*TSTFALTIVVGALLFERAF
  TSTFALTIVVGALLFERAF*DQ*G*A*DAIYEH*IN*

The protein of pdb code 1BGY6

The protein of pdb code 1BRX chain A (bacteria)
TM Helix 1
   WLALGTALMGLGTLYFL
EWIWLALGTALMGLGTLYFLVKG
EWIWLALGTALMGLGTLYFLVKG

TM Helix 2
          TTLVPAIAFTMYLSMLLGYGLTMV
          TTLVPAIAFTMYLSMLLG*YGLTMV*
*D*PDAKKFYAITTLVPAIAFTMYLSMLLG

TM Helix 3
         TPLLLLDL
 RYADWLFTTPLLLLDLALLVD*AD*Q
*A*RYADWLFTTPLLLLDLALLV

TM Helix 4
        VGADGIMIGTGLVGALTKVYS
       LVGADGIMIGTGLVGAL*TKVY*S
*D*QGTILALVGADGIMIGTGLVGAL

TM Helix 5
     WAISTAAMLYILYVLFF
 RFVWWAISTAAMLYILYVLFFG*FT*
YRFVWWAISTAAMLYILYVLFFG

TM Helix 6
               VVLWSAYPVVWLIGSEGAGI
       FKVLRNVTVVLWSAYPVVWLIG*S*
*R*PE*VAST*FKVLRNVTVVLWSAYPVVWLIG

TM Helix 7
 VPLNIETLLFMVLDVSAKVG
 *V*PLNIETLLFMVLDVSAKVGF
  PLNIETLLFMVLDVSAKVGFGLIL*LR*

The protein of pdb code 1EUL chain A (rabbit)
TM Helix 1
WELVIEQFEDLLVRILLLAACISFVLAWFEEG
 *WELV*IEQFEDLLVRILLLAACISFVLAWFEE***G***
     IEQFEDLLVRILLLAACISFVLAWFEE

TM Helix 2
 AFVEPFVILLILIANAIVGVWQERN
 AFVEPFVILLILIANAIVGVWQERN
   VEPFVILLILIANAIVGVWQERNA*E*NAIE*A*L

TM Helix 3 *
  LQQKLDEFGEQLSKVISLICVAVWLINIGHFNDPVHG
  LQQKLDEFGEQLSKVISLICVAVWLI***NIGHFNDPV*****H*****G***
TPLQQKLDEFGEQLSKVISLICVAVWLI

TM Helix 6
 ALIPVQLLWVNLVTDGLPATALGFNPP
 *ALI*PVQLLWVNLVTDGLPATALGFN*PP*
   IPVQLLWVNLVTDGLPATALGF

TM Helix 7 *
  WLFFRYMAIGGYVGAATVGAAAWWFMYAEDGPGVTYH
  WLFFRYMAIGGYVGAATVGAA***AWWFMYAE*****D*****GPGVTYH***
*S*GWLFFRYMAIGGYVGAATVGAA

The protein of pdb code 1OCC1 chain A (bovine mito)
TM Helix 1
     TLYLLFGAWAGMVGTALSLLI
    GTLYLLFGAWAGMVGTALSLLIR
HKDIGTLYLLFGAWAGMVGTALSLLIRAE

TM Helix 4
 GASVDLTIFSLHLAGVSSILGAINFITTIIN
 *G*ASVDLTIFSLHLAGVSSILGAINFITTIIN*M*
  ASVDLTIFSLHLAGVSSILGAINFITTIIN

TM Helix 5
 QTPLFVWSVMITAVLLLLSLPVLAAGITML
 *QTP*LFVWSVMITAVLLLLSLPVLAAGITML
    LFVWSVMITAVLLLLSLPVLAAGITMLLTD

TM Helix 6
 ILYQHLFWFFGHPEVYILILPGFGMISHI
 *ILYQHLFWF*FGHPEVYILILPGFGMISHI
          FGHPEVYILILPGFGMISHIVTY*Y*

TM Helix 8
      FTSATMIIAIPTGVKVFSWLATLH
   RAYFTSATMIIAIPTGVKVFSWLATLH
VDTRAYFTSATMIIAIPTGVKVFSWLATL

TM Helix 9
 PAMMWALGFIFLFTVGGLTGIVLANSSLD

PAMMWALGFIFLFTVGGLTGIV*LAN*SS*LD*
PAMMWALGFIFLFTVGGLTGIV

TM Helix 10
YYVVAHFHYVLSMGAVFAIMGGFVHWFPL
*T*YYVVAHFHYVLSMGAVFAIMGGFVHWFPL
YYVVAHFHYVLSMGAVFAIMGGFVHWFPLF

TM Helix 11
TWAKIHFAIMFVGVNMTFFPQHFL
TWAKIHFAIMFVGVNMTFFPQHFL
DTWAKIHFAIMFVGVNMTFFPQHFL*G*L

TM Helix 12
PDAYTMWNTISSMGSFISLTAVMLMVFIIWEAF
*PDA*YTMWNTISSMGSFISLTAVMLMVFIIWEAF
YTMWNTISSMGSFISLTAVMLMVFIIWEAF*A*S

The protein of pdb code 1OCC3 chain C (bovine mito)
TM Helix 2
FNSMTLLMIGLTTNML
*FNSMTL*LMIGLTTNMLTMYQWW*R*
TLLMIGLTTNMLTMYQWWRDVIRE*ST*

TM Helix 3
ILFIISEVLFFTGFFWAFYHSSLAPTPELG
GMILFIISEVLFFTGFFWAFYHSS*LAPTPELG*
PAVQKGLRYGMILFIISEVLFFTGFFWAFYHSS

TM Helix 4
PLLNTSVLLASGVSITW
*LE*VPLLNTSVLLASGVSITWAHH
VPLLNTSVLLASGVSITWAHHSL*M*

TM Helix 5
MLQALFITITLGVYFTLLQ
MLQALFITITLGVYFTLLQASE
RKHMLQALFITITLGVYFTLLQASEYY*E*

TM Helix 6
DGVYGSTFFVATGFHGLHVIIGSTFLIVCF
*D*GVYGSTFFVATGFHGLHVIIGSTFLIVCFFR
GVYGSTFFVATGFHGLHVIIGSTFLIVCFFRQL

TM Helix 7
AGAWYWHFVDVVWLFLYVSIYWWGS
AGAWYWHFVDVVWLFLYVSIY*WWG*
FG*F*EAGAWYWHFVDVVWLFLYVSI

The protein of pdb code 1OCC4 chain D (bovine)
TM Helix 1

   TVVGAAMFFIGFTALLLIWEKHYVYG
   TVVGAAMFFIGFTALLLIWEKHYV*YG*
EWKTVVGAAMFFIGFTALLLIWEKHYV

The protein of pdb code 1OCC5 chain G (bovine)
TM Helix 1
  TWRFLTFGLALPSVALCTLNS
  TWRFLTFGLALPSVALCTLNSWL***H***
ARTWRFLTFGLALPSVALCTLNSWL

The protein of pdb code 1OCC6

The protein of pdb code 1OCC7 chain J (bovine)
TM Helix 1
             CLGGTLYSLYCLG
       RVTMTLCLGGTLYSLYCLGWAS*F*
*A*TDNILYRVTMTLCLGGTLYSLYCLGWAS

The protein of pdb code 1OCC8 chain K (bovine)
TM Helix 1
         LASGATFCVAVWVYMATQIGIE
         LASGATFCVAVWVYMATQ*IGIE*
FHD*KYG*NAVLASGATFCVAVWVYMATQ

The protein of pdb code 1OCC9 chain L (bovine)
TM Helix 1
       MTLFFGSGFAAPFF
  RLLAMMTLFFGSGFAAPFFIVRH
K*W*RLLAMMTLFFGSGFAAPFFIVRHQ*L*

The protein of pdb code 1OCC10 chain M (bovine)
TM Helix 1
       LSVTFLSFLLPAGWVL
    AIGLSVTFLSFLLPAGWVLYHL
PKEQAIGLSVTFLSFLLPAGWVLYH

The protein of pdb code 1AR11 chain A (bacteria)
TM Helix 1
      ILYLFTAGIVGLISVCFTVY
    IGILYLFTAGIVGLISVCFTVYMR
NHKDIGILYLFTAGIVGLISVCFTVYMRME*L*Q*H*

TM Helix 3
      WMYVCGVALGVASLL

LNNLSYWMYVCGVALGVASLLA***PGG***
LNNLSYWMYVCGVALGVASLLA

TM Helix 4
SMDLAIFAVHVSGASSILGAINIIT
*S*MDLAIFAVHVSGASSILGAINIITTFLN***MR***
SMDLAIFAVHVSGASSILGAINIITTFLN

TM Helix 5
KVPLFAWSVFITAWLILLSLPVLAGAIT
K*V*PLFAWSVFITAWLILLSLPVLAGAIT
PLFAWSVFITAWLILLSLPVLAGAIT*M*L***L***M**D*****RNF***

TM Helix 8
MLATMTIAVPTGIKVFSWIAT
QAYFMLATMTIAVPTGIKVFSWIATMW
LTQQAYFMLATMTIAVPTGIKVFSWIATM

TM Helix 9
KTPMLWAFGFLFLFTVGGVTGVVLSQAPLD
KTPMLWAFGFLFLFTVGGVTGVVLSQAP
KTPMLWAFGFLFLFTVGGVTGVVLSQ

TM Helix 10
YHDTYYVVAHFHYVMSLGAVFGIFAGVYYWIGKM
DTYYVVAHFHYVMSLGAVFGIFAGVYY*W*IGKM
DTYYVVAHFHYVMSLGAVFGIFAGVYY

TM Helix 11
AGQLHFWMMFIGSNLIFFPQHFL
AGQLHFWMMFIGSNLIFFPQHFL
*P*EWAGQLHFWMMFIGSNLIFFPQHFLGR

TM Helix 12
NISSIGAYISFASFLFFIG
*EF*AYWNNISSIGAYISFASFLFFIG
AYWNNISSIGAYISFASFLFFIGIVFYT***L****F****A***

The protein of pdb code 1AR12 chain B (bacteria)
TM Helix 1
QWLDHFVLYIITAVTIFVCLLLLIC
QWLDHFVLYIITAVTIFVCLLLLICIVR
LAHDQQWLDHFVLYIITAVTIFVCLLLLICIVR

TM Helix 2
IEVIWTLVPVLILVAIGAFSLPIL
IEVIWTLVPVLILVAIGAFSLPIL
NTPIEVIWTLVPVLILVAIGAFSLPILFR*SQE*

The protein of pdb code 1QLE chain C (bacteria)
TM Helix 1
FGAIGAFVMLTGAVAWMKGITFFGL
IWPFFGAIGAFVMLTGAVAWM*KG*

IWPFFGAIGAFVMLTGAVAWM

TM Helix 2
 LPVEGPWMFLIGLVGVLY
 *LPVEG*PWMFLIGLVGVLYVMFGW
      PWMFLIGLVGVLYVMFGWWA*D*VVNEGE*TG*

TM Helix 4
    INTLILLLSGVAVTW
HLPLINTLILLLSGVAVTWAHHA
HLPLINTLILLLSGVAVTWAHHAFV*LE*

TM Helix 5
        IVAVILGVCFTG
  TTINGLIVAVILGVCFTGLQAYEY
*R*KTTINGLIVAVILGVCFTGLQAYEYS*HA*

TM Helix 6
      FYMATGFHGAHVIIGTIFLFV
    GAFYMATGFHGAHVIIGTIFLFVCLIR
*TVYA*GAFYMATGFHGAHVIIGTIFLFVCLIR*LLK*

TM Helix 7
       AWYWHFVDVVWLFLFVVIYIWGR
       AWYWHFVDVVWLFLFVVIYIWG
*H*VGFEAAAWYWHFVDVVWLFLFVVIYIWG

The protein of pdb code 1EHK1 chain A (bacteria)
TM Helix 1
       FLVLGFLALIVGSLFGPF
  KATLYFLVLGFLALIVGSLFGPFQAL
*EK*KATLYFLVLGFLALIVGSLFGPFQALNY*GN*

TM Helix 2
      LHGVLNAIVFTQLFAQAI
   GLTLHGVLNAIVFTQLFAQAIMVYLP
YYQGLTLHGVLNAIVFTQLFAQAIMVYLPAREL

TM Helix 3
    WLSWWMAFIGLVVAALPLLANEAT
 GLMWLSWWMAFIGLVVAALPLLA*NEAT*
*MG*LMWLSWWMAFIGLVVAALPLL

TM Helix 4
   AFYLGASVFVLSTWVSIY
*GHW*AFYLGASVFVLSTWVSIYIVLD
  WAFYLGASVFVLSTWVSIYIVLDLWRRWKAA

TM Helix 5 *
 LVTYMAVVFWLMWFLASLGLVLEAVLFLLPWSFGLVEG
 LVTYMAVVFWLMWFLASLGLVLEAVLFLLPWSF*GLVEG*
 LVTYMAVVFWLMWFLASLGLVLEAVLFLLPWS

TM Helix 6
          TGHPIVYFWLLPAYAIIYT

WWTGHPIVYFWLLPAYAIIYTILP
PLVARTLFWWTGHPIVYFWLLPAYAIIYTILPKQ

TM Helix 7
RLAFLLFLLLSTPVG
DPMARLAFLLFLLLSTPVGFHH*Q*
DPMARLAFLLFLLLSTPVGFHH

TM Helix 8
HSVLTLFVAVPSLMTAFTVAA
IHSVLTLFVAVPSLMTAFTVAASLE
PTWKMIHSVLTLFVAVPSLMTAFTVAASLEFA**GRL**

TM Helix 9
AFVAPVLGLLGFIPGGAGGIVNASF
AFVAPVLGLLGFIPGGAGGIVN*ASF*T*L*D
PAFVAPVLGLLGFIPGGAGGIVN

TM Helix 11
VWLWFLGMMIMAVGLHWA
GLAVVWLWFLGMMIMAVGLHWAGL*L*
***D*AQR**RLGLAVVWLWFLGMMIMAVGLHWAGL

TM Helix 12
VPMVFNVLAGIVLLVALLLFIYGLF
VPMVFNVLAGIVLLVALLLFIYGLF
P*HA*AVPMVFNVLAGIVLLVALLLFIYGLF**S*VLL***

TM Helix 13
RIGFWFAVAAILVVLAYGPTLV
*R*IGFWFAVAAILVVLAYGPTLV
IGFWFAVAAILVVLAYGPTLVQL*F*

The protein of pdb code 1EHK2 chain B (bacteria)
TM Helix 1
EKGWLAFSLAMLFVFIALIAYTLATHTAGV
EKGWLAFSLAMLFVFIALIAYTLA***THT***AG*V*
EHKAHK*A*I*L*A***Y***EKGWLAFSLAMLFVFIALIAYTLA

The protein of pdb code 1FUM1 chain D (bacteria)
TM Helix 1
LFGAGGMWSAIIAPVMILLVGILLPLGLFP
LFGAGGMWSAIIAPVMILLV*GI*L**L***PLGLFP*
*DEP*VFWGLFGAGGMWSAIIAPVMILLV

TM Helix 2
FIGRVFLFLMIVLPL
*RVLAFAQSF*IGRVFLFLMIVLPLWCGL
FIGRVFLFLMIVLPLWCGLHRMHHAMHD*L*

The protein of pdb code 1FUM2 chain C (bacteria)

TM Helix 1
AVWFSIELIFGLF
GTAVPAVWFSIELIFGLFAL***KNG***
YRFYMLR*E*GTAVPAVWFSIELIFGLFAL

TM Helix 2
DFLQNPVIVIINLITLAA
*GFVDFLQN*PVIVIINLITLAAALLH
PVIVIINLITLAAALLHTKTWFELA

TM Helix 3
KSLWAVTVVATIVIL
PIIKSLWAVTVVATIVILFVAL*Y*
*P*EPIIKSLWAVTVVATIVILFVAL

The protein of pdb code 1QLA chain C (?)
TM Helix 1
WQSATGLFLGLFMIGHMFFVSTILLGDN
DWWQSATGLFLGLFMIGHMFFVSTIL*LGDN*
MPAKLDWWQSATGLFLGLFMIGHMFFVSTIL

TM Helix 2
FEGGKPIVVSFLAAFVFAVFIAHAFLAM
*FEGGKP*IVVSFLAAFVFAVFIAHAFLAMR
IVVSFLAAFVFAVFIAHAFLAMR*K*

TM Helix 3
WIQAMTGFAMFFLGSVHLYIMMTQPQ
DTTLWWIQAMTGFAMFFLGSVHLYIMMT*QPQ*
GDTTLWWIQAMTGFAMFFLGSVHLYIMMT

TM Helix 4
EWMWPLYLVLLFAVELHGSV
*VSEWM*WPLYLVLLFAVELHGSVGL
WPLYLVLLFAVELHGSVGLYRLA*V*K*W*

The protein of pdb code 1FX8 chain A (bacteria)
TM Helix 1
EFLGTGLLIFFGVGCVAALKVA
IAEFLGTGLLIFFGVGCVAALKV
TLKGQCIAEFLGTGLLIFFGVGCVAALKV*A*

TM Helix 2 **
GQWEISVIWGLGVAMAIYLTAGV
QWEISVIWGLGVAMAIYLTA
G*Q*WEISVIWGLGVAMAIYLTA

TM Helix 4
VSQVAGAFCAAALVYGLYY
VSQVAGAFCAAALVYGLY*Y*NLFFD
*K*VIPFIVSQVAGAFCAAALVYGLY

TM Helix 5

```
     QAFAVEMVITAILMGLILA
HINFVQAFAVEMVITAILMGLILALTD
  NFVQAFAVEMVITAILMGLILALTDD

TM Helix 8
    LVPLFGPIVGAIV
IPYFLVPLFGPIVGAIVGAFA
  YFLVPLFGPIVGAIVGAFAYRKL



The protein of pdb code 1E12 chain A (bacteria)
TM Helix 1
       SLWVNVALAGIAILVFV
ENALLSSSLWVNVALAGIAILVFVYMG
 NALLSSSLWVNVALAGIAILVFVYMG

TM Helix 2
       ATLMIPLVSISSYLGLLSGLTVGM
      GATLMIPLVSISSYLGLLSGLTVGM
RPRLIWGATLMIPLVSISSYLGLL

TM Helix 3
         STPMILLALGLLADVDLGS
  RYLTWALSTPMILLALGLLADVDLGS
WGRYLTWALSTPMILLALGLLA

TM Helix 4
      VIAADIGMCVTGLAAAM
      VIAADIGMCVTGLAAAMTTSA
LGSLFTVIAADIGMCVTGLAAAMT

TM Helix 5
      FYAISCAFFVVVLSALVTD
      FYAISCAFFVVVLSALVTDWAAS
LLFRWAFYAISCAFFVVVLSALV

TM Helix 6
      LRVLTVVLWLGYPIVWAVGVEGLALVQSVGVT
      LRVLTVVLWLGYPIVWAVGVEGLALVQSVGVT
AEIFDTLRVLTVVLWLGYPIVWAV

TM Helix 7
           DVFAKYVFAFI
 GVTSWAYSVLDVFAKYVFAFILLRWV
VGVTSWAYSVLDVFAKYVFAFILLRWVAN



The protein of pdb code 1BL8 chain A (bacteria)
TM Helix 1
   AGAATVLLVIVLLAGSYL
WRAAGAATVLLVIVLLAGSYLAVLAE
  AAGAATVLLVIVLLAGSYLAVLAER

TM Helix 2
```

WGRCVAVVVMVAGITSFGLVTAALA
WGRCVAVVVMVAGITSFGLVTAALATW***FVG***
WGRCVAVVVMVAGITSFGLVTAALATW

The protein of pdb code 1KZU1 chain A (bacteria)
TM Helix 1
ALLGSVTVIAILVHLAIL
IPALLGSVTVIAILVHLAILS*H*
*NPAI*GIPALLGSVTVIAILVHLAILS

The protein of pdb code 1KZU2 chain B (bacteria)
TM Helix 1
RVFLGLALVAHFLAFS
DGTRVFLGLALVAHFLAFSA*TP*
L*H*KYVIDGTRVFLGLALVAHFLAFSA

The protein of pdb code 1LGH1 chain A (?)
TM Helix 1
PSTWLPVIWIVATVVAIAVHAAV
PSTWLPVIWIVATVVAIAVHAAV
*N*PSTWLPVIWIVATVVAIAVHAAV*LAA*

The protein of pdb code 1LGH2 chain B (?)
TM Helix 1
IILAAVAHVLV
KTTFSAFIILAAVAHVLVWVWK*P*
*T*EEEAIAVHDQFKTTFSAFIILAAVAHVLVWVWK

The protein of pdb code 1MSL chain B (bacteria TB)
TM Helix 1
IVDLAVAVVIGTAFTAL
*RGNI*VDLAVAVVIGTAFTALVTK
VDLAVAVVIGTAFTALVTKFT*DSI*IT*PLI*

TM Helix 2
QTIDLNVLLSAAINFFLIAFAVYFLVVL
*QTID*LNVLLSAAINFFLIAFAVYFLVVLP
LNVLLSAAINFFLIAFAVYFL

The protein of pdb code 2RCR1 chain H (bacteria)
TM Helix 1
FGNFDLASLAIYSFWIFLAGLIYYL
*FGNFDL*ASLAIYSFWIFLAGLIYYLQTE
ASLAIYSFWIFLAGLIYYLQTE*NMRE*G*Y*

The protein of pdb code 2RCR2 chain L (bacteria)
TM Helix 1 *
GGNLFDFWVGPFYVGFFGVATFFFAALGIILIAWSAVL
*GGNLFDFWVGPFY*VGFFGVATFFFAALGIILIAWSAV*L*
VGFFGVATFFFAALGIILIAWSAV

TM Helix 2
APLAKGGLWQIITICATGA
*GGAPLAK*GGLWQIITICATGAFV
GGLWQIITICATGAFVSWALREVEI*CRK*

TM Helix 3
HIPFAFAFAILAYLTLVLFRPVMMG
HIPFAFAFAILAYLTLVLFRPVMM*G*
*GY*HIPFAFAFAILAYLTLVLFRPVMM

TM Helix 4
HMIAISFFFTNALALALHGALV
AHMIAISFFFTNALALALHGALVL
*YNP*AHMIAISFFFTNALALALHGALVLS*AA*

TM Helix 5
TLGIHRLGLLLSLSAVFFSALCMIITGTIW
TLGIHRLGLLLSLSAVFFSALCMII*TGTIW*
*IG*TLGIHRLGLLLSLSAVFFSALCMII

The protein of pdb code 2RCR3 chain M (bacteria)
TM Helix 1 *
NAQLGPIYLGSLGVLSLFSGLMWFFTIGIWFWYQAGW
*NAQLGPI*YLGSLGVLSLFSGLMWFFTIGIWFWY*QAGW*
YLGSLGVLSLFSGLMWFFTIGIWFWY

TM Helix 3
HTAWAFLSAIWLWMVLGFIRPILMG
HTAWAFLSAIWLWMVLGFIRPI*LMG*
*GMGK*HTAWAFLSAIWLWMVLGFIRPI

TM Helix 4
PFHGLSIAFLYGSALLFAMHGATIL
PFHGLSIAFLYGSALLFAMHGATI*L*
*LFYN*PFHGLSIAFLYGSALLFAMHGATI

TM Helix 5
RWAIWMAVLVTLTGGIGILLSGTVVD
RWAIWMAVLVTLTGGIGI*LLSGTVVD*
*FNATME*GI*H*RWAIWMAVLVTLTGGIGI

The protein of pdb code 1PRC2 chain L (bacteria)
TM Helix 1 *

GDLFDFWVGPYFVGFFGVSAIFFIFLGVSLIGYAASQ
*GDLFDFWVGPYF*VGFFGVSAIFFIFLGVSLIGYAAS*Q*
VGFFGVSAIFFIFLGVSLIGYAAS

TM Helix 2
APLLEGGFWQAITVCALGAF
*GAAPLLE*GGFWQAITVCALGAFIS
GGFWQAITVCALGAFISWMLREVEISRK*L*

TM Helix 3
HVPLAFCVPIFMFCVLQVFRPLLLG
HVPLAFCVPIFMFCVLQVFRPLLL*G*
*GW*HVPLAFCVPIFMFCVLQVFRPLLL

TM Helix 4
HMSSVSFLFVNAMALGLHGGLI
GHMSSVSFLFVNAMALGLHGGLIL
*YN*PGHMSSVSFLFVNAMALGLHGGLILSV*A*

TM Helix 5
ALSIHRLGLFLASNIFLTGAFGTIASGPFW
ALSIHRLGLFLASNIFLTGAFGTIA*SG*P*FW*
I*G*ALSIHRLGLFLASNIFLTGAFGTIA

The protein of pdb code 1PRC3 chain M (bacteria)
TM Helix 1 *
AQIGPIYLGASGIAAFAFGSTAILIILFNMAAEVHF
*AQIGPIY*LGASGIAAFAFGSTAILIILFNMAAEV*H*F
LGASGIAAFAFGSTAILIILFNMAAEV

TM Helix 2
PLHDGGWWLMAGLFMTLSLGSW
*PLH*DGGWWLMAGLFMTLSLGSW
DGGWWLMAGLFMTLSLGSWWIRVYS*RA*RA*L*

TM Helix 3
HIAWNFAAAIFFVLCIGCIHPTLVG
HIAWNFAAAIFFVLCIGCIHPTLV*G*
*GT*HIAWNFAAAIFFVLCIGCIHPTLV

TM Helix 4
PWHGFSIGFAYGCGLLFAAHGATIL
PWHGFSIGFAYGCGLLFAAHGATIL
*YC*PWHGFSIGFAYGCGLLFAAHGATIL*AV*

TM Helix 5
RWGWFFSLMVMVSASVGILLTGTFVD
RWGWFFSLMVMVSASVGILL*TGT*F*VD*
*AT*IESVHRWGWFFSLMVMVSASVGILL

The protein of pdb code 1F88 chain A (bovine)
TM Helix 1
EPWQFSMLAAYMFLLIMLGFPINFLTL

EPWQFSMLAAYMFLLIMLGFPINFLTLYVTVQH
EPWQFSMLAAYMFLLIMLGFPINFLTLYVTVQH

TM Helix 2
LLNLAVADLFMVFGGFTTTLYTSLHGYFV
LNYILLNLAVADLFMVFGGFTTTLYTSLHG
*T*PLNYILLNLAVADLFMVFGGFTTTLYTSLHG

TM Helix 4
IMGVAFTWVMALACAAPPLV
HAIMGVAFTWVMALACAAPPLV*GW*
*G*ENHAIMGVAFTWVMALACAAPPLV

TM Helix 5
SFVIYMFVVHFIIPLIVIFFCYGQLV
ESFVIYMFVVHFIIPLIVIFFCYGQL***V***
NNESFVIYMFVVHFIIPLIVIFFCYGQL

TM Helix 6
IIMVIAFLICWLPYAGVAFYIFTHQGSDF
RMVIIMVIAFLICWLPYAGVAFYIFTH*Q*
KAEK*EV*TRMVIIMVIAFLICWLPYAGVAFYIFTH

TM Helix 7
PIFMTIPAFFA
PIFMTIPAFFAKTSAVYNPVI
*G*PIFMTIPAFFAKTSAVYNPVIY*IMMN*

The protein of pdb code 1EHK chain C (bacteria)
TM Helix 1
LVLTLTILVFWLGV
GALAVILVLTLTILVFWLGVYAVFFAR
KPKGALAVILVLTLTILVFWLGVYAVFFAR

Chapter 4: The prediction of the transmembrane hydrophobic center

Abstract

In the prediction of structure of G-protein-coupled receptors (GPCR's) using the MembStruk, the accurate determination of the hydrophobic center of each helix is one of the critical steps. This paper demonstrates how this calculated value corresponds to the physical membrane bilayer center and evaluates its accuracy and generality to bovine rhodopsin and well as three other bacterial 7-helical proteins of known structure. The predicted hydrophobic centers (HC) for bovine rhodopsin fit to a plane with a 0.78 Å RMS error. In addition, experiments corroborate this HC as the actual physical center of the membrane bilayer. These results support the use of this HC to initially orient helices and to provide the center for the calculation of the hydrophobic moment used in rotational orientation of the helices.

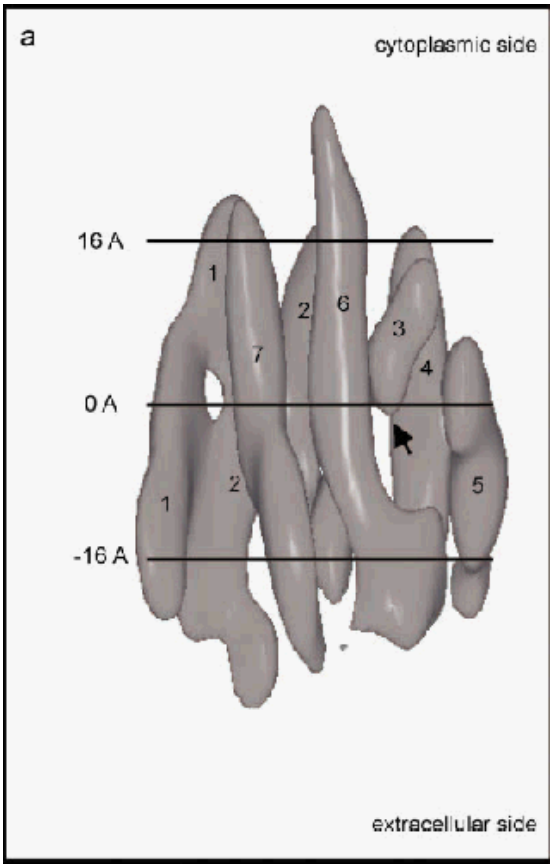
Introduction

The MembStruk protocol (Floriano et al., 2000; Vaidehi et al., 2002; Trabanino et al., 2004) uses mostly first principles in determining the structures of GPCRs. It has been applied successfully in various cases (Floriano et al., 2004; Freddolino et al., 2004; Kalani et al., 2004), yielding good correlation to binding, mutation, activation, and structural data. In various critical steps, the accurate independent prediction of structural features such as the TM (transmembrane) helical extent, the hydrophobic center, and the helix rotations distinguish this method from homology modeling or structure building with distance constraints from experiments (Strader et al., 1994; Herzyk et al., 1995). Such first principles determination of these values may be a large part of the reason for its success, since GPCRs of low homology to rhodopsin (10-20%) can be built.

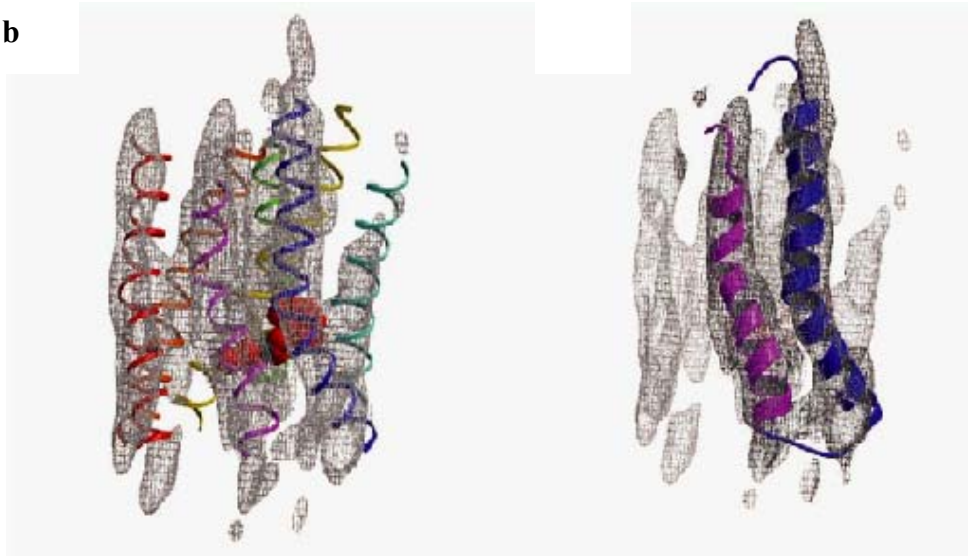
A transmembrane helix may be thought of as a buoy within a sea of lipid bilayer. Probably one of the most important driving forces for orienting this helix within this bilayer is the hydrophobic effect, positioning the most hydrophobic residues at or near the lipid bilayer center.

Usually, experimental TM protein structures do not directly yield information as to the membrane interfaces or the membrane center. However, recently the electron cryomicroscopy structure of bovine rhodopsin was determined from two-dimensional crystals of $p22_1\ 2_1$ symmetry (Krebs et al., 2003). In this crystal, the adjacent rhodopsin molecules were oriented upside down with respect to one another. This allowed the determination of the orientation of the molecule with respect to the membrane bilayer. Thus the hydrophobic center plane of the molecule can be roughly described, as shown in Figure 1.

Figure 1: The membrane plane relative to the bovine rhodopsin cryo-EM structure. Figure: (Krebs et al., 2003)



b



In addition, the non-crystallographic method of site-directed spin labeling has been useful in accurately determining the hydrophobic center residue in transmembrane proteins such as bacteriorhodopsin, KcsA, and the mechanosensitive channel (Altenbach et al., 1994; Hubbell et al., 1994; Gross et al., 1999; Perozo et al., 2001). This was achieved by determining the accessibility of a spin-labeled protein site to a soluble paramagnetic species as previously described (Altenbach et al., 1994).

The theoretical determination of the helical orientation relative to the membrane bilayer has been attempted (Tseitin et al., 1999) by moving TM fragments across a water-octanol-water system and assessing the energy (intramolecular and solvation) of this rigid helix as it crosses the membrane. They report a -0.15 +- 3.12 residue error in 2/3 of the helix cases and a 2.17 +- 3.07 residue error in the rest of the cases. One issue, though, is that their definition of the membrane interface is not directly correlated with experiment, but is based largely on the position of the ends of the TM helical segments, which does not usually correspond accurately to the membrane boundary.

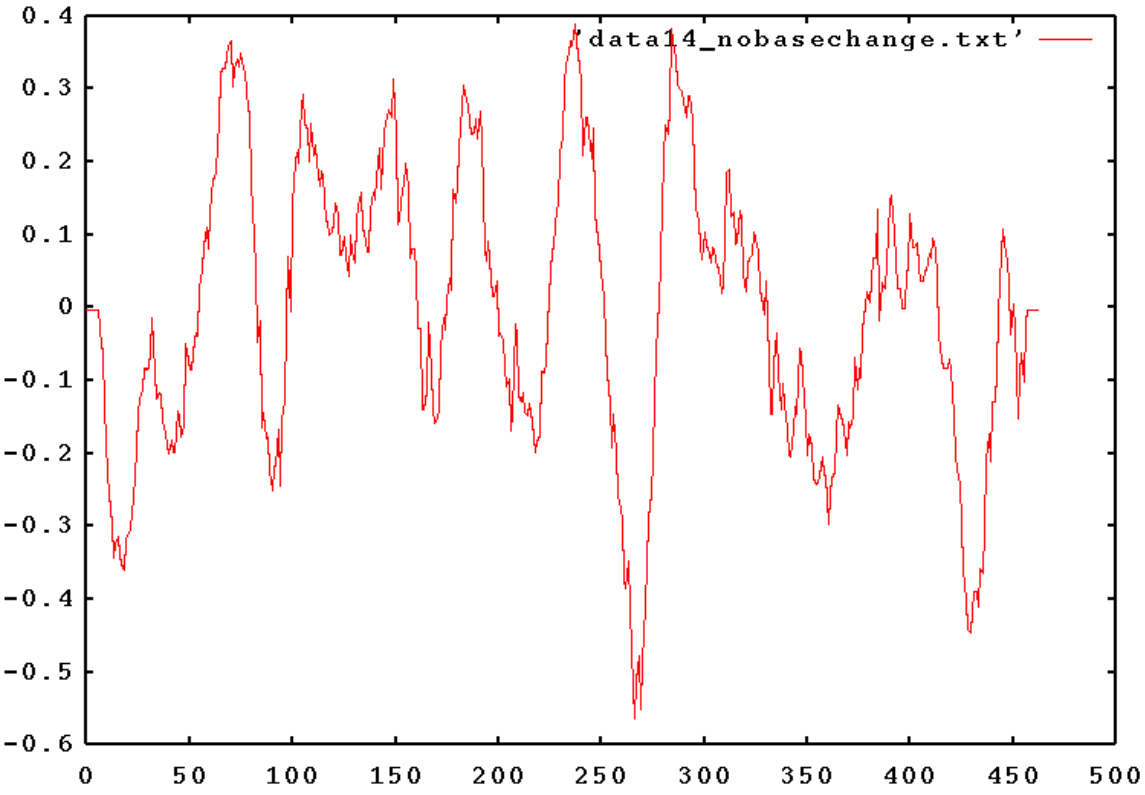
The TM2ndS protocol of MembStruk not only predicts the extent of the TM helices but also the hydrophobic centers of these helices, as previously described (Trabanino et al., 2004). This protocol has been used to obtain a bovine rhodopsin structure which deviates by 2.84 Å CRMS (coordinate root mean-square) in the TM region. This paper describes this method of hydrophobic center prediction and evaluates its accuracy and usefulness in building 3D structures of membrane proteins.

Methods

The procedure for the determination of the hydrophobic center by the get_centers protocol in TM2ndS has been previously described in detail (Trabanino et al., 2004). It will now currently be described in greater detail in the context of the TM proteins of bovine rhodopsin (pdb code 1F88) and bacteriorhodopsin (pdb code 1CSM).

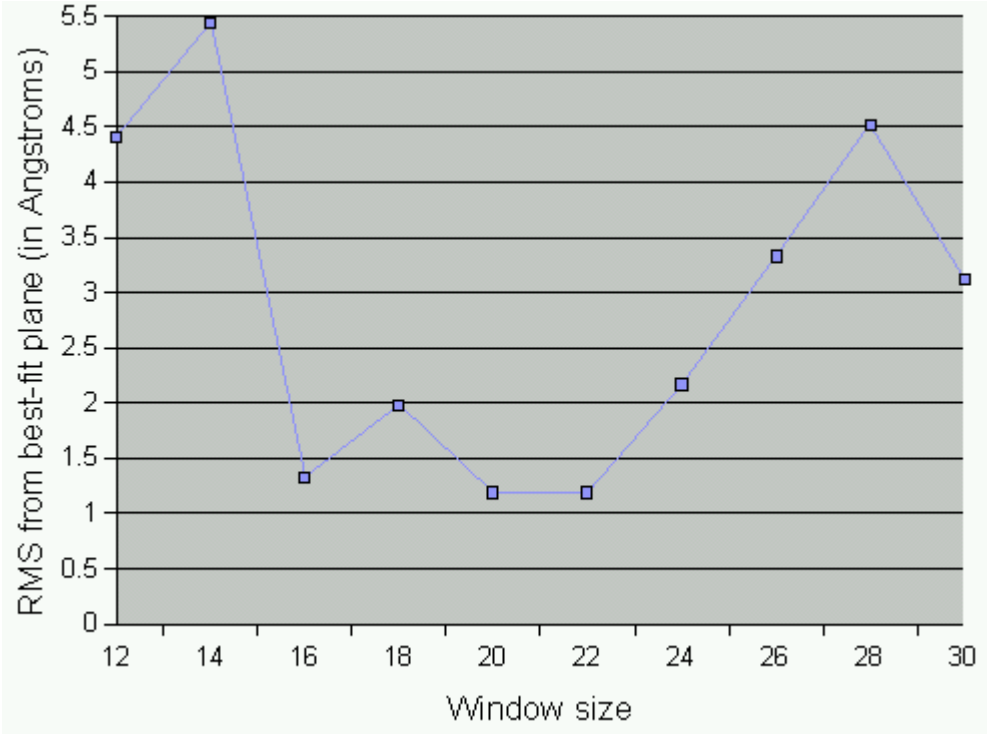
The hydrophobic profiles for window sizes 12-30 are generated by TM2ndS as described previously (Trabanino et al., 2004). Essentially, the profile is a plot of the average hydrophobicity (using the Eisenberg scale in this case) over a certain number of residues, designated as the window size. The profile for window size 14 for bovine rhodopsin is shown in Figure 2.

Figure 2: The TM2ndS hydrophobic profile at window size 14.



The maxima of these 7 peaks are used to determine the hydrophobic center. Since the maxima are actually averages over a window size, these indicate the maximum hydrophobic buoys for each helix. The positions vary across window size, as shown in Trabanino et al, 2004 and the fit to a common plane seems to be best for window size close to 20 (Figure 3).

Figure 3: The value of the fit to a plane for the predicted HCs at different window sizes in bovine rhodopsin.



The hydrophobic profiles for window size 12-24 are shown in Figure 4. The variations in the peak maxima are readily seen. The variations of these maxima from window size 20 are analyzed by TM2ndS to determine a stable region of stability for the

HC, as shown in Table 1. These values are averaged to yield one HC for each helix. This protocol was run on bovine rhodopsin and three bacterial 7-helical proteins.

Figure 4: The overlayed hydrophobic profiles for window sizes 12-24.

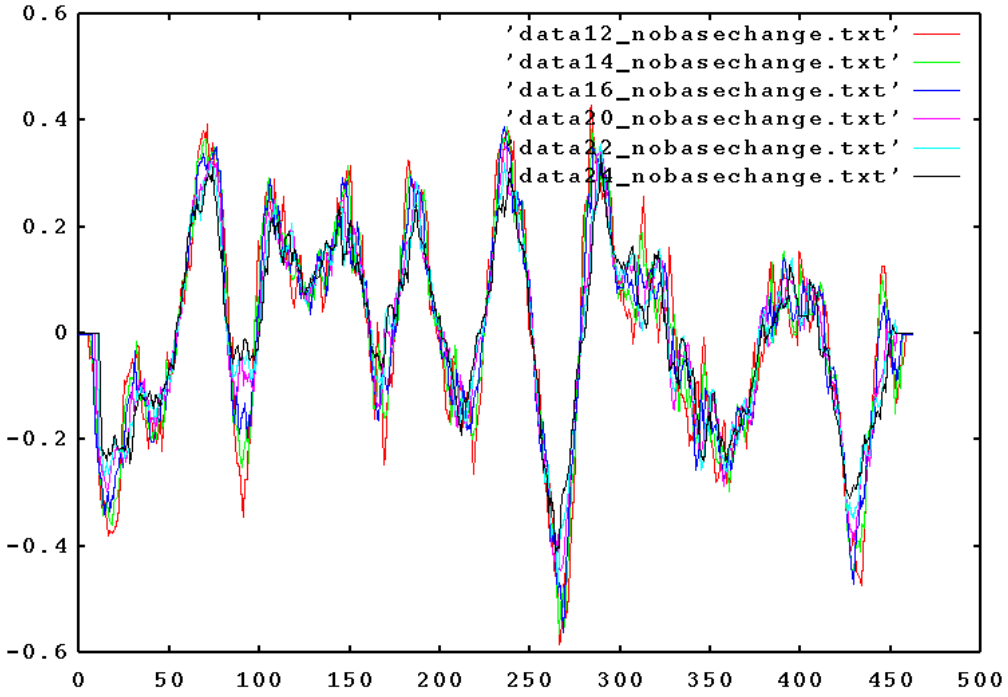


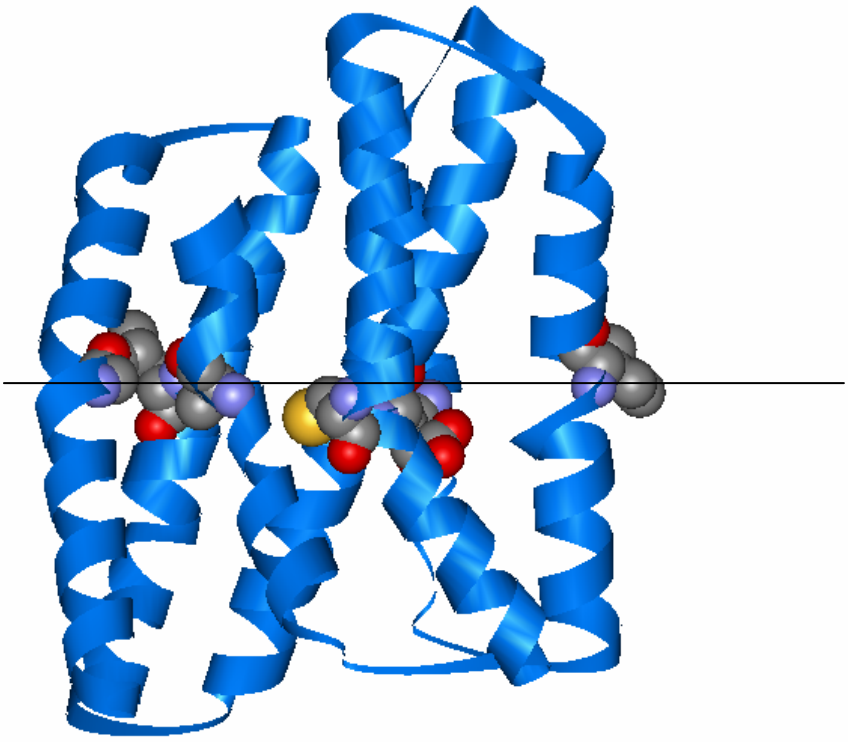
Table 1: The raw predicted HC's at each window size for bovine rhodopsin.

Helix number	Window size										
	12	14	**16**	**18**	**20**	**22**	**24**	26	28	30	HC
1	15	13	20	18	18	17	18	15	16	13	18.2
2	20	12	12	14	15	15	14	22	19	20	14.0
3	19	20	17	18	15	16	15	12	11	12	16.2
4	9	9	10	15	12	13	13	12	11	17	12.6
5	15	19	13	12	14	16	16	17	16	15	14.2
6	8	9	11	11	13	14	14	15	16	17	12.6
7	19	4	17	15	14	14	13	12	11	10	14.6

Results and discussion

The positions of the predicted HC for each helix in bovine rhodopsin are shown in Figure 5. Interestingly, these HC residues lie on a common plane with a fit of 0.79 Å. And when comparing to the experimentally determined membrane bilayer plane by cryo-EM in Figure 1, the planes actually correspond quite closely.

Figure 5: The residue positions of the HC's relative to the experimental cryo-EM bilayer center in bovine rhodopsin.



Fit to plane=0.78 Å

The HC variations across window sizes are different in the case of bacteriorhodopsin as shown in Table 2.

Table 2: The raw values of the HC's at each window size for bacteriorhodopsin. 1CSM fit to plane = 1.78 Å

Helix	Window sizes 12-30									
	9	10	11	*12*	*11*	14	14	13	7	6
	3	**6**	**10**	*12*	*9*	10	9	11	13	12
	12	12	13	*11*	*12*	22	**8**	18	24	24
	4	8	9	*7*	*7*	7	12	11	12	13
	17	13	12	*11*	*13*	13	9	7	7	8
	15	**14**	**15**	***16***	*20*	19	20	21	20	22
	6	6	5	*17*	*15*	2	1	13	12	11

The bacteriorhodopsin structure consists generally of shorter helices and of more charged residues. Also, unlike the bovine rhodopsin case, the variations across window sizes are localized into more than one window range. For example, for helix 6, the window sizes 12-18 comprise a stable region for the HC value quite different than the range for bovine rhodopsin (16-24 as in Table 1). So for cases where there is more than one range of HC stability, the range with the lower value is chosen. The fit to a common plane was 1.78 Å. This was applied in the cases of helix 2,3, and 6. Interestingly, a similar trend was observed for the other two bacterial 7-helical proteins halorhodopsin and sensory rhodopsin II (pdb codes 1E12 and 1H68), as shown in Tables 3 and 4.

Table 3: The raw values of the HC's at each window size for halorhodopsin. 1E12 (mutated protein) fit to plane = 2.7 Å

Helix	Window sizes 12-30
	19 20 17 18 14 15 15 14 13 10
	5 5 6 7 8 9 10 11 12 13
	13 12 11 11 21 22 23 24 25 26
	12 5 6 7 8 6 7 11 3 2
	12 11 9 8 9 10 6 8 4 3
	12 10 11 15 16 14 15 16 17 18
	17 16 15 14 12 11 11 9 8 8

Table 4: The raw values of the HC's at each window size for sensory rhodopsin. 1H68 Fit to plane= 2.4 Å

Helix	Window sizes 12-30
	11 14 14 15 11 12 12 13 14 15
	14 15 15 11 13 14 14 13 12 17
	15 11 13 11 11 12 **8 8** 18 17
	13 14 6 9 8 10 9 21 21 22
	4 3 6 6 7 7 9 1 1 2
	11 10 11 14 15 14 17 16 20 20
	19 11 10 10 9 12 13 5 6 5

The presence of charged and non-charged polar residues within the TM regions of these proton pumps and ion channels lead to more difficulty in finding a hydrophobic center than in the case of bovine rhodopsin. The presence of these residues also leads to interactions with other helices and with the phosphate heads of lipids which may

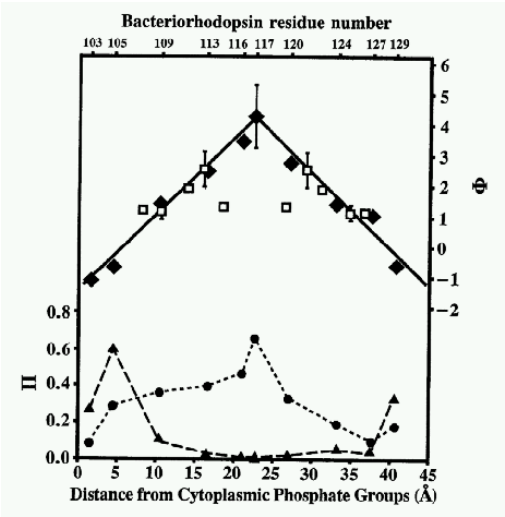
constrain the helices in a vertical orientation that deviates from the predicted orientation (which assumes independent helix properties). In addition, the 1E12 protein is a bioengineered protein with a mutation in a critical position at the beginning of helix 7, so results for it are difficult to interpret. One other thing to note is that these bacterial 7-helical proteins bind to all-trans-retinal and thus may correspond to "activated" states of the receptor when comparing to bovine rhodopsin. This may lead to shifts in some helices from the bilayer center.

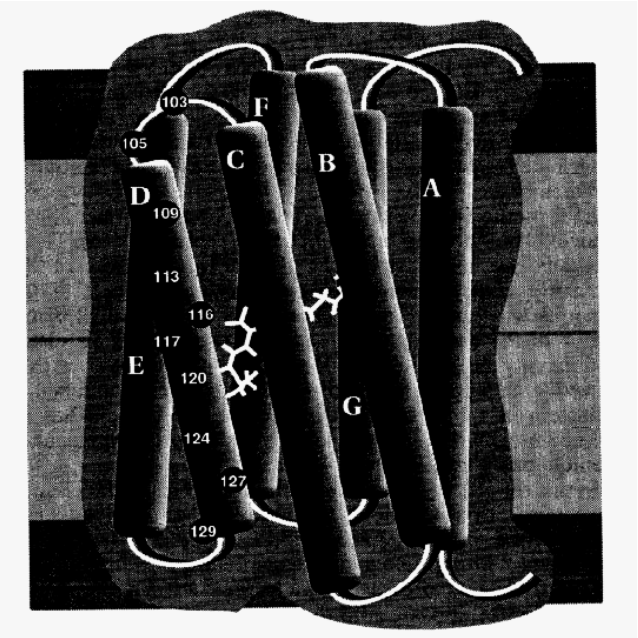
How exactly does the predicted HC correspond to the bilayer center? The output of the get_centers program gives the HC variation with window size in Figure 5, and the default predicted HC as compared with that determined by site-directed spin labeling studies (Altenbach et al., 1994) . Thus, the predicted vertical orientation of at least the helix 4 of bacteriorhodopsin agrees well with the experimental HC (results of experiment shown in Figure 6).

Figure 5: get_centers output showing the HC values across window sizes and the values chosen for the final HC calculation in bacteriorhodopsin. Also, the experimental HC (bold) and the predicted HC (underlined) are compared.

```
4 8 9 7 7 7 12 11 12 13

7 7
For helix 4 the hydrophobic center index is 7


DQGTILALVGADG I MIGTGLVGAL
```

Figure 6: The experimental results for HC determination for bacteriorhodopsin (a) with spin labeled sites indicated (b). Figure: (Altenbach et al., 1994)





How well can a hydrophobic center be predicted in an ion channel? As previously mentioned, the immersion depth of spin labeled sites was predicted for KcsA (Gross et al., 1999) and the mechanosensitive channel (Perozo et al., 2001). The data for KcsA, as shown in Figure 7, is incomplete due to poor labeling of some sites near the center of the membrane. The get_centers output is shown in Figure 8, together with the predicted and

experimental HC determinations. In this case, there is a 3 residue difference, although this exact value is not discernible due to the incomplete experimental data.

However, for the case of the MscL (mechanosensitive channel), the data is complete, as shown in Figure 9. There is a 3-residue difference between the predicted HC and the experimental immersion depth maximum, as shown in Figure 10. There are two ranges of HC stability, although not as distinct as in the previous cases of bacterial 7-helical proteins.

It should be noted at this point that this experimental immersion depth determination labels the protein on the membrane-exposed face of the helix. Thus, the maximum value of the plot may not correspond exactly to the residue position at the center of the membrane. That maximum value may be up to a turn away from the actual HC position. In addition, unlike the case of bacteriorhodopsin, the experiments were carried out in dodecyl maltoside (as opposed to phosphatidylcholine in the case of bacteriorhodopsin).

Since the current focus of MembStruk is to predict structures of GPCRs, the accurate determination of the HC in the case of bovine rhodopsin is encouraging. Even the uncertainties and errors of up to 3 residues in the cases of the other bacterial ion transport proteins would be low enough to initially position the helices for optimization in bundle and explicit lipid bilayer, an important fine optimization within the MembStruk protocol.

Figure 7: Experimental results for the HC determination in KcsA. Figure: (Gross et al., 1999)

Incomplete data

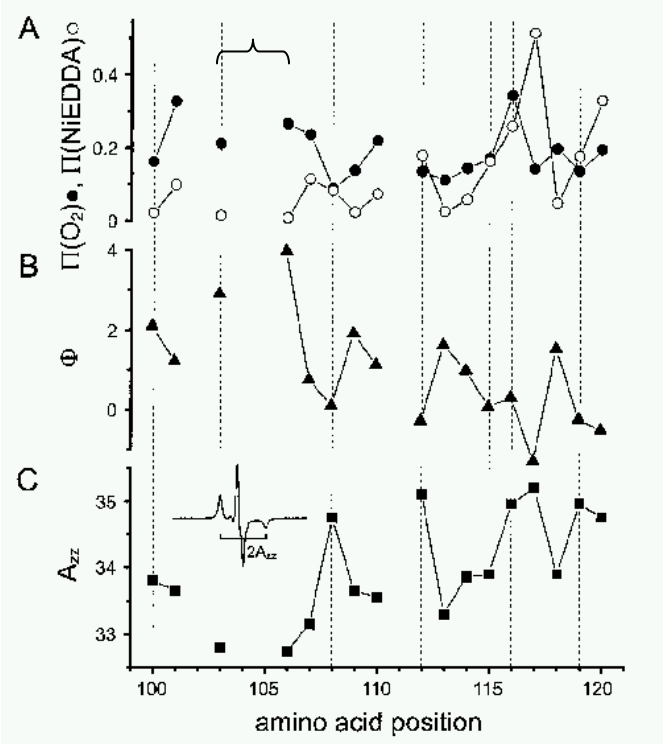


Figure 8: get_centers output showing the HC values across window sizes and the values chosen for the final HC calculation in KcsA. Also, the experimental HC (bold) and the predicted HC (underlined) are compared.

```
11 12 12 14 15 18 17 17 15 15

11 12 12 14 15 18 17 17 15 15
For helix 2 the hydrophobic center index is 14.6


WGRCVAVVVMVAGITSF GLV TAALATW
```

Figure 9: Experimental results for the HC determination in MscL. Figure: (Perozo et al., 2001)

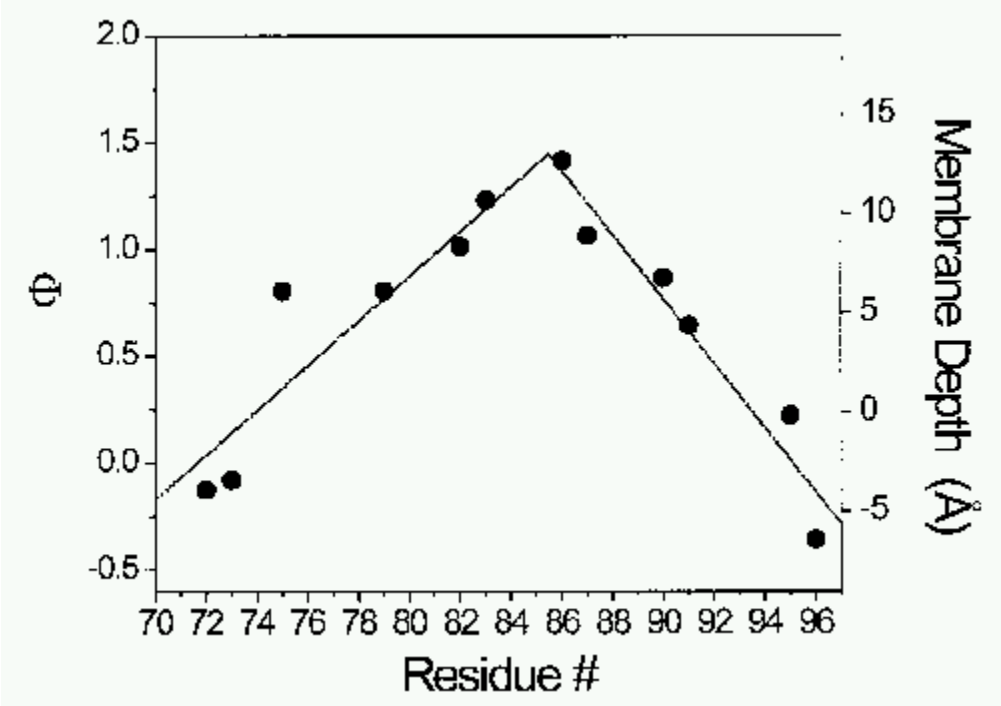


Figure 10: get_centers output showing the HC values across window sizes and the values chosen for the final HC calculation in MscL. Also, the experimental HC (bold) and the predicted HC (underlined) are compared.

```
21 22 20 20 17 18 17 16 16 16

20 20 17
For helix 2 the hydrophobic center index is 19


LNVLLSAAINFFLIAFAVYFL
```

Conclusion

The get_centers protocol for obtaining the hydrophobic center positions based on sequence information alone provides the ability to build structure independent structures of GPCRs with low homology to bovine rhodopsin. Its excellent agreement with experiment in the case of bovine rhodopsin increases confidence in its application in the case of GPCRs, which have fewer number of polar residues than ion channels and proton pumps. Nevertheless, the fact that the method still performs reasonably well for such cases allows for the method's applicability in building structures of such non-GPCR membrane proteins, in conjunction with the fine optimizations of the MembStruk protocol.

References

Altenbach, C.; Greenhalgh, D. A.; Khorana, H. B. and Hubbell, W. L. (1994). A collision gradient method to determine the immersion depth of nitroxides in lipid bilayers: Application to spin-labeled mutants of bacteriorhodopsin. *Proc. Natl. Acad. Sci. USA.* 91, 1667-1671.

Floriano, W. B.; Vaidehi, N. and Goddard III, W. A. (2004). Making sense of olfaction through molecular structure and function prediction of olfactory receptors. *Chem. Senses*, (in press).

Floriano, W. B.; Vaidehi, N.; Singer, M.; Shepherd, G. and Goddard III, W. A. (2000). Molecular mechanisms underlying differential odor responses of a mouse olfactory receptor. *Proc. Natl Acad. Sci. USA.* 97, 10712-10716.

Freddolino, P. L.; Kalani, M. S. Y.; Vaidehi, N.; Floriano, W. B.; Hall, S. E.; Trabanino, R. J.; Kam, W. T. and Goddard III, W. A. (2004). Predicted 3D structure for the human beta2 adrenergic receptor and its binding site for agonists and antagonists. *Proc. Natl. Acad. Sci. USA.* 101, 2736-2741.

Gross, A.; Columbus, L.; Hideg, K.; Altenbach, C. and Hubbell, W. L. (1999). Structure of the KcsA Potassium Channel from Streptomyces lividans: A site-directed spin labeling study of the second transmembrane segment. *Biochemistry* 38, 10324-10335.

Herzyk, P. and Hubbard, R. E. (1995). Automated method for modeling seven-helix transmembrane receptors from experimental data. *Biophys J.* 69., 2419-2442.

Hubbell, W. L. and Altenbach, C. (1994). Investigation of structure and dynamics in membrane proteins using site-directed spin labeling. *Currect Opinion in Structural Biology* 4, 566-573.

Kalani, M. Y. S.; Vaidehi, N.; Hall, S. E.; Trabanino, R. J.; Freddolino, P. L.; Kalani, M. A.; Floriano, W. B.; Kam, V. W. and Goddard III, W. A. (2004). The predicted 3D structure of the human D2 dopamine receptor and the binding site and binding affinities for agonists and antagonists. *Proc. Natl. Acad. Sci. USA.* 101, 3815-3820 (in press).

Krebs, A.; Edwards, P. C.; Villa, C.; Li, J. and Schertler, G. F. X. (2003). The three-dimensional structure of bovine rhodopsin determined by electron cryomicroscopy. *The Journal of Biological Chemistry* 278, 50217-50225.

Perozo, E.; Kloda, A.; Cortes, D. M. and Martinac, B. (2001). Site-directed spin-labeling analysis of reconstituted Mscl in the closed state. *J. Gen. Physiol.* 118, 193-205.

Strader, C. D.; Fong, T. M.; Tota, M. R.; Underwood, D. and Dixon, R. A. (1994). Structure and Function of G-Protein Coupled receptors. *Annu. Rev. Biochem.* 63, 101-132.

Trabanino, R. J.; Hall, S. E.; Vaidehi, N.; Floriano, W. B.; Kam, V. W. T. and Goddard III, W. A. (2004). First principles predictions of the structure and function of G-protein-coupled receptors: validation for bovine rhodopsin. *Biophys. J.* 86, 1904-1921.

Tseitin, V. M. and Nikiforovich, G. V. (1999). Isolated transmembrane helices arranged across a membrane: computational studies. *Protein Engineering* 12, 305-311.

Vaidehi, N.; Floriano, W. B.; Trabanino, R.; Hall, S. E.; Freddolino, P.; Choi, E. J.; Zamanakos, G. and Goddard, W. A. (2002). Prediction of structure and function of G protein-coupled receptors. *Proc. Natl Acad. Sci. USA.* 99, 12622-12627.

Chapter 5: Data mining of GPCRs and classification of human GPCRs

Abstract

The field of bioinformatics in the context of G-protein-coupled receptors (GPCRs) is of utmost pharmaceutical interest as the discovery and classification of GPCRs would improve the search for new targets and analyze drug cross-reactivity. Using TM2ndS, the search for 7-helical proteins was conducted on the Riken mouse cDNA collection. Virtually all the GPCRs in the database were detected, and some unclassified sequences were also found. In addition, the TM2ndS program was used to predict the TM helical regions and hydrophobic centers (HC) for all the human GPCRs which are SwissProt entries. A graphical user interface and database scheme has been devised to later be able to search this database for possible homologues. Using this database, the 7 core TM (transmembrane) regions of the human GPCR were used for classifying the GPCRs bases on individual TM homology. Some interesting translocations of family members across families in different TMs were noted.

Introduction

As GPCRs are important pharmaceutical targets, their informational analysis is necessary for determining cross-reactivity and potential new targets for a certain drug. A recent bioinformatics study of endogenous GPCRs were classified (Vassilatis et al., 2003) based on a combination of ligand type and class (A,B,D,F/S based on shared sequence motifs), as well as sequence homology. It was found that a majority (60%) of the endoGPCRs clustered by ligand type (neurotransmitter, lipid, peptide), which allows possible prediction of ligands for orphan GPCRs. The orphan GPCRs were determined by homology to known GPCRs.

Various other schemes for classification of GPCRs have been tried, such as binary loop topology patterns (Inoue et al., 2004), "fingerprint" motifs amongst families (Fredriksson et al., 2003), and combinations of sequence similarity and "fingerprint" pattern analysis (Papasaikas et al., 2003). Nevertheless, bioinformatics for GPCRs remains an important field of research (Gaulton et al., 2003; Yanbin et al., 2003).

This current study was aimed at data-mining of unknown GPCRs by direct secondary structure prediction using TM2ndS (Trabanino et al., 2004) TM helix prediction program. For initial testing, the annotated mouse cDNA Riken library (Consortium, 2001) was used to determine known and previously unknown GPCRs based on the presence of 7 helices.

In addition, the TM helices and hydrophobic centers were determined for endogenous human GPCRs in the high-quality SwissProt database. The family

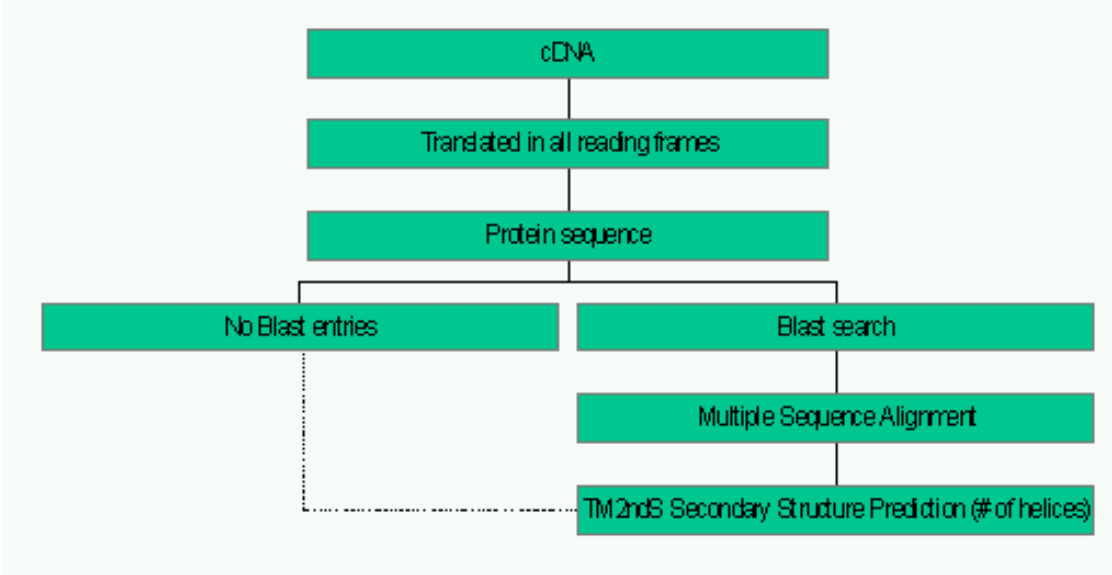
relationships for TM core sequence similarity were determined and yield interesting changes to the traditionally accepted relationships.

Methods

Data mining of GPCRs by secondary structure prediction

The annotated cDNA mouse collection by Riken (from Riken Mouse Gene Encyclopedia Project) was translated to protein in all six reading frames. The largest coding sequence found was used as input into TM2ndS. This program either predicts TM helical regions based primarily on hydrophobicity using one sequence or an ensemble of homologues to the sequence of interest. This sequence ensemble is obtained using Blast (Altschul et al., 1990; Altschul et al., 1997). The use of the sequence ensemble is important for accurate TM helical prediction, not TM identification. Further, because this study was aimed at analyzing 20,000 clones, the first option was chosen. In this case, the TM2ndS program searched for 7 helices over a range of window sizes. A flow chart of the procedure is shown in Figure 1.

Figure 1: A flow chart of the data mining procedure using TM2ndS.



GPCR TM helix and hydrophobic center database creation

For the purpose of obtaining TM core (HC +- 10 residues) information, the set of all human GPCRs with SwissProt database entries excluding the olfactory receptors was obtained from GPCR.ORG (www.gpcr.org). The SwissProt sequence database was chosen because it consists of higher quality well-annotated sequences. Then TM2ndS v2.0 was run on this set to obtain TM helix and hydrophobic center predictions. This run used an ensemble of sequences from a Blast search with bit score 200 or over with respect to the query sequence.

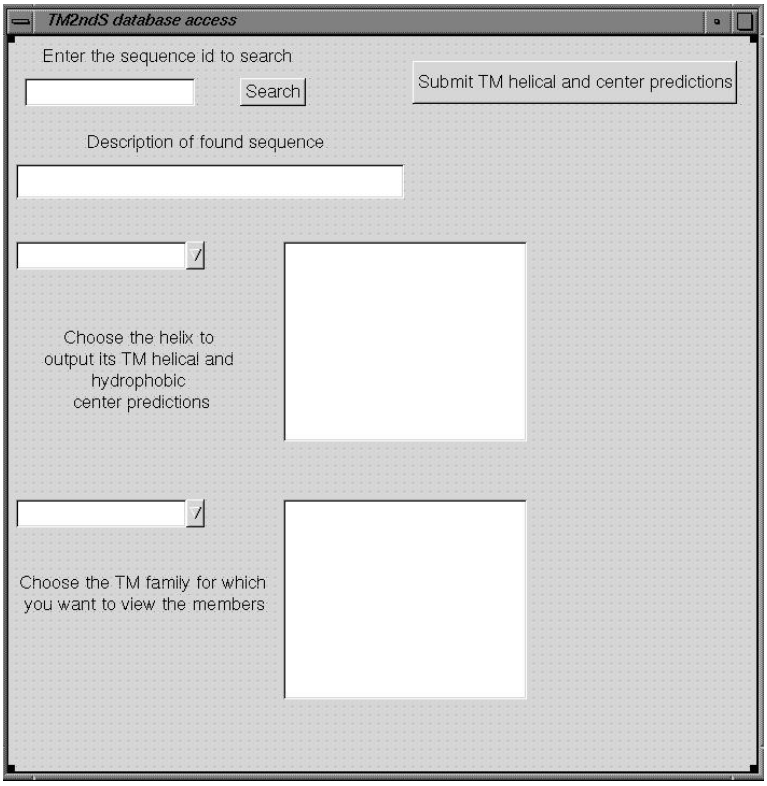
Using those TM cores obtained from the first round of TM2ndS, the individual TM cores were each used as queries for individual Blast searches for homologues to these sequences. This new ensemble was then used to obtain refined TM2ndS helix and hydrophobic center predictions. All the files for the first round and second round of TM2ndS were saved for subsequent incorporation into a database. The initial code for the SQL tables is shown below:

```
//saves predictions and centers after one round of TM2ndS for one entry
create table t1
( id number,
  swissprot_id number not null,
  description varchar2(1000),
  simple_desc varchar2(100),
  tm1_c varchar2(200),
  tm2_c varchar2(200),
  tm3_c varchar2(200),
  tm4_c varchar2(200),
  tm5_c varchar2(200),
  tm6_c varchar2(200),
  tm7_c varchar2(200),
  hc1_c number,
  hc2_c number,
  hc3_c number,
  hc4_c number,
  hc5_c number,
```

```
  hc6_c number,
  hc7_c number,
  tm1_i varchar2(200),
  tm2_i varchar2(200),
  tm3_i varchar2(200),
  tm4_i varchar2(200),
  tm5_i varchar2(200),
  tm6_i varchar2(200),
  tm7_i varchar2(200),
  hc1_i number,
  hc2_i number,
  hc3_i number,
  hc4_i number,
  hc5_i number,
  hc6_i number,
  hc7_i number,
  constraint t1_pk primary key (id),
  constraint t1_swissprot_uk unique(swissprot_id)
);

create index t1_swissprot_idx on t1(swissprot_id);


//saves pairwise alignment after first round of TM2ndS
create table t2
(
  id1 number,
  id2 number,
  seq varchar2(200),
  constraint t2_pk primary key (id1, id2),
  constraint t2_fk1 foreign key (id1) references t1 (id),
  constraint t2_fk2 foreign key (id2) references t1 (id)
);

//saves alignment and other info after iterative step on all 7 helices
create table t3
(
  id1 number,
  idx number,
  id2 number,
  seq varchar2(200),
  constraint t3_pk primary key (id1, idx, id2),
  constraint t3_fk1 foreign key (id1) references t1 (id),
  constraint t3_fk2 foreign key (id2) references t1 (id),
--  constraint t3_idx check (idx between 1 and 7)
);
```

Some snapshots of the graphical user interface which may later access this database is shown in Figure 2.

Figure 2: The GUI design for accessing the TM helix and HC database.



Classification of human GPCRs based on TM core homology

The TM helical and hydrophobic center predictions for D2 dopamine receptor were used to align the 7 TM cores with the rest of the 267 endogenous human GPCRs (descriptions are shown in Figure S1). The profile alignments were performed using Clustalw (Thompson et al., 1994) with high gap penalties (9) for the TM core regions.

The tree diagram was generated for each TM core using RETREE (Felsenstein, 1989; BioNavigator, by Entigen Corporation) using the output from Clustalw.

Results and discussion

Data mining GPCRs on cDNA library

Some of the true positive hits obtained are shown in Figure 3. Over 90% of the GPCRs in the cDNA library were found. Those which were not detected were actually fragments of the full genes, a problem with the cDNA method. This may arise from error in reverse transcription which may have introduced premature “stop” codons or from partially degraded mRNA from which cDNA was obtained.

Unfortunately of the >20,000 clones they obtained, they only classified 361 as signal transduction proteins. Of these only ~30-40 are classified as GPCR's. The other GPCR's are either missing from the library or are unclassified. This database is thus not complete.

In addition, a large number of false positives was obtained, owing probably to the window scanning of TM2ndS. In the newer version 2.0, the predictions are conducted only at window 14, and as discussed in Chapter 3, this leads to over 90% scoring for TM region distinction. This newer method may be applied on the complete human genome (Celera or GenBank).

Classification of human GPCRs

The 7 D2 dopamine TM core regions were aligned to the rest of the 267 sequences for non-olfactory human GPCRs. The relationship tree diagrams are shown in Figure 4. Higher resolution versions are in the Appendix.

The GPCRs have previously been classified into four families based on the classification scheme of GPCRDB (www.gpcr.org). These are Class A,B,C,F/S (Frizzled/Smoothened). In this study, this classification is not assumed so even proteins across these classes are classified according to their TM core homology. Representative of each class are indicated in Figure 4 (muscarinic or ACM, EDG, chemokine, and rhodopsin for class A; BAI, PTT, and VIP for class B; MGR or metabotropic glutamate for class C; FZ or frizzled for class F/S).

The diagrams have indicated large family divisions and smaller family subdivisions for each of the seven TM cores. One thing to note is that with the exception of helices 2, 3, 4, the proteins of class B, C, and F/S fall into the same families or subfamilies.

In the case of the class A proteins, the muscarinic receptors (ACM) fall in the same subfamily as the aminergic receptors for all TM's except 5, 6, 7 indicating possible residues which may be able to distinguish binding of drugs in these proteins. This information may be useful in reducing cross-reactivity of drugs.

The chemokine receptors fall into separate subfamilies as the aminergic receptors, indicating that there are structural differences in the receptors which may be taken advantage of when designing drugs with reduced cross-reactivity across these types. The chemokine receptors interestingly fall in the same subfamily as the muscarinic receptors for TM core 6, which correlates with the cross-reactivity seen in BX471 (Hesselgesser et al., 1998), an antagonist for which the binding site (which involves helix 6) will be

discussed in Chapter 9. The EDG lipid receptors fall into different subclasses as the aminergic receptors for TM's 1,4,5,7.

Interestingly, for TM 4 some of the adrenergics, serotonin, and dopamines, fall into separate families as the main aminergic group, indicating variation in this TM even in functionally related proteins.

In addition to these insights into cross-selectivity, endogenous ligands for orphan receptors may be determined based on the known TM interactions of classified GPCRs. In addition, new data mined GPCRs may be classified by a similar method.

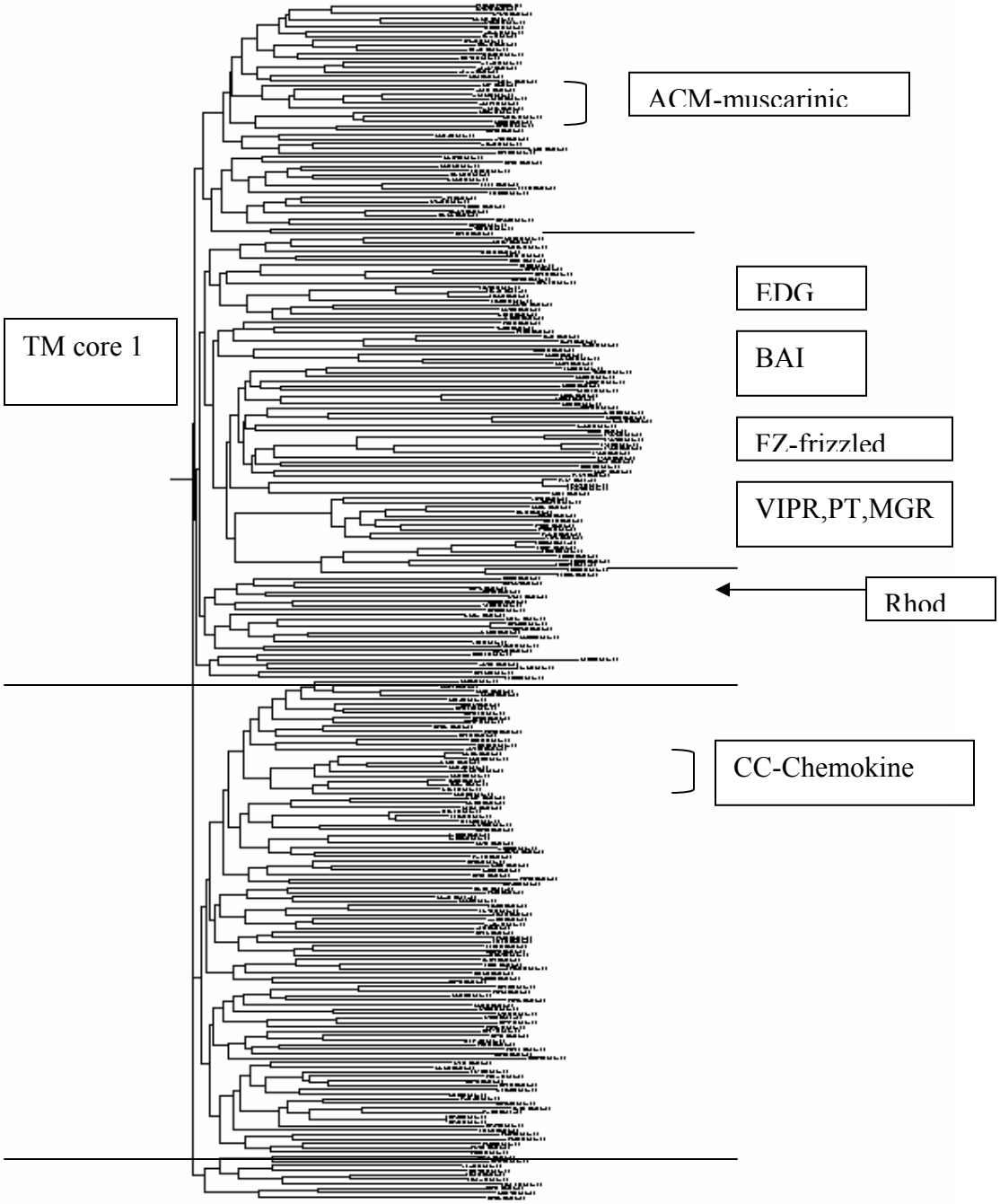
Conclusions

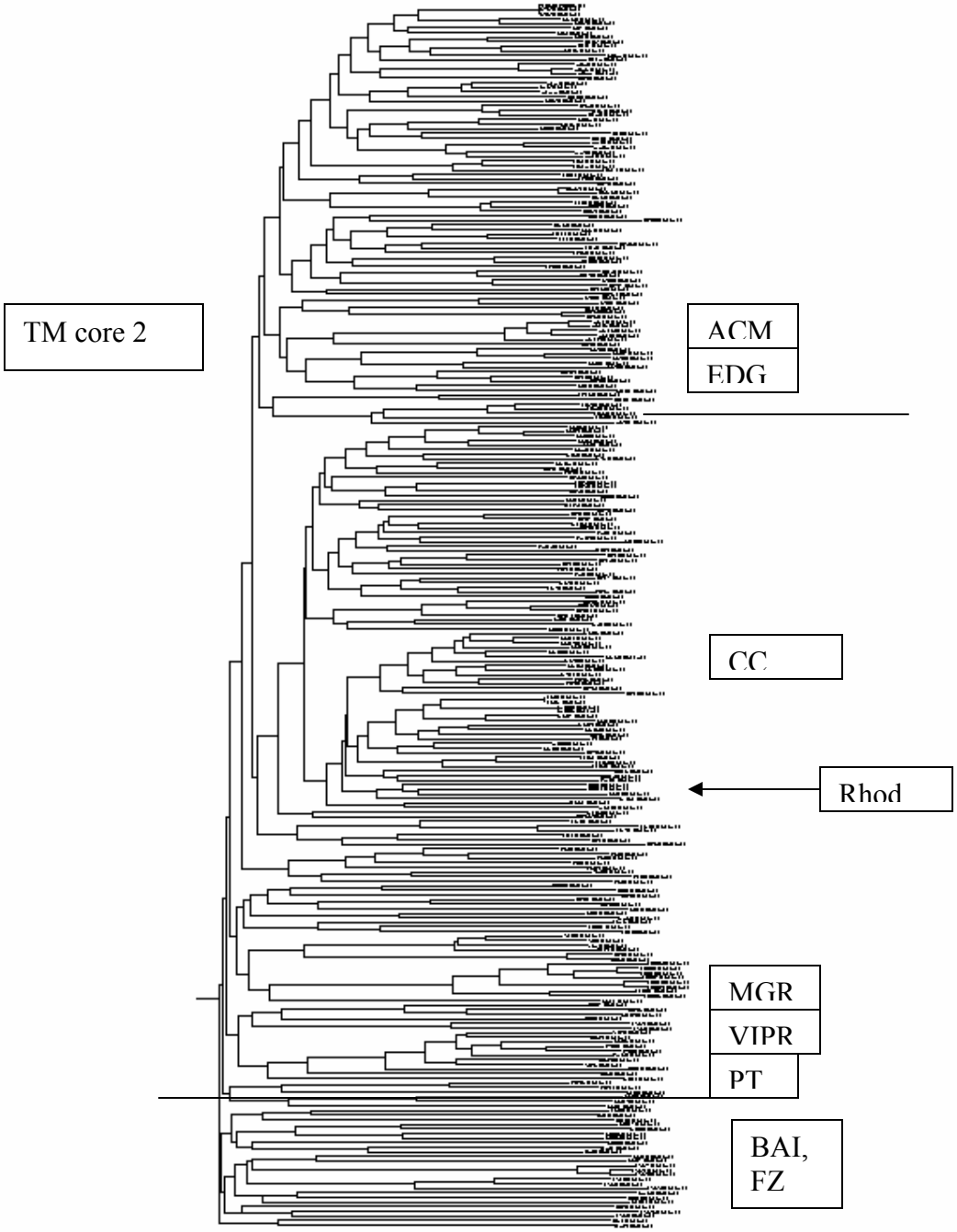
Because of their general pharmaceutical importance, the bioinformatics field as applied to GPCRs is of utmost importance in ascertaining possible cross-selectivity and discovering new GPCRs. This current study has delved into methods for mining GPCRs based on direct secondary structure prediction. In addition, since our group is also currently working on the determination of the 3D structures of GPCRs, the organization of this data together with TM helical and hydrophobic center predictions has been initiated. As an integral process in this information organization for relevant data extraction, the known human GPCRs have been classified based on homology in the true TM core regions as defined by the hydrophobic center (discussed in Chapter 4) predictions. This, together with binding site information (i.e. which helices are involved) for drugs, would provide an abundance of information for designing new drugs with controlled cross-reactivities for many receptors at once. In addition, the deorphanization of new data mined GPCRs may be achieved by determining their relationships with classified GPCRs.

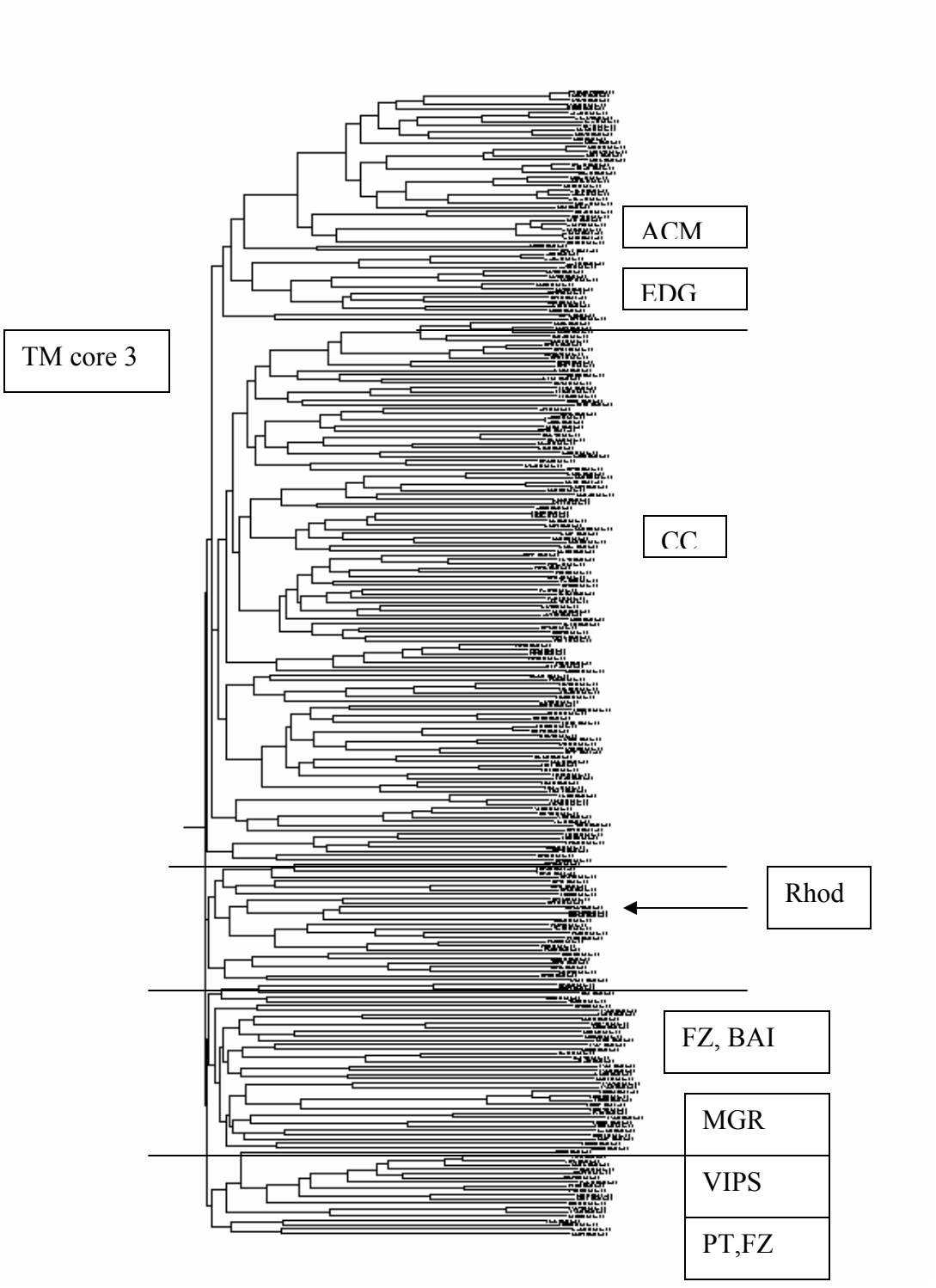
Figure 3: Some true positive hits from the data mining search using TM2ndS.

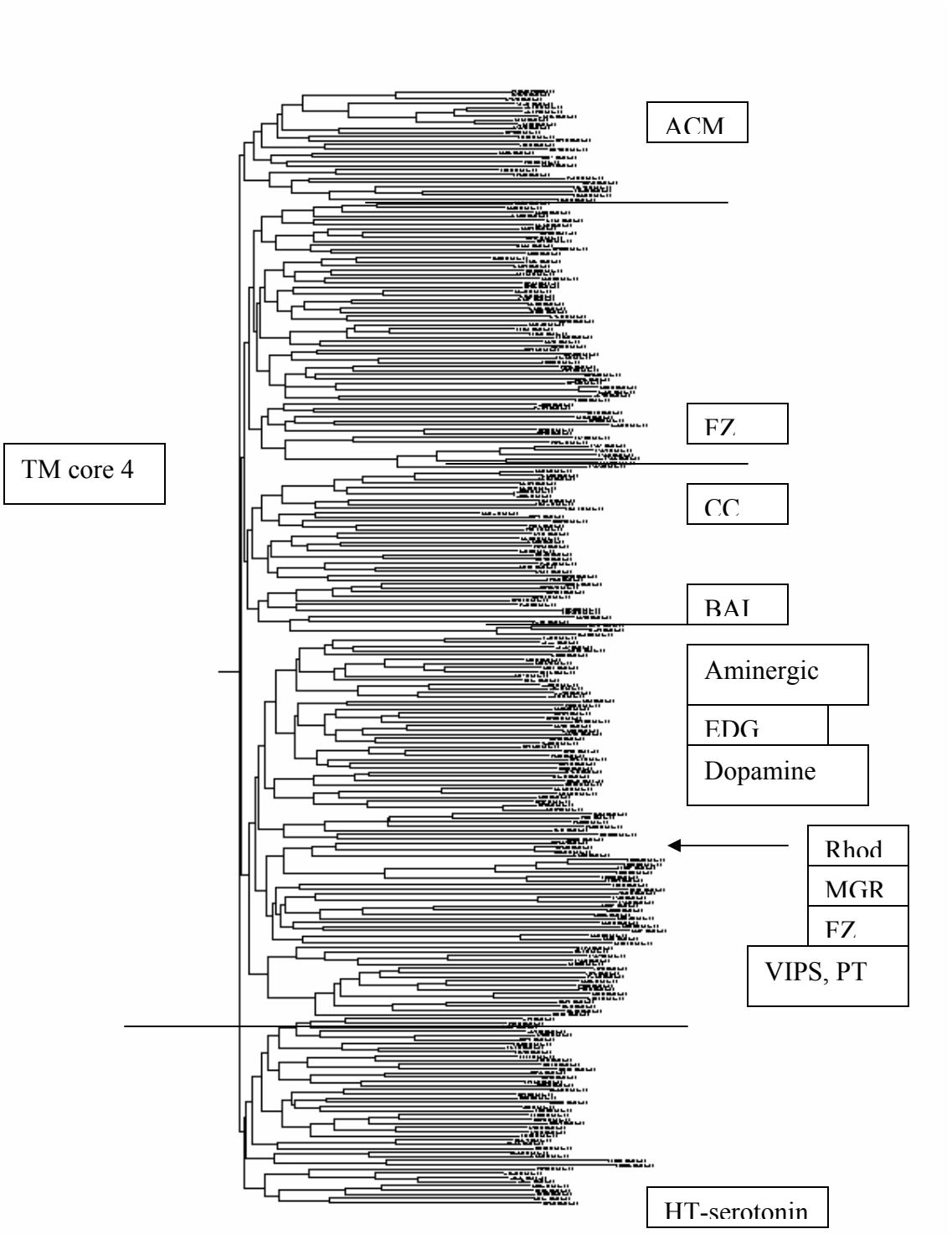
```
*1200004B16     endothelial differentiation sphingolipid G-protein-
coupled receptor 1   MGD     MGI:1096355     ISS
*1300015C04     purinergic receptor P2Y, G-protein coupled 2    MGD
MGI:105107      ISS
*1300018H12     toll-like receptor 2    MGD     MGI:1346060     ISS
*1700025D19     homolog to putative G protein-coupled receptor  NCBI-nr
7657136 ISS
*1810047I05     L-CCR.  SPTR    O70171  ISS
*2900079B22     7 transmembrane receptor (rhodopsin family) containing
protein  Pfam    PF00001 ISS
*4833409N04     prostaglandin F receptor        MGD     MGI:97796
ISS
*4921504D23     homolog to KIAA0001 gene product; putative G-protein-
coupled receptor; G protein coupled receptor for UDP-glucose
LocusLink       9934    ISS
*4933424L13     IQ motif containing GTPase activating protein 1 MGD
MGI:1352757     ISS
*4933433E02     similar to OLFACTORY RECEPTOR.  SPTR    Q9QZ18  ISS
*1200012O13     cholecystokinin A receptor      MGD     MGI:99478
ISS
*2010001L06     7 transmembrane receptor containing protein     Pfam
PF00001 ISS
*2210420B03     endothelial differentiation, G-protein-coupled receptor
6       MGD     MGI:1333809     ISS
*3732413I11     homolog to MULTIPLE MEMBRANE SPANNING RECEPTOR TRC8
(PATCHED RELATED PROTEIN TRC8).     SPTR    O75485  ISS
*4632401H02     chemokine (C-C) receptor 10     MGD     MGI:1341902
ISS
*4930401I05     MAS1 oncogene   MGD     MGI:96918       ISS
*4930500J03     related to NMDA1 PROTEIN (N-METHYL-D-ASPARTATE
RECEPTOR-ASSOCIATED PROTEIN).    SPTR    Q9V6H7  ISS
*4932441H21     similar to OLFACTORY RECEPTOR.  SPTR    Q9QZ18  ISS
*4933403I07     follicle stimulating hormone receptor   MGD
MGI:95583       ISS
*5430432J15     purinergic receptor P2Y, G-protein coupled 2    MGD
MGI:105107      ISS
*6330420K13     homolog to CHEMOKINE RECEPTOR-LIKE 2 (G-PROTEIN-COUPLED
RECEPTOR GPR41).        SPTR    O08878  ISS
```

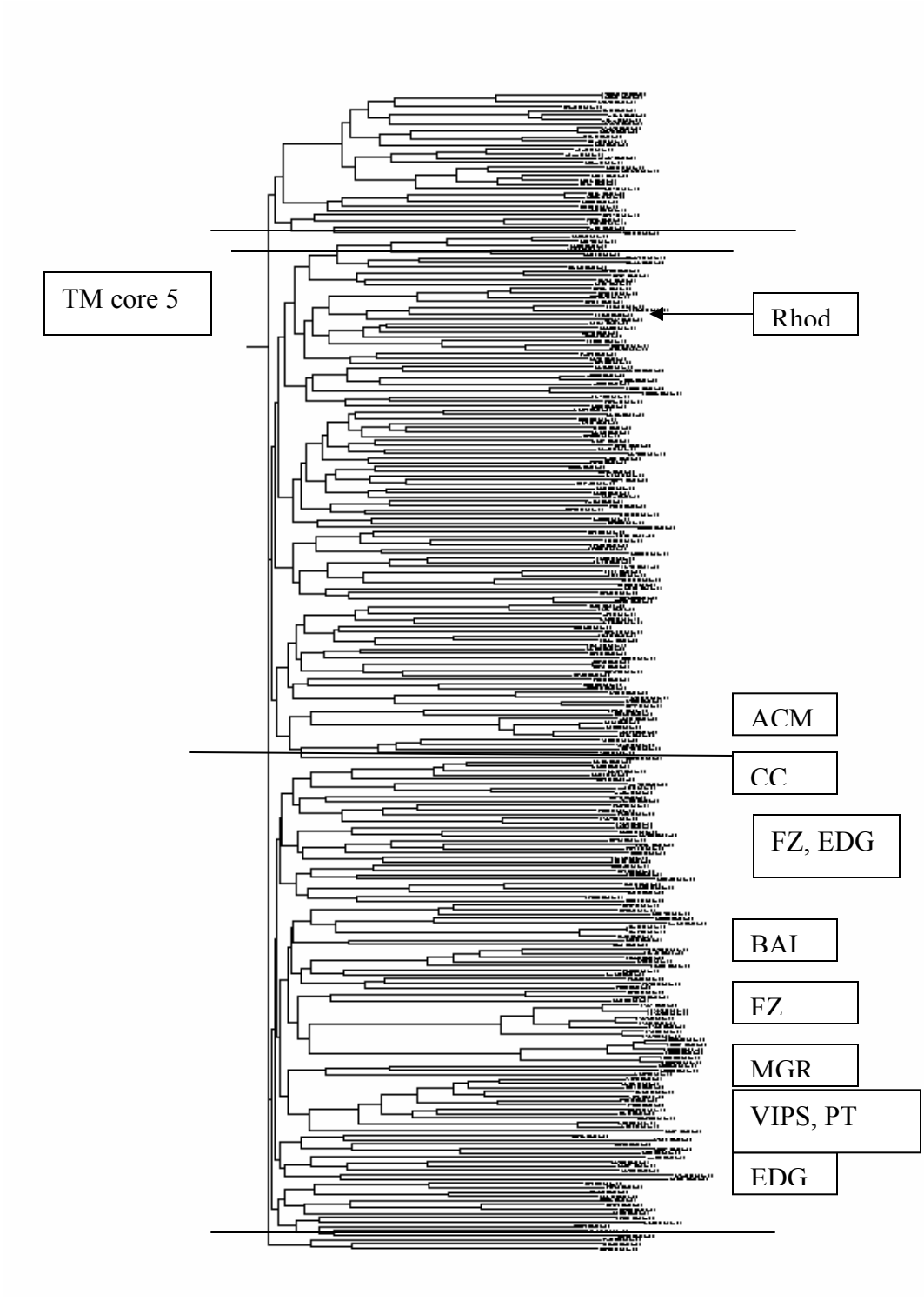
Figure 4: Tree diagrams displaying the relationships between TM cores of human GPCRs. The longer horizontal lines indicate families while the shorter lines indicate sub-families. Some GPCR types are named as well, in order to monitor the changes in the relationships for the different TM cores.

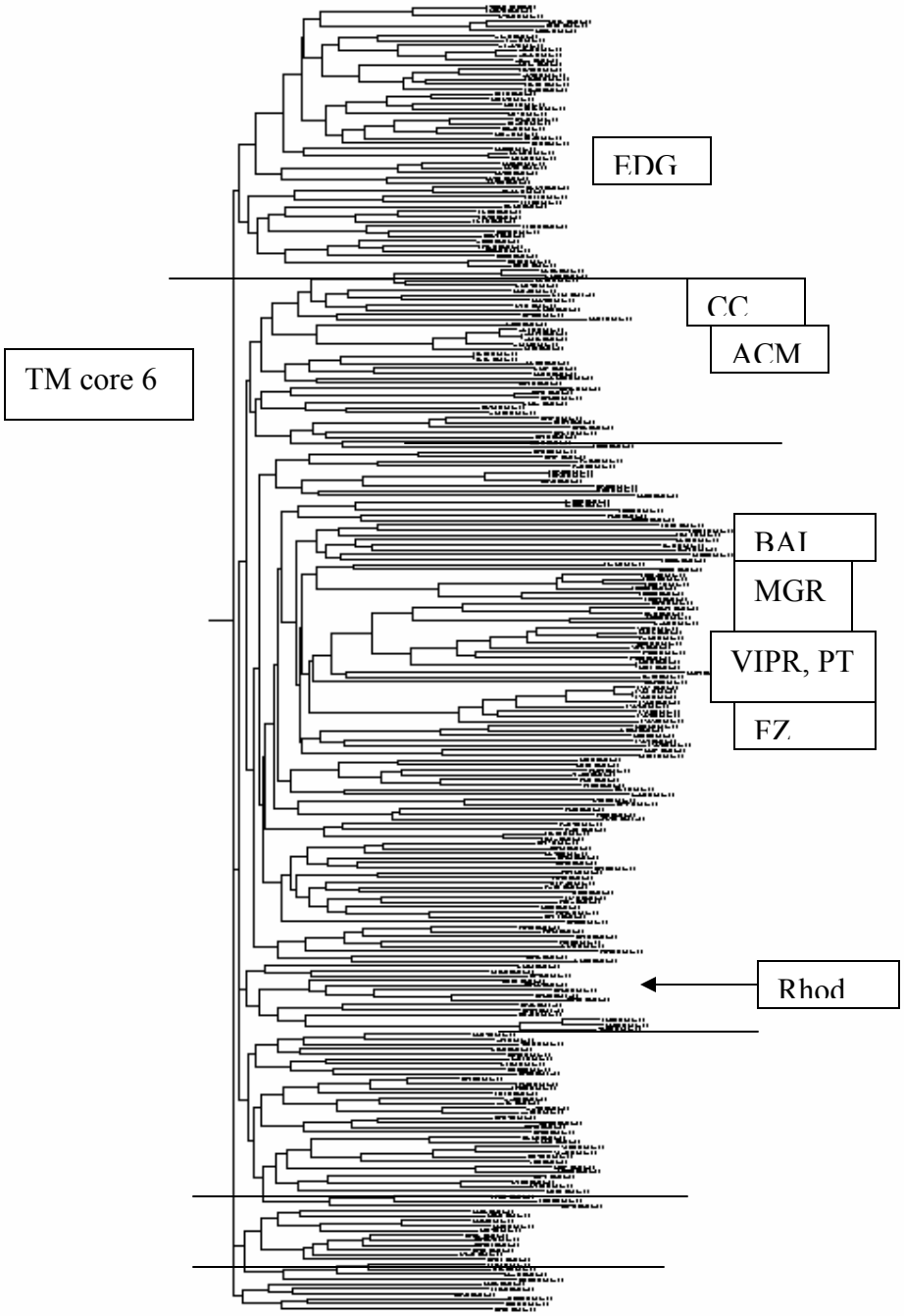

ACM-muscarinic
EDG
TM core 1
BAI
FZ-frizzled
VIPR,PT,MGR
Rhod
CC-Chemokine


TM core 2
ACM
EDG
CC
Rhod
MGR
VIPR
PT
BAI,
FZ


TM core 3
ACM
EDG
CC
Rhod
FZ, BAI
MGR
VIPS
PT,FZ


TM core 4
ACM
FZ
CC
BAI
Aminergic
EDG
Dopamine
Rhod
MGR
FZ
VIPS, PT
HT-serotonin


TM core 5
Rhod
ACM
CC
FZ, EDG
BAI
FZ
MGR
VIPS, PT
EDG


EDG
CC
ACM
TM core 6
BAI
MGR
VIPR, PT
FZ
Rhod

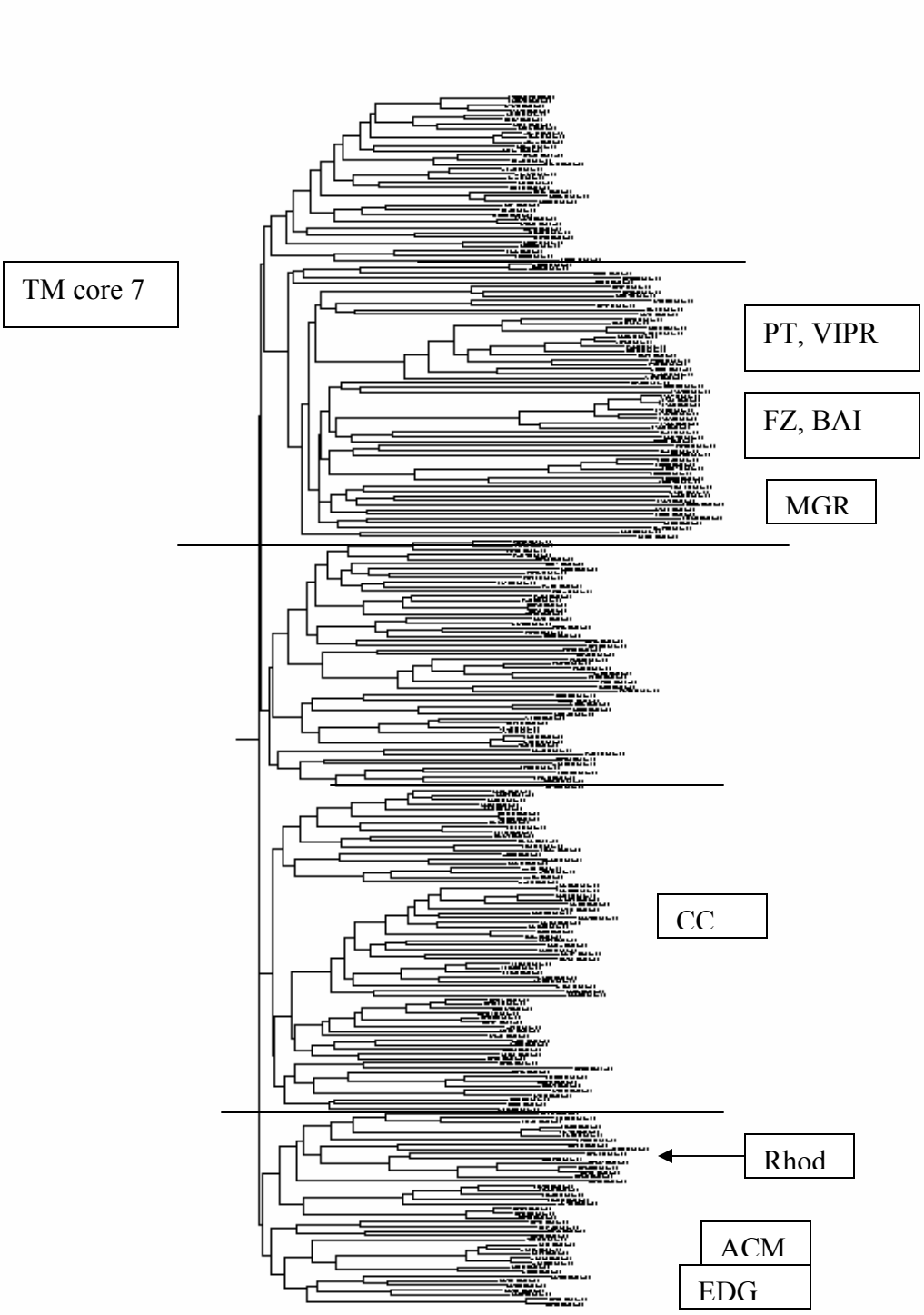

TM core 7
PT, VIPR
FZ, BAI
MGR
CC
Rhod
ACM
EDG

Figure S1: The gi's, sp ID's, and descriptions of the classified proteins. The empty lines are sequences which were not analyzed because they are non-SwissProt.

0 12644225 | sp | P35346 | SSR5_HUMAN Somatostatin receptor type 5 (SS5R)
1 401130 | sp | P31391 | SSR4_HUMAN Somatostatin receptor type 4 (SS4R)
2 417815 | sp | P32745 | SSR3_HUMAN Somatostatin receptor type 3 (SS3R) (SSR-28)

7 1705896 | sp | P51681 | CKR5_HUMAN C-C chemokine receptor type 5 (C-C CKR-5) (CC-CKR-5) (CCR-5) (CCR5) (HIV-1 fusion co-receptor) (CHEMR13) (CD195 antigen)

9 1352454 | sp | P25025 | IL8B_HUMAN High affinity interleukin-8 receptor B (IL-8R B) (CXCR-2) (GRO/MGSA receptor) (IL-8 receptor type 2) (CDw128b)

16 401124 | sp | P30872 | SSR1_HUMAN Somatostatin receptor type 1 (SS1R) (SRIF-2)
17 416802 | sp | P32246 | CKR1_HUMAN C-C chemokine receptor type 1 (C-C CKR-1) (CC-CKR-1) (CCR-1) (CCR1) (Macrophage inflammatory protein-1 alpha receptor) (MIP-1alpha-R) (RANTES-R) (HM145) (LD78 receptor)

19 124356 | sp | P25024 | IL8A_HUMAN High affinity interleukin-8 receptor A (IL-8R A) (IL-8 receptor type 1) (CXCR-1) (CDw128a)
20 2829400 | sp | P49682 | CCR3_HUMAN C-X-C chemokine receptor type 3 (CXC-R3) (CXCR-3) (CKR-L2) (CD183 antigen)

22 1168245 | sp | P35368 | A1AB_HUMAN Alpha-1B adrenergic receptor (Alpha 1B-adrenoceptor)

24 730229 | sp | P41145 | OPRK_HUMAN Kappa-type opioid receptor (KOR-1)

27 1168246 | sp | P35348 | A1AA_HUMAN Alpha-1A adrenergic receptor (Alpha 1A-adrenoceptor) (Alpha-1C adrenergic receptor)

31 1705894 | sp | P51679 | CKR4_HUMAN C-C chemokine receptor type 4 (C-C CKR-4) (CC-CKR-4) (CCR-4) (CCR4) (K5-5)

32 1346165 | sp | P48146 | GPR8_HUMAN Probable G protein-coupled receptor GPR8
33 1168243 | sp | P25100 | A1AD_HUMAN Alpha-1D adrenergic receptor (Alpha 1D-adrenoceptor) (Alpha-1A adrenergic receptor)
34 401126 | sp | P30874 | SSR2_HUMAN Somatostatin receptor type 2 (SS2R) (SRIF-1)
35 1168965 | sp | P41597 | CKR2_HUMAN C-C chemokine receptor type 2 (C-C CKR-2) (CC-CKR-2) (CCR-2) (CCR2) (Monocyte chemoattractant protein 1 receptor) (MCP-1-R) (CCR2)
36 3123242 | sp | P41143 | OPRD_HUMAN Delta-type opioid receptor (DOR-1)

38 2851402 | sp | P35372 | OPRM_HUMAN Mu-type opioid receptor (MOR-1)

41 1705892 | sp | P51677 | CKR3_HUMAN C-C chemokine receptor type 3 (C-C CKR-3) (CC-CKR-3) (CCR-3) (CCR3) (CKR3) (Eosinophil eotaxin receptor)
42 21264488 | sp | P41231 | P2Y2_HUMAN P2Y purinoceptor 2 (P2Y2) (P2U purinoceptor 1) (P2U1) (ATP receptor) (Purinergic receptor)
43 3023883 | sp | O00155 | GP25_HUMAN Probable G protein-coupled receptor GPR25
44 400654 | sp | P30991 | CCR4_HUMAN C-X-C chemokine receptor type 4 (CXC-R4) (CXCR-4) (Stromal cell-derived factor 1 receptor) (SDF-1 receptor) (Fusin) (Leukocyte-derived seven transmembrane domain receptor) (LESTR) (LCR1) (FB22) (NPYRL) (HM89) (CD184 antigen)
45 1346163 | sp | P48145 | GPR7_HUMAN Probable G protein-coupled receptor GPR7
46 543823 | sp | P35414 | APJ_HUMAN Apelin receptor (G protein-coupled receptor APJ) (Angiotensin receptor-like 1) (HG11)

48 1351394 | sp | P49238 | C3X1_HUMAN CX3C chemokine receptor 1 (C-X3-C CKR-1) (CX3CR1) (Fractalkine receptor) (GPR13) (V28) (Beta chemokine receptor-like 1) (CMK-BRL-1) (CMKBLR1)
49 6016096 | sp | O60755 | GALT_HUMAN Galanin receptor type 3 (GAL3-R) (GALR3)
50 1707884 | sp | P51685 | CKR8_HUMAN C-C chemokine receptor type 8 (C-C CKR-8) (CC-CKR-8) (CCR-8) (GPR-CY6) (GPRCY6) (Chemokine receptor-like 1) (CKR-L1) (TER1) (CMKBRL2) (CC-chemokine receptor CHEMR1)

52 1352335 | sp | P32248 | CKR7_HUMAN C-C chemokine receptor type 7 precursor (C-C CKR-7) (CC-CKR-7) (CCR-7) (MIP-3 beta receptor) (EBV-induced G protein-coupled receptor 1) (EBI1) (BLR2)

54 6016094 | sp | O43603 | GALS_HUMAN Galanin receptor type 2 (GAL2-R) (GALR2)
55 3041713 | sp | Q15722 | L4R1_HUMAN Leukotriene B4 receptor 1 (LTB4-R) (P2Y purinoceptor 7) (P2Y7) (Chemoattractant receptor-like 1)
56 416718 | sp | P32302 | CCR5_HUMAN C-X-C chemokine receptor type 5 (CXC-R5) (CXCR-5) (Burkitt'S lymphoma receptor 1) (Monocyte-derived receptor 15) (MDR15)
57 1709524 | sp | P51582 | P2Y4_HUMAN P2Y purinoceptor 4 (P2Y4) (Uridine nucleotide receptor) (UNR) (P2P)
58 231519 | sp | P30556 | AG2R_HUMAN Type-1 angiotensin II receptor (AT1) (AT1AR)

60 1346093 | sp | P47211 | GALR_HUMAN Galanin receptor type 1 (GAL1-R) (GALR1)
61 399504 | sp | P25090 | FML1_HUMAN FMLP-related receptor I (FMLP-R-I) (Lipoxin A4 receptor) (LXA4 receptor) (RFP) (HM63)

65 2495002 | sp | Q13725 | AG2S_HUMAN Type-1B angiotensin II receptor (AT1B) (AT1BR)
66 20141395 | sp | P49683 | GP10_HUMAN Prolactin-releasing peptide receptor (PrRP receptor) (PrRPR) (G protein-coupled receptor GPR10) (hGR3)
67 1703214 | sp | P50052 | AG22_HUMAN Type-2 angiotensin II receptor (AT2)
68 112821 | sp | P28222 | 5H1B_HUMAN 5-hydroxytryptamine 1B receptor (5-HT-1B) (Serotonin receptor) (5-HT-1D-beta) (Serotonin 1D beta receptor) (S12)

70 417029 | sp | P32239 | GASR_HUMAN Gastrin/cholecystokinin type B receptor (CCK-B receptor) (CCK-BR)
71 2506481 | sp | P30411 | BRB2_HUMAN B2 bradykinin receptor (BK-2 receptor) (B2R)
72 1730237 | sp | P51686 | CKR9_HUMAN C-C chemokine receptor type 9 (C-C CKR-9) (CC-CKR-9) (CCR-9) (GPR-9-6)

74 115262 | sp | P21730 | C5AR_HUMAN C5a anaphylatoxin chemotactic receptor (C5a-R) (CD88 antigen)
75 730230 | sp | P41146 | OPRX_HUMAN Nociceptin receptor (Orphanin FQ receptor) (Kappa-type 3 opioid receptor) (KOR-3)
76 21264435 | sp | Q13304 | GP17_HUMAN Probable P2Y purinoceptor GPR17 (P2Y-like receptor) (R12)

78 19857032 | sp | P21452 | NK2R_HUMAN Substance-K receptor (SKR) (Neurokinin A receptor) (NK-2 receptor) (NK-2R)

82 112819 | sp | P28221 | 5H1D_HUMAN 5-hydroxytryptamine 1D receptor (5-HT-1D) (Serotonin receptor) (5-HT-1D-alpha)
83 13878599 | sp | Q9GZQ6 | NFF1_HUMAN Neuropeptide FF receptor 1 (RFamide-related peptide receptor OT7T022)
84 3024351 | sp | O00254 | PAR3_HUMAN Proteinase activated receptor 3 precursor (PAR-3) (Thrombin receptor-like 2) (Coagulation factor II receptor-like 2)

86 120427 | sp | P21462 | FMLR_HUMAN fMet-Leu-Phe receptor (fMLP receptor) (N-formyl peptide receptor) (FPR) (N-formylpeptide chemoattractant receptor)

89 1346170 | sp | P49685 | GP15_HUMAN G protein-coupled receptor GPR15 (BOB)
90 20139074 | sp | Q96RI0 | PAR4_HUMAN Proteinase activated receptor 4 precursor (PAR-4) (Thrombin receptor-like 3) (Coagulation factor II receptor-like 3)
91 2495018 | sp | Q15077 | P2Y6_HUMAN P2Y purinoceptor 6 (P2Y6)

92 2506480 | sp | P46663 | BRB1_HUMAN B1 bradykinin receptor (BK-1 receptor) (B1R)
93 1170008 | sp | P46094 | CXC1_HUMAN Chemokine XC receptor 1 (XC chemokine receptor 1) (Lymphotactin receptor) (G protein-coupled receptor GPR5)
94 8247919 | sp | P46092 | CKRA_HUMAN C-C chemokine receptor type 10 (C-C CKR-10) (CC-CKR-10) (CCR-10) (G-protein-coupled receptor 2)

97 132207 | sp | P25106 | RDC1_HUMAN G protein-coupled receptor RDC1 homolog
98 17380487 | sp | Q99788 | CML1_HUMAN Chemokine receptor-like 1 (G-protein-coupled receptor DEZ) (G protein-coupled receptor ChemR23)
99 8488960 | sp | P34969 | 5H7_HUMAN 5-hydroxytryptamine 7 receptor (5-HT-7) (5-HT-X) (Serotonin receptor) (5HT7)
100 461604 | sp | P13945 | B3AR_HUMAN Beta-3 adrenergic receptor
101 3121816 | sp | O00574 | CCR6_HUMAN C-X-C chemokine receptor type 6 (CXC-R6) (CXCR-6) (G protein-coupled receptor bonzo) (G protein-coupled receptor STRL33)

103 3023884 | sp | O00270 | GP31_HUMAN Probable G protein-coupled receptor GPR31

105 20138087 | sp | Q9Y271 | CLT1_HUMAN Cysteinyl leukotriene receptor 1 (CysLTR1) (Cysteinyl leukotriene D4 receptor) (LTD4 receptor) (HG55) (HMTMF81)
106 3041709 | sp | P43657 | P2Y5_HUMAN P2Y purinoceptor 5 (P2Y5) (Purinergic receptor 5) (RB intron encoded G-protein-coupled receptor)

110 14285406 | sp | Q9NPB9 | CKRB_HUMAN C-C chemokine receptor type 11 (C-C CKR-11) (CC-CKR-11) (CCR-11) (Chemokine receptor-like 1) (CCRL1) (CCX CKR)

113 1352692 | sp | P47900 | P2YR_HUMAN P2Y purinoceptor 1 (ATP receptor) (P2Y1) (Purinergic receptor)
114 8928474 | sp | Q9UKP6 | UR2R_HUMAN Urotensin II receptor (UR-II-R)

116 23821812 | sp | Q9NPC1 | L4R2_HUMAN Leukotriene B4 receptor 2 (LTB4-R2) (Seven transmembrane receptor BLTR2) (Leukotriene B4 receptor BLT2) (LTB4 receptor JULF2)
117 2851567 | sp | P51684 | CKR6_HUMAN C-C chemokine receptor type 6 (C-C CKR-6) (CC-CKR-6) (CCR-6) (LARC receptor) (GPR-CY4) (GPRCY4) (Chemokine receptor-like 3) (CKR-L3) (DRY6)

119 118228 | sp | P21728 | DADR_HUMAN D(1A) dopamine receptor
120 18206259 | sp | Q9Y5Y4 | GP44_HUMAN Putative G protein-coupled receptor GPR44 (Chemoattractant receptor-homologous molecule expressed on Th2 cells)
121 398967 | sp | P30939 | 5H1F_HUMAN 5-hydroxytryptamine 1F receptor (5-HT-1F) (Serotonin receptor)

123 114765 | sp | P07550 | B2AR_HUMAN Beta-2 adrenergic receptor
124 543727 | sp | P28223 | 5H2A_HUMAN 5-hydroxytryptamine 2A receptor (5-HT-2A) (Serotonin receptor) (5-HT-2)

125 120425 | sp | P25089 | FML2_HUMAN FMLP-related receptor II (FMLP-R-II)

127 6225807 | sp | O43613 | OX1R_HUMAN Orexin receptor type 1 (Ox1r) (Hypocretin receptor type 1)

129 20138034 | sp | Q9NS75 | CLT2_HUMAN Cysteinyl leukotriene receptor 2 (CysLTR2) (PSEC0146) (HG57) (HPN321) (hGPCR21)
130 128359 | sp | P25103 | NK1R_HUMAN Substance-P receptor (SPR) (NK-1 receptor) (NK-1R)

132 1346295 | sp | P49019 | HM74_HUMAN Probable G protein-coupled receptor HM74
133 2495042 | sp | Q99677 | P2Y9_HUMAN P2Y purinoceptor 9 (P2Y9) (Purinergic receptor 9) (G protein-coupled receptor GPR23) (P2Y5-like receptor)
134 20178318 | sp | P25116 | PAR1_HUMAN Proteinase activated receptor 1 precursor (PAR-1) (Thrombin receptor) (Coagulation factor II receptor)

138 1709580 | sp | P55085 | PAR2_HUMAN Proteinase activated receptor 2 precursor (PAR-2) (Thrombin receptor-like 1) (Coagulation factor II receptor-like 1)
139 1708027 | sp | P46093 | GPR4_HUMAN Probable G protein-coupled receptor GPR4 (GPR19)
140 416772 | sp | P32238 | CCKR_HUMAN Cholecystokinin type A receptor (CCK-A receptor) (CCK-AR)
141 128364 | sp | P29371 | NK3R_HUMAN Neuromedin K receptor (NKR) (Neurokinin B receptor) (NK-3 receptor) (NK-3R)
142 112816 | sp | P28335 | 5H2C_HUMAN 5-hydroxytryptamine 2C receptor (5-HT-2C) (Serotonin receptor) (5HT-1C)

144 123120 | sp | P25021 | HH2R_HUMAN Histamine H2 receptor (H2R) (Gastric receptor I)
145 128393 | sp | P28336 | NMBR_HUMAN Neuromedin-B receptor (NMB-R) (Neuromedin-B-preferring bombesin receptor)
146 114752 | sp | P08588 | B1AR_HUMAN Beta-1 adrenergic receptor
147 232185 | sp | P30550 | GRPR_HUMAN Gastrin-releasing peptide receptor (GRP-R) (GRP-preferring bombesin receptor)
148 12644029 | sp | Q13639 | 5H4_HUMAN 5-hydroxytryptamine 4 receptor (5-HT-4) (Serotonin receptor) (5-HT4)

151 10719861 | sp | O15218 | ADMR_HUMAN Adrenomedullin receptor (AM-R)

156 1168220 | sp | P41595 | 5H2B_HUMAN 5-hydroxytryptamine 2B receptor (5-HT-2B) (Serotonin receptor)

158 6831552 | sp | O75388 | GP32_HUMAN Probable G protein-coupled receptor GPR32
159 1703010 | sp | P50406 | 5H6_HUMAN 5-hydroxytryptamine 6 receptor (5-HT-6) (Serotonin receptor)

160 400628 | sp | P30989 | NTR1_HUMAN Neurotensin receptor type 1 (NT-R-1) (High-affinity levocabastine-insensitive neurotensin receptor) (NTRH)

162 118214 | sp | P21918 | DBDR_HUMAN D(1B) dopamine receptor (D(5) dopamine receptor) (D1beta dopamine receptor)
163 2495039 | sp | Q99678 | GP20_HUMAN Probable G protein-coupled receptor GPR20
164 21263703 | sp | Q9Y5Y3 | GP45_HUMAN Probable G protein-coupled receptor GPR45 (PSP24-alpha) (PSP24-1)
165 21263618 | sp | Q9BXC1 | FK79_HUMAN Putative P2Y purinoceptor FKSG79
166 3913747 | sp | O43193 | MTLR_HUMAN Motilin receptor (G protein-coupled receptor GPR38)

169 2494998 | sp | Q92847 | GHSR_HUMAN Growth hormone secretagogue receptor type 1 (GHS-R) (GH-releasing peptide receptor) (GHRP) (Ghrelin receptor)

175 1345607 | sp | P47898 | 5H5A_HUMAN 5-hydroxytryptamine 5A receptor (5-HT-5A) (Serotonin receptor) (5-HT-5)
176 3122159 | sp | O15529 | GP42_HUMAN Probable G protein-coupled receptor GPR42
177 21263819 | sp | O00398 | P2YA_HUMAN Putative P2Y purinoceptor 10 (P2Y10) (P2Y-like receptor)

179 1169206 | sp | P35462 | D3DR_HUMAN D(3) dopamine receptor

181 13632136 | sp | Q9UHM6 | OPN4_HUMAN Opsin 4 (Melanopsin)
182 1352610 | sp | P49146 | NY2R_HUMAN Neuropeptide Y receptor type 2 (NPY2-R) (NPY-Y2 receptor)
183 27151763 | sp | P18089 | A2AB_HUMAN Alpha-2B adrenergic receptor (Alpha-2B adrenoceptor) (Subtype C2)
184 14194819 | sp | Q9H3N8 | HH4R_HUMAN Histamine H4 receptor (HH4R) (GPRv53) (G protein-coupled receptor 105) (GPCR105) (SP9144) (AXOR35)

186 13878604 | sp | Q9Y5X5 | NFF2_HUMAN Neuropeptide FF receptor 2 (Neuropeptide G protein-coupled receptor) (G-protein-coupled receptor HLWAR77)
187 3023539 | sp | Q99527 | CML2_HUMAN Chemokine receptor-like 2 (IL8-related receptor DRY12) (Flow-induced endothelial G protein-coupled receptor) (FEG-1) (G protein-coupled receptor GPR30) (GPCR-BR)

190 112822 | sp | P28566 | 5H1E_HUMAN 5-hydroxytryptamine 1E receptor (5-HT-1E) (Serotonin receptor) (5-HT1E) (S31)
191 28380053 | sp | Q96P88 | GRR2_HUMAN Gonadotropin-releasing hormone II receptor (Type II GnRH receptor) (GnRH-II-R)

194 3122158 | sp | O14843 | GP41_HUMAN Putative G protein-coupled receptor GPR41

195 119622 | sp | P24530 | ETBR_HUMAN Endothelin B receptor precursor (ET-B) (Endothelin receptor Non-selective type)

197 416926 | sp | P32249 | EBI2_HUMAN EBV-induced G protein-coupled receptor 2 (EBI2)

200 416726 | sp | P32247 | BRS3_HUMAN Bombesin receptor subtype-3 (BRS-3)
201 128997 | sp | P25929 | NY1R_HUMAN Neuropeptide Y receptor type 1 (NPY1-R)
202 1351392 | sp | P47901 | V1BR_HUMAN Vasopressin V1b receptor (V1bR) (AVPR V1b) (Vasopressin V3 receptor) (AVPR V3) (Antidiuretic hormone receptor 1b)
203 21263685 | sp | Q9BZJ6 | GP63_HUMAN Probable G protein-coupled receptor GPR63 (PSP24-beta) (PSP24-2)

205 119606 | sp | P25101 | ET1R_HUMAN Endothelin-1 receptor precursor (ET-A)
206 1709423 | sp | P50391 | NY4R_HUMAN Neuropeptide Y receptor type 4 (NPY4-R) (Pancreatic polypeptide receptor 1) (PP1)
207 129557 | sp | P25105 | PAFR_HUMAN Platelet activating factor receptor (PAF-R)

211 20455469 | sp | O00590 | CKD6_HUMAN Chemokine binding protein 2 (Chemokine-binding protein D6) (C-C chemokine receptor D6) (Chemokine receptor CCR-9) (CC-Chemokine receptor CCR10)

213 1170002 | sp | P46091 | GPR1_HUMAN Probable G protein-coupled receptor GPR1
214 118206 | sp | P14416 | D2DR_HUMAN D(2) dopamine receptor
215 6225810 | sp | O43614 | OX2R_HUMAN Orexin receptor type 2 (Ox2r) (Hypocretin receptor type 2)

217 15214047 | sp | Q9NS67 | GP27_HUMAN Probable G protein-coupled receptor GPR27 (Super conserved receptor expressed in brain 1)
218 21263687 | sp | Q9BZJ8 | GP61_HUMAN Probable G protein-coupled receptor GPR61 (Biogenic amine receptor-like G-protein-coupled receptor)
219 231454 | sp | P08908 | 5H1A_HUMAN 5-hydroxytryptamine 1A receptor (5-HT-1A) (Serotonin receptor) (5-HT1A) (G-21)

222 586197 | sp | P37288 | V1AR_HUMAN Vasopressin V1a receptor (V1aR) (Vascular/hepatic-type arginine vasopressin receptor) (Antidiuretic hormone receptor 1a) (AVPR V1a)
223 464921 | sp | P34981 | TRFR_HUMAN Thyrotropin-releasing hormone receptor (TRH-R) (Thyroliberin receptor)

225 1345939 | sp | P21917 | D4DR_HUMAN D(4) dopamine receptor (D(2C) dopamine receptor)
226 112938 | sp | P29275 | AA2B_HUMAN Adenosine A2b receptor
227 3122160 | sp | O15552 | GP43_HUMAN Probable G protein-coupled receptor GPR43

230 20141211 | sp | P18825 | A2AC_HUMAN Alpha-2C-adrenergic receptor (Alpha-2C adrenoceptor) (Subtype C4)

235 20137533 | sp | Q9P296 | C5L2_HUMAN C5a anaphylatoxin chemotactic receptor C5L2
236 266719 | sp | P30559 | OXYR_HUMAN Oxytocin receptor (OT-R)

239 113118 | sp | P11229 | ACM1_HUMAN Muscarinic acetylcholine receptor M1

241 15214046 | sp | Q9NPD1 | GP85_HUMAN Probable G protein-coupled receptor GPR85 (Super conserved receptor expressed in brain 2) (PKrCx1)

243 1351831 | sp | P33765 | AA3R_HUMAN Adenosine A3 receptor
244 20141559 | sp | P41968 | MC3R_HUMAN Melanocortin-3 receptor (MC3-R)

246 15214315 | sp | Q9NS66 | SRB3_HUMAN Super conserved receptor expressed in brain 3
247 231473 | sp | P30542 | AA1R_HUMAN Adenosine A1 receptor
248 26393399 | sp | Q9HBW0 | EDG4_HUMAN Lysophosphatidic acid receptor Edg-4 (LPA receptor 2) (LPA-2)
249 3024312 | sp | O14718 | OPSX_HUMAN Visual pigment-like receptor peropsin
250 119130 | sp | P21453 | EDG1_HUMAN Probable G protein-coupled receptor EDG-1

252 17380172 | sp | Q9H1Y3 | OPN3_HUMAN Opsin 3 (Encephalopsin) (Panopsin)
253 399002 | sp | Q01718 | ACTR_HUMAN Adrenocorticotropic hormone receptor (ACTH receptor) (ACTH-R) (Melanocortin-2 receptor) (MC2-R) (Adrenocorticotropin receptor)
254 267256 | sp | P30518 | V2R_HUMAN Vasopressin V2 receptor (Renal-type arginine vasopressin receptor) (Antidiuretic hormone receptor) (AVPR V2)
255 12643337 | sp | Q9UPC5 | GP34_HUMAN Probable G protein-coupled receptor GPR34

259 1346548 | sp | P49286 | ML1B_HUMAN Melatonin receptor type 1B (Mel-1B-R)
260 1170006 | sp | P46089 | GPR3_HUMAN Probable G protein-coupled receptor GPR3 (ACCA orphan receptor)
261 3913748 | sp | O43194 | GP39_HUMAN Putative G protein-coupled receptor GPR39

263 129207 | sp | P08100 | OPSD_HUMAN Rhodopsin (Opsin 2)
264 3122157 | sp | O14842 | GP40_HUMAN Putative G protein-coupled receptor GPR40
265 1346544 | sp | P48039 | ML1A_HUMAN Melatonin receptor type 1A (Mel-1A-R)

266 417280 | sp | P32245 | MC4R_HUMAN Melanocortin-4 receptor (MC4-R)

271 23503039 | sp | P08173 | ACM4_HUMAN Muscarinic acetylcholine receptor M4
272 26393418 | sp | Q9UBY5 | EDG7_HUMAN Lysophosphatidic acid receptor Edg-7 (LPA receptor 3) (LPA-3)
273 115562 | sp | P21554 | CB1R_HUMAN Cannabinoid receptor 1 (CB1) (CB-R) (CANN6)

279 12644376 | sp | Q01726 | MSHR_HUMAN Melanocyte stimulating hormone receptor (MSH-R) (Melanotropin receptor) (Melanocortin-1 receptor) (MC1-R)
280 2495036 | sp | Q15760 | GP19_HUMAN Probable G protein-coupled receptor GPR19 (GPR-NGA)
281 729996 | sp | P33032 | MC5R_HUMAN Melanocortin-5 receptor (MC5-R) (MC-2)
282 113125 | sp | P20309 | ACM3_HUMAN Muscarinic acetylcholine receptor M3
283 543740 | sp | P29274 | AA2A_HUMAN Adenosine A2a receptor

289 1346168 | sp | P47775 | GP12_HUMAN Probable G protein-coupled receptor GPR12

292 543761 | sp | P08912 | ACM5_HUMAN Muscarinic acetylcholine receptor M5

295 17367264 | sp | Q9Y5N1 | HH3R_HUMAN Histamine H3 receptor (HH3R) (G protein-coupled receptor 97)

297 129203 | sp | P03999 | OPSB_HUMAN Blue-sensitive opsin (Blue cone photoreceptor pigment)
298 2495032 | sp | Q99500 | EDG3_HUMAN Lysosphingolipid receptor (EDG-3)

302 129215 | sp | P04001 | OPSG_HUMAN Green-sensitive opsin (Green cone photoreceptor pigment)

304 548476 | sp | P35408 | PE24_HUMAN Prostaglandin E2 receptor, EP4 subtype (Prostanoid EP4 receptor) (PGE receptor, EP4 subtype)
305 2495041 | sp | Q99680 | GP22_HUMAN Probable G protein-coupled receptor GPR22
306 547645 | sp | P35367 | HH1R_HUMAN Histamine H1 receptor

307 1351829 | sp | P08913 | A2AA_HUMAN Alpha-2A adrenergic receptor (Alpha-2A adrenoceptor) (Alpha-2AAR subtype C10)

309 129219 | sp | P04000 | OPSR_HUMAN Red-sensitive opsin (Red cone photoreceptor pigment)

311 3023411 | sp | Q16581 | C3AR_HUMAN C3a anaphylatoxin chemotactic receptor (C3a-R) (C3AR)

313 26454626 | sp | Q92633 | EDG2_HUMAN Lysophosphatidic acid receptor Edg-2 (LPA receptor 1) (LPA-1)
314 21263830 | sp | Q96G91 | P2YB_HUMAN P2Y purinoceptor 11 (P2Y11)

316 461697 | sp | P34972 | CB2R_HUMAN Cannabinoid receptor 2 (CB2) (CB-2) (CX5)

319 1350592 | sp | P47804 | RGR_HUMAN RPE-retinal G protein-coupled receptor

321 15214044 | sp | Q9HC97 | GP35_HUMAN Probable G protein-coupled receptor GPR35

324 1172071 | sp | P43115 | PE23_HUMAN Prostaglandin E2 receptor, EP3 subtype (Prostanoid EP3 receptor) (PGE receptor, EP3 subtype)

328 1170009 | sp | P46095 | GPR6_HUMAN Probable G protein-coupled receptor GPR6

331 21263686 | sp | Q9BZJ7 | GP62_HUMAN Probable G protein-coupled receptor GPR62 (hGPCR8)
332 2494993 | sp | Q15761 | NY5R_HUMAN Neuropeptide Y receptor type 5 (NPY5-R) (NPY-Y5 receptor) (Y5 receptor) (NPYY5)
333 544350 | sp | P23945 | FSHR_HUMAN Follicle stimulating hormone receptor precursor (FSH-R) (Follitropin receptor)
334 135920 | sp | P04201 | MAS_HUMAN MAS proto-oncogene

336 1174572 | sp | P21731 | TA2R_HUMAN Thromboxane A2 receptor (TXA2-R) (Prostanoid TP receptor)

338 113122 | sp | P08172 | ACM2_HUMAN Muscarinic acetylcholine receptor M2

340 464360 | sp | P34995 | PE21_HUMAN Prostaglandin E2 receptor, EP1 subtype (Prostanoid EP1 receptor) (PGE receptor, EP1 subtype)
341 136448 | sp | P16473 | TSHR_HUMAN Thyrotropin receptor precursor (TSH-R) (Thyroid stimulating hormone receptor)

344 12230026 | sp | O60883 | ETB2_HUMAN Endothelin B receptor-like protein-2 precursor (ETBR-LP-2)

346 10720152 | sp | O95665 | NTR2_HUMAN Neurotensin receptor type 2 (NT-R-2) (Levocabastine-sensitive neurotensin receptor) (NTR2 receptor)

349 1172070 | sp | P43116 | PE22_HUMAN Prostaglandin E2 receptor, EP2 subtype (Prostanoid EP2 receptor) (PGE receptor, EP2 subtype)
350 2494977 | sp | Q13585 | ML1X_HUMAN MELATONIN-RELATED RECEPTOR (H9)

353 1172500 | sp | P43119 | PI2R_HUMAN Prostacyclin receptor (Prostanoid IP receptor) (PGI receptor)
354 24638055 | sp | Q96AM1 | MRGF_HUMAN Mas-related G protein-coupled receptor MRGF (Mas-related gene F protein)

364 7674059 | sp | O14626 | H963_HUMAN Probable G protein-coupled receptor H963
365 28381373 | sp | P22888 | LSHR_HUMAN Lutropin-choriogonadotropic hormone receptor precursor (LH/CG-R) (LSH-R) (Luteinizing hormone receptor) (LHR)

371 12643545 | sp | O15354 | GP37_HUMAN Probable G protein-coupled receptor GPR37 precursor (Endothelin B receptor-like protein-1) (ETBR-LP-1)

374 1172442 | sp | P43088 | PF2R_HUMAN Prostaglandin F2-alpha receptor (Prostanoid FP receptor) (PGF receptor) (PGF2 alpha receptor)
375 547920 | sp | P35410 | MRG_HUMAN Mas-related G protein-coupled receptor MRG

377 21362643 | sp | Q9HBX9 | LGR7_HUMAN Relaxin receptor 1 (Leucine-rich repeat-containing G protein-coupled receptor 7)
378 2495009 | sp | Q13258 | PD2R_HUMAN Prostaglandin D2 receptor (Prostanoid DP receptor) (PGD receptor)

381 21362625 | sp | Q8WXD0 | LGR8_HUMAN Relaxin receptor 2 (Leucine-rich repeat-containing G protein-coupled receptor 8) (G protein-coupled receptor affecting testicular descent)

386 21263835 | sp | Q9H244 | P2YC_HUMAN P2Y purinoceptor 12 (P2Y12) (P2Y12 platelet ADP receptor) (P2Y(ADP)) (ADP-glucose receptor) (ADPG-R) (P2Y(AC)) (P2Y(cyc)) (P2T(AC)) (SP1999)

390 21542118 | sp | O75473 | LGR5_HUMAN Leucine-rich repeat-containing G protein-coupled receptor 5 precursor (Orphan G protein-coupled receptor HG38) (G protein-coupled receptor 49)

393 1346133 | sp | P48546 | GIPR_HUMAN Gastric inhibitory polypeptide receptor precursor (GIP-R) (Glucose-dependent insulinotropic polypeptide receptor)
394 399777 | sp | P30968 | GRHR_HUMAN Gonadotropin-releasing hormone receptor (GnRH receptor) (GnRH-R)

398 2495040 | sp | Q99679 | GP21_HUMAN Probable G protein-coupled receptor GPR21

402 3122322 | sp | Q15391 | P2YX_HUMAN UDP-glucose receptor (G protein-coupled receptor GPR105)

409 17432988 | sp | O75084 | FZD7_HUMAN Frizzled 7 precursor (Frizzled-7) (Fz-7) (hFz7) (FzE3)
410 17433018 | sp | Q13467 | FZD5_HUMAN Frizzled 5 precursor (Frizzled-5) (Fz-5) (hFz5) (FzE5)
411 17433091 | sp | Q9ULW2 | FZ10_HUMAN Frizzled 10 precursor (Frizzled-10) (Fz-10) (hFz10) (FzE7)
412 418253 | sp | P32241 | VIPR_HUMAN Vasoactive intestinal polypeptide receptor 1 precursor (VIP-R-1) (Pituitary adenylate cyclase activating polypeptide type II receptor) (PACAP type II receptor) (PACAP-R-2)

416 21542119 | sp | Q9BXB1 | LGR4_HUMAN Leucine-rich repeat-containing G protein-coupled receptor 4 precursor (G protein-coupled receptor 48)
417 12643950 | sp | Q9Y2T5 | GP52_HUMAN Probable G protein-coupled receptor GPR52

419 2506489 | sp | P47872 | SCRC_HUMAN Secretin receptor precursor (SCT-R)
420 2495077 | sp | Q14833 | MGR4_HUMAN Metabotropic glutamate receptor 4 precursor (mGluR4)

426 27734323 | sp | Q96P67 | GP82_HUMAN Probable G protein-coupled receptor GPR82

430 3041685 | sp | Q02643 | GRFR_HUMAN Growth hormone-releasing hormone receptor precursor (GHRH receptor) (GRF receptor) (GRFR)
431 17433092 | sp | Q9UP38 | FZD1_HUMAN Frizzled 1 precursor (Frizzled-1) (Fz-1) (hFz1) (FzE1)

434 417555 | sp | Q03431 | PTRR_HUMAN Parathyroid hormone/parathyroid hormone-related peptide receptor precursor (PTH/PTHR receptor)
435 17432964 | sp | O00144 | FZD9_HUMAN Frizzled 9 precursor (Frizzled-9) (Fz-9) (hFz9) (FzE6)

437 1171986 | sp | P41586 | PACR_HUMAN Pituitary adenylate cyclase activating polypeptide type I receptor precursor (PACAP type I receptor) (PACAP-R-1)
438 6226847 | sp | Q13324 | CRF2_HUMAN Corticotropin releasing factor receptor 2 precursor (CRF-R 2) (CRF2) (Corticotropin-releasing hormone receptor 2) (CRH-R 2)
439 12644040 | sp | O00222 | MGR8_HUMAN Metabotropic glutamate receptor 8 precursor (mGluR8)
440 1169956 | sp | P43220 | GLP1_HUMAN Glucagon-like peptide 1 receptor precursor (GLP-1 receptor) (GLP-1-R) (GLP-1R)

442 17433090 | sp | Q9ULV1 | FZD4_HUMAN Frizzled 4 precursor (Frizzled-4) (Fz-4) (hFz4) (FzE4)
443 12230071 | sp | O95838 | GLP2_HUMAN Glucagon-like peptide 2 receptor precursor (GLP-2 receptor) (GLP-2-R) (GLP-2R)

448 2495058 | sp | Q16602 | CGRR_HUMAN Calcitonin gene-related peptide type 1 receptor precursor (CGRP type 1 receptor)

450 399180 | sp | P30988 | CALR_HUMAN Calcitonin receptor precursor (CT-R)
451 17433019 | sp | Q14332 | FZD2_HUMAN Frizzled 2 precursor (Frizzled-2) (Fz-2) (hFz2) (FzE2)

455 2495078 | sp | Q14831 | MGR7_HUMAN Metabotropic glutamate receptor 7 precursor (mGluR7)

457 3219999 | sp | P51810 | OA1_HUMAN Ocular albinism type 1 protein

459 1346906 | sp | P49190 | PTR2_HUMAN Parathyroid hormone receptor precursor (PTH2 receptor)

461 461836 | sp | P34998 | CRF1_HUMAN Corticotropin releasing factor receptor 1 precursor (CRF-R) (CRF1) (Corticotropin-releasing hormone receptor 1) (CRH-R 1)

464 21362642 | sp | Q9HBX8 | LGR6_HUMAN Leucine-rich repeat-containing G protein-coupled receptor 6

467 1168781 | sp | P41180 | CASR_HUMAN Extracellular calcium-sensing receptor precursor (CaSR) (Parathyroid Cell calcium-sensing receptor)

476 3024134 | sp | O15303 | MGR6_HUMAN Metabotropic glutamate receptor 6 precursor (mGluR6)

483 1346144 | sp | P47871 | GLR_HUMAN Glucagon receptor precursor (GL-R)

489 6226566 | sp | P48960 | CD97_HUMAN Leucocyte antigen CD97 precursor

501 2506490 | sp | P41587 | VIPS_HUMAN Vasoactive intestinal polypeptide receptor 2 precursor (VIP-R-2) (Pituitary adenylate cyclase activating polypeptide type III receptor) (PACAP type III receptor) (PACAP-R-3) (Helodermin-preferring VIP receptor)

507 2495033 | sp | Q14330 | GP18_HUMAN Probable G protein-coupled receptor GPR18

512 22095550 | sp | Q9HCU4 | CLR2_HUMAN Cadherin EGF LAG seven-pass G-type receptor 2 precursor (Epidermal growth factor-like 2) (Multiple epidermal growth factor-like domains 3) (Flamingo 1)

514 2495075 | sp | Q14416 | MGR2_HUMAN Metabotropic glutamate receptor 2 precursor (mGluR2)

517 22095551 | sp | Q9NYQ6 | CLR1_HUMAN Cadherin EGF LAG seven-pass G-type receptor 1 precursor (Flamingo homolog 2) (hFmi2)
518 17433071 | sp | Q9NPG1 | FZD3_HUMAN Frizzled 3 precursor (Frizzled-3) (Fz-3) (hFz3)

520 10719900 | sp | O14514 | BAI1_HUMAN Brain-specific angiogenesis inhibitor 1 precursor

522 2495072 | sp | Q14246 | EMR1_HUMAN Cell surface glycoprotein EMR1 precursor (EMR1 hormone receptor)

524 27151770 | sp | Q16570 | DUFF_HUMAN Duffy antigen/chemokine receptor (Fy glycoprotein) (GpFy) (Glycoprotein D) (Plasmodium vivax receptor) (CD234 antigen)

527 22095552 | sp | Q9NYQ7 | CLR3_HUMAN Cadherin EGF LAG seven-pass G-type receptor 3 precursor (Flamingo homolog 1) (hFmi1) (Multiple epidermal growth factor-like domains 2) (Epidermal growth factor-like 1)
528 12643618 | sp | O60242 | BAI3_HUMAN Brain-specific angiogenesis inhibitor 3 precursor

531 12643641 | sp | O75899 | GBR2_HUMAN Gamma-aminobutyric acid type B receptor, subunit 2 precursor (GABA-B receptor 2) (GABA-B-R2) (Gb2) (GABABR2) (G protein-coupled receptor 51) (GPR 51) (HG20)

533 10719903 | sp | O60241 | BAI2_HUMAN Brain-specific angiogenesis inhibitor 2 precursor
534 6226142 | sp | Q99835 | SMO_HUMAN Smoothened homolog precursor (SMO) (Gx protein)
535 2495074 | sp | Q13255 | MGR1_HUMAN Metabotropic glutamate receptor 1 precursor (mGluR1)

537 1709020 | sp | P41594 | MGR5_HUMAN Metabotropic glutamate receptor 5 precursor (mGluR5)
538 17433053 | sp | Q9H461 | FZD8_HUMAN Frizzled 8 precursor (Frizzled-8) (Fz-8) (hFz8)

540 2495076 | sp | Q14832 | MGR3_HUMAN Metabotropic glutamate receptor 3 precursor (mGluR3)

545 17432985 | sp | O60353 | FZD6_HUMAN Frizzled 6 precursor (Frizzled-6) (Fz-6) (hFz6)

549 12643873 | sp | Q9UBS5 | GBR1_HUMAN Gamma-aminobutyric acid type B receptor, subunit 1 precursor (GABA-B receptor 1) (GABA-B-R1) (Gb1)

References

Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W. and D.J., L. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.

Altschul, S. F.; W.;, G.; Miller, W.; Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol.* 215, 403-410.

BioNavigator (by Entigen Corporation). by Entigen Corporation (http://www.entigen.com).

Consortium (2001). Functional annotation of a full-length mouse cDNA collection. *Nature* 409, 685-690.

Felsenstein, J. (1989). Phylip-Phylogeny Inference Package (version 3.2). *Cladistics* 5, 164-166.

Fredriksson, R.; Lagerstrom, M. C.; Lundin, L.-G. and Schioth, H. B. (2003). The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Molecular Pharmacology* 63, 1256-1272.

Gaulton, A. and Attwood, T. K. (2003). Bioinformatics approaches for the classification of G-protein-coupled receptors. *Current Opinion in Pharmacology* 3, 114-120.

Hesselgesser, J.; Ng, H. P.; Liang, M.; Zheng, W.; May, K.; Bauman, J. G.; Monahan, S.; Islam, I.; Wei, G. P.; Ghannam, A.; Taub, D. D.; Rosser, M.; Snider, R. M.; Morrissey, M. M.; Perez, H. D. and Horuk, R. (1998). *J. Biol. Chem.* 273, 15687-15692.

Inoue, Y.; Ikeda, M. and Shimizu, J. (2004). Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern. *Computational Biology and Chemistry* 28, 39-49.

Papasaikas, P. K.; Bagos, P. G.; Litou, Z. I. and Hamodrakas, S. J. (2003). A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models. *SAR and QSAR in environmental research* 14, 413-420.

Thompson, J. D.; Higgins, D. G. and Gibson, T. J. (1994). Clustal-W - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673-4680.

Trabanino, R. J.; Hall, S. E.; Vaidehi, N.; Floriano, W. B.; Kam, V. W. T. and Goddard III, W. A. (2004). First principles predictions of the structure and function of G-protein-coupled receptors: validation for bovine rhodopsin. *Biophys. J.* 86, 1904-1921.

Vassilatis, D. K.; Hohmann, J. G.; Zeng, H.; Li, F.; Ranchalis, J. E.; Mortrud, M. T.; Brown, A.; Rodriguez, S. S.; Weller, J. R.; Wright, A. C.; Bergmann, J. E. and Gaitanaris, G. A. (2003). The G protein-coupled receptor repertoires of human and mouse. *Proc. Natl. Acad. Sci. USA.* 100, 4903-4908.

Yanbin, Y.; Jingchu, L. and Ying, J. (2003). Advances in G-protein coupled receptor research and related bioinformatics study. *Chinese Science Bulletin* 48, 511-516.

Chapter 6: Simulation of EC-II loop closure

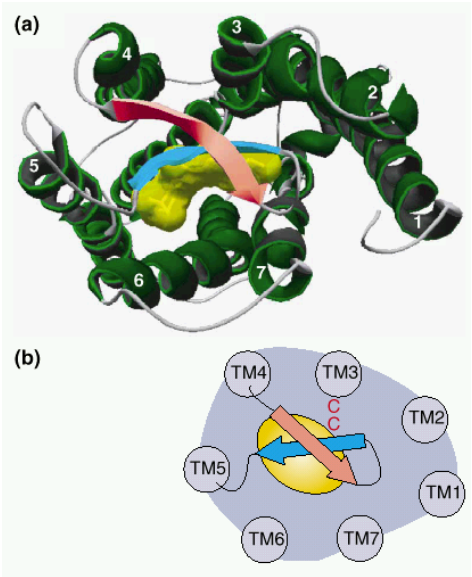
Abstract

In G-protein-coupled receptors (GPCRs), the second extracellular loop (EC-II) between helices 4 and 5 has been known to be important in binding ligands and in the activation process. In the bovine rhodopsin crystal structure, it assumes a closed form, which interacts closely with the cis-retinal in the ground state of the protein. Thus, the closing of this loop is of critical importance in creating structures for GPCRs. The EC_LOOP_SIM protocol of MembStruk was created to simulate the folding process of this loop from an "open" to a "closed" form. This method has been validated for the case of bovine rhodopsin, with the loop's final CRMS from the crystal structure of 6.88 Å in main chain atoms. When the loops were aligned to each other, the error was 4.71 Å in the internal structure. The residues important in binding in the B strand of the EC-II loop beta sheet were positioned correctly in this simulated "closed" loop structure. This novel approach of folding the EC-II from first principles is important in building structures of GPCRs with distant homology to bovine rhodopsin (precluding the usage of homology modeling here). In addition, the role of the "opening" and "closing" of this loop on retinal binding in rhodopsin is elucidated in this paper.

Introduction

In the crystal structure of bovine rhodopsin (Palczewski et al., 2000), the EC-II loop forms an anti-parallel beta-sheet with one strands (the B or second strand) interacting more closely with the bound cis-retinal ligand, as shown in Figure 1. Since this crystal structure had retinal in the cis form, it corresponds to the ground state inactive form of rhodopsin. It is interesting to note that this loop is uncharacteristic among loops in that the crystallographic B-factors in this region are comparable to those in helical region. This temperature stability is probably due in a great part to the presence of a disulfide linkage between two conserved (across GPCRs) cysteines in the N-terminal side of helix 3 and the EC-II loop. This constraint provides the stability needed for such an important loop in ligand binding.

Figure 1: Structure of the EC-II loop over the bound retinal in the bovine rhodopsin crystal structure. Figure: (Onuffer et al., 2002)

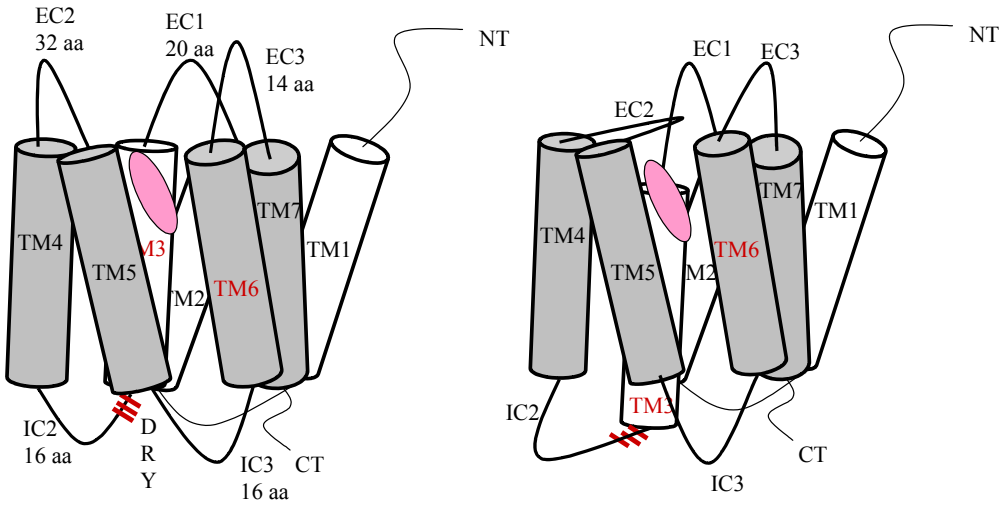


Various experiments have corroborated the importance of this loop in ligand binding and activation in other GPCRs. In chemokine receptors (another class of GPCRs), the binding of antibodies directed to the B strand of the EC-II loop was inhibiting by the presence of antagonist (Onuffer et al., 2002). This was demonstrated for CCR5 binding to TAK779 and E913 (Dragic et al., 2000; Maeda, 2001) and for CXCR4 binding to AMD3100 (Gerlach, 2001).

In fact, many of the mutations causing blindness in retinitis pigmentosa (RP) are in the ECII loop (Okada et al., 2001) and mutations of the cysteines which form the disulfide also cause RP and change ligand binding and receptor activation (Schoneberg et al., 2002).

The coupling between this loop and helix 3 allows for a transfer of movement from helix 3 to the loop. A helix 3 movement upon activation is expected from the following observation: trans retinal contacts in the rhodopsin crystal structure (Borhan et al., 2000) can only match cross-linking experiments if the ligand clashes with the helix 3 in the ground state conformation (Bourne et al., 2000). A proposed mechanism for loop movement upon activation was described previously (Trabanino et al., 2004). In this mechanism, the isomerization of retinal leads to a translation of helix 3, which because of its coupling to the EC-II loop, would “open” this loop and allow exit of the retinal ligand (Figure 2).

Figure 2: Schematic of proposed mechanism for loop closure.



Thus, the "open" form of ECII may correspond to the activated form of the helix bundle by this coupling, as supported by the fact that GPCRs in which this disulfide is not present (melanocortin, cannabinoid, and sphingolipid receptors), the agonist-independent activity (constitutive activity) tends to be high (Milligan, 2003).

This paper describes the method used by MembStruk for EC-II loop closing and validates its use in the only available GPCR crystal structure, bovine rhodopsin. In addition, the changes in retinal binding observed in the "open" loop form of the receptor are shown and discussed in the context of a potential retinal binding mechanism which involved partitioning from the membrane environment into the receptor interior.

Methods

Loop addition and pre-closing optimization

In the MembStruk protocol, the loops for GPCRs are added using either Whatif (Vriend, 1990) or Modeller 6v2. These loops are derived from a database of crystal structure fragments and are thus not well representative of the loops which may be present in GPCRs. The EC_LOOP_SIM program was created to fold this loop in the environment of the particular protein of interest. Thus annealing dynamics using Biograf software is performed on the EC-II loop for 100 ps with cycles from 300 to 600 degrees. After every two cycles, the loop's potential energy is minimized. This step ensures that internal structures in the added loop are randomized, to avoid bias in later folding due to a particular internal structure. This step is optional if one wants to begin the next step with an internal structure in the loop (beta-sheet, etc).

Disulfide bond formation

The disulfide bond is formed between the cysteines in extracellular 2 (EC-II) loop (which are conserved across many GPCRs) and the N-terminal edge of TM3 or EC3. In the case of bovine rhodopsin, the alignment of 44 sequences indicates only one pair of fully conserved cysteines which are on the same side of the membrane (extracellular side). The disulfide bond was formed and optimized (with one cycle of annealing dynamics from 300 to 600 degrees after each decrement) with equilibrium distances lowered in decrements of 2 Å until the bond distance was 2 Å. Then the loop was optimized (one cycle of annealing dynamics) with the default equilibrium disulfide bond distance of 2.07 Å.

Room temperature molecular dynamics

At this point, it was found that the backbone of the helices needed to be allowed movable to accommodate the closing EC-II loop. Thus, TVN molecular dynamics with half the bundle movable was performed. The half of the bundle was defined as all atoms towards the extracellular side from the hydrophobic center of each helix. After 100 ps, the EC-II assumed a "closed" conformation in the simulation of the loop in the crystal structure of bovine rhodopsin. The snapshots of this closing are seen in Figure 3. In other cases (like human CCR1), however, this MD step was not sufficient to fold the EC-II loop into a closed conformation.

Annealing dynamics with movable side chains

In such cases, 300 cycles of simulated annealing dynamics was performed on the EC-II loop with surrounding side chains in the protein movable. The temperature variation was 300 to 600 degrees with a temperature increment of 200 degrees, and a potential energy minimization was performed after every two cycles. After every two cycles, the residues around the loop which are designated as movable are updated, in order to provide a movable time-dependent path for the loop's closing.

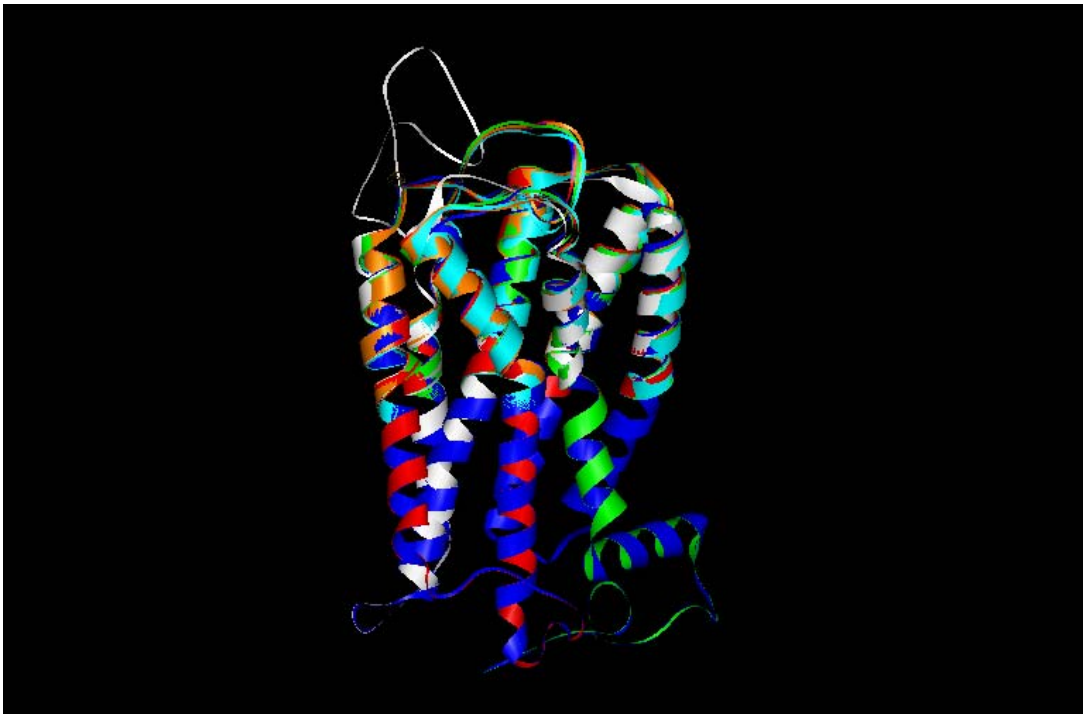
Results and discussion

Validation for bovine rhodopsin

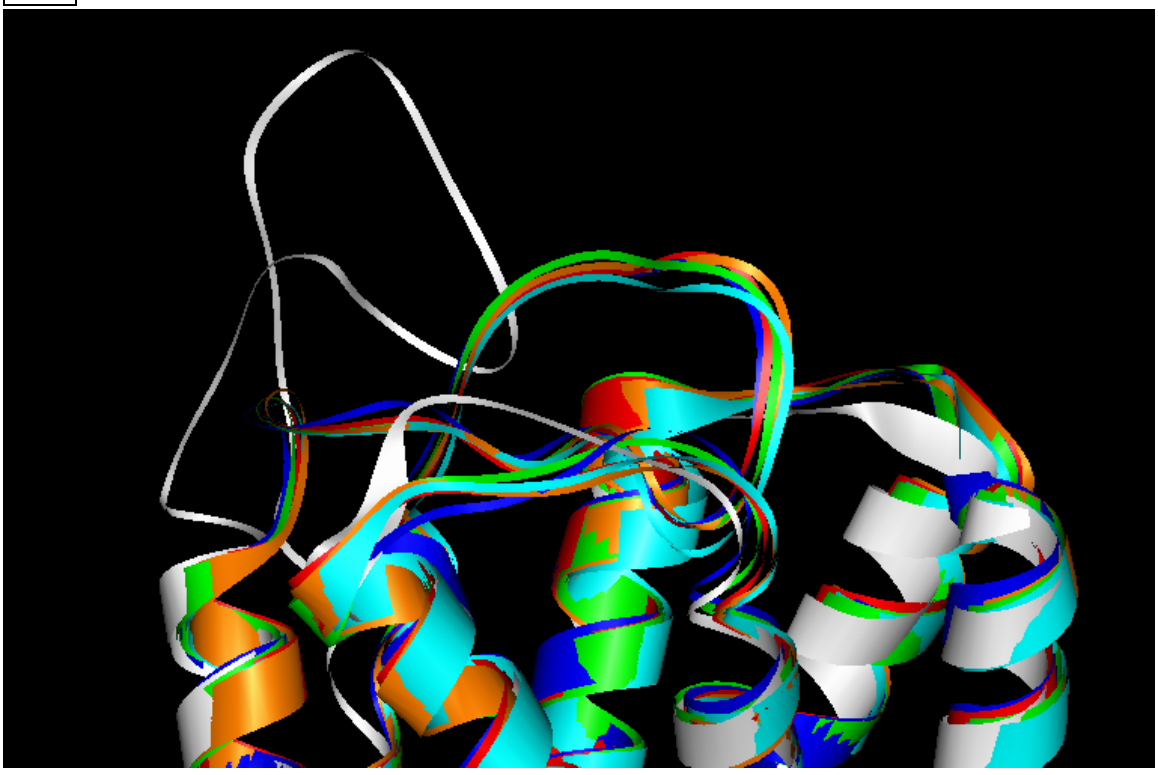
In the case of bovine rhodopsin, the closing of the EC-II loop is depicted in Figure 3ab with snapshots every 20 ps (sequential colors of white, cyan, orange, green, red, blue corresponding to those snapshots). These snapshots are taken throughout the course of the room temperature molecular dynamics described in the Methods section. The bundle opens slightly and allows the entry of the EC-II loop structure into a "closed" conformation. This simulation could also be conducted with ligand bound once the binding site is determined.

Figure 3: The snapshots of loop closure every 20 ps in bovine rhodopsin. (Progression is white, cyan, orange, green, red, blue)

a



b



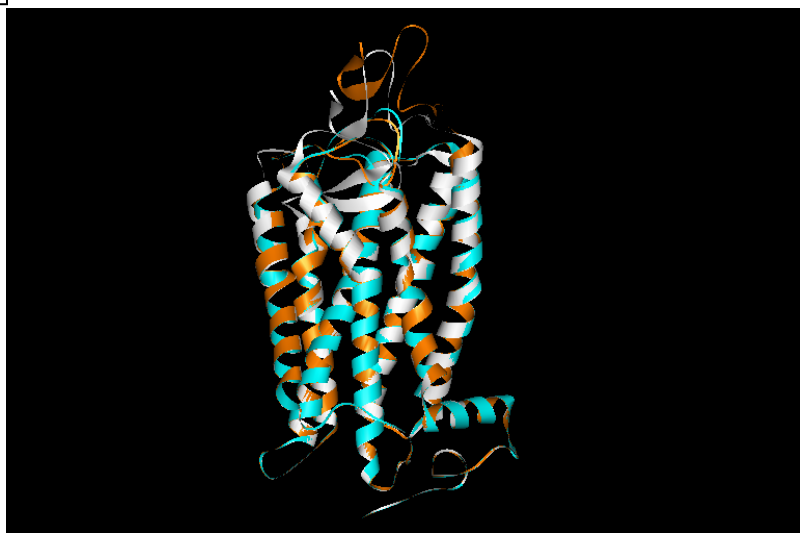
The comparison between the simulated (cyan) bovine rhodopsin EC-II loop and the crystal structure (white) of this loop is shown in Figure 4ab. The CRMS error in the main chain atoms of this loop is 6.77 Å. This error is partly due to some translation error which may arise from the fact that the simulated protein is not packed in a crystal as for the experimental structure. In addition, it may be partly due to the absence of the N-

terminus (since this is not currently modeled in the MembStruk procedure, it was not included here) which packs under the EC-II loop in the crystal structure as shown in white in Figure 4. If the loops are aligned by a best-fit algorithm, the error in the internal structure is 3.67 Å. This arises from the absence of the hydrogen bonding needed for beta-sheet formation. The hydrogen bonding energy wells in the force field parameters could be deepened to increase the likelihood of the simulated structure forming a beta-sheet.

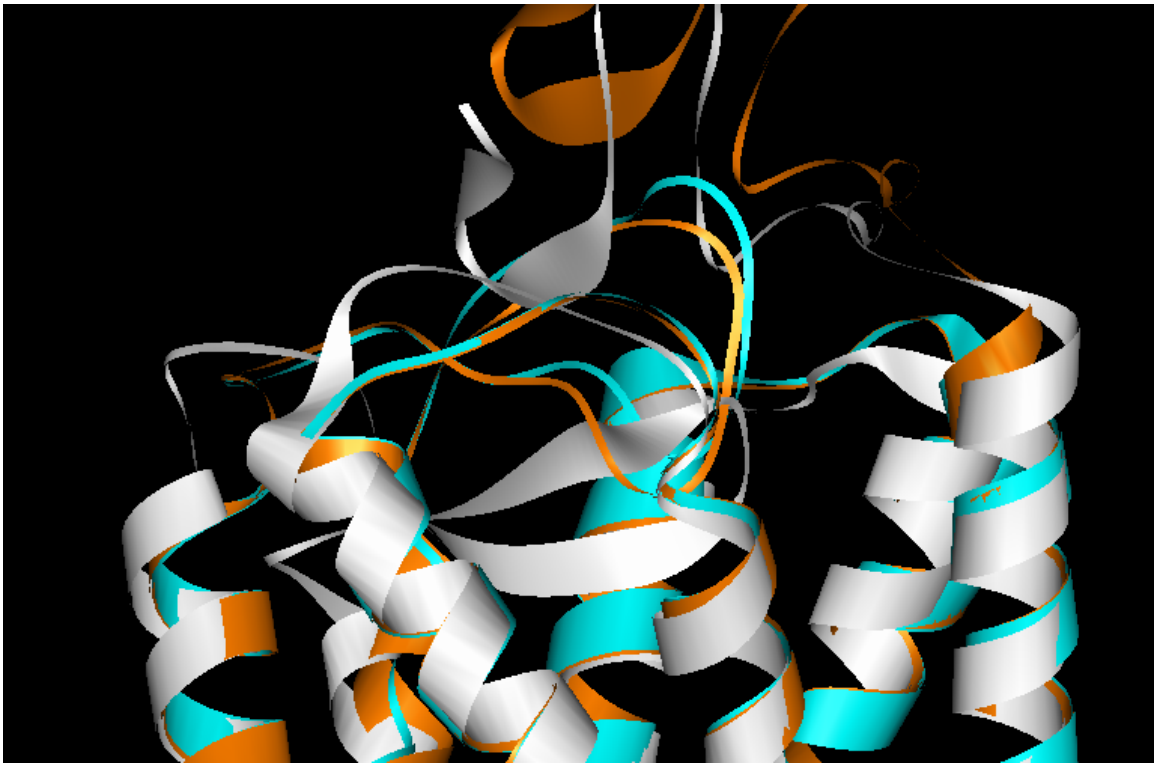
If the N-terminus is included (orange in Figure 4) in the simulation, the error in CRMS improves. The error is 5.73 Å, and after alignment of the loops, the internal structure has an error of 3.56 Å. Thus, as expected, the translation of the loop improved but the beta-sheet structure still was not formed. The translation may be improved even further by packing the N-terminus exactly as in the crystal structure (currently the N-terminus was displaced slightly from its position in the crystal structure in order to reduce clashes with the simulated loop).

Figure 4: Comparison of simulated loop without (cyan) and with (orange) the N-terminus with the crystal structure of the loop (white).

a



b



Role of EC-II loop in ligand binding

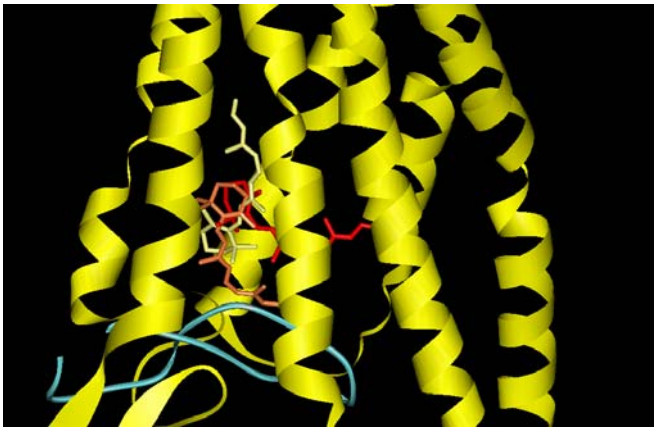
A structure of bovine rhodopsin was previously built using the MembStruk version 2.0 with loop open. The binding of the retinal ligand ligand was predicted using HierDock (Trabanino et al., 2004). There were three binding modes of retinal found amongst the top scoring in this open loop structure, as shown in Figure 5. One binding mode (red) corresponded with the mode observed in the crystal structure. The other modes were vertical in terms of their orientation and may correspond to intermediate binding modes. This is supported by the studies in retinal binding (Isralewitz et al., 1997; Schadel et al., 2003) which conclude that retinal binds by partitioning from membrane to

the protein interior. Because of its amphipathicity, retinal would be expected to partition in a vertical orientation, aligning with the lipids of the bilayer.

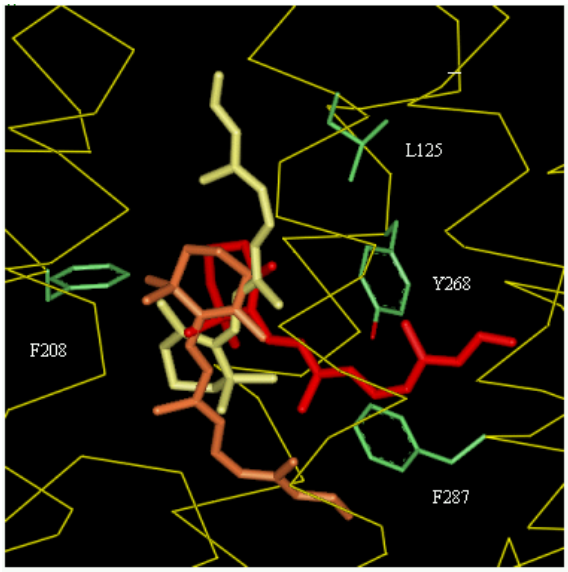
In addition, if the EC-II loop (cyan) from the crystal structure is incorporated, the two vertical binding modes are eliminated from the top scoring conformations of HierDock. Thus, the closing of the EC-II loop favors the horizontal retinal orientation which is able to form the Schiff base linkage and subsequently absorb incident radiation in the visible range required for activation and finally vision.

Figure 5: Retinal conformation in an "open" loop predicted rhodopsin structure. The position of the EC-II loop in the crystal structure is shown (cyan).

a

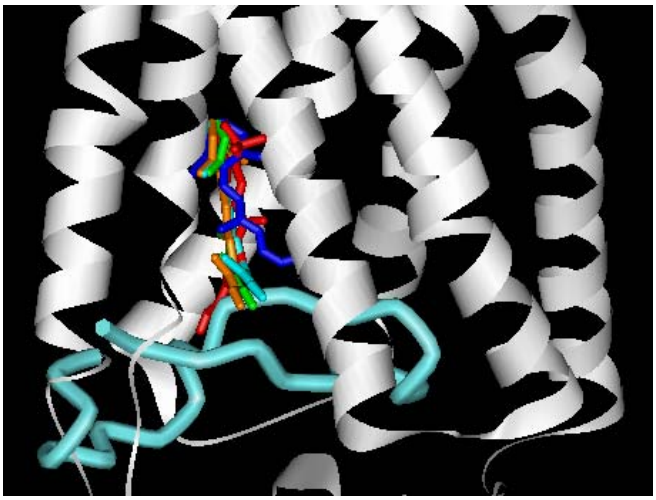


b



In fact, if the EC-II is "opened" in the bovine rhodopsin crystal structure, the vertical conformations are observed amongst the top HierDock conformations (Figure 6). The closest in CMRS of these conformations to the crystal is 2.14 Å (as opposed to the 1.2 Å CRMS of the best scoring HierDock retinal conformation in the closed EC-II form of the crystal structure). In addition, the vertical conformations found from the HierDock procedure are oriented with the polar aldehyde tail towards the direction of the extracellular region, and thus aligned with the amphipathic orientation of the lipids in the membrane bilayer.

Figure 6: Retinal conformation in the rhodopsin crystal structure with an "open" loop. The position of the EC-II loop in the "closed" crystal structure is shown (cyan).



Conclusion

The role of the EC-II loop in binding and activation in GPCRs has been verified by various experiments. In addition, its role in determining retinal binding, as reported here, has been studied as well. It is therefore crucial to simulate the opening and closing of this loop for function prediction. This method has been validated in the case of bovine rhodopsin and has been used in the simulation of the EC-II loop in human CCR1, where is has proven to yield insights into binding of various antagonists (as discussed later in Chapter 9).

References

Borhan, B.; Souto, M. L.; Imai, H.; Schichida, Y. and Nakanashi, K. (2000). Movement of retinal along the visual transduction path. *Science* 288, 2209-2212.

Bourne, H. R. and Meng, E. C. (2000). Structure - Rhodopsin sees the light. *Science* 289, 733-734.

Dragic, T.; Trkola, A.; Thompson, D. A. D.; Cormier, E. G.; Kajumo, F. A.; Maxwell, E.; Lin, S. W.; Ying, W.; Smith, S. O.; Sakmar, T. P. and Moore, J. P. (2000). A binding pocket for a small molecule inhibitor of HIV-1 entry within the transmembane helices of CCR5. *Proc. Natl Acad. Sci. USA.* 97, 5639-5644.

Gerlach, L. O., et al (2001). Molecular interactionsof cyclam and bicyclam non-peptide antagonists with the CXCR4 chemokine receptor. *J. Biol. Chem.* 276, 14153-14160.

Isralewitz, B.; Izrailev, S. and Schulten, K. (1997). Binding pathway of retinal to bacterio-opsin: A prediction by molecular dynamics simulations. *Biophys J.* 73, 2972-2979.

Maeda, K., et al. (2001). Novel low molecular weight spirodiketopiperazine derivatives potently inhibit R5 HIV-1 infection through their antagonistic effects on CCR5. *J. Biol. Chem.* 276, 35194-35200.

Milligan, G. (2003). Constitutive activity and inverse agonists of G Protein coupled receptors: a current perspective. *Molecular Pharmacology* 64, 1271-1276.

Okada, T.; Ernst, O. P.; Palczewski, K. and Hofmann, K. P. (2001). Activation of rhodopsin: new insights from structural and biochemical studies. *TRENDS in Biochemical Sciences* 26, 318-324.

Onuffer, J. and Horuk, R. (2002). Chemokines, chemokine receptors and small-molecule antagonists: recent developments. *TRENDS in Biochemical Sciences* 23, 459-467.

Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C.; Motoshima, H.; Fox, B.; Trong, I.; Teller, D.; Okada, T.; Stenkamp, R.; Yamamoto, M. and Miyano, M. (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* 289, 739-745.

Schadel, S. A.; Heck, M.; Maretzki, D.; Filipek, S.; Teller, D. C.; Palczewski, K. and Hofmann, K. P. (2003). Ligand channeling within a G-protein-coupled receptor. *J Biol. Chem.* 27, 24896-24903.

Schoneberg, T.; Schulz, A. and T., G. (2002). The structural basis of G-protein-coupled receptor function and dysfunction in human diseases. *Rev. Phys. Biochem. Pharm.* 144, 145-227.

Trabanino, R. J.; Hall, S. E.; Vaidehi, N.; Floriano, W. B.; Kam, V. W. T. and Goddard III, W. A. (2004). First principles predictions of the structure and function of G-protein-coupled receptors: validation for bovine rhodopsin. *Biophys. J.* 86, 1904-1921.

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

Part 2: Applications

Chapter 7: “Let there be sight”: Molecular mechanism for color distinction in humans

Abstract

The role of proteins in the modulation of the absorption and fluorescence properties of small molecules is a fundamental scientific problem with useful applications in other fields such as medicine. This study uses QM and hydrid QMMM techniques to ascertain how the molecular design of the three human opsin proteins has made possible the modulation of the absorbed frequency of a bound chromophore, in the crucial step of vision. It is found that the green to red shift can be attributed to the presence of dipolar residues. In addition, molecular dynamics of a QM-fitted retinal protonated Schiff base (PSB) within the opsins has allowed the calculation of absorption spectra for the three color opsins. It is found that the twisting of the retinal PSB plays the predominant role in the green to blue opsin shift, whereas the polarizable aromatic side chains play a surprising role of red-shifting the blue opsin with respect to the green opsin, as a fine adjustment to the opsin shift.

Introduction

The role of proteins in the modulation of the absorption and fluorescence properties of small molecules is a fundamental scientific problem with useful applications in medicine (immunochemistry and histology). An example of this role is the modulation by antibodies of trans-stilbene fluorescence (23). But probably the most interesting example is the modulation of retinal absorption by opsin proteins in the crucial step of vision, in which electromagnetic radiation is converted into mechanical motion of the retinal and subsequently of the protein.

Retinal is the chromophore responsible for vision in humans. This molecule can absorb radiation corresponding to the energy gap of the electronic transition from the closed shell singlet to the open shell singlet state. In response, the molecule may undergo an isomerization about the 11-12 bond (Figure 1).

In addition, the terminal aldehyde of 11cis-retinal can form a Schiff base bond with amines. The unprotonated form of 11cis-retinal absorbs radiation at a wavelength maximum of ~380 nm in organic solvents such as ethanol (1). But this Schiff base can be protonated at physiologic pH's to form a protonated Schiff base (PSB). This resulting PSB absorbs radiation at a wavelength maximum of 440 nm.

Such a covalent linkage occurs via a specific lysine (Lys296) (Figure 2) in opsin proteins localized in the retina of humans. These proteins modulate the absorption maximum of 11cis-retinal even further. The rod cells of the retina, for example, contain rhodopsin, an opsin/retinal complex which absorbs light at ~500 nm. The cone cells

contain 3 types of opsins in humans which (in complexes with retinal) absorb light maximally in the blue (~425 nm), green (~530 nm), and red (~560 nm) regions of the spectrum (2).

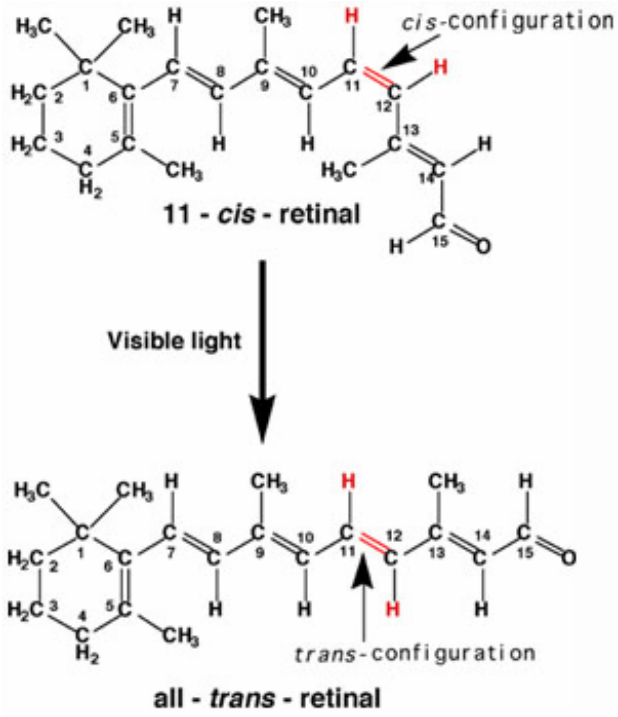


Figure taken from (22)

Figure 1: Photoisomerization of the free retinal molecule.

Figure 2: Schiff base bond of retinal ligand with protein via lysine side chain in opsins.

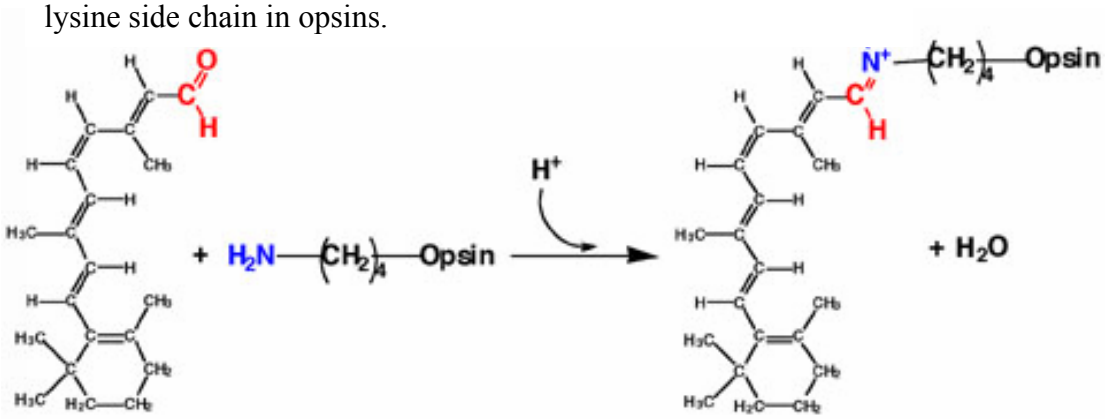


Figure taken from (22)

What are the factors which contribute to this spectral tuning of the retinal chromophore? This has been a fundamental question which has motivated decades of research. It is has generally been determined that 4 factors may contribute to this spectral shift in opsins:

1) The degree of coplanarization of the retinal polyene chain and the ionone ring (3)
2) Interaction of the retinal Schiff base with its counterion (acidic side chain) (4)
3) Influence of charged and dipolar residues around the retinal conjugated system (5)
4) Effect of polarizable side chains (such as aromatic side chains) (6)

The first factor was studied in bacteriorhodopsin (bR) by using retinal analogs which have the ionone ring "locked," preventing it from twisting (3). So this analog locks the C6-C7 bond (the bond adjacent to ionone ring), which has been known to modulate spectroscopic properties of the protein/retinal complex (14). It was found that the spectral shift due to this ionone ring's twisting was ~1200 cm^{-1} (in comparison to the total 5100 cm^{-1} shift for bR).

To understand how factors 2, 3, and 4 may result in spectral shifts, one must look at the possible resonance structures of the protonated and deprotonated 11cis-retinal schiff bases (Figure 3). There are two important facts confirmed by experiment in relation to these structures: 1) the excited state wavefunction in theoretical calculations has a larger weight put on the resonance structures on the right than the ground state wavefunction (7), 2) the electrons are more delocalized (and thus there is less bond alternation) in the excited state than in the ground state (7). Thus, the excited state of the PSB retinal has a larger dipole moment than the ground state. The positive charge is distributed to the

ionone ring upon excitation. The relative stability of these structures, and thus the excitation energy, may be influenced by the electric field (or a polarizable medium) in the environment of the chromophore.

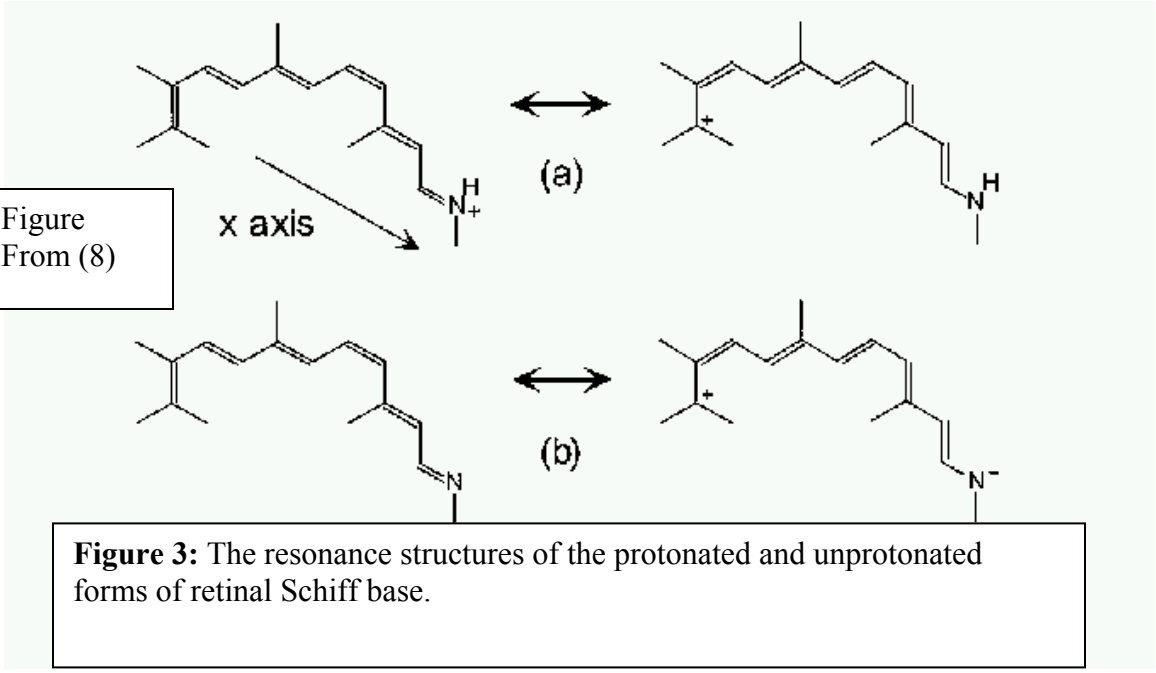


Figure 3: The resonance structures of the protonated and unprotonated forms of retinal Schiff base.

The role of the second factor in spectral tuning was studied in rhodopsin by Sakmar et al (4). The residue Glu-113 in rhodopsin is known to be the counterion for the schiff base of the conjugated retinal (13). This residue was thus mutated to Gln, Asp, Asn, and Ala. As expected, the Schiff base pKa's were reduced. In addition, the absorption maxima were blue-shifted as a result of replacing the anionic Glu residue. This blue-shift, corresponding to an increase in the excitation energy, is due to the role of the negatively charged counterion in stabilizing the charge distribution of the ground state (which, as explained above, has a larger positive distribution around the schiff base region) and destabilizing the excited state distribution (which has a larger negative distribution around the schiff base region).

Experiments have been conducted to determine the contribution to the spectral shift in the absence of the first two factors. This was carried out on bacteriorhodopdin by Yan et al using retinal analogs which did not have an ionone ring and in which the conjugated polyene system was interrupted by hydrogenating the double bond adjacent to the schiff base bond. Thus, a spectral shift of 2080 cm^{-1} was determined in the absence of the coplanarization/counterion effects.

The role of the third factor was further elucidated in a study by Kochendoerfer et al (2). They used Raman resonance spectroscopy on the opsin/retinal complexes to determine that the peak for the ethylenic stretch gradually was shifted (blue,1559 cm^{-1} to green,1531 cm^{-1} to red,1526 cm^{-1}) for the opsins. This is a result of the increasing electronic delocalization in the red opsin environment (and thus the reduction of bond alteration). Furthermore, models of the human opsins were built. From the amino acids around the chomophore, it was shown that the distribution of dipolar residues around the ionone ring was increased from the blue to red opsin model. Qualitatively, the role of these residues in differentially stabilizing the ground and excited state electron distributions in retinal (much the same mechanism as the counterion above) could be inferred.

The role of the last factor was studied by Houjou et al (6) using the ab initio (quantum mechanical) method of self-consistent reaction field (SCRF), in which the charge distribution of the retinal polarizes the medium, which in turn produces a reaction field which acts back on the retinal. The calculated opsin shift arising from the effect of the polarizable medium was ~1000 cm^{-1}.

The recent crystal structure determination of bovine rhodopsin has helped reveal the specific residues which may be involved in spectral tuning in opsins. The sequence identities and similarities between bovine rhodopsin and the human opsins are as follows: 88,91 human rhodopsin, 40,60 human blue opsin, 40,57 human green opsin, 39,56 human red opsin. The values are probably higher for the transmembrane region alone. This high similarity justifies the use of homology modeling and threading techniques to determine the opsin structures (12).

Using such structures, one can perform quantum mechanical calculations to determine the spectral tuning of retinal by the protein residues. Since the computational costs of such calculations increase rapidly with the number of atoms, it is necessary to reduce the system in question or apply approximate calculations in combination with accurate quantum mechanical (QM) techniques.

In this study, the human opsins are constructed using bovine rhodopsin as a template. The spectral shift associated with each opsin are ascertained by 1) performing quantum mechanical calculations using the Jaguar suite (10) on selected residues in the retinal binding site, and 2) running a combined QMMM (Quantum mechanics/ molecular mechanics) calculation treating the retinal molecule's ground and excited states by quantum mechanics and the protein environment by molecular mechanics. In this way, residues and environmental factors (membrane, dielectric) responsible for spectral tuning are determined.

Methods

Opsin structure building

A pairwise multiple sequence alignment of the sequences for bovine rhodopsin together with those of the four human opsins (rhodopsin and the three color opsins) was carried out using Clustalw (15). Based on this alignment, the sequence of each human opsin was threaded through the bovine rhodopsin structural template as follows. Side chain replacement using SCWRL (16) of the bovine rhodopsin structure was carried out in order to replace those in the bovine rhodopsin structure with the correct sequence of the respective opsin. SCWRL obtains optimized rotamers for the side chain using a backbone-dependent library of rotamers. As seen in Figure 4, the only gaps in the alignment occurred in the N and C terminus. These regions with gaps were excluded from the created structures, with the reasoning that they are too far from the bound retinal in order to have a significant role in spectral tuning. The protein charges were assigned from CHARM22. The structures were optimized using conjugate gradient minimization of the protein in vacuum as implemented in MPSim (20). Then the retinal protonated Schiff base was placed into the opsins in the same orientation as in the bovine rhodopsin crystal structure. SCWRL was used again and the complex minimized. SCWRL was used once more and the complex re-minimized. A side chain replacement program called SCREAM, which uses an explicit potential energy calculation to score side chain rotamers, was subsequently used to finely optimize side chains in the 5 A binding site of the retinal PSB.

Quantum mechanical calculations on retinal and derivatives

As a validation for determining the excitation energy for retinal within the opsins, some calculations are performed on free 11cis-retinal and protonated Schiff Base (PSB) 11cis-retinal. The 6-31g** basis set was used at the Hartree-Fock level using the self-consistent field (SCF) method within the Jaguar suite (10). The molecules were gemometry optimized at the ground state electronic configuration (closed shell singlet). Then the excited state with the triplet configuration was obtained with the geometry optimized structure above to simulate vertical excitation. Finally, using the initial wavefunction guesses from the triplet state, the open shell singlet configuration was obtained.

Fig 4: The alignment of the bovine rhodopsin sequence to the other 3 color opsins and human rhodopsin.

```
gi|129215|sp|P04001|OPSG_HUMAN      VAKQQKESESTQKAEKEVTRMVVVMVLAFCFCWGPYAFFACFAAANPGYP
gi|129219|sp|P04000|OPSR_HUMAN      VAKQQKESESTQKAEKEVTRMVVVMIFAYCVCWGPYTFFACFAAANPGYA
 gi|129204|sp|P02699|OPSD_BOVIN     ------------------MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLA
 gi|129207|sp|P08100|OPSD_HUMAN     ------------------MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLA
 gi|129215|sp|P04001|OPSG_HUMAN     MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNS--NSTRGPFEGPNYHIA
 gi|129219|sp|P04000|OPSR_HUMAN     MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNS--NSTRGPFEGPNYHIA
 gi|129203|sp|P03999|OPSB_HUMAN     ------------------MRKMSEEEFYLFKN---ISSVGPWDGPQYHIA
                                                       .     .::   .       .*:: *:*::*

gi|129204|sp|P02699|OPSD_BOVIN      EPWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVA
gi|129207|sp|P08100|OPSD_HUMAN      EPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVA
gi|129215|sp|P04001|OPSG_HUMAN      PRWVYHLTSVWMIFVVIASVFTNGLVLAATMKFKKLRHPLNWILVNLAVA
gi|129219|sp|P04000|OPSR_HUMAN      PRWVYHLTSVWMIFVVTASVFTNGLVLAATMKFKKLRHPLNWILVNLAVA
gi|129203|sp|P03999|OPSB_HUMAN      PVWAFYLQAAFMGTVFLIGFPLNAMVLVATLRYKKLRQPLNYILVNVSFG
                                      * : : :.:*  :.  ..  * :.* .*::.**** ***:**:*::..

gi|129204|sp|P02699|OPSD_BOVIN      DLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLGGEIALWSLVVLA
gi|129207|sp|P08100|OPSD_HUMAN      DLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLGGEIALWSLVVLA
gi|129215|sp|P04001|OPSG_HUMAN      DLAETVIASTISVVNQVYGYFVLGHPMCVLEGYTVSLCGITGLWSLAIIS
gi|129219|sp|P04000|OPSR_HUMAN      DLAETVIASTISIVNQVSGYFVLGHPMCVLEGYTVSLCGITGLWSLAIIS
gi|129203|sp|P03999|OPSB_HUMAN      GFLLCIFSVFPVFVASCNGYFVFGRHVCALEGFLGTVAGLVTGWSLAFLA
                                    .:   . .    .  .  ****:*   * ***:  :: *    ***..::

gi|129204|sp|P02699|OPSD_BOVIN      IERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIPEG
gi|129207|sp|P08100|OPSD_HUMAN      IERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLAGWSRYIPEG
gi|129215|sp|P04001|OPSG_HUMAN      WERWMVVCKPFGNVRFDAKLAIVGIAFSWIWAAVWTAPPIFGWSRYWPHG
gi|129219|sp|P04000|OPSR_HUMAN      WERWLVVCKPFGNVRFDAKLAIVGIAFSWIWSAVWTAPPIFGWSRYWPHG
gi|129203|sp|P03999|OPSB_HUMAN      FERYIVICKPFGNFRFSSKHALTVVLATWTIGIGVSIPPFFGWSRFIPEG
                                     **::*:***:.*.**. : *:  :  :*  .   : **: ****: *.*

gi|129204|sp|P02699|OPSD_BOVIN      MQCSCGIDYYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLVFTVKE
gi|129207|sp|P08100|OPSD_HUMAN      LQCSCGIDYYTLKPEVNNESFVIYMFVVHFTIPMIIIFFCYGQLVFTVKE
gi|129215|sp|P04001|OPSG_HUMAN      LKTSCGPDVFSGSSYPGVQSYMIVLMVTCCITPLSIIVLCYLQVWLAIRA
gi|129219|sp|P04000|OPSR_HUMAN      LKTSCGPDVFSGSSYPGVQSYMIVLMVTCCIIPLAIIMLCYLQVWLAIRA
gi|129203|sp|P03999|OPSB_HUMAN      LQCSCGPDWYTVGTKYRSESYTWFLFIFCFIVPLSLICFSYTQLLRALKA
                                    :: *** * ::       :*:   :::     *: :* :.* *:  :::
```

QM/MM calculation on the opsin/retinal complex (theory)

Since the computational cost of quantum mechanical calculations increases by the power of the number of atoms, it is necessary to use a combines QM/MM method for large systems. In order to be able to reproduce the complete spectral shift across the opsins, the combined QM/MM technique described by Murphy, Philipp et al. (9) was used. The system is divided into a quantum mechanical (QM) portion (retinal PSB) and a molecular mechanics (MM) portion (opsin); the respective energies are calculated with the following expressions:

$$E_{\mathrm{QM}} = \sum_{\mu\nu} P_{mn} H_{\mu\nu}^{\mathrm{core}*} + \frac{1}{2}\sum_{\mu\nu} P_{\mu\nu}[2J_{\mu\nu} - K_{\mu\nu}] + \sum_{AC} \frac{Z_a Z_c}{R_{ac}} + \sum_{AM} \frac{Z_A q_M}{R_{AM}} \quad (1)$$

From (9)

$$E_{\mathrm{MM}} = \sum_{i=\text{stretches}} k_i(r_i - r_0)^2 + \sum_{j=\text{bends}} k_j(\theta_j - \theta_0)^2 + \sum_{k=\text{torsions}} \frac{V_{k,1}}{2}(1+\cos\phi_k) + \frac{V_{k,2}}{2}(1-\cos 2\phi_k) + \frac{V_{k,3}}{2}(1+\cos 3\phi_k) + \sum_{MN} \frac{q_M q_N}{R_{MN}} + \sum_{MN} 4\varepsilon_{mn}\left[\left(\frac{\sigma_{MN}}{R_{MN}}\right)^{12} - \left(\frac{\sigma_{MN}}{R_{MN}}\right)^{6}\right] \quad (2)$$

In equation [1], the first term included the Hamiltonian corresponding to the field of MM point charges (CHARMM charges) which will influence the wavefunction and electronic energy of the QM region. The second term includes the set of Hartree-Fock equations for the QM region as if it were not interacting with the protein environment. The third term is the nuclear interactions between atoms in the QM region and the last

term is the interaction of the QM nuclei with the MM charges. The QM energy will be evaluated by Jaguar suite of programs (10).

In equation [2], the energy of the MM region is given by the typical MM expression, with stretches, bends, torsions, electrostatics, and van der Waals terms. This MM energy will be evaluated using MPSim (20).

In certain cases, the QM system is minimized in the field of the protein. To do so, an adiabatic minimization procedure was developed (9) in which the QM region is minimized (for one geometry step) across its potential energy gradient (the gradient of the above energy expressions). This is done with a frozen MM region. Then the MM region is minimized (down to a certain rms force tolerance) across its potential energy gradient (with QM region fixed). This MM region minimization is done treating the QM atoms as partial charges with a certain van der Waals radius (between the radius from QM calculation and that from the MM force field). The QM charges are obtained by fitting to the electrostatic potential described in the wavefunction. This cycle is repeated until the QM region converges (rms reaches a certain tolerance).

1.1 QM/MM calculation on the opsin/retinal complex (application)

The above QMMM method has been implemented to perform the quantum mechanical calculations for the ground state (as given in equation [2]) as well as the triplet excited state. In this case, the energy gap for exciting the electron would be just the subtraction of the ground and excited state energies for the QM regions after QMMM minimization.

This energy could be related to the frequency of light needed for the excitation, and thus the spectral shifts due to the opsin environments could be reproduced. The precise roles of the first 3 factors described section 1.0 was determined.

The last factor probably arises from the close aromatic side chains which may polarize in response to the retinal excitation (and its subsequent increase in dipole moment). Since the side chains in molecular mechanics are non-polarizable, this factor will only be tested with the pure QM calculations as described above. The procedure is as follows. The QM region is chosen to be the PSB retinal together with the aromatic polarizable residue (tyrosine or tryptophan) at position 265 in the protein. The unpaired electrons remain on the retinal chain in the excited state calculation in the field of a polarizable ground state aromatic side chain. This calculation is performed both in vacuum and in the field of an MM-treated protein environment in order to ascertain any additional influence the rest of the protein may have in the polarizability of the aromatic side chain.

1.2 Molecular dynamics using a QM-fitted force field

Proteins are of course dynamical systems at room temperature. The distributions evident in absorption spectra are due to structural fluctuations which occur in the excited molecule and affected by fluctuations of surrounding molecules. Since it is currently not feasible to perform QM dynamics on large systems, our force field parameters were fit more closely to the quantum mechanics for the PSB retinal. After applying these general parameters for torsions and bond distance, the PSB retinal structure minimized to a

structure which was ~2 kcal/mol from the QM geometry optimized structure. This validated the use of the force field for our opsin shift studies.

Subsequently, the PSB 11cis-retinal/ ospin complex was simulated with molecular dynamics in a constant temperature heat bath (TVN) at 300 K for 100 ps. Then, single point QMMM energies as described above were obtained for snapshots obtained every 2 ps in the dynamics in order to obtain a calculated absorption spectrum reflective of dynamical changes in the protein and ligand.

Results and discussion

QM on retinal and derivatives

The QM geometry optimized structure of retinal is shown in Figure 5. Calculating the closed shell singlet (CSS) and open shell singlet energy (OSS) from this geometry leads to a 69.3 kcal/mol energy gap. This corresponds to 413 nm photon wavelength which corresponds well with the 380 nm determined experimentally. The calculated closed shell singlet to triplet energy was 66.8 kcal/mol or 428 nm. This validates the use of QM at the hartree-fock level for use in these calculations. Since our method of QMMM does currently account for the OSS case, the CSS to triplet transition is studied for the purpose of obtaining relative opsin shifts in the case of retinal in the protein.

The QM geometry optimized structure of a retinal protonated Schiff base (PSB) structure is shown in Figure 6. It consists of the bound retinal and three carbons of the lysine side chain. The CSS to triplet energy was 42.9 kcal/mol (and CSS to OSS was 47.7 kcal/mol). When a QMMM calculation is performed with the PSB structure (the same conformation as above) treated with QM and the Glu113 counterion treated with MM, the CSS to triplet energy is 44.6 kcal/mol. And when the same retinal PSB is placed into the bovine rhodopsin crystal structure (matching the ligand orientation with that in the crystal structure), the CSS to triplet energy is 43.5 kcal/mol. The relative differences indicate the role of the counterions in blue-shifting the energy gap and the role of the rest of the bovine rhodopsin protein in red-shifting from there. Later, the dynamics of the ligand in protein will show how changes in the ligand conformation in response to the protein also modulate this energy gap.

QMMM on opsin complexes

QMMM one point calculations were carried out with the retinal PSB in the same conformation as above. The calculated energy gaps were 43.48, 44.90, 42.86 kcal/mol for red, green, and blue opsins, respectively. The energy gaps shift are 1.42 for red to green and –2.04 kcal/mol for green to blue. The experimental values for these opsin shifts are 2.9 for red to green and 13.3 kcal/mol for green to blue.

The residues close to the retinal PSB which are present in the red opsin and not in the green opsin are shown in Figure 7. These residues are non-polar in the green opsin. In particular, the conformation of the Tyr261 (which is F in the green opsin) was thought to be important in the opsin shift. When the OH dipole of the tyrosine was rotated 180, the calculated energy gap was 43.50, virtually equivalent to the other conformer.

As can be noted, there are no differences in the region close to the retinal PSB which may cause twisting. Thus, the assumption that the ligand is in the same conformation in both opsins may be valid. Nevertheless, the role of dynamics on this small energy shift was also observed and will be reported later.

The presence of dipolar residues near the Schiff base side of the retinal PSB was thought to play the predominant role in the green to blue shift (2). But that is not the case in these calculations. In fact, an opposite shift of ~2 kcal/mol is predicted. Thus the polarizable side chains or the twisting of the retinal must play a much more predominant role in this particular shift.

Role of polarizable side chains on the opsin shift

As mentioned before, polarizable side chains have been implicated in modulating the absorption frequency of retinal. Within the binding site of the retinal PSB within the opsins, there is a TRP265 in the green and red opsin, whereas in the blue opsin, this residue is a TYR265 (Figure 8). So possibly the different effects of these polarizable residues may be responsible for the 13 kcal/mol shift in excitation energy between the green and blue opsins.

First, SCREAM was used to obtain rotamers of these residues which were at least 25 kcal/mol different in energy score. This Monte Carlo side chain sampling would be equivalent to a long-time dynamics run, since there are barriers which would need to be overcome between these conformations. Then, a QM calculation was performed on the retinal PSB and the residue at 265 only. This gives the polarizability contribution independent of the rest of the protein. The energy gaps are shown in Table 1. The energy gaps are virtually equivalent. But how does this effect change in the protein?

Table 1: The energy gaps of bound retinal with a QM treatment of residue 265.

Green opsin Conformation number	Energy gap (kcal/mol)	Blue opsin Conformation number	Energy gap (kcal/mol)
1	42.68	1	43.14
2	42.47	2	43.24
3	43.51	3	43.12
4	44.08	4	42.73
5	42.95	5	43.01
6	42.76	6	42.87
		7	42.63
		8	42.45
		9	42.77
		10	43.04
average	43.08		42.90

The aromatic side chain and retinal PSB were treated as QM in a QMMM calculation. The results are shown in Table2. An opposite shift of ~3 kcal/mol is predicted. It is thus obvious that twisting of the retinal PSB must cause this shift and this was tested.

Table 2: The energy gaps of bound retinal with a QM treatment of residue 265 and MM treatment of the rest of the protein.

Green opsin Conformation number	Energy gap (kcal/mol)	Blue opsin Conformation number	Energy gap (kcal/mol)
1	44.67	1	42.15
2	44.41	2	42.24
3	45.21	3	42.01
4	45.52	4	41.71
5	44.89	5	42.08
6	44.62	6	41.95
		7	41.76
		8	41.62
		9	41.78
		10	42.00
average	44.89		41.93

QM-fitted molecular dynamics on opsin complexes

It is currently not feasible to perform ab initio QM dynamics within protein. But the dreiding force field parameters of torsion and bond stretching could be changed to fit the QM values, as discussed in Sec. 2.5. Subsequently, molecular dynamics was run for 100 ps. A one point QMMM calculation was performed on snapshots every 2 ps.The calculated CSS to triplet energy gaps from these 50 structures were used to create the histograms for all 3 opsins as shown in Figure 9.

The peaks of these histograms correspond to energy gaps of 41.00, 42.72, and 58.73 kcal/mol for the red, green, and blue opsins respectively. The calculated red to green shift is 1.72 which corresponds with the 2.9 kcal/mol known from experiment. The calculated green to blue shift is 16.01, which correspond with the 13.3 kcal/mol known from experiment. This shift is in fact overpredicted by 3.3 kcal/mol. This is where the polarizable side chains may play a role - the TRP265 in the green opsin was found to blue shift the energy gap by an average of 2.96 kcal/mol (Table 2) as compared with the TYR265 in the blue opsin. Thus 16.01 – 2.96 = 13.05 corresponds very well with the green to blue opsin shift of 13.3 kcal/mol known from experiment.

Figure 5: QM geometry optimized structure of retinal.

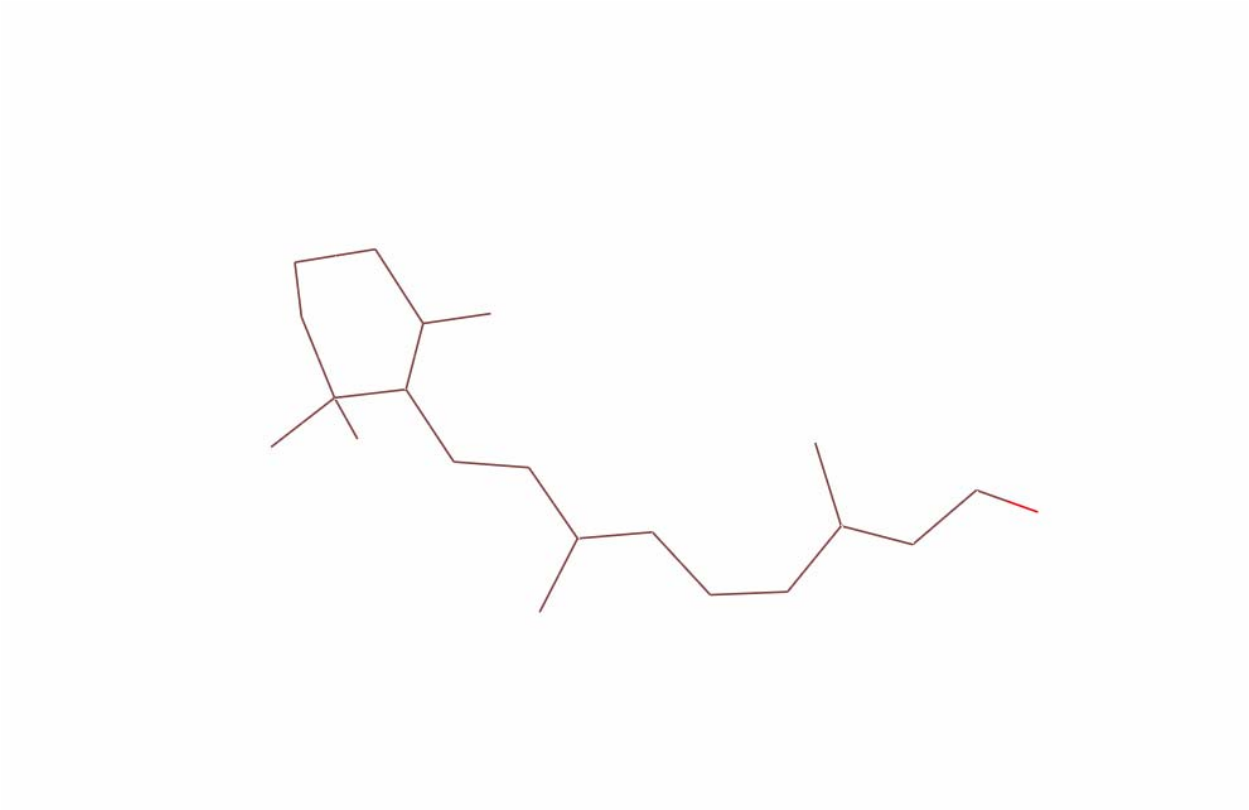


Figure 6: QM geometry optimized structure of a retinal protonated Schiff base (PSB).

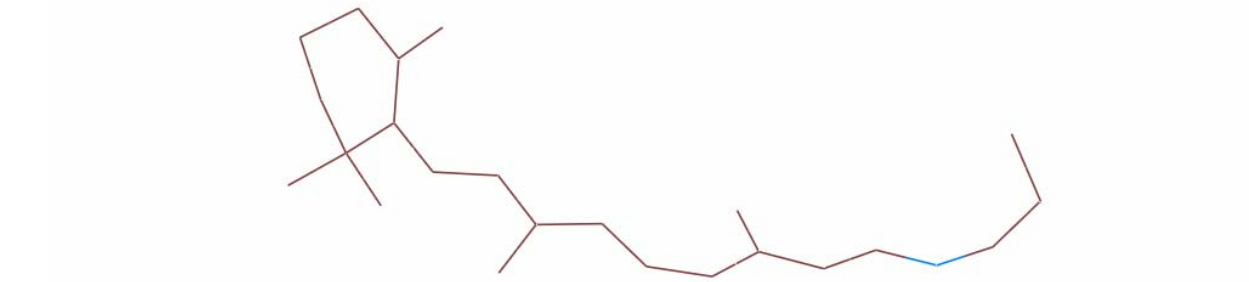


Figure 7: Residues close to the retinal PSB which are present in the red opsin and not in the green opsin.

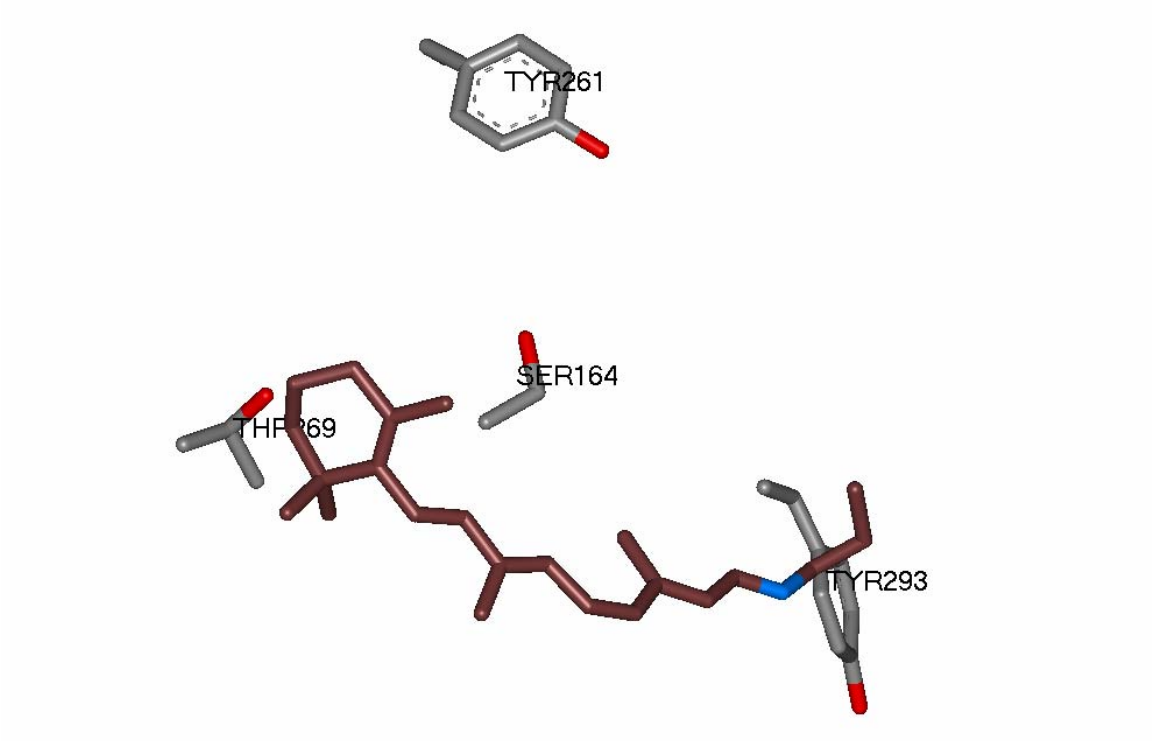
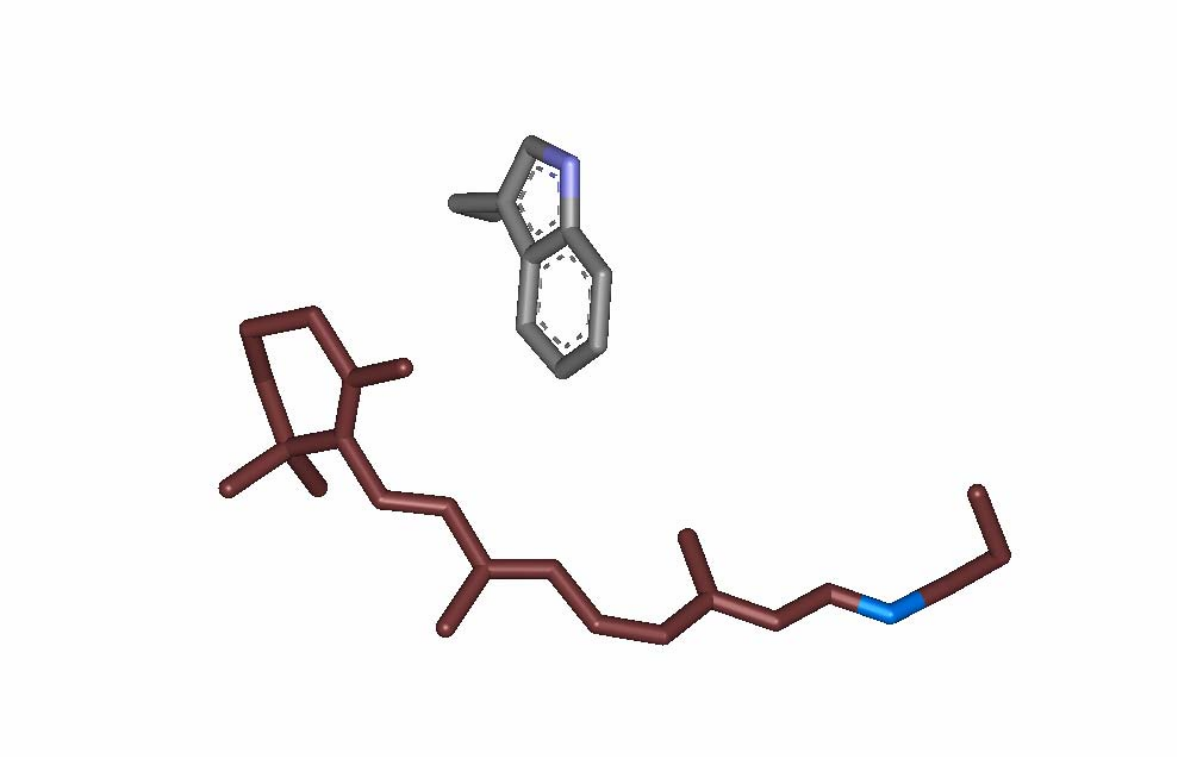


Figure 8: The residue 265 is Trp in the green opsin but Tyr in the blue opsin.

a)



b)

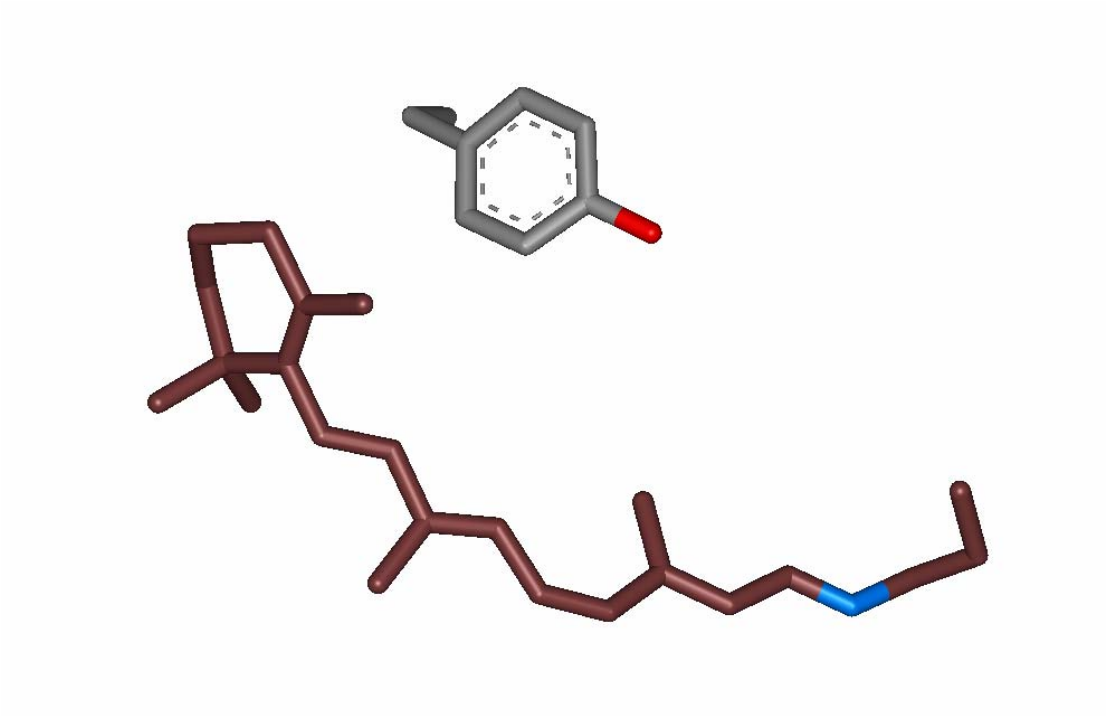
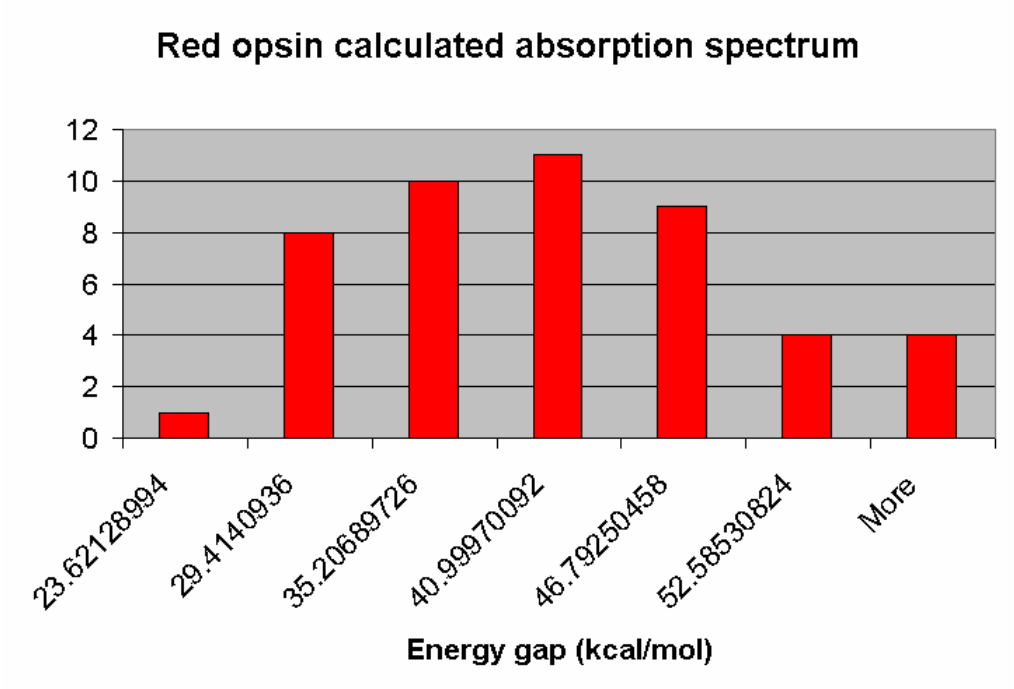
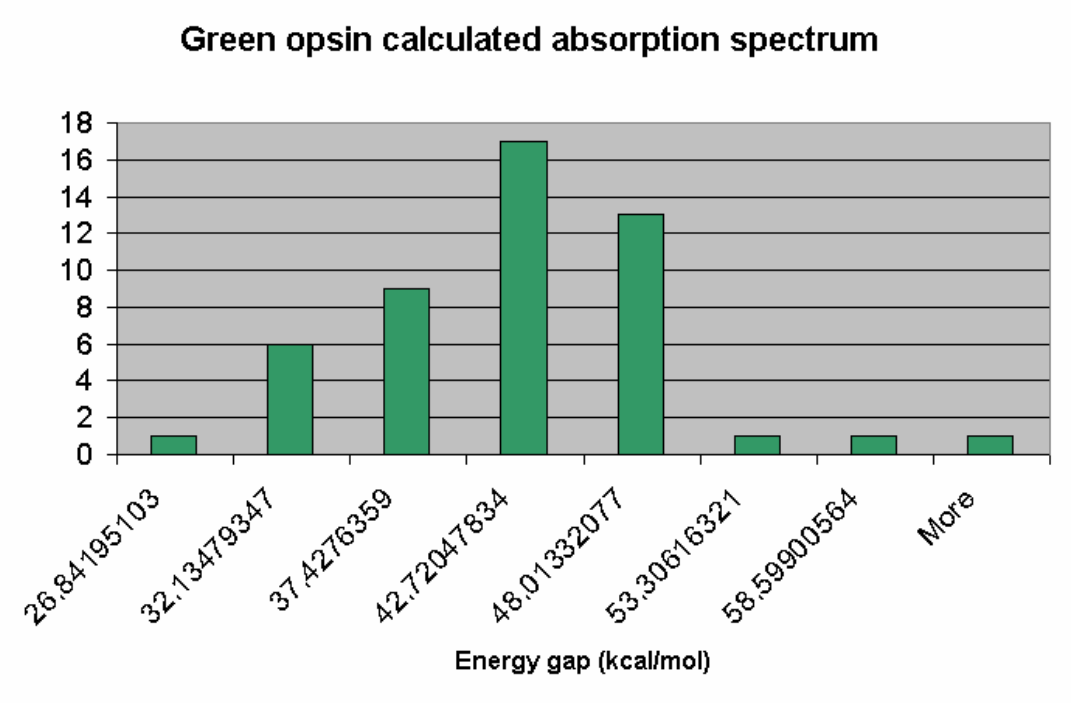


Figure 9: Calculated absorption spectra for the 3 opsins.

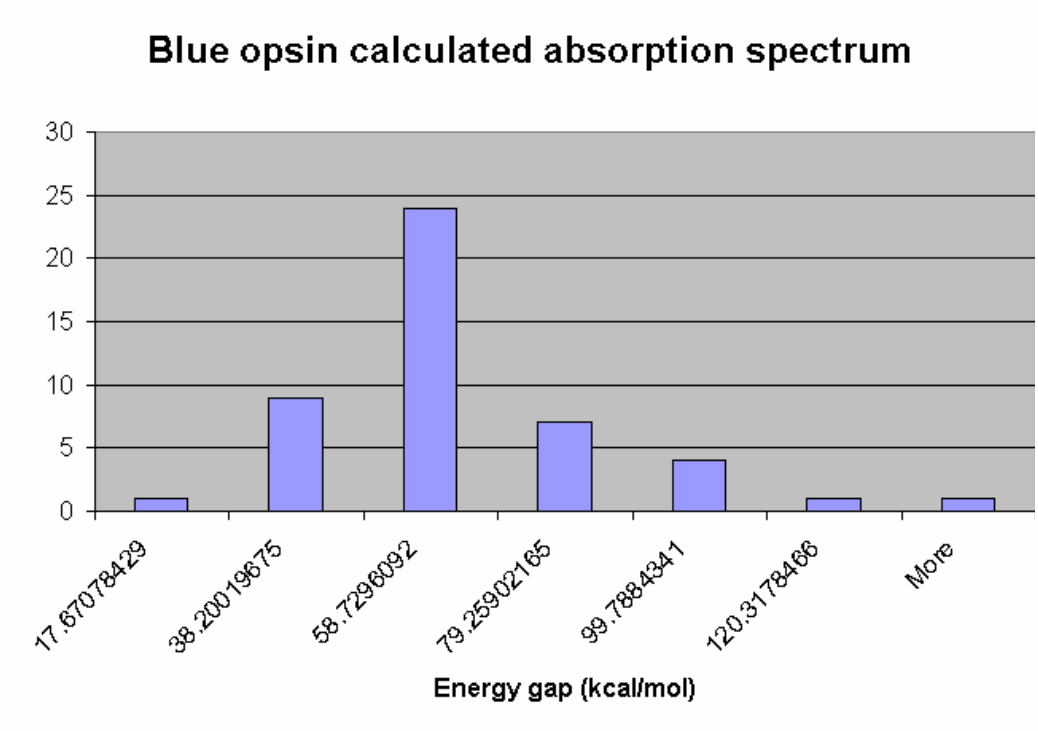
a)



b)



c)


Blue opsin calculated absorption spectrum
30
25
20
15
10
5
0
17.67078429
38.20019675
58.7296092
79.25902165
99.7884341
120.3178466
More
Energy gap (kcal/mol)

Conclusion

The fundamental question of how opsins modulate the frequency of light absorbed by a bound chromophore has been studied for decades. This study used the most updated techniques of structure building and optimization to build the human opsin structures. In addition, the roles of these opsin structures in modulating the frequency of maximal absorption of light by bound retinal PSB were determined using QM and hybrid QMMM techniques.

This fundamental understanding of this mechanism may lead to design application in which proteins (or a synthetic analog of a peptide) may be designed to modulate the frequency of light absorbed by any bound chromophore. This would serve well in the field medical imaging diagnostic tests. In addition, the chromophore may be modified to interact specifically with a native protein and thus absorb at a certain frequency of light.

References

(Style for PNAS submission)

1) Lin, S. W.; Kochendoerfer, G. G.; Carroll, K. S.; Wang, D.; Mathies, R. A.; Sakmar, T. P. *J. Biol. Chem.* **1998**, *273*, 24583

2) Kochendoerfer, G. G.; Lin, S. W.; Sakmar, T. P.; Mathies, R. A. *Trends Biochem. Sci.* **1999**, *24*, 300

3) van der Steen, R.; Biesheuvel, P. L.; Mathies, R. A.; Lugtenburg, J. *J. Am. Chem. Soc.* **1986**, *108*, 6410

4) Sakmar, T. P.; Franke, R. R.; Khorana, H. G. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 3079

5) Yan, B.; Spudich, J. L.; Mazur, P.; Vunnam, S.; Derguini, F.; Nakanishi, K. *J. Biol. Chem.* **1995**, *270*, 29668

6) Houjou, H.; Inoue, Y.; Sakurai, M. *J. Am. Chem. Soc.* **1998**, *120*, 4459

7) Mathies, R.; Stryer, L. *Proc. Natl. Acad. Sci. U.S.A.* **1976**, *73*, 2169

8) Torii, H. ; *J. Am. Chem. Soc.* **2002,** 124, 9272

9) Murphy, R. B.; Philipp, D. M., Friesner, R. A. *J. Comput. Chem.* **2000**, 21, 1442

10) Jaguar v4.0, Schrodinger Inc.

11) Ueyama, H., Kuwayama, S., Imai, H., Tanabe, S., Oda, S., Nishida, Y., Wada, A., Shichida, Y., Yamade, S. *Biochem. Biophys. Res. Commun.* **2002,** 294, 205

12) Stenkamp, R. E., Filipek, S., Driessen, C.A.G.G., Teller, D.C., Palczewski, K. *Biochimica et Biophysica Acta.* **2002,** 1565, 168

13) Sakmar, T.P., Franke, R.R., Khorana, H.G. *Proc. Natl. Acad. Sci.USA.* **1991,** 88, 3079

14) Lugtenburg, J.; Muradin-Szweykowska, M.; Harbison,G.S.; Smith, S.O.; Heeremans, C.; Pardoen, J.A.; Herzfeld, J.; Griffin, R.G.; Mathies, R.A. *J. Am. Chem. Soc.* **1986**, 108,3104

15) Higgins, D., Thompson, J., Gibson, T., Thompson, J.D., Higgins, D.G., Gibson, T.J. (1994), *Nucleic Acids Res*. **22**, 4673-4680.

16) Bower, M., Cohen, F.E., and Dunbrack, Jr. R.L., (1997), *J. Mol. Biol*. **267**, 1268-1282.

17) Mayo, S. L., Olafson, B.D. & Goddard III, W.A. (1990). *J. Phys. Chem.* **94**, 8897-8909.

18) MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998). *J. Phys. Chem. B* **102**, 3586-3616.

19) Rappé, A.K. & Goddard III, W.A. (1991). *J. Phys. Chem.* **95**, 3358-3363.

20) Lim, K-T, Brunett, S., Iotov, M., McClurg, R.B., Vaidehi, N., Dasgupta, S., Taylor, S. & Goddard III, W.A. (1997). *J. Comput. Chem.* **18**, 501-521.

21) Nussbaum,R.L., McInnes, R.R., Willard, H.F. (6^{th} ed.) 2001, pg 130

22) http://wunmr.wustl.edu/EduDev/LabTutorials/Vision/Vision.html

23) Simeonov, A., Matcushita, M., Juban, E.A., Thompson, E.H.Z., Hoffman, T.Z., Beuscher, A.E., Taylor, M.J., Wirsching, P., Rettig, W., McCusker, J.K., Stevens, R.C., Millar, D.P., Schultz, P.G., Lerner, R.A., Janda, K.D. (2000). Science. 290, 307-313.

Chapter 8: Ab initio simulation of photoisomerization for full retinal chromophore in free and rhodopsin-bound states

Abstract

The photoisomerization event of retinal is the crucial step in the process of vision in which the energy of a photon of light is converted to a signal transduction cascade. It is the interaction of retinal with protein within rhodopsin, a G-protein-coupled receptors, which increases the efficiency of this process greatly. This study reports the ab initio potential surfaces of the ground and excited states for the full retinal chromophore in the free state as well as in the opsin-bound state and demonstrates the conical intersection (CI) in these molecules. In addition, the observed difference in the extent and form of this CI in the protein-bound state may be crucial in increasing the quantum yield of isomerization.

Introduction

The primary event in vision is the conversion of electromagnetic energy into mechanical energy of protein movement. This occurs when the chromophore 11cis-retinal absorbs a photon of light and subsequently isomerizes to yield all-trans-retinal (Figure 1). This isomerization process occurs with an experimental quantum yield of 0.2 (Kropf et al., 1970; Bensasson et al., 1978). The protein rhodopsin (of the G-protein-coupled receptor superfamily), however, increases the efficiency of this process to .65 (Schoenlein et al., 1991) which completes in 200 fs (Boucher et al., 1985).

Quantum mechanically, the isomerization may occur as a two-state or three-state process. Schematics of the two processes are shown in Figure 2. In the two-state process, the molecule is excited to the FC (Frank-Codon) point on the S1 (first excited state) surface after which there is a descent to the CI (conical intersection) where there is intersystem crossing to the S0 (ground state) surface. In the three-state process, the difference is that there is a small barrier on the S1 surface corresponding to an avoided crossing with the S2 surface. The support (experimental and theoretical) for a three-state process has focused on the retinal/bacteriorhodopsin system (Hasson et al., 1996; Humphrey et al., 1998; Kobayashi et al., 2001), in which the protein catalyzes the formation of 13-cis retinal from all-trans-retinal. This bacterial process is a reverse isomerization (trans->cis) in which bond selectivity for the rotation is also a significant issue. The two-state model remains possible for the retinal/rhodopsin system and in fact

there is support for such a path for the 11cis->all-trans isomerization of free retinal (Gonzalez-Luque et al., 2000).

Conical intersections have been found in many photochemical singlet reactions (Bernardi et al., 1996) and have been used to explain the observed phenomenon of weak fluorescence for various photochemical processes in ethylene, butadiene, and stilbene (Zerbetto et al., 1990; Olivucci et al., 1993; Quenneville et al., 2003). In the case of free retinal, rapid isomerization with weak fluorescence has been observed experimentally (Hochstrasser et al., 1976; Doukas et al., 1983; Alex et al., 1992).

The relaxation process from FC has been studied using quantum dynamics techniques on retinal models (smaller molecules which incorporate crucial elements of the retinal structure) in the free state (Vreven et al., 1997; Bub et al., 2000) as well as more recently in bacteriorhodopsin (Warshel et al., 2001; Hayashi et al., 2003) using hybrid quantum mechanics/molecular mechanics (QMMM) methodology. In addition, the isomerization process and associated conical intersection have been explained by graphing the potential surfaces for smaller free retinal models along the minimum energy path (MEP) coordinates (Gonzalez-Luque et al., 2000; Migani et al., 2003). In fact, it was found that there is an extended intersection space (IS) and not a single CI which is responsible for the intersystem crossing (Migani et al., 2003).

Mapping the isomerization along actual geometric coordinates has been applied to ethylene and stilbene (Quenneville et al., 2003) using as the nuclear coordinates the torsion of the isomerizing bond and the pyramidalization of associated atoms of this bond. This study demonstrated the presence of CI in both molecules. Because of computational costs, such a study has not yet been applied to the complete retinal

molecule. There has been published work on mapping the isomerization using only the coordinate of torsion (Lee et al., 2002), but this did not demonstrate the CI of retinal.

This paper reports the ab initio calculation of So and S1 potential surfaces with respect to the C11-C12 torsion angle of 11cis-retinal and demonstrates the conical intersections in stilbene, retinal, retinal protonated Schiff base (RPSB), and RPSB within rhodopsin using QM or QMMM methods. Insights into the role of the potein in catalyzing the isomerization process emerge from this study.

Material and methods

Small molecule molecular mechanics optimizations

The potential energies of all small molecules were initially minimized using conjugate gradients. Stilbene and retinal (Figure 3ab) were rotated about their isomerizing bond to torsion angles from 0 to 180 degrees in increments of 10 degrees. The larger RPSB (Figure 3c) form of retinal was rotated from 0 to 180 degrees in increments of 30 degrees. Keeping the atoms of the torsion fixed, the potential energies of the different torsional forms of the molecules were minimized with the FF description of DREIDING (Mayo et al., 1990) using the MPSim (Lim et al., 1997) program. This was carried out to an RMS force of 0.05 kcal/mol/Å using conjugate gradients.

Protein MM optimization

The crystal structure of bovine rhodopsin (resolution 2.80 Å) was downloaded from the protein database (pdb entry 1F88). The Hg ions, sugars, and waters were deleted from this structure. This crystal structure is missing 10 complete residues in loop regions and the side chain atoms for 15 additional residues. We added the missing residues and side chains using WhatIf (Vriend, 1990). Then we added hydrogens to all the residues using the PolyGraf software. We then fixed the TM helices and minimized (using conjugate gradients) the structure of the loop region to an RMS force of 0.1 kcal/mol/Å. The potential energy of the entire structure of rhodopsin was then minimized (using conjugate gradients) to an RMS force of 0.1 kcal/mol/Å . This minimized structure deviates from the x-ray crystal structure by 0.29 Å coordinate root mean square (CRMS) error over all atoms in the crystal structure.

Quantum mechanical calculation

Subsequently, the different torsional forms of the molecules were geometry optimized in vacuum with the closed shell singlet (CSS) electronic configuration using Jaguar (Jaguar, v4.0) at the Hartree-Fock level with the basis set 6-31g**. Only the C11-C12 torsions were kept fixed during this geometry optimization (C11-C12 bond length was allowed to change). Using this same geometry for each torsional form of the molecules, a single point QM calculation of the triplet configuration was performed and the resulting orbital guess was used to perform another single point calculation of the open shell singlet (OSS). In addition, the molecules were geometry optimized in the triplet or the OSS electronic configurations.

QMMM calculation

The QM region for this calculation was the entire RPSB which consisted of the retinal and lysine side chain capped by a hydrogen before the protein backbone (Figure 4). The MM region was the rest of the protein with counterions (Na and Cl) for the charged side chains (except for Glu113 which remained salt-bridged to the Schiff base proton of the RPSB). At each torsional form of the RPSB, the RPSB/protein complex was minimized adiabatically (Appendix) with a QM CSS description of the RPSB. Using this geometry, a single point calculation of the triplet configuration was performed. Subsequently, a geometry optimization was performed in the triplet configuration using the same type of QMMM adiabatic minimization. The OSS configuration was not obtained since it has not been implemented in the QMMM program. The assumption (as supported by the energy profiles of the small molecules) is that the triplet and OSS are sufficiently close in energy to study the general excited state trends.

Results and discussion

The potential energy plots for the geometry optimized closed shell singlet (GO CSS), triplet, open shell singlet (OSS), geometry optimized open shell singlet (GO OSS), and geometry optimized triplet (GO triplet) from the torsional forms of the different molecules are shown in Figure 5a-6c.

Stilbene QM calculations and symmetry issue

Stilbene is known to have two characteristic lower excited states S1 and S2 characterized by Bu and Ag symmetry respectively (Allen et al., 1989). As such, there are two nearby peaks in the experimental absorption spectrum at 290 and 230 nm (corresponding to energies of 98.59 and 124.31 kcal/mol respectively). The calculated potential energy graph for stilbene is shown in Figure 5a. For structurally symmetric molecules like stilbene, the symmetric Ag is solved for most of the graph's excited state surface. At 90, 0, and 10 degree torsions, however, the lower Bu is solved. Nevertheless, the energy gap (from GO CSS to OSS) at 180 degree torsion is 126.98 kcal/mol, corresponding very well with the peak at 230 nm (the S2 Ag state). It is expected that the Bu state would be clearly solved for an asymmetric molecule. To test this, a methyl was added to stilbene. The energy profile for this molecule is shown in Figure 5b. The S1 Bu state is now solved. It is thus expected that calculations with asymmetric molecules like retinal would yields the asymmetric Bu state.

The intersections at 90 degrees torsion may correspond to the CI of the S1 state with the S0 state. It is conceivable that the molecule in the FC region may relax in terms of its nuclear geometry to reach the CI region, where it may undergo intersystem crossing without radiation. This proposed path is indicated in dashed lines on the energy profiles of this study. This is consistent with the experimental observation of low fluorescence quantum yield (0.05) for stilbene (Allen et al., 1989).

Retinal QM calculation

Retinal is known to have S1 and S2 states corresponding to the Bu and Ag symmetries (Gonzalez-Luque et al., 2000). The experimental absorption peak (Wolken, 1966; Dowling, 1987) of 11cis-retinal in ethanol is at 380 nm (corresponding to an energy gap of 75.24 kcal/mol). The potential energy graph for retinal is shown in Figure 6a. As expected, for this structurally asymmetric molecule, the asymmetric Bu is solved for all of the graph's excited state surface of S1. The energy gap (from GO CSS to OSS) at 0 degree torsion is 69.29 kcal/mol, corresponding with the absorption peak at 380 nm (S1 Bu state). In addition, a near intersection (~1 kcal/mol gap) of the S0 and S1 surfaces occurs at 90 degrees, indicating a possible conical intersection. This is consistent with the experimental observation of weak fluorescence.

Retinal PSB calculation

The potential energy graph for RPSB is shown in Figure 6b. The calculated energy gap (from GO CSS to OSS) at 0 degree torsion is 47.43 kcal/mol (S1 Bu state). The graph reveals qualitative trends in the energy profiles of the relaxation process of the molecule. Interestingly, the profile indicates the lack of the CI point at 90 which was observed for the retinal and stilbene cases. How is this different in the protein, where this Schiff base linkage actually occurs?

RPSB/opsin QMMM calculation

The potential energy graph for RPSB in bovine rhodopsin is shown in Figure 6c. The energy gaps are generally large, which is due to the fact that the geometries were minimized using QM and MM forces alternatively as described in the Appendix. This type of adiabatic minimization with both kinds of forces may lead to steeper gradients than with separate forces. Nevertheless, in the case of RPSB, insight emerges from a qualitative analysis of the graph. The most striking feature is the extensive region of overlap between the GO triplet and CSS, which indicates a larger region of intersection space within the protein as compared with free retinal. This would explain the large increase in isomerization quantum yield from retinal in the free state to retinal in the protein-bound state.

The overlap between the S1 and S0 surfaces begins at around 60 degrees and ends at around 120 degrees. In addition, the CSS energy profiles has a “plateau” shape instead of a “hill” shape as in free-retinal, which may lead to less excited retinal reverting to 0 degrees upon intersecting the CSS surface at ~60 degrees. These features of the potential surfaces may explain the more than 3-fold increase in isomerization quantum yield of the retinal within the protein.

Conclusion

The interesting phenomenon of conical intersections is thought to be the best explanation for the low fluorescence yields of many small molecules. This study observes these conical intersections using ab initio potential energy mapping with the torsion of the isomerizing bond as the nuclear coordinate. Thus torsion scans were obtained for stilbene as well as the full retinal molecule, as well as the protonated Schiff base of retinal in free and opsin-bound state. The demonstration of an extensive intersection space of the RPSB/opsin complex points to the protein's role in increasing the surface crossing probability through the CI.

This phenomenon may be used in applications of increasing isomer yields in photochemical reactions. Mutant proteins with capabilities of increasing the photoisomerization quantum yield or speeding the process by steepening the gradient for isomerization may prove to be applicable for various such processes. In addition, the fluorescence properties of chemicals can be modified by either modifying the chromophore or by exposing it to certain proteins. This may be very useful in immunohistochemistry and medical diagnostics. The theoretical understanding of this isomerization process may thus have various application beyond the fact that it is of fundamental scientific interest.

Appendix

The QMMM adiabatic minimization procedure was developed (Murphy et al., 2000) in which the QM region is minimized (for one geometry step) along its potential energy gradient (the gradient of the QM energy expressions). This is done with a frozen MM region. Then the MM region is minimized (down to a certain rms force tolerance) across its potential energy gradient (with QM region fixed). This MM region minimization is done treating the QM atoms as partial charges with a certain van der Waals radius (between the radius from QM calculation and that from the MM force field). The QM charges are obtained by fitting to the electrostatic potential described in the wavefunction. This cycle is repeated until the QM region converges (rms reaches a certain tolerance).

References

Jaguar v4.0, Schrodinger Inc. Portland, Oregon.

Alex, S.; Le Thanh, H. and Vocelle, D. (1992). Studies of the effect of hydrogen bonding on the absorption and fluorescence spectra of all-trans-retinal at room temperature. *Can. J. Chem.* 70, 880-887.

Allen, M. T. and Whitten, D. G. (1989). The Photophysics and Photochemistry of a,w-Diphenylpolyene Singlet States. *Chem. Rev.* 89, 1691-1702.

Bensasson, R. and Land, E. J. (1978). *Nouv. J. Chim.* 2, 503-507.

Bernardi, F.; Olivucci, M. and Robb, M. A. (1996). *Chem. Soc. Rev.* 25, 321-328.

Boucher, F. and Leblanc, R. (1985). *Photochem. Photobiol.* 41, 459.

Bub, V.; Weingart, O. and Sugihara, M. (2000). Fast Photoisomerization of a Rhodopsin Model-- An Ab Initio Molecular Dynamics Study. *Angew. Chem. Int. Ed.* 39, 2784-2786.

Doukas, A. G.; Junnarkar, M. R.; Chandra, D.; Alfano, R. R. and Callender, R. H. (1983). *Chem. Phys. Lett.* 100, 420-424.

Dowling, J. E. (1987). The Retina. An Approachable Part of the Brain. Cambridge, Massachusetts, The Belknap University Press.

Gonzalez-Luque, R.; Garavelli, M.; Bernardi, F. O., M.; Robb, M.A.; Merchan, M.; Robb, M. A. and Olivucci, M. (2000). Computational evidence in favor of a two-state, two-mode model of the retinal chromophore photoisomerization. *Proc. Natl. Acad. Sci. USA.* 97, 9379-9384.

Hasson, K. C.; Gai, F. and Anfinrud, P. A. (1996). The photoisomerization of retinal in bacteriorhodopsin: Experimental evidence for a three-state model. *Proc. Natl. Acad. Sci. USA.* 93, 15124-15129.

Hayashi, A.; Tajkhorshid, E. and Schulten, K. (2003). Molecular Dynamics Simulation of Bacteriorhodopsin's Photoisomerization Using Ab Initio Forces for the Excited Chromophore. *Biophys. J.* 85, 1440-1449.

Hochstrasser, R. M.; Narva, D. L. and Nelson, A. C. (1976). *Chem. Phys. Lett.* 43, 15.

Humphrey, W.; Lu, H.; Logunov, I.; Werner, H.-J. and Schulten, K. (1998). Three Electronic State Model of the Primary Phototransformation of Bacteriorhodopsin. *Biophys. J.* 75, 1689-1699.

Kobayashi, T.; Saito, T. and Ohtani, H. (2001). *Real-time spectroscopy of transition states in bacteriorhodopsin during retinal isomerization* 414, 531-534.

Kropf, A. and Hubbard, R. (1970). *Photochem. Photobiol.* 12, 249.

Lee, H. M.; Kim, J.; Kim, C.-J. and Kim, K. S. (2002). Ab initio study of the isomerization of retinal chromophore and its derivatives. *J. Chem. Phys.* 116, 6549-6559.

Lim, K.-T.; Brunett, S.; Iotov, M.; McClurg, R. B.; Vaidehi, N.; Dasgupta, S.; Taylor, S. and Goddard III, W. A. (1997). Moelcular Dynamics for very large systems on massively parallel computers: The MPSim program. *J. Comput. Chem.* 18, 501-521.

Mayo, S. L.; Olafson, B. D. and Goddard III, W. A. (1990). Dreiding - A generic force-field for molecular simulations. *J. Phys. Chem* 94, 8897-8909.

Migani, A.; Robb, M. A. and Olivucci, M. (2003). Relationship between Photoisomerization Path and Intersection Space in a Retinal Chromophore Model. *J. Am. Chem. Soc.* 125, 2804-2808.

Murphy, R. B.; Philipp, D. M. and Friesner, R. A. (2000). *J. Comput. Chem.* 21, 1442.

Olivucci, M.; Ragazos, I. N.; Bernardi, F. and Robb, M. A. (1993). A Conical Intersection Mechanism for the Photochemistry of Butadiene. A MC-SCF Study. *J. Am. Chem. Soc.* 115, 3710-3721.

Quenneville, J. and Martinez, T. J. (2003). Ab Initio Study of Cis-Trans Photoisomerization in Stilbene and Ethylene. *J. Phys. Chem. A* 107, 829.

Schoenlein, R. W.; Petenau, L. A.; Matthies, R. A. and Shank, C. V. (1991). *Science* 254, 412.

Vreven, T.; Bernardi, F.; Garavelli, M.; Olivucci, M.; Robb, M. A. and Schlegel, H. B. (1997). Ab initio Photoisomerization Dynamics of a Simple Retinal Chromophore Model. *J. Am. Chem. Soc.* 119, 12687-12688.

Vriend, G. (1990). *J. Mol. Graph.* 8, 52-56.

Warshel, A. and Chu, Z. T. (2001). Nature of the Surface Crossing Process in Bacteriorhodopsin: Computer Simulations of the Quantum Dynamics of the Primary Photochemical Event. *J. Phys. Chem. B.* 105, 9857-9871.

Wolken, J. J. (1966). Vision. Biophysics and Biochemistry of the Retinal Photoreceptors. Springfield, Illinois, Thomas.

Zerbetto, F. and Zgierski, M. Z. (1990). The missing fluorescence of s-trans butadiene. *J. Am. Chem. Soc.* 115, 1235.

Figure 1: Photoisomerization of the free retinal molecule and retinal ligand Schiff base bound to the protein via lysine side chain in opsins.

a)

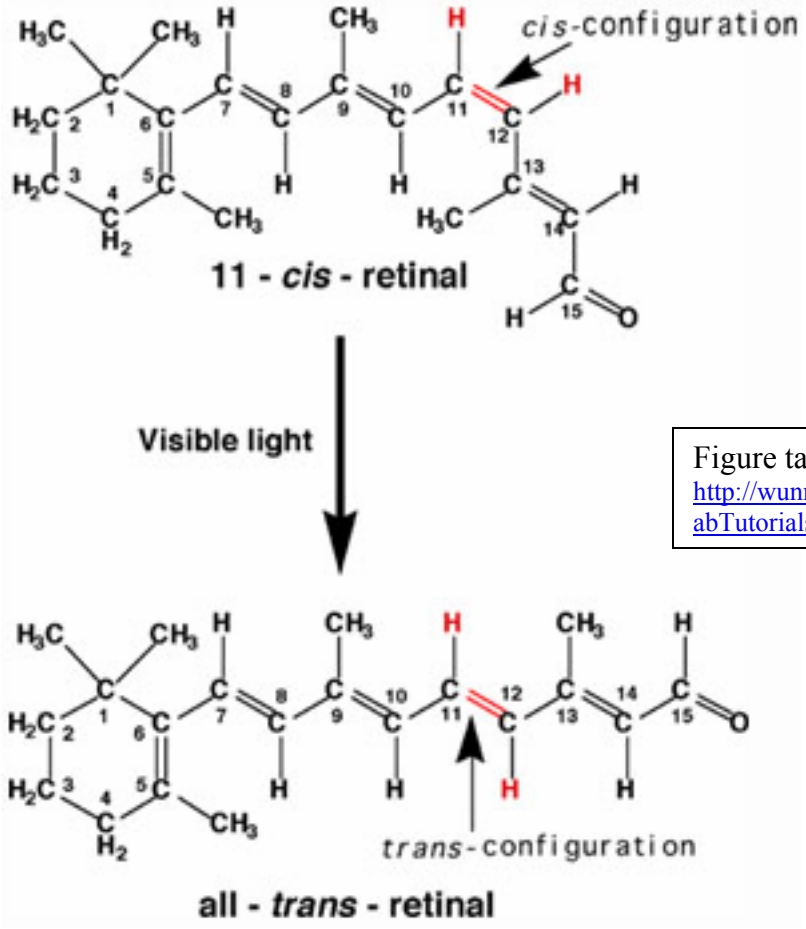


Figure taken from http://wunmr.wustl.edu/EduDev/LabTutorials/Vision/Vision.html

b)

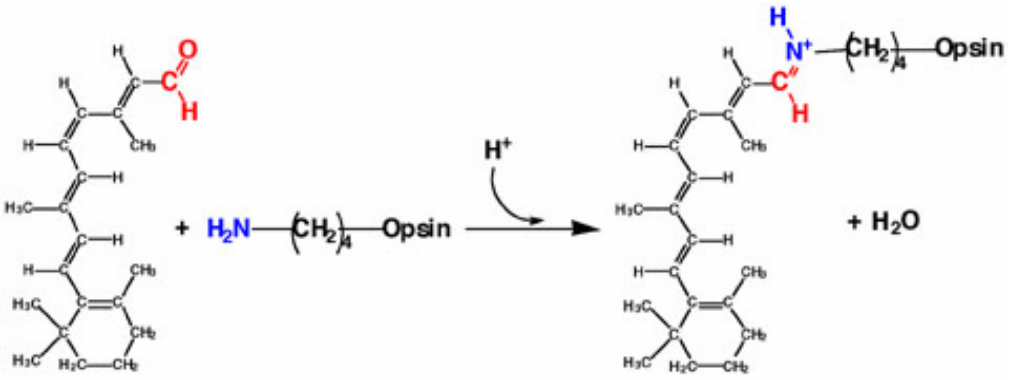


Figure taken from
http://wunmr.wustl.edu/EduDev/LabTutorials/Vision/Vision.html

Figure 2: Schemes of two and three state isomerization models.

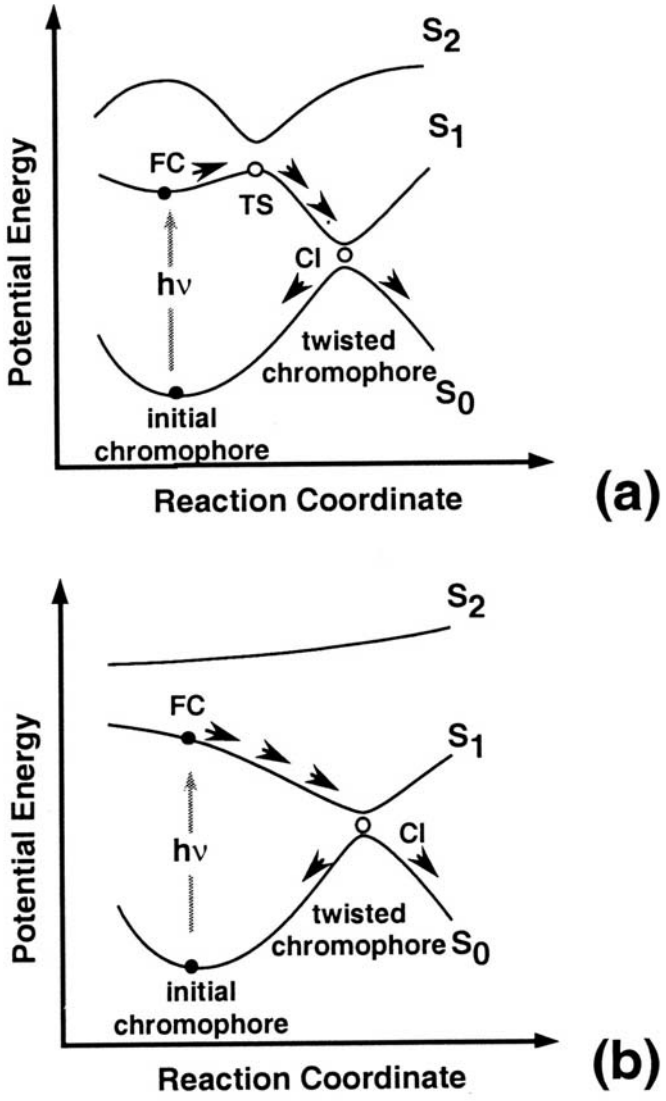
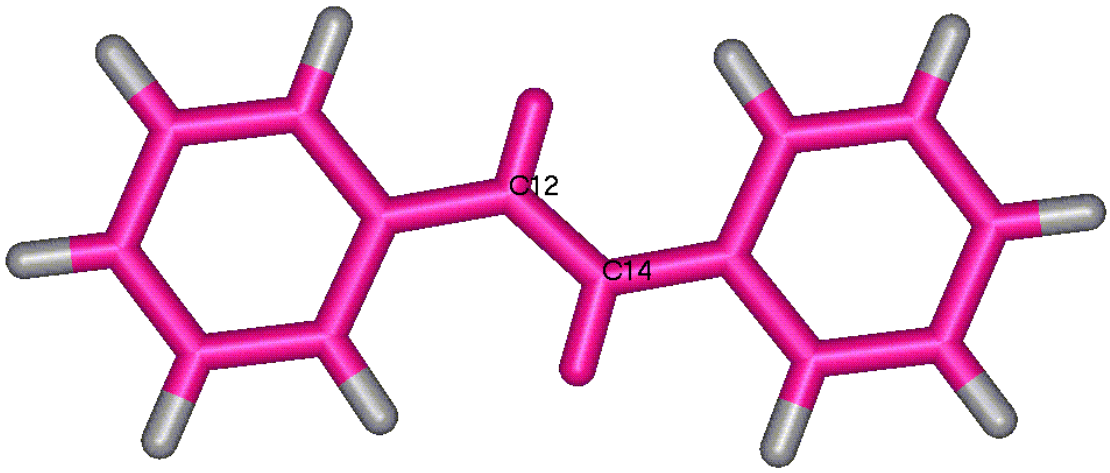


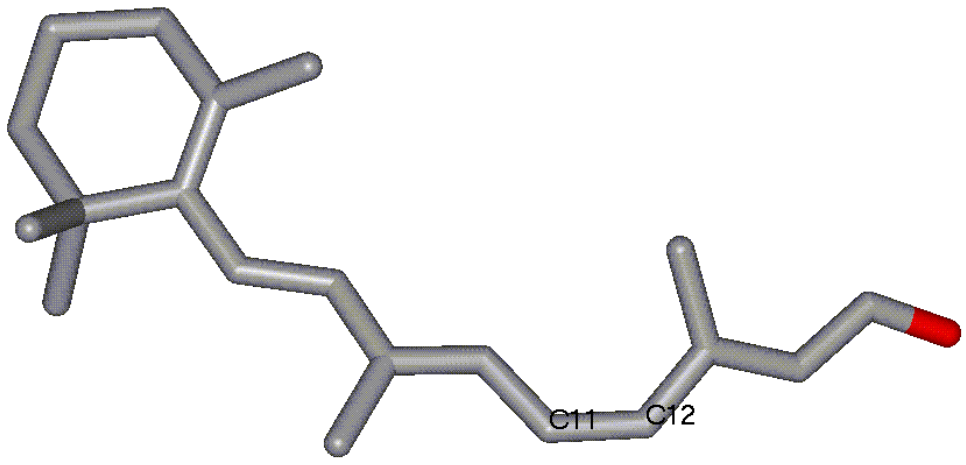
Figure from (Gonzalez-Luque et al., 2000)

Figure 3: Stilbene, retinal, and RPSB (retinal protonated Schiff base).

a)



b)



c)

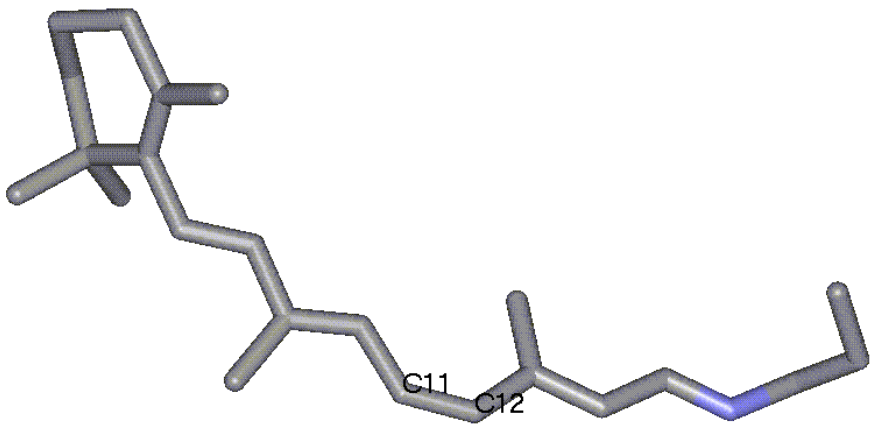


Figure 4: RPSB in protein.

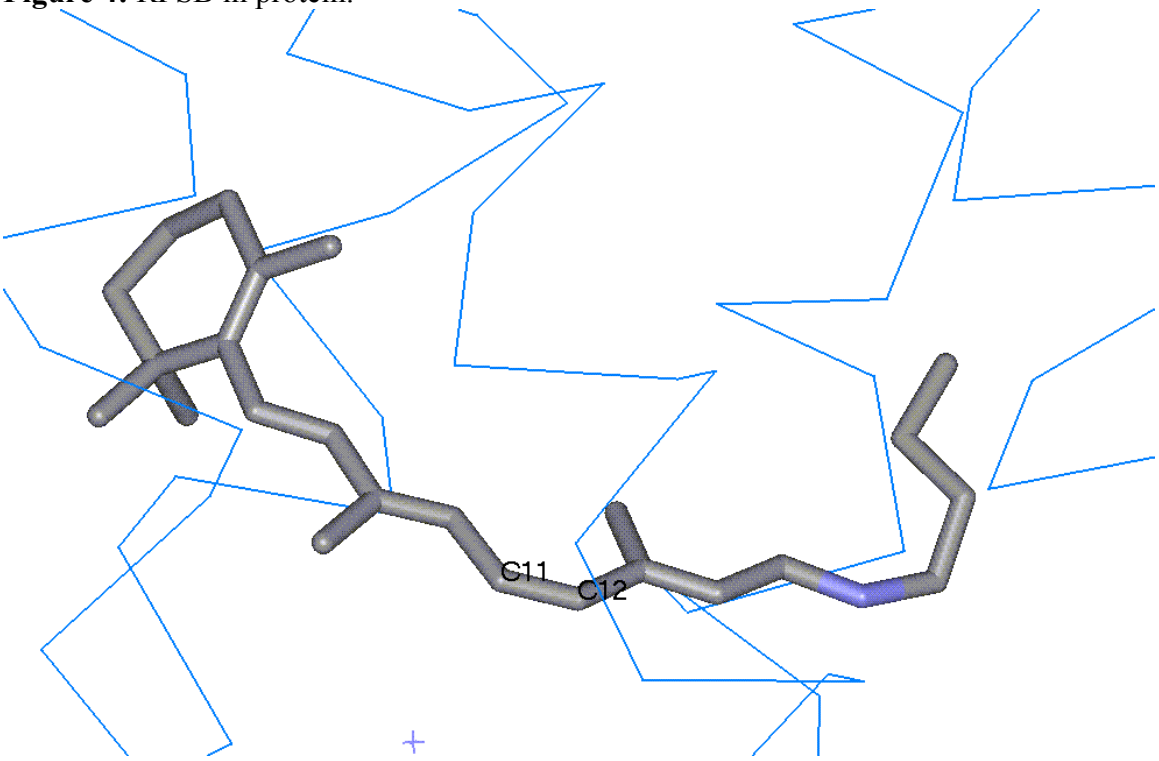
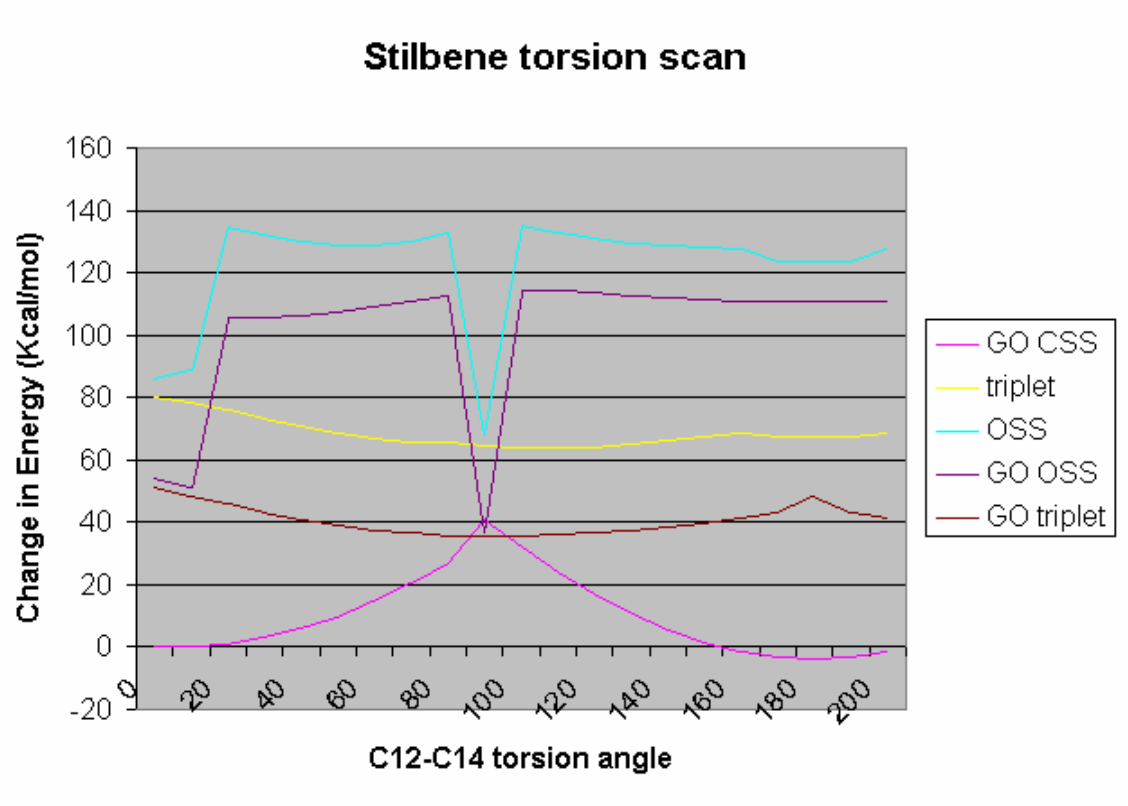


Figure 5: Plots of potential energy for stilbene and met-stilbene. Energy profiles are shown for the geometry optimized closed shell singlet (GO CSS), triplet, open shell singlet (OSS), geometry optimized open shell singlet (GO OSS), and geometry optimized triplet (GO triplet).

a)



b)

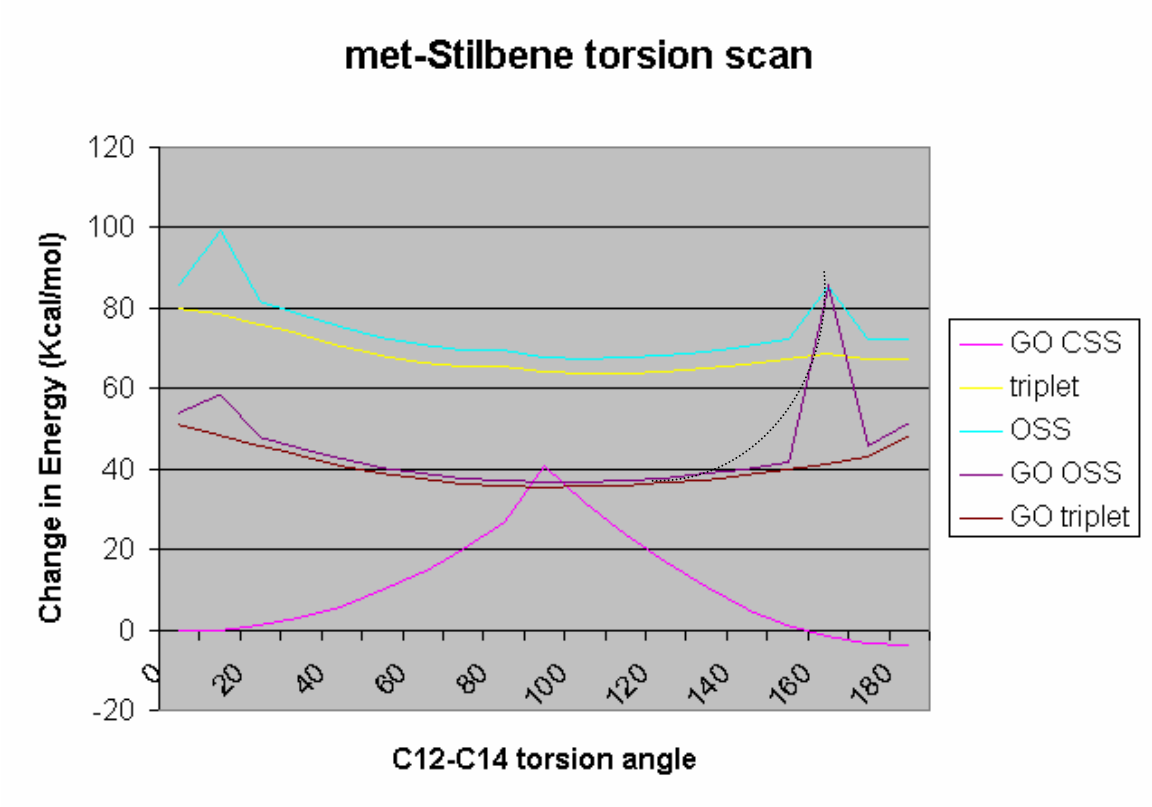
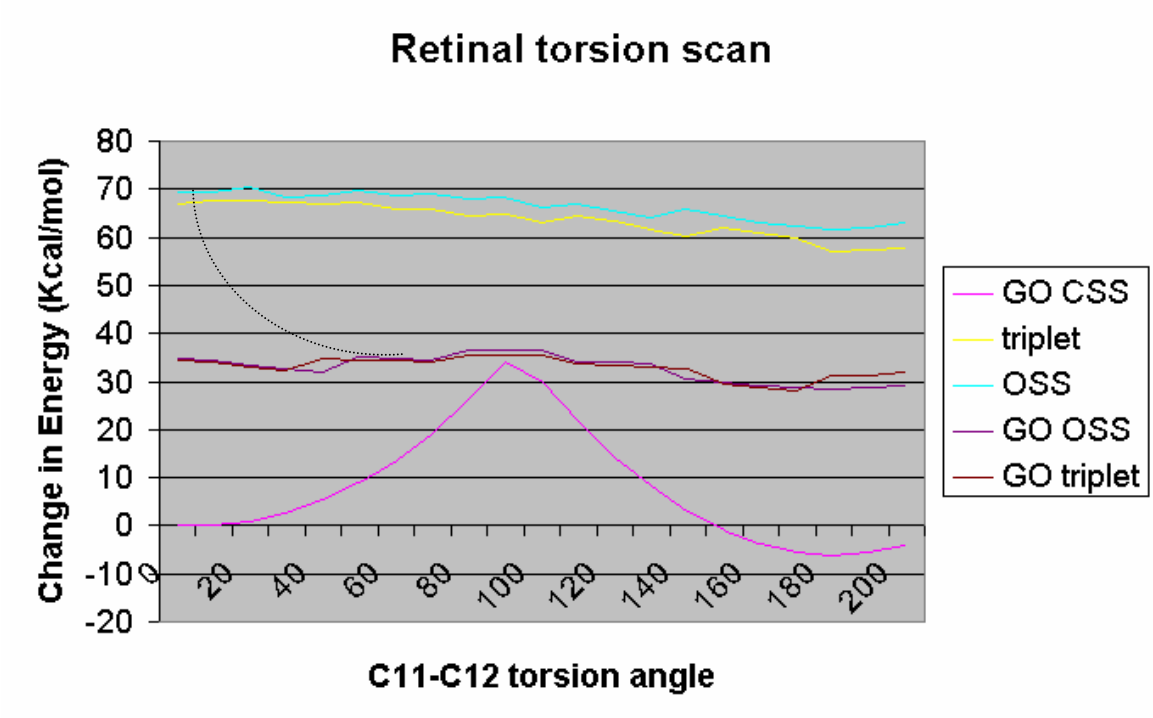
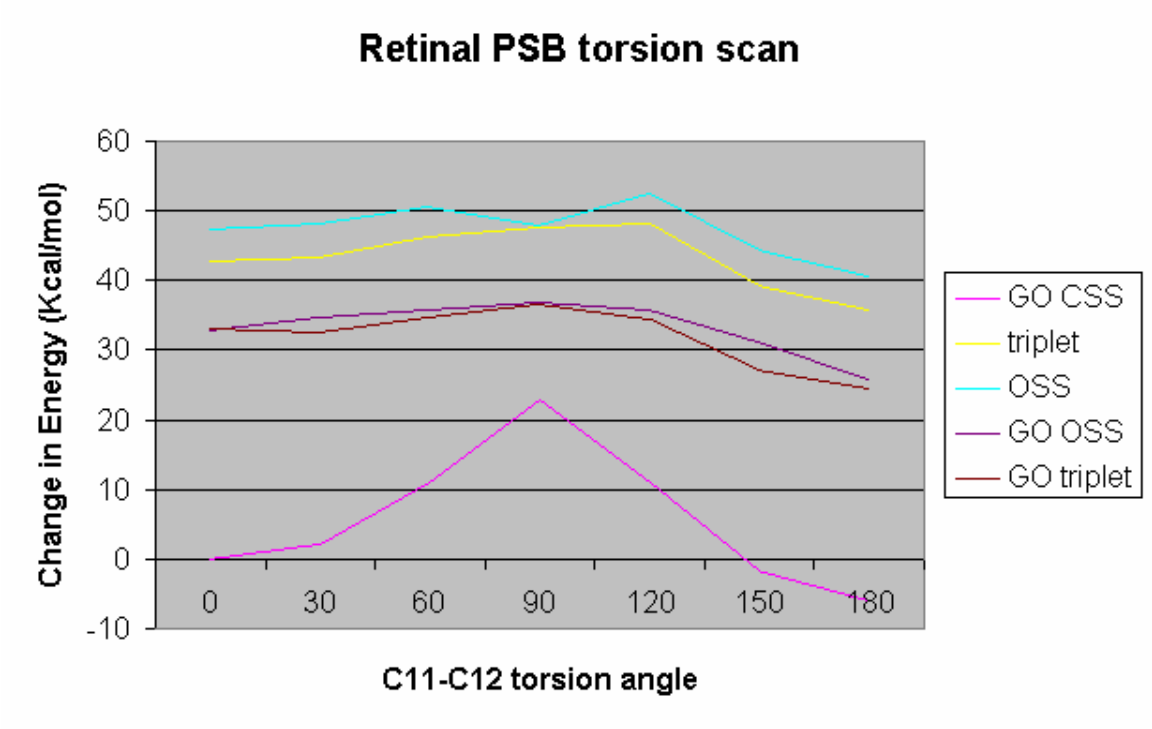

met-Stilbene torsion scan
Change in Energy (Kcal/mol)
120
100
80
60
40
20
0
-20
0
20
40
60
80
100
120
140
160
180
C12-C14 torsion angle
GO CSS
triplet
OSS
GO OSS
GO triplet

Figure 6: Energy profiles for retinal, RPSB, and RPSB within protein.

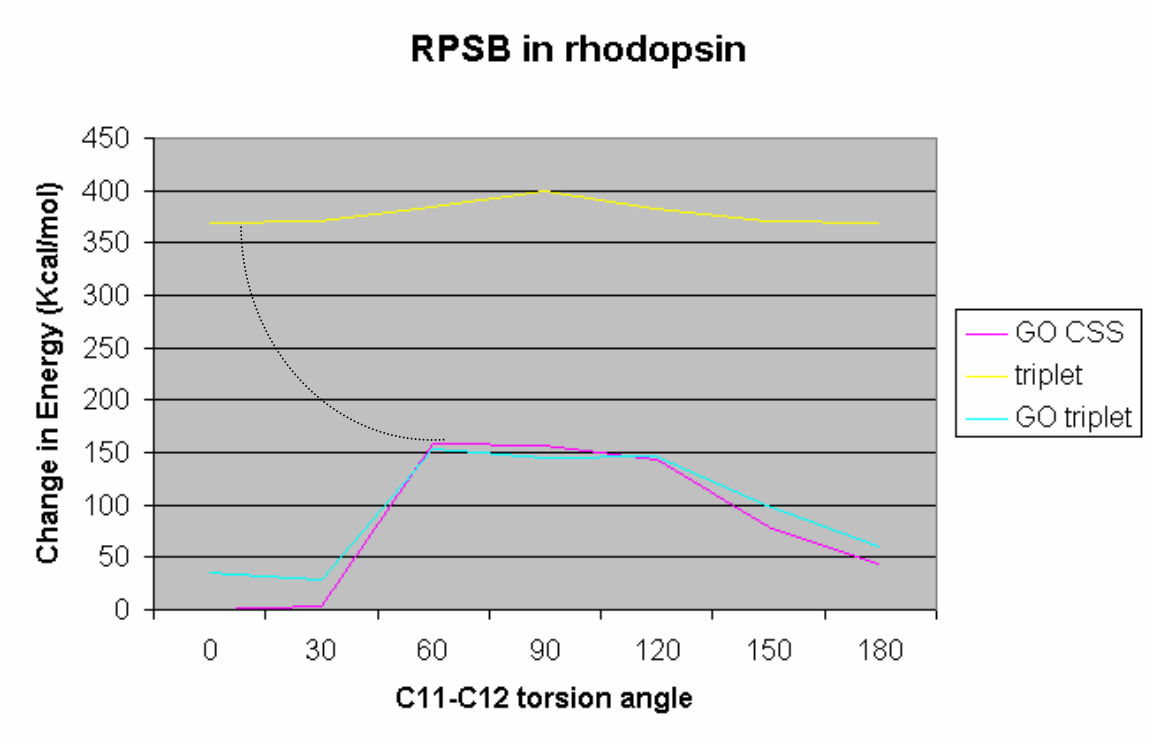
a)



b)


Retinal PSB torsion scan
Change in Energy (Kcal/mol)
60
50
40
30
20
10
0
-10
0
30
60
90
120
150
180
C11-C12 torsion angle
GO CSS
triplet
OSS
GO OSS
GO triplet

c)


RPSB in rhodopsin
Change in Energy (Kcal/mol)
450
400
350
300
250
200
150
100
50
0
0
30
60
90
120
150
180
C11-C12 torsion angle
GO CSS
triplet
GO triplet

Chapter 9: CCR1 and CCR5 structure and function prediction

Abstract

The G-protein-coupled receptor (GPCR) superfamily has been the target for many pharmaceutical drugs and thus the 3D structure of these proteins is essential to future rational drug design. Specifically the chemokine receptor family is emerging as a class of particular important for the possible treatment of immunological pathologies. The CCR1 receptor has been implicated in multiple sclerosis and organ transplant rejection whereas the homologous receptor CCR5 has been implicated in HIV infection. This study uses the MembStruk protocol to predict the structure and the HierDock protocol to predict the function of the CCR1 receptor using BX471, as well as 5 other compounds. The CCR5 receptor was constructed as a homology model from the predicted CCR1 structure and the binding of Tak-779 was determined, as well as 4 other compounds. In both receptors, the binding and mutation studies correlate well with the predicted values based on the structures. The MembStruk protocol was thus validated for these cases and has yielded new insights into the binding and important general structural features of these receptors.

Introduction

Chemokines have been generally accepted as the primary factors for leukocyte migration (Mackay, 2001; Thelen, 2001). They are small peptides with a characteristic number of conserved cysteine pairs; they act as antagonists or agonists on specific receptors in the GPCR superfamily (Onuffer et al., 2002). As such, they are involved in specific chemotactic axes (specific receptor-chemokine pairs) for leukocyte recruitment to specific sites (Houshmand et al., 2003). In addition to their leukocyte attractant role, chemokines are also involved in cell growth and HIV infection (Mackay, 2001). Resultingly, they are involved in a variety of diseases related to inflammatory cell localization: asthma, multiple sclerosis, atherosclerosis, arthritis, organ transplant rejection, and cancer (Gerard, 2001). The specific role of chemokines in cancer may be negative or positive, acting directly as tumor growth factors or indirectly in attracting leukocytes which may release tumor growth and angiogenetic factors. Positively, the presence of leukocytes may increase tumor rejection. Finally, the specific expression of certain chemokine may attract specific chemokine receptor expressing cells to metastasize. This case of cancer demonstrates the intrinsic roles of chemokines in complex immunological processes.

Thus, the development of small antagonists to chemokine receptors is an important challenge. The main challenges include antagonist cross-reactivity with other GPCRs and reduced affinity in animal models compared to humans (Horuk, 2003). In particular, the CCR1 inhibitors of Berlex have demonstrated reactivity to dopaminergic and muscarinic receptors (Hesselgesser et al., 1998). In addition, the antagonist BX471

has much reduced affinity to the rodent receptors (Liang et al., 2000), thus requiring much higher doses to induce the required effect. The reduced affinity in the mouse receptor was also found for other similar 4-hydroxypiperidine antagonists (Onuffer et al., 2003). Such similar problems have been found in the case of CCR5 inhibitors of Shering-Plough, where reactivity to muscarinic receptors was found (Tagat et al., 2001). Also, the antagonist SHC C of Shering Plough was found to have poor rodent receptor affinity (Horuk, 2003).

To aid in the design of antagonists with reduced cross-reactivity and with better affinity to animal receptors, the availability of 3D structures for the receptors is critical. Currently, the only 3D experimental structure available is for bovine rhodopsin (Palczewski et al., 2000). Since the sequence homology of rhodopsin to most GPCRs is less than 20%, homology modeling is not a feasible avenue for building accurate structures (Archer et al., 2003). The develop of the MembStruk procedure (Floriano et al., 2000; Vaidehi et al., 2002; Trabanino et al., 2004) has provided an alternative for constructing the 3D structures of GPCRs from mostly first principles. This current study utilizes the MembStruk procedure to construct the 3D structure of human CCR1 and provides validation for the structure based on binding data, specific protein structural features, and based on correlation to mutation data for a homology model of CCR5 from CCR1.

Methods

Prediction of TM helical regions

The prediction of TM (transmembrane) segments from sequence is achieved using the TM2ndS protocol (as discussed in Chapters 2 and 3). This protocol relies on hydrophobicity profiles generated for different window sizes and general helix capping rules to obtain accurate TM helical predictions. These hydrophobic profiles are by default obtained from a Clustalw multiple sequence alignment (Thompson et al., 1994) of sequences with over 200 bit score to the query sequence obtained using Blast (Altschul et al., 1990; Altschul et al., 1997). The sequences used are shown in Figure 1. The hydrophobic profile for window size 12 and the chosen baseline are shown in Figure 2.

Figure 1: Sequences obtained from Blast for alignment.

sp|P32246|CKR1_HUMAN C-C chemokine receptor type 1 (C-C CKR-1) (...	560	e-159
sp|P56482|CKR1_MACMU C-C chemokine receptor type 1 (C-C CKR-1) (...	487	e-137
sp|P51675|CKR1_MOUSE C-C chemokine receptor type 1 (C-C CKR-1) (...	471	e-132
sp|P51678|CKR3_MOUSE Probable C-C chemokine receptor type 3 (C-C...	358	7e-99
sp|P51677|CKR3_HUMAN C-C chemokine receptor type 3 (C-C CKR-3) (...	357	2e-98
sp|P56483|CKR3_MACMU C-C chemokine receptor type 3 (C-C CKR-3) (...	355	6e-98
sp|P56492|CKR3_CERAE C-C CHEMOKINE RECEPTOR TYPE 3 (C-C CKR-3) (...	353	2e-97
sp|O54814|CKR3_RAT C-C CHEMOKINE RECEPTOR TYPE 3 (C-C CKR-3) (CC...	346	3e-95
sp|P51676|CKRV_MOUSE C-C chemokine receptor 1-like protein 1 (Ma...	344	2e-94
sp|Q9Z2I3|CKR3_CAVPO C-C CHEMOKINE RECEPTOR TYPE 3 (C-C CKR-3) (...	340	2e-93
sp|O97883|CKR5_HYLLE C-C chemokine receptor type 5 (C-C CKR-5) (...	313	4e-85
sp|O97882|CKR5_PYGNE C-C chemokine receptor type 5 (C-C CKR-5) (...	312	7e-85
sp|P56441|CKR5_PAPHA C-C chemokine receptor type 5 (C-C CKR-5) (...	310	2e-84
sp|P79436|CKR5_MACMU C-C chemokine receptor type 5 (C-C CKR-5) (...	310	2e-84
sp|O62743|CKR5_CERTO C-C chemokine receptor type 5 (C-C CKR-5) (...	310	2e-84
sp|P51682|CKR5_MOUSE C-C chemokine receptor type 5 (C-C CKR-5) (...	310	4e-84
sp|O97879|CKR5_TRAPH C-C chemokine receptor type 5 (C-C CKR-5) (...	309	5e-84
sp|P56440|CKR5_PANTR C-C chemokine receptor type 5 (C-C CKR-5) (...	309	6e-84
sp|O97881|CKR5_PONPY C-C chemokine receptor type 5 (C-C CKR-5) (...	309	6e-84
sp|O97880|CKR5_PYGBI C-C chemokine receptor type 5 (C-C CKR-5) (...	308	8e-84
sp|P51681|CKR5_HUMAN C-C chemokine receptor type 5 (C-C CKR-5) (...	308	8e-84
sp|O97878|CKR5_TRAFR C-C chemokine receptor type 5 (C-C CKR-5) (...	308	1e-83
sp|P56439|CKR5_GORGO C-C chemokine receptor type 5 (C-C CKR-5) (...	308	1e-83
sp|P56493|CKR5_CERAE C-C chemokine receptor type 5 (C-C CKR-5) (...	307	2e-83
sp|O55193|CKR2_RAT C-C chemokine receptor type 2 (C-C CKR-2) (CC...	305	7e-83
sp|P51683|CKR2_MOUSE C-C chemokine receptor type 2 (C-C CKR-2) (...	300	2e-81
sp|O08556|CKR5_RAT C-C CHEMOKINE RECEPTOR TYPE 5 (C-C CKR-5) (CC...	300	3e-81
sp|O18793|CKR2_MACMU C-C chemokine receptor type 2 (C-C CKR-2) (...	293	3e-79
sp|P41597|CKR2_HUMAN C-C chemokine receptor type 2 (C-C CKR-2) (...	268	9e-72
sp|P51679|CKR4_HUMAN C-C chemokine receptor type 4 (C-C CKR-4) (...	254	1e-67
sp|P51680|CKR4_MOUSE C-C chemokine receptor type 4 (C-C CKR-4) (...	253	5e-67
sp|O97665|CKR8_MACMU C-C chemokine receptor type 8 (C-C CKR-8) (...	207	2e-53
sp|P51685|CKR8_HUMAN C-C chemokine receptor type 8 (C-C CKR-8) (...	207	3e-53
sp|P56484|CKR8_MOUSE C-C chemokine receptor type 8 (C-C CKR-8) (...	204	3e-52

Prediction of the hydrophobic center of the helices

As described previously (Trabanino et al., 2004), the hydrophobic center (HC) of each helix is determined by analyzing the hydrophobic profiles at various window sizes. A range of window sizes for which the maxima are stable is used to calculate a final residue number which is taken to be at the center of the membrane bilayer, as discussed in Chapter 4. The positions of these HC's are shown on the hydrophobic profile of Figure 2. In addition, the TM helical sequences (before and after capping) and HC positions (underlined) are shown in Figure 3. Thus, the individual helices are translated along their helical axes so that these residues are aligned to a common plane of the bundle, as shown in Figure 4.

Figure 2: The hydrophobic profile for CCR1 at window size 12 with HC's indicated.

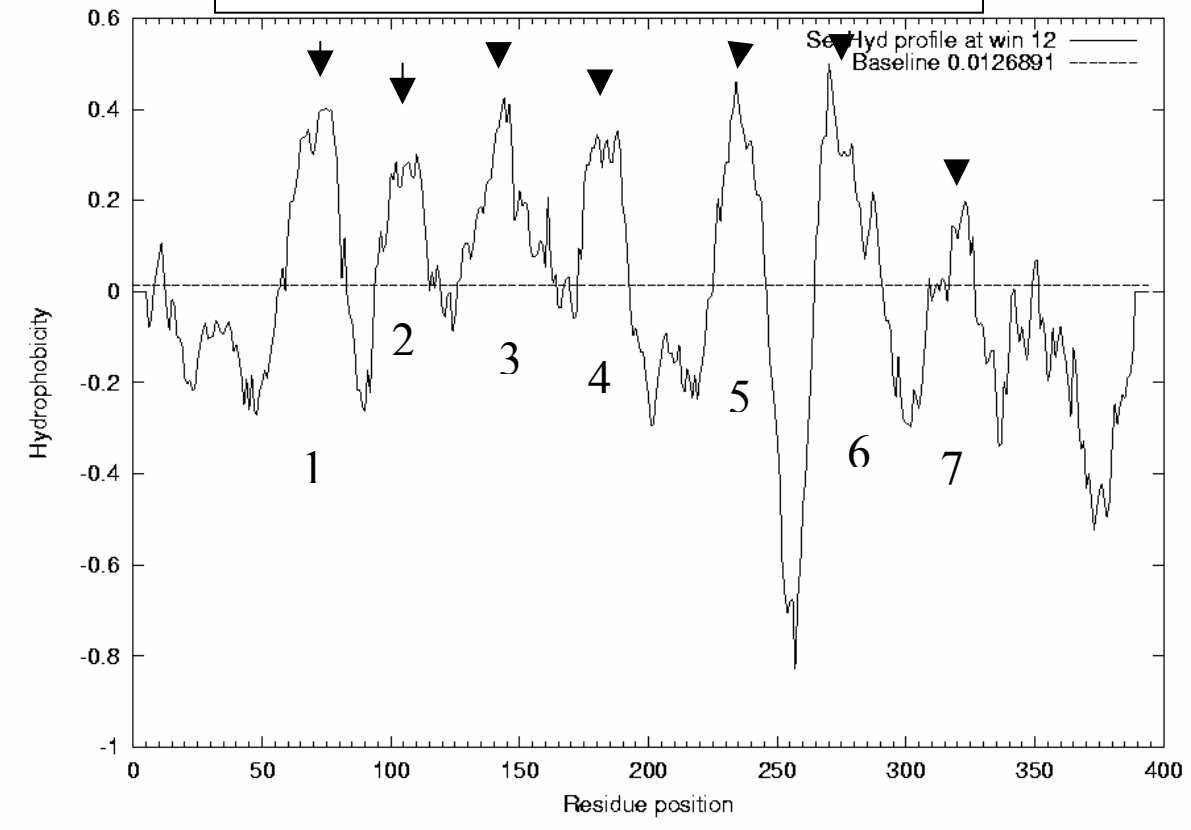
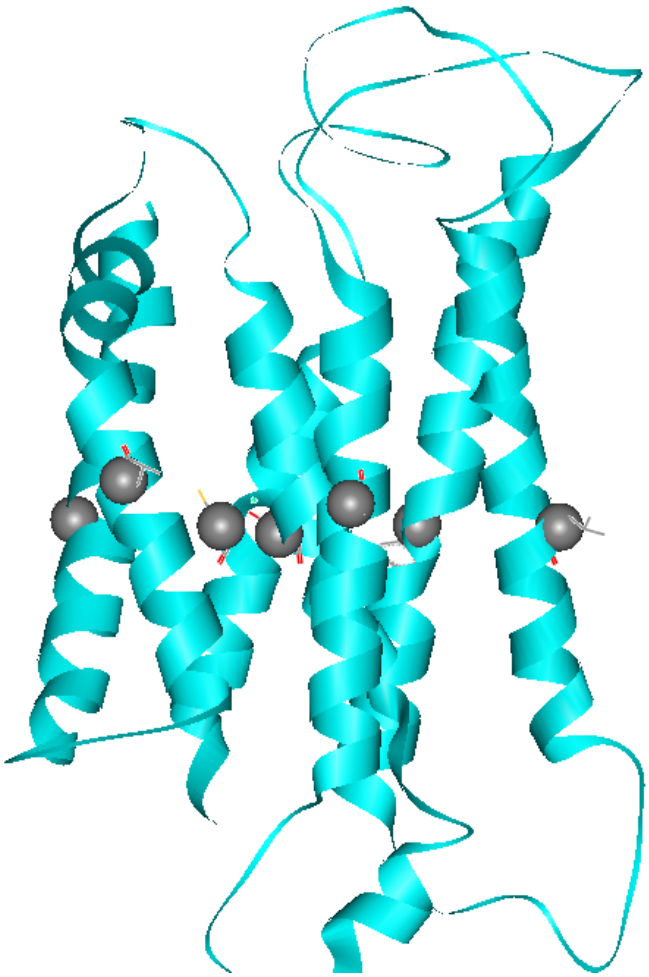


Figure 3: The CCR1 TM helix predictions with HC's indicated.

```
The TM prediction 1

QLLPPLYSLVFVIGLVGNILVVLVL
QLLPPLYSLVFVIGLVGNILVVLVLVQYK

The TM prediction 2

   LLNLAISDLLFLFTLPFWIDYKLK
SIYLLNLAISDLLFLFTLPFWIDYKLK

The TM prediction 3

AMCKILSGFYYTGLYSEIFFIILLTIDRYLAIVH
AMCKILSGFYYTGLYSEIFFIILLTIDRYLAIVH

The TM prediction 4

  GVITSIIIWALAILASMPGL
TFGVITSIIIWALAILASMPGLYF

The TM prediction 5

 LKLNLFGLVLPLLVMIICYTGI
ALKLNLFGLVLPLLVMIICYTGII

The TM prediction 6

  IFVIMIIFFLFWTPYNLTILISVFQDF
RLIFVIMIIFFLFWTPYNLTILISVFQDF

The TM prediction 7

     YTHCCVNPVIYA
TEVIAYTHCCVNPVIYAFVGER
```

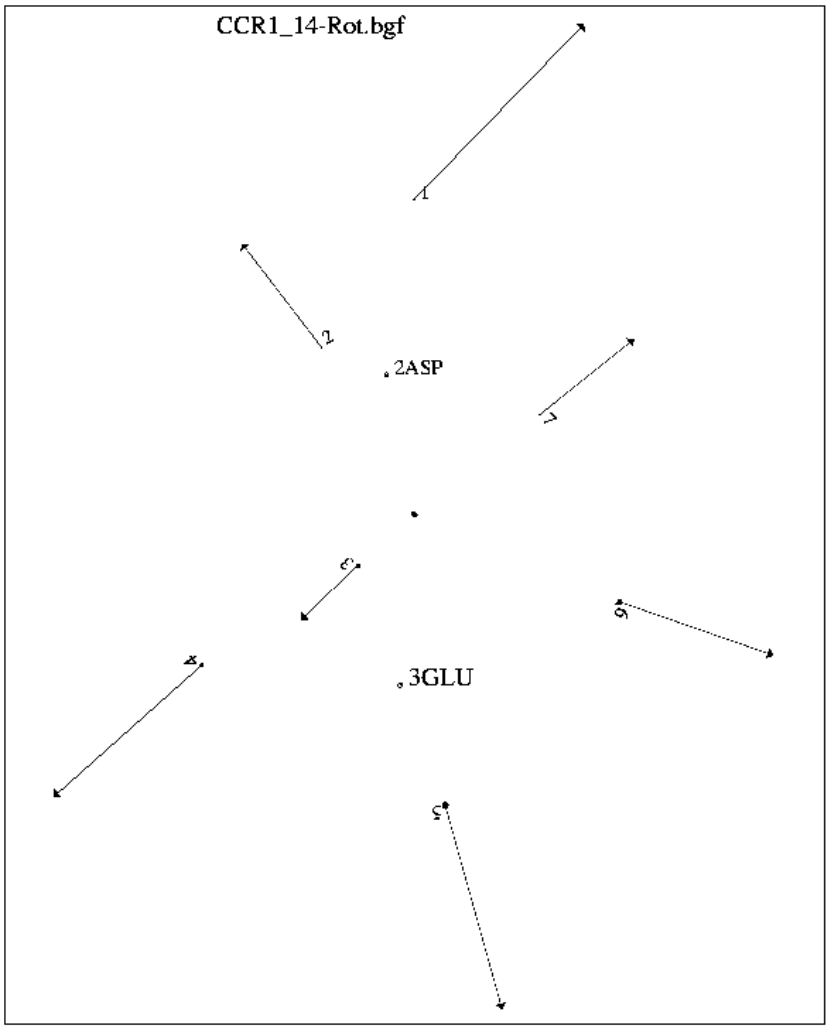
Figure 4: The residue c-alphas predicted to correpond to the HC's aligned to a common plane in CCR1.



Initial Rotation

Once the hydrophobic center is predicted, the sector angle of the face between pairs of helices (about a rotatable helix) is determined (Trabanino et al., 2004).Then the helical face of the rotatable helix with a maximum hydrophobic moment within that sector angle is pointed between these two helices. The orientations of these calculated hydrophobic moments are shown in Figure 5.

Figure 5: Rotational orientations of the hydrophobic moments of the helices in CCR1 after the initial rotation.



Optimization of helix backbones

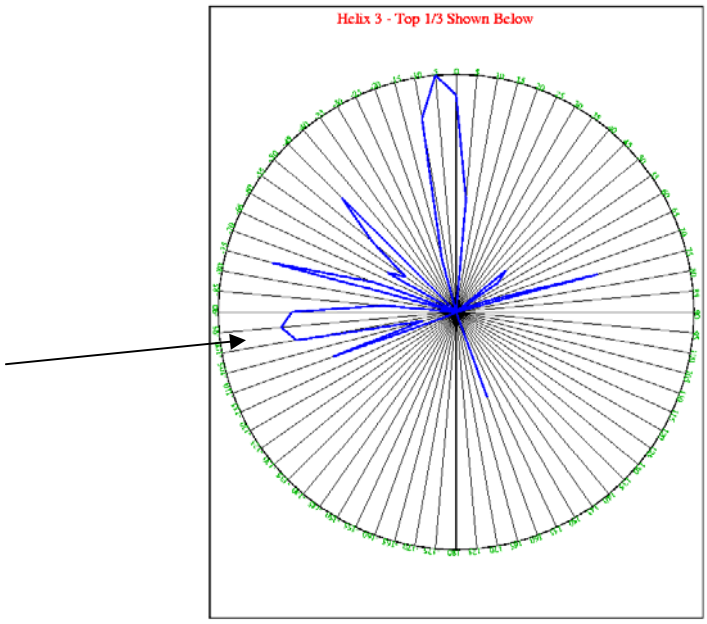
After the helices are initially orientated, molecular dynamics at 300 degrees on the individual helices is carried out as an initial backbone optimization. MPSim (Lim et al., 1997), which is used for all energy calculations and molecular dynamics in subsequent steps, was used. Helix 6 exhibited the most bending of the 7 helices and helix 1 to a lesser extent.

Optimization of helical rotations

Hel 3

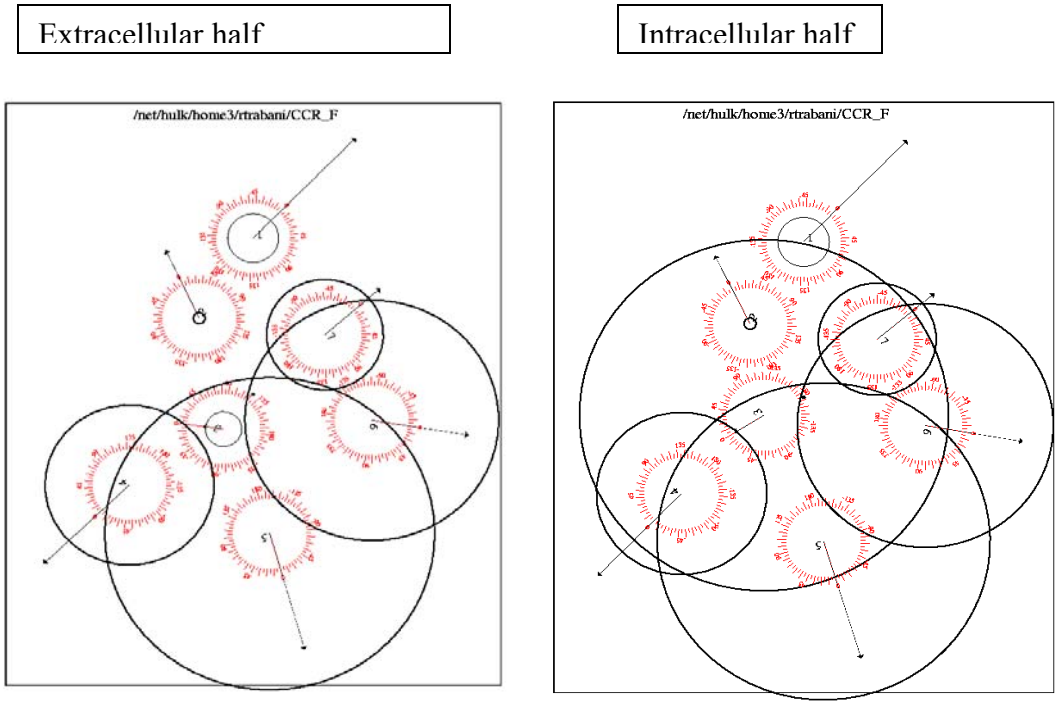
The rotation of Helix 3 requires more analysis, as it is a helix without one well-define hydrophobic face. In addition to this hydrophobic moment analysis, a 360 degree energy scan of the helix 3 using Rotmin (Trabanino et al., 2004) was performed. The results of this scan are shown in Figure 6.

Figure 6: 360 degree energy scan of helix 3 without Hel7 HSP.



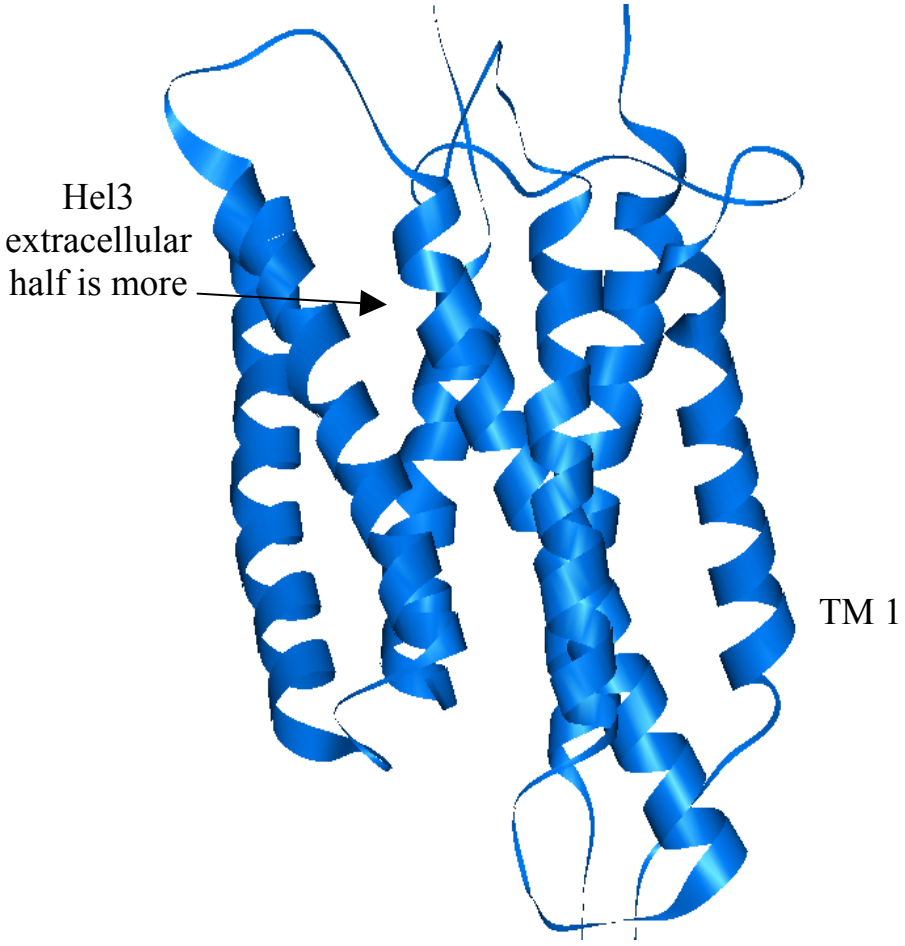
As previously mentioned, helix 3 in GPCRs is a special case in that it, because of its position in the bundle, actually has two faces which face the membrane and these faces are approximately 90 degrees from each other. As a second check to the rotation by energy, two hydrophobic moments could be calculated for the two halves of helix 3 about the hydrophobic center. This was done for CCR1 using 7 residues displaced at least one residue from the hydrophobic center. The resulting moment orientations for each half are shown in Figure 7. This moment orientation corresponds to the rotation indicated by the arrow in Figure 6. Thus, this helix 3 rotation was used for later optimization.

Figure 7: Helix 3 hydrophobic moment orientations for the extracellular and intracellular halves of the helix.



Although it was expected that the moments for the two halves would be 90 degrees to each other, this case demonstrates that often the extracellular moment is more reliable than the intracellular moment. This is because the extracellular moment is defined for a more exposed (to membrane) helical face as shown in Figure 8.

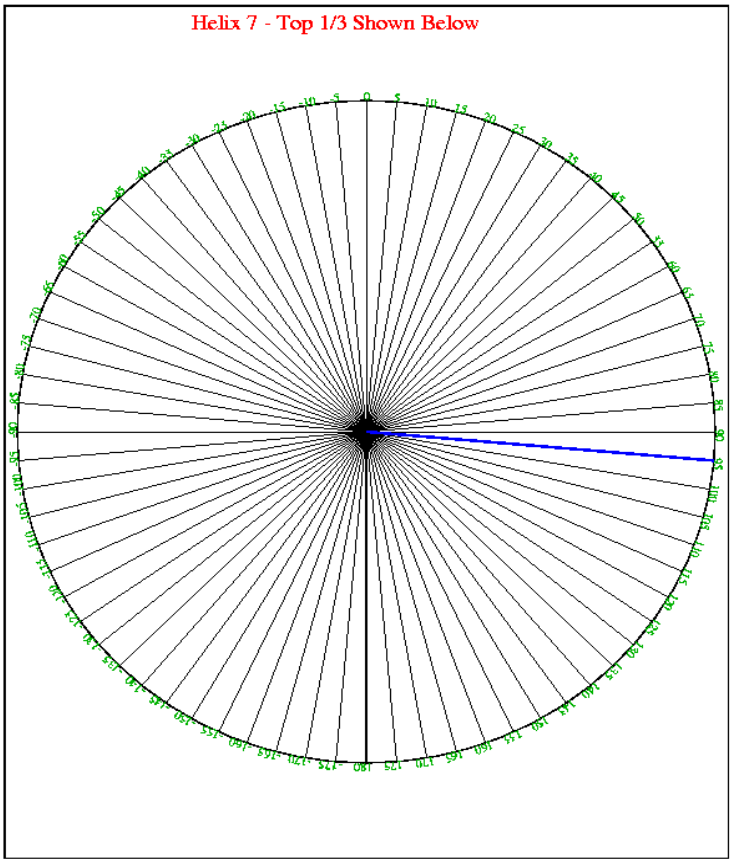
Figure 8: CCR1 structure with the two helix 3 membrane-exposed faces shown.



Hel 7 rotation – Hel3/7 salt bridge formation

Even before the rotation of helix 3, the Glu120 of helix 3 was within 7 Å of the His293 of helix 7. It should be noted that the previous energy scan of helix 3 was performed with a neutral form of His293. At this point, the His was protonated (to Hsp) and a 360 degree energy scan was performed on helix 7. The results are shown in Figure 9. The rotation of helix 7 at +95 degrees was thus chosen for further optimization.

Figure 9: 360 degree energy scan for helix 7 with HSP.

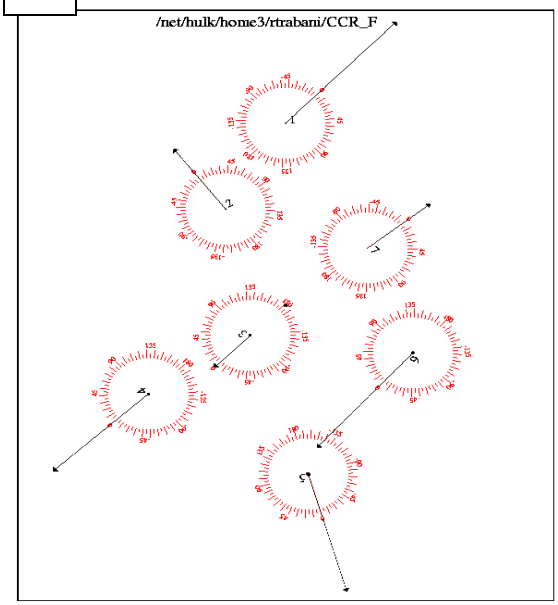


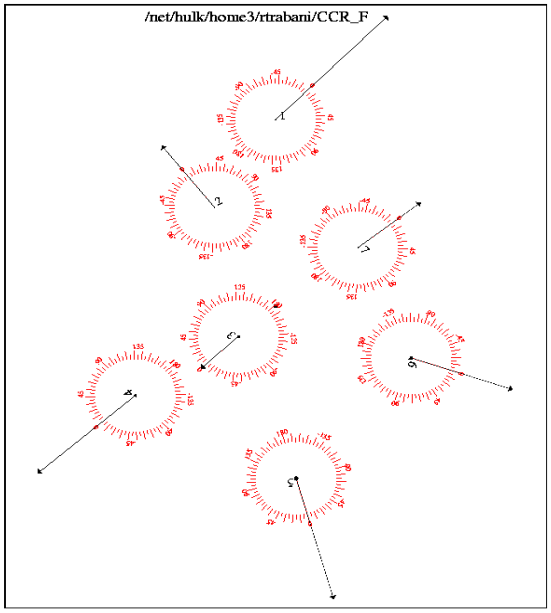
Hel 6 rotation

Helix 6 is also typically a difficult helix to rotate owing to the ill-defined hydrophobic moment. This may have a role, since helix 6 has been implicated in rotating upon GPCR activation in rhodopsin (Farrens et al., 1996). When validating the MembStruk protocol for rhodopsin (Trabanino et al., 2004), helix 6 had the largest error in rotation as compared with the ground state of rhodopsin. A recent advance indicates an improvement in helix 6 can be obtained by scanning various numbers of middle residues (about the HC) for the calculation of the hydrophobic moment. In bovine rhodopsin, this lead to two predominant rotations, which correlated with the experimental rotation for the ground and activated state of the receptor. The number of middle residues used to obtain the helix 6 rotation as in the ground state crystal structure of rhodopsin was 19 or 21. Thus, for CCR1, the number of middle residues used was 21 to reproduce this ground state rotation. The rotations using 15 and 21 as the number of middle residues are shown in Figure 10.

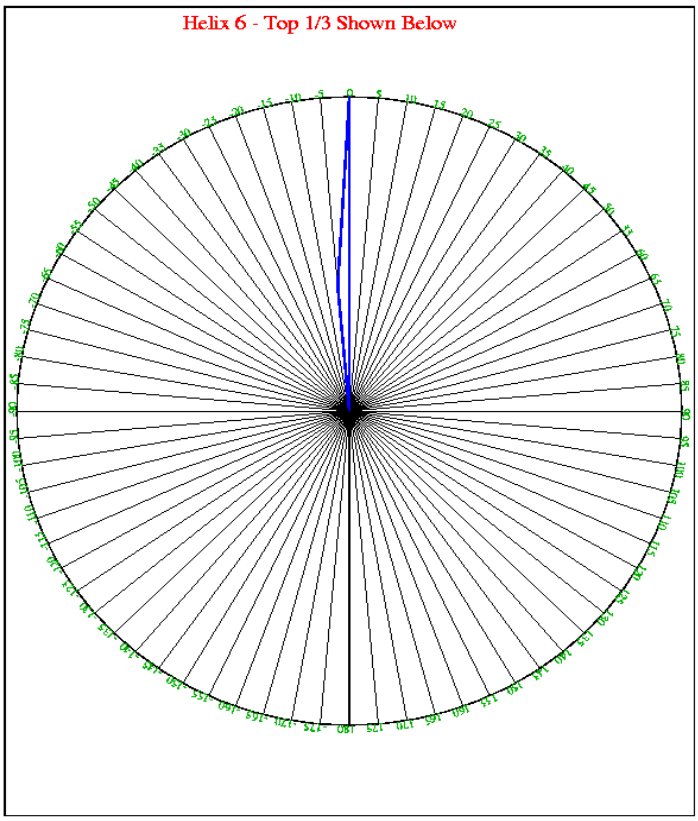
Figure 10: a) Helix 6 alternate hydrophobic moment orientations using different number of middle residues for calculation. b) 360 degree energy scan for helix 6

a)





b)



As a double check for the rotation, a 360 degree energy scan of helix 6 in the structure before adding loop revealed this to be the best rotation, as seen in Figure 10b.

Rigid body molecular dynamics (RBMD)

As discussed previously (Trabanino et al., 2004), molecular dynamics was performed on the bundle at this point in order to optimize rotations, tilts, and translations, treating each helix as a rigid body. This is done for 50 ps in explicit lipid.

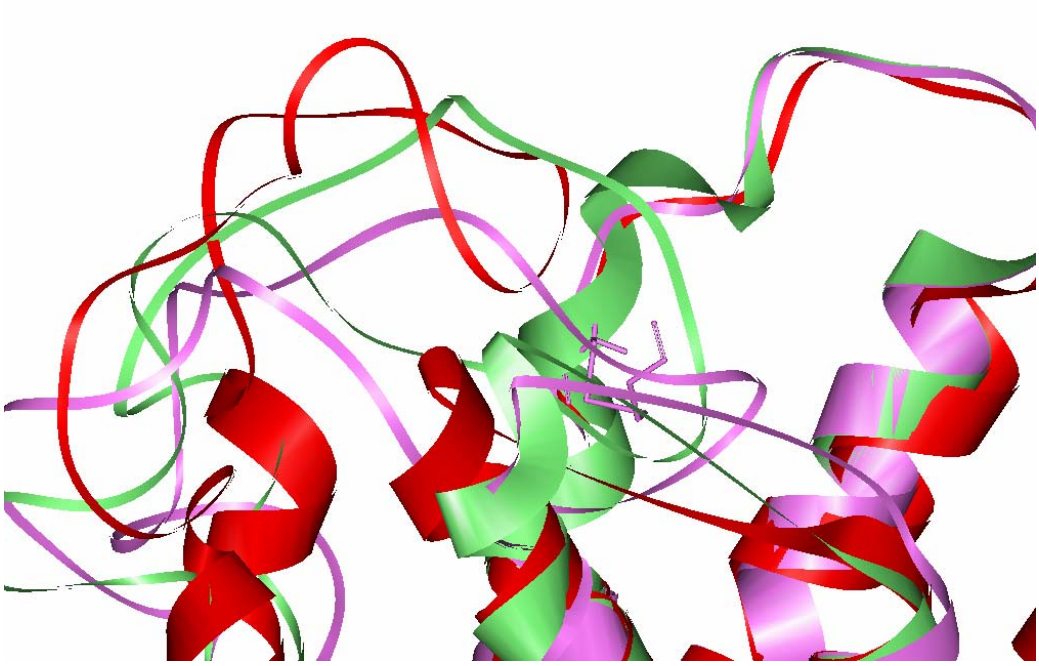
Loop addition

Loops were added using Modeller 6v2, and the structure's loop regions were minimized in potential energy using the conjugate gradient method. Then the complete structure was minimized in potential energy.

EC-II loop simulation

The EC-II loop was folded into a closed form using the procedure outline in Chapter 6. The disulfide bond between the conserved cysteines on helix 3 and ECII was formed and the loop was annealed from 1000 to 2000 degrees. Then TVN 300 degree molecular dynamics was conducted on half of the bundle to allow optimization of loop with movable TM backbone. Finally, the loop was again annealed with movable side chains in the vicinity of the loop for ~120 ps. The progression of the loop closing at each step is shown in Figure 11.

Figure 11: Snapshots of the EC-II loop closing in CCR1.



Start 100 ps TVN Dyn 140 ps annealing

Homology model of CCR5

From the final human CCR1 structure, the homology model of CCR5 was constructed. Since the whole sequence homology is 54% and gaps were not present consecutively more than once, the model building was straightforward. The Clustalw sequence alignment between these two proteins is shown in Figure 12.

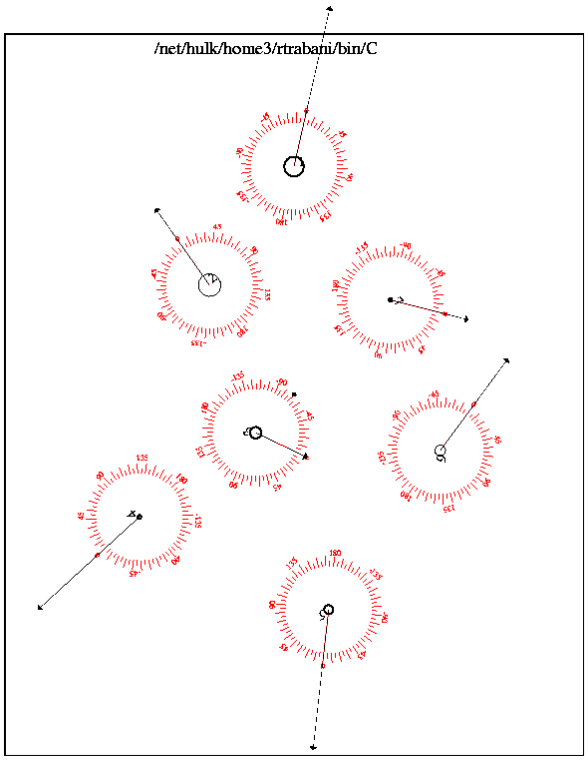
Figure 12: Alignment of human CCR1 and CCR5 sequences.

```
gi|4502639|ref|NP_000570.1|      MDYQVSSPIYDIN----YYTSEPCQKINVKQIAARLLPPLYSLVFIFGFV
sp|P32246|CKR1_HUMAN             METPNTTEDYDTTTEFDYGDATPCQKVNERAFGAQLLPPLYSLVFVIGLV
                                 *:   ::  ** .    *  : ****:* : :.*:**********::*:*

gi|4502639|ref|NP_000570.1|      GNMLVILILINCKRLKSMTDIYLLNLAISDLFFLLTVPFWAHY-AAAQWD
sp|P32246|CKR1_HUMAN             GNILVVLVLVQYKRLKNMTSIYLLNLAISDLLFLFTLPFWIDYKLKDDWV
                                 **:**:*:*:: ****.**.***********:**:*:*** .*    :*

gi|4502639|ref|NP_000570.1|      FGNTMCQLLTGLYFIGFFSGIFFIILLTIDRYLAVVHAVFALKARTVTFG
sp|P32246|CKR1_HUMAN             FGDAMCKILSGFYYTGLYSEIFFIILLTIDRYLAIVHAVFALRARTVTFG
                                 **::**::*:*:*: *::* **************:*******:*******

gi|4502639|ref|NP_000570.1|      VVTSVITWVVAVFASLPGIIFTRSQKEGLHYTCSSHFPYSQYQFWKNFQT
sp|P32246|CKR1_HUMAN             VITSIIIWALAILASMPGLYFSKTQWEFTHHTCSLHFPHESLREWKLFQA
                                 *:**:* *.:*::**:**: *:::* *  *:*** ***:.. : ** **:

gi|4502639|ref|NP_000570.1|      LKIVILGLVLPLLVMVICYSGILKTLLRCRNEKKRHRAVRLIFTIMIVYF
sp|P32246|CKR1_HUMAN             LKLNLFGLVLPLLVMIICYTGIIKILLRRPNEKK-SKAVRLIFVIMIIFF
                                 **: ::*********:***:**:* ***  ****  :******.***::*

gi|4502639|ref|NP_000570.1|      LFWAPYNIVLLLNTFQEFFGLNNCSSSNRLDQAMQVTETLGMTHCCINPI
sp|P32246|CKR1_HUMAN             LFWTPYNLTILISVFQDFLFTHECEQSRHLDLAVQVTEVIAYTHCCVNPV
                                 ***:***:.:*:..**:*:  ::*..*.:** *:****.:. ****:**:

gi|4502639|ref|NP_000570.1|      IYAFVGEKFRNYLLVFFQKHIAKRFCKCCSIFQQEAPERASSVYTRSTGE
sp|P32246|CKR1_HUMAN             IYAFVGERFRKYLRQLFHRRVAVHLVKWLPFLSVDRLERVSST-SPSTGE
                                 *******:**:**  :*:::::* :: *  .::. :  **.**. : ****

gi|4502639|ref|NP_000570.1|      QEISVGL
sp|P32246|CKR1_HUMAN             HELSAGF
                                 :*:*.*:
```

Another issue was the helix 7 rotation, which was rotated with His in the protonated form in CCR1 to form a salt bridge with the helix 3 Glu120. In CCR5, this Glu is not present,

so for the case of CCR5, the helix 7 rotation was determined by the hydrophobic moment orientation of helix 7 alone. The orientation before rotation is shown in Figure 13.

Figure 13: Orientation of helix 7, which was rotated -45 degrees for rotation by hydrophobicity.

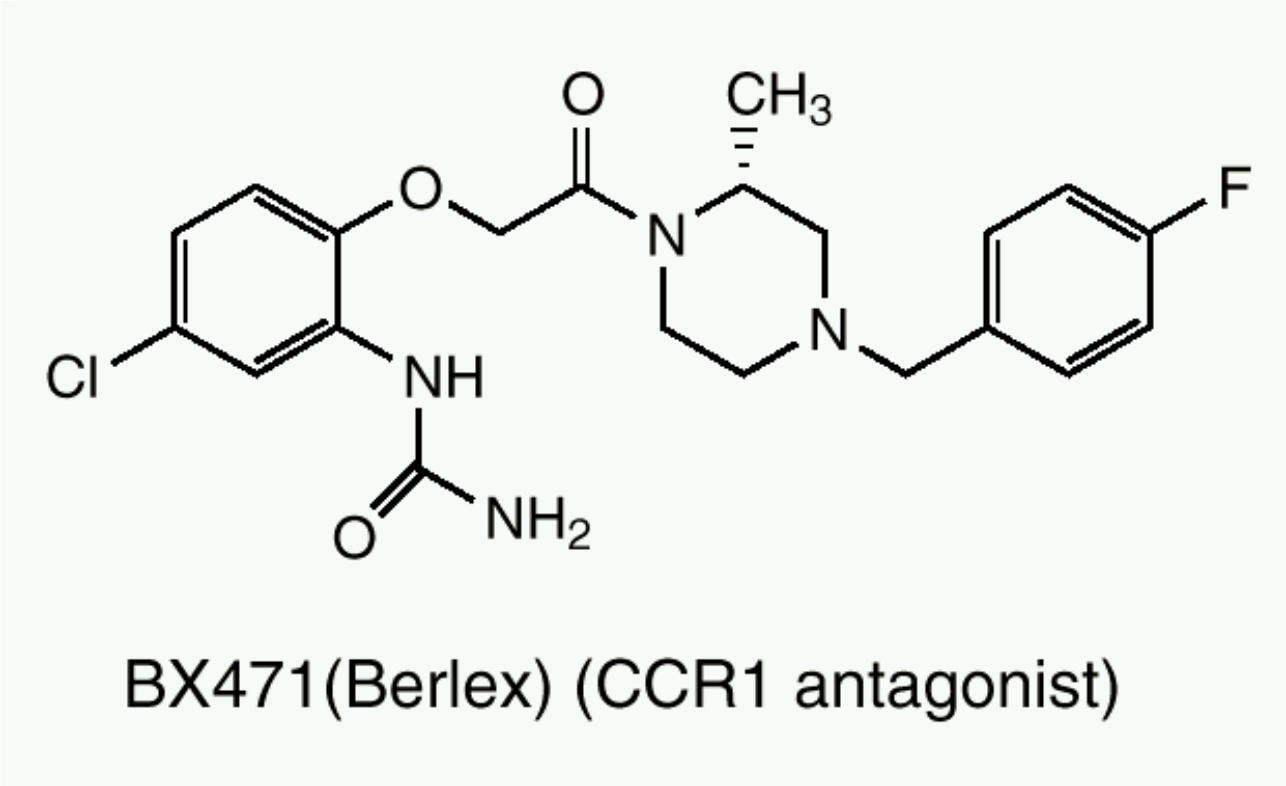


The helix 7 was rotated –45 degrees. The final structure was minimized in potential energy.

HierDock function prediction

As described previously (Trabanino et al., 2004), the molecular surface of the protein was determined and spheres were mapped onto this surface. All the spheres within the protein were divided into 14 boxes of volume 7 $Å^3$. The molecule BX471 (shown in Figure 14) was docked to these spheres using DOCK 4.0. The box with the best conformation energetically was used for later fine HierDock.

Figure 14: Structure of BX471 CCR1 antagonist. Figure: (Onuffer et al., 2002)



The spheres of the best box were then used to perform the fine-grain HierDock. This method was modified to look for conformations which anchored to the Glu190 of ECII with a distance within 4 Å. Also, HierDock without anchoring was also performed on all the five ligands (proprietary) and BX471. The rank scoring of these bound ligands was obtained by either perturbing from the BX471 structure or from the HierDock structure in the same binding mode (whichever was less in energy).

For finer optimization of the BX471 ligand, the EC-II loop was then annealed (temperatures in cycles 300 to 600 for 96 ps) with ligand present and His189 (adjacent to

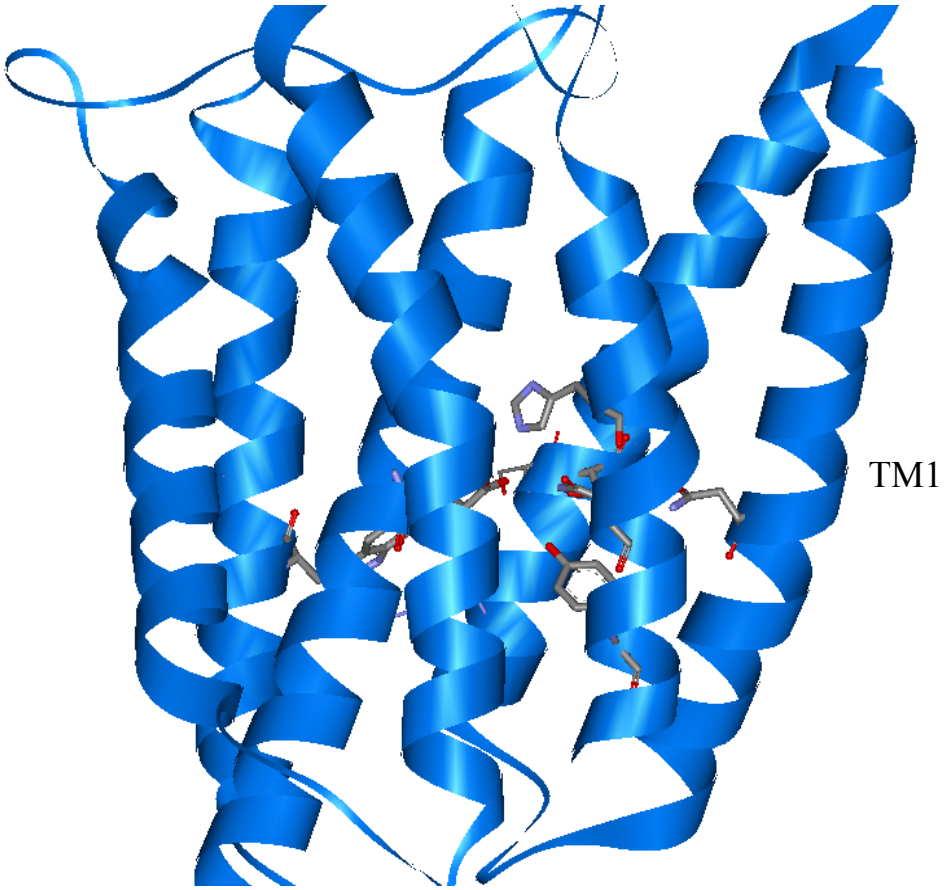
Glu) in a neutral state. The ligand was then annealed (temperatures in cycles 300 to 600 for 120 ps). The loop was annealed again for 120 ps and then the ligand for 120 ps. The distance between the protonated amine of the ligand and the oxygen of the Glu190 was 3.8 Å at this point. The side chain of Glu190 was minimized with a constraint of 2.5 Å between the proton of the ligand amine and the Glu oxygen. Then the entire complex was minimized in potential energy with no constraints. The final distance between the protonated amine of ligand and the oxygen of the Glu190 was 3.3 Å.

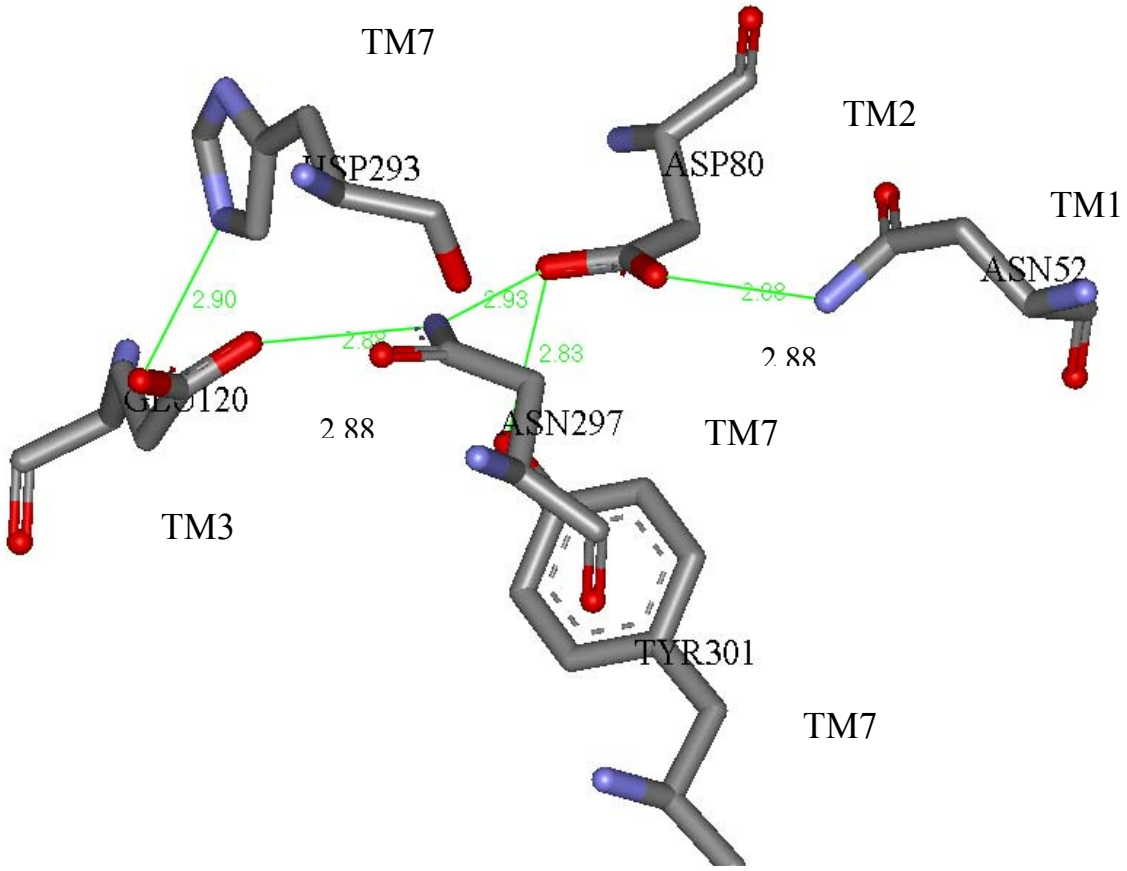
Results and discussion

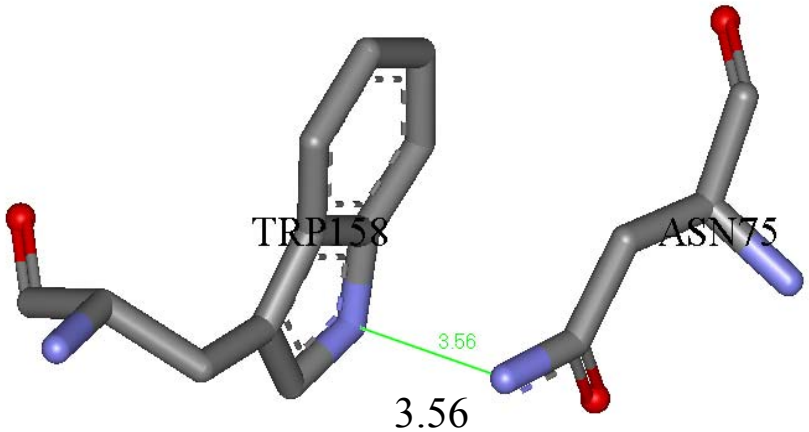
Structural features of human CCR1

There are a variety of interhelical interactions which were obtained using the MembStruk protocol between helices 1,2,3,4,7 as shown in Figure 15. In addition hel 7/3 salt bridge is shown in Figure 16.

Figure 15: General and detailed views of the CCR1 interhelical interactions.




TM7
TM2
ASP80
TM1
ASN52
2.90
2.93
2.88
2.88
2.88
ASN297
TM7
TM3
TYR301
TM7


TRP158
ASN75
3.56

3.56

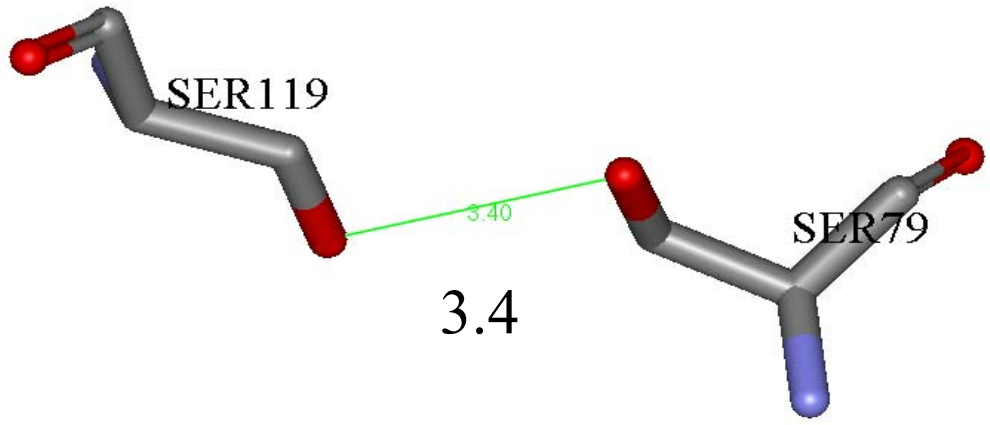
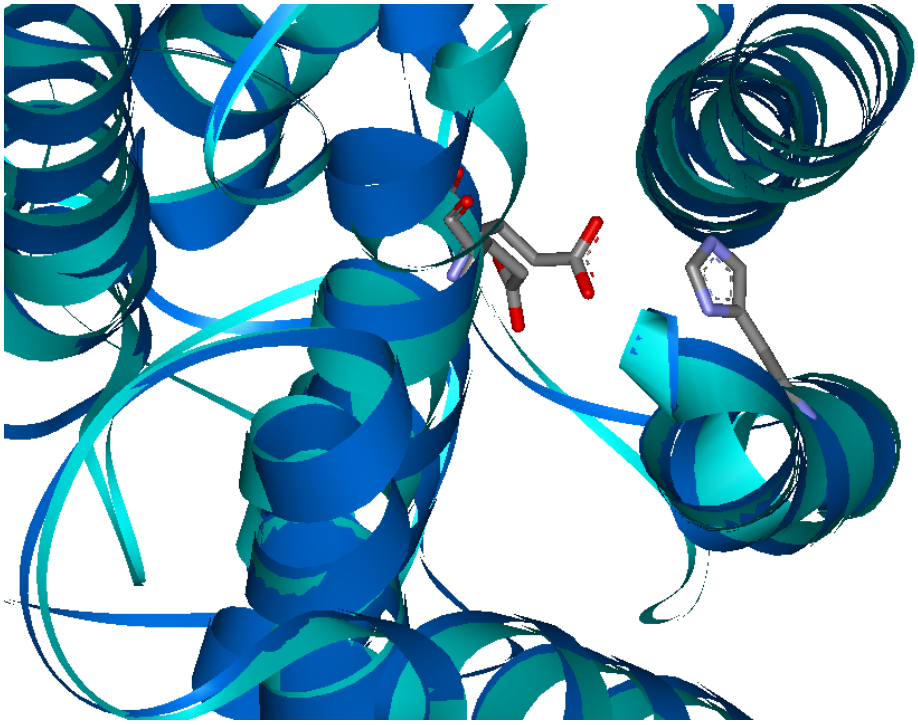


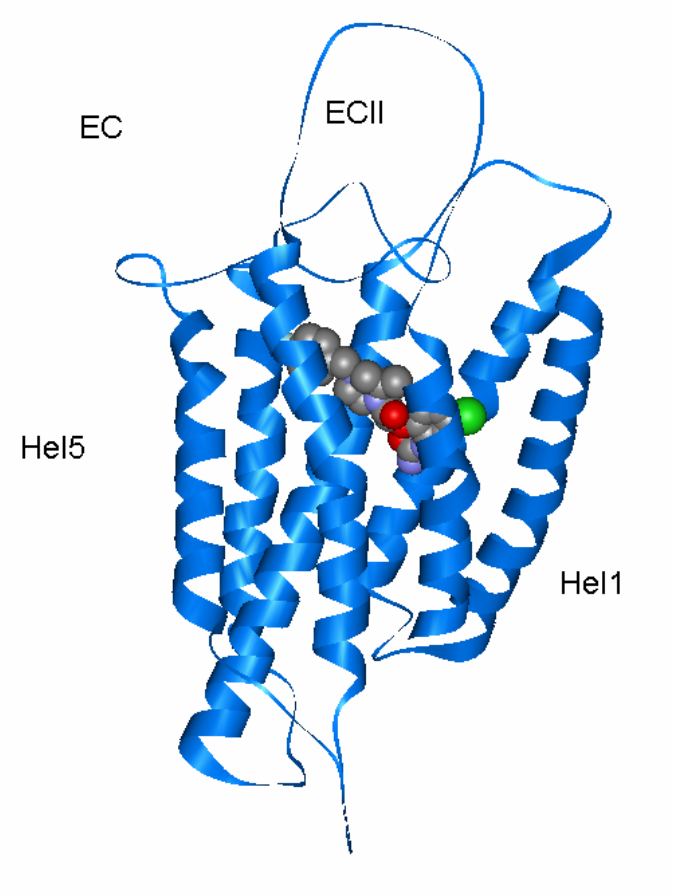
Figure 16: The helix 3/7 salt bridge formed between the Hel3 Glu120 and Hel7 Hsp293.



Binding site of BX471 in human CCR1

The optimized binding site of BX471 within the human CCR1 structure is shown in Figure 17. The details of the binding site are shown in Figure 18. The Glu 120 interacts with the tertiary amine of the ligand although this interaction may be weaker and is probably a water-mediated interaction. Of the other polar interactions, the Hsp293 interacts with the carbonyl and ether oxygens of the ligand very well while the amide nitrogen interacts with the Asp80 of helix 2 through a longer distance interaction. Various hydrophobic residues interact via hydrophobic interactions with the rings of the ligand. As can be seen, the ligand does not disrupt the polar interactions with keep the helices fixed relative to one another, a good feature for a high affinity antagonist.

Figure 17: General front and top views of the BX471 binding site in human CCR1.



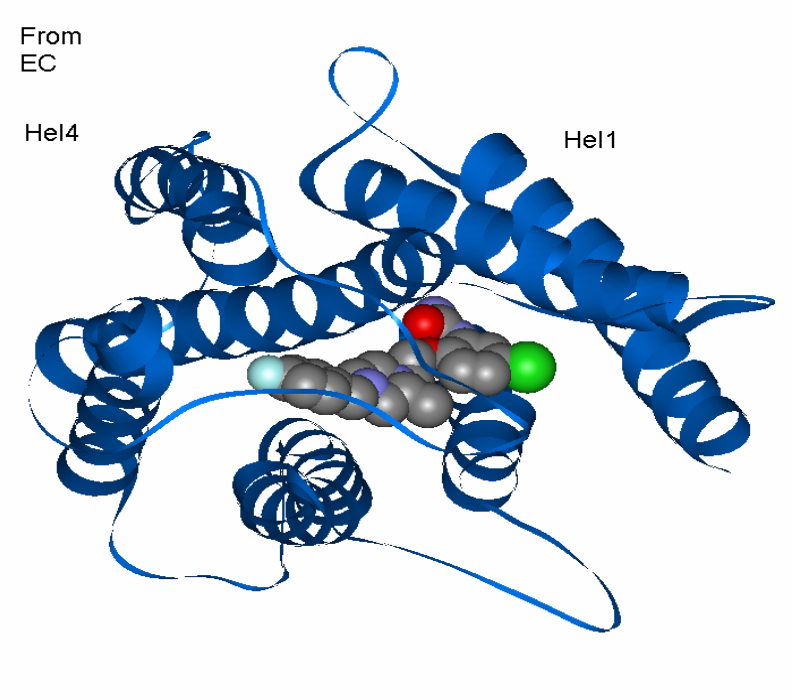
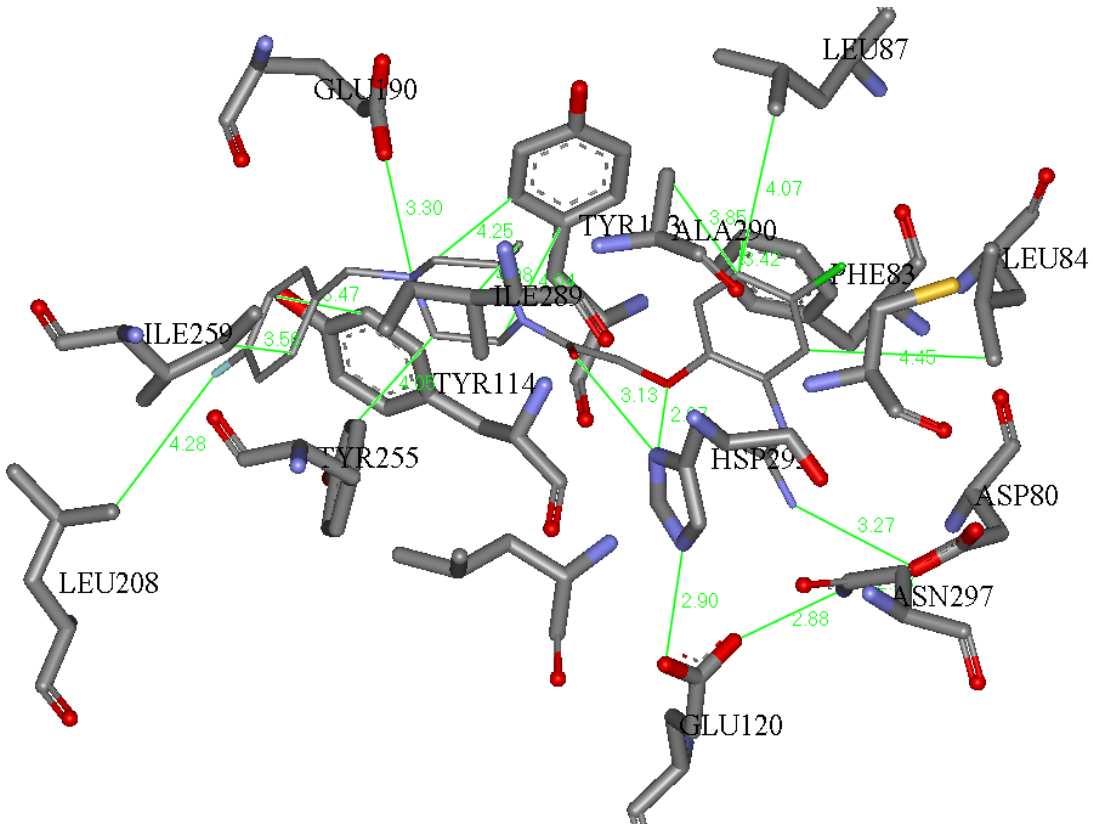


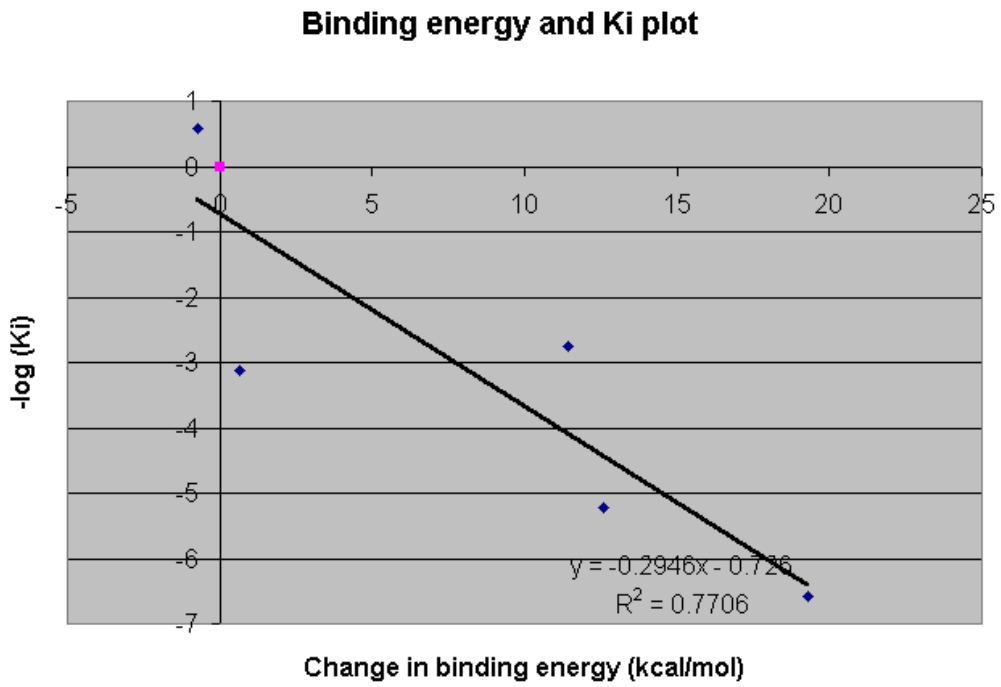
Figure 18: Detailed view of the BX471 binding site in human CCR1.



Rank ordering of the 6 compounds

The rank ordering of the 6 compounds (five are proprietary) is shown in Figure 19. The worst binder actually had internal strain and it is as of yet unclear whether it may prefer another different binding mode (a common problem with the bad binders).

Figure 19: Rank ordering for the 6 compounds in CCR1.



Mouse/Human structural differences

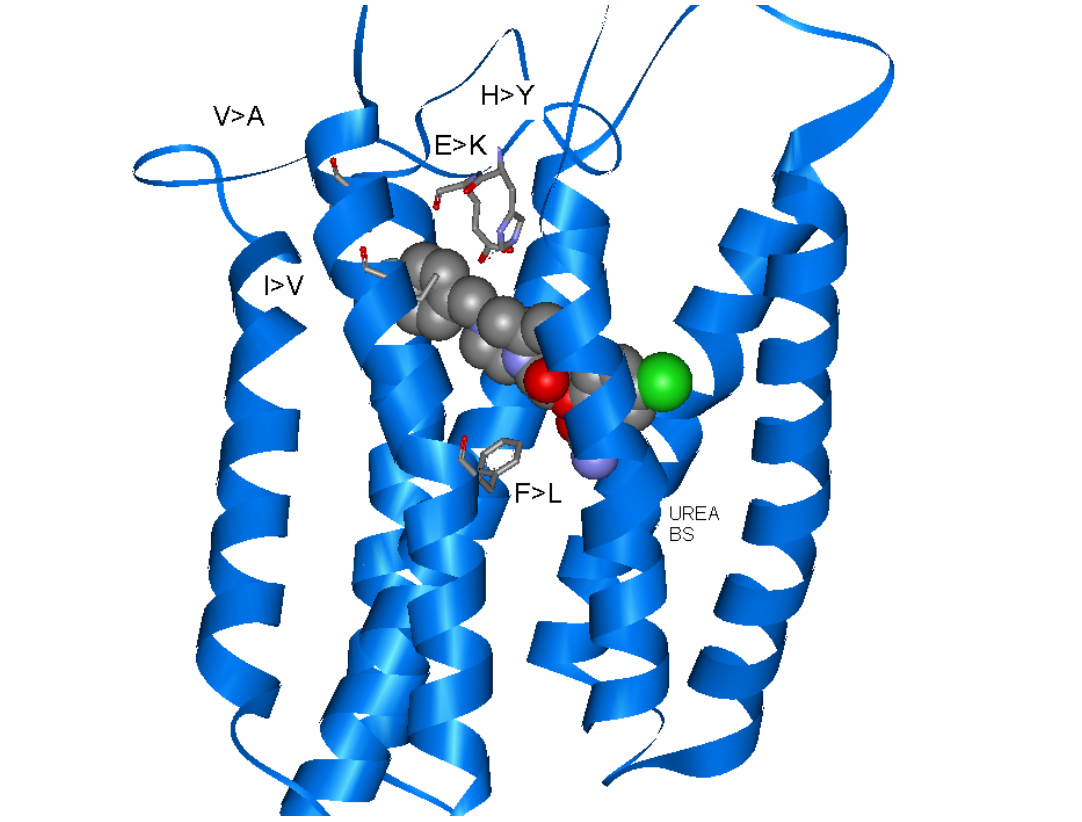
As discussed previously, the binding affinity of BX471 is greatly reduced in the mouse receptor. The MembStruk structure provides insight into the structural cause of the difference. The alignment of Human/Mouse sequences is shown in Figure 20. In Figure 21, the differences in the 5 Å binding site are shown in detail. The greatest difference is in the ECII loop, with the Glu120 as a Lys in the Mouse, which may have electrostatic repulsion with the bound ligand and thus may explain the low affinity for the mouse

receptor. Another notable difference is the Valine to Ala at the end of helix 6, although this residue does not interact greatly with the ligand energetically.

Figure 20: Alignment of human and mouse CCR1 sequences. The residues in bold are within 5 Å of the bound ligand, while those residues in red are different between mouse and human. The TM helices are underlined.

```
sp|P32246|CKR1_HUMAN            METPNTTEDYDTTTEFDYGDATPCQKVNERAFGAQLLPPLYSLVFVIGLV
gi|6857771|ref|NP_034042.1|     MEISDFTEAYPTTTEFDYGDSTPCQKTAVRAFGAGLLPPLYSLVFIIGVV
                                ** .: ** * *********:*****.  ***** **********:**:*

sp|P32246|CKR1_HUMAN            GNILVVLVLVQYKRLKNMTSIYLLNLAISDLLFLFTLPFWIDYKLKDDWV
gi|6857771|ref|NP_034042.1|     GNVLMILVLMQHRRLQSMTSIYLFNLAVSDLVFLFTLPFWIDYKLKDDWI
                                **:*::***:*::**:.******:***:***:*****************:

sp|P32246|CKR1_HUMAN            FGDAMCKILSGFYYTGLYSEIFFIILLTIDRYLAIVHAVFALRARTVTFG
gi|6857771|ref|NP_034042.1|     FGDAMCKLLSGFYYLGLYSEIFFIILLTIDRYLAIVHAVFALRARTVTLG
                                *******:****** **********************************:*

sp|P32246|CKR1_HUMAN            VITSIIIWALAILASMPGLYFSKTQWEFTHHTCSLHFPHESLREWKLFQA
gi|6857771|ref|NP_034042.1|     IITSIITWALAILASMPALYFFKAQWEFTHRTCSPHFPYKSLKQWKRFQA
                                :***** **********.*** *:******:*** ***::**::** ***

sp|P32246|CKR1_HUMAN            LKLNLFGLVLPLLVMIICYTGIIKILLRRPNEKKSKAVRLIFVIMIIFFL
gi|6857771|ref|NP_034042.1|     LKLNLLGLILPLLVMIICYAGIIRILLRRPSEKKVKAVRLIFAITLLFFL
                                *****:**:**********:***:******.*** *******.* ::***

sp|P32246|CKR1_HUMAN            FWTPYNLTILISVFQDFLFTHECEQSRHLDLAVQVTEVIAYTHCCVNPVI
gi|6857771|ref|NP_034042.1|     LWTPYNLSVFVSAFQDVLFTNQCEQSKHLDLAMQVTEVIAYTHCCVNPII
                                :******::::*.***.***::****:*****:***************:*

sp|P32246|CKR1_HUMAN            YAFVGERFRKYLRQLFHRRVAVHLVKWLPFLSVDRLERVSSTSPSTGEHE
gi|6857771|ref|NP_034042.1|     YVFVGERFWKYLRQLFQRHVAIPLAKWLPFLSVDQLERTSSISPSTGEHE
                                *.****** *******:*:**: *.**********:***.** ********

sp|P32246|CKR1_HUMAN            LSAGF
gi|6857771|ref|NP_034042.1|     LSAGF
                                *****
```

Figure 21: The residues within 5 Å of the bound ligand which are different between mouse and human CCR1. The mouse residue is shown second.



CCR5 function prediction

The putative binding sites of Tak-779 (shown in Figure 22) in the homology model of CCR5 were determined as above, with the exception that the spheres were divided into sphere clusters using Pass instead of dividing into separate boxes. The two best binding sites are shown in Figure 23. The second binding site conformation was used to extract the spheres (in the vicinity of the ligand) for fine HierDock. The final HierDock conformation for Tak-779 and four other published ligands (sch350634, sch351125, trans-pyrrolidine1, cis-pyrrolidine1) are shown in Figure 24. A comparison of the binding sites of BX471 in CCR1 and Tak-779 in CCR5 is shown in Figure 25. It can be seen that BX471 interacts with TMs 2, 3, 5, 6 and 7 while Tak-779 interacts with TMs 1, 2, 3, and 7 only. A detailed picture of the Tak-779 binding site is shown in Figure 26. The correlation with mutation studies (Dragic et al., 2000; Kazmierski et al., 2003) is good. The residues which have been shown to interact strongly by mutation are colored red. Tyr 108 forms a hydrogen bond with the carbonyl oxygen of the ligand and also has van der Waals interactions with the ring. The Trp86 and Tyr37 interact via pi-cation interactions with the ligand's quaternary amine and also provide hydrophobic interactions with the carbon units of the ligand. Of the medium experimental interactors, shown in green in green, Glu 283 interacts via a long distance interaction with the quaternary amine, as is expected for a quaternary amine.

The important differences between CCR1 and CCR5 are the presence of more polar residues on TM's 4 and 5 [Y170->I164 (TM4); N204->V199 (TM5)] and the presence of more nonpolar residues on TM's 6 and 7 [V263->T259 (TM6); V288->T284 (TM7)] for CCR1. To reduce cross-reactivity to CCR5 receptor, polar group may be

added onto the ligand where it interacts with helices 4 and 5. This demonstrates the use of the 3D structures to reduce cross-reactivity.

Figure 22: CCR5 antagonist Tak-779.

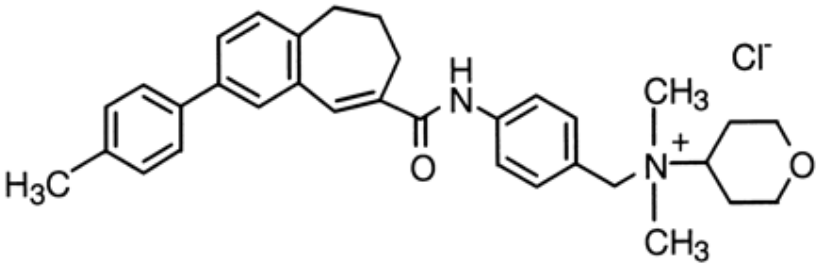


Figure 23: Two possible binding sites of Tak-779 within human CCR5.

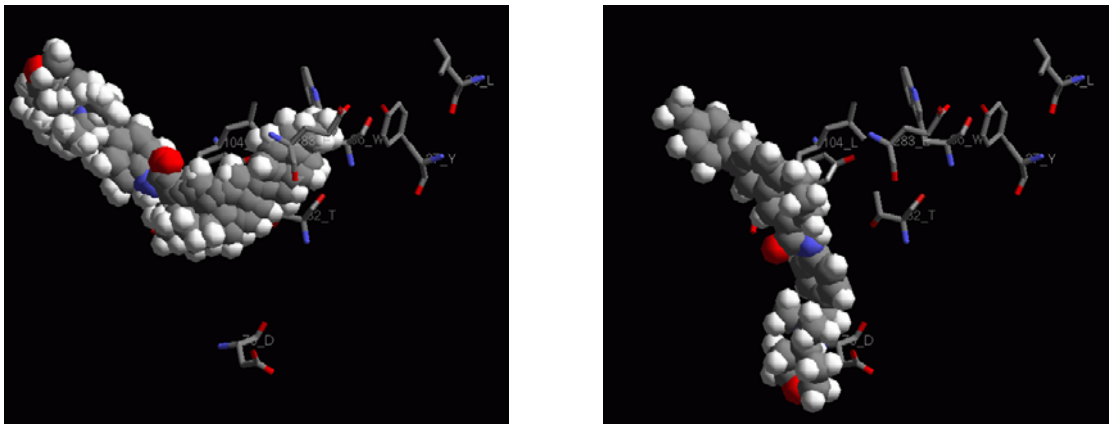


Figure 24: Binding sites of Tak-779 and 4 other compounds after HierDock.

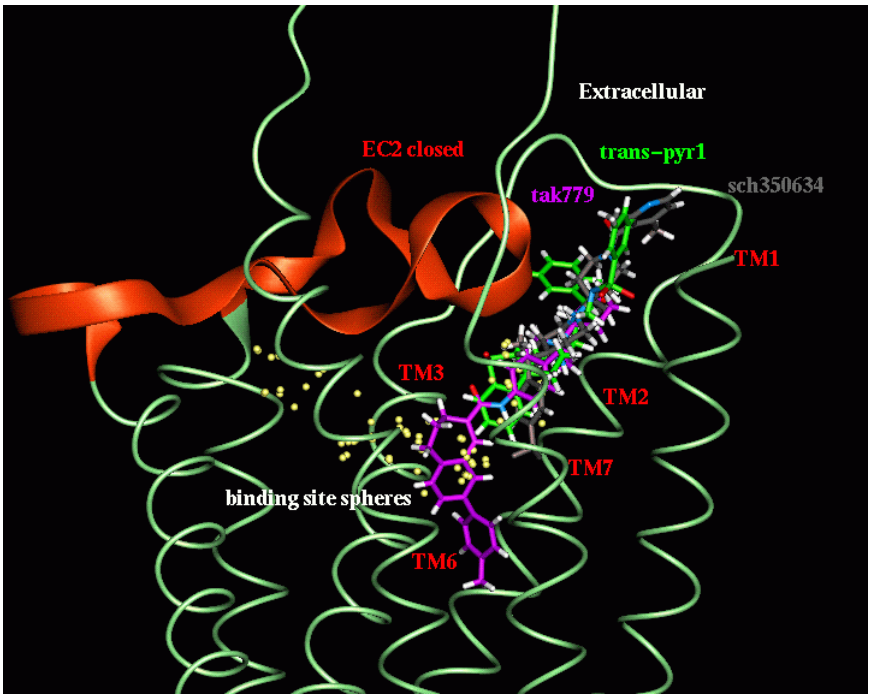


Figure 25: A comparison between the binding sites of BX471 and Tak-779.

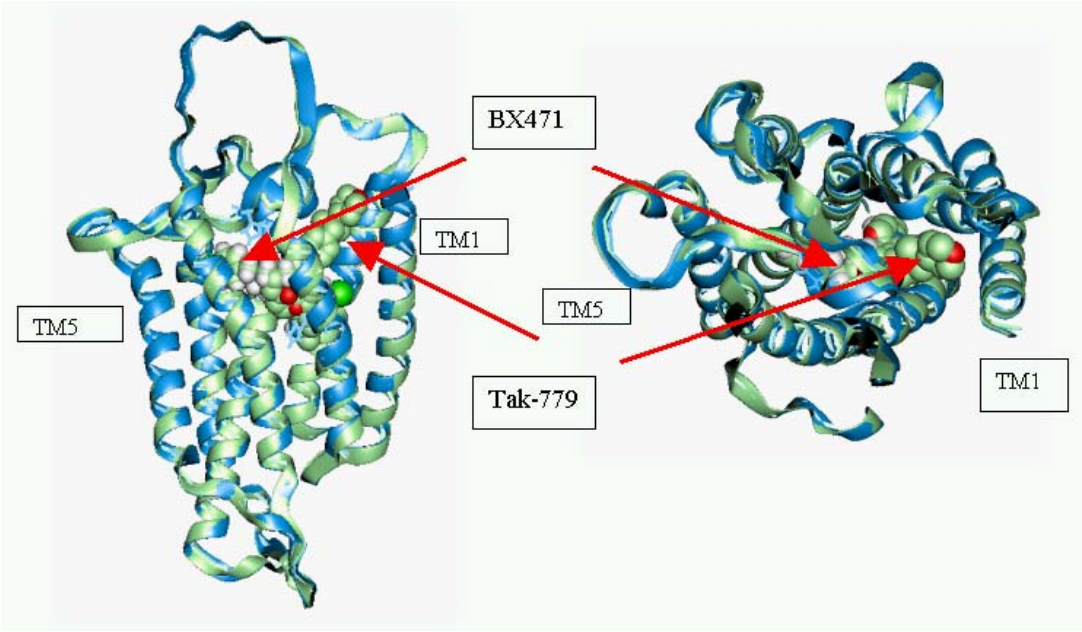
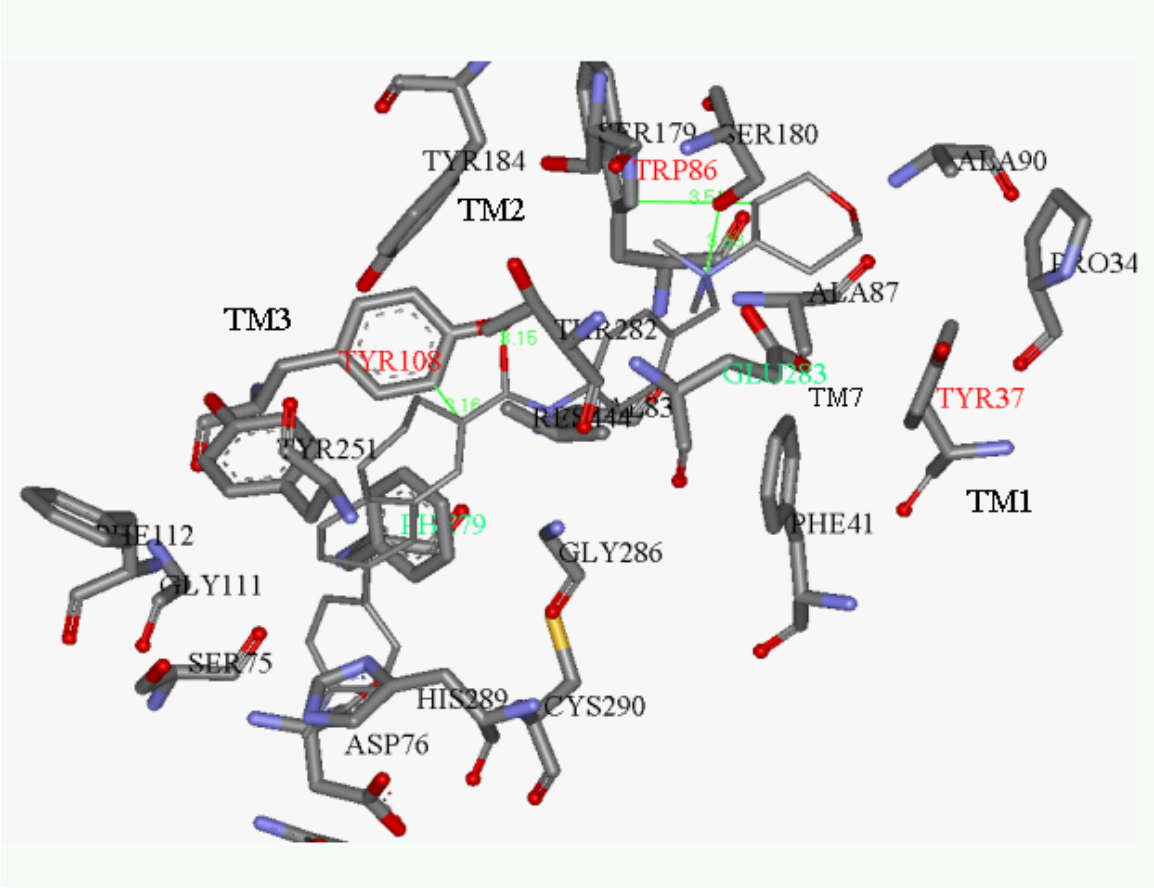


Figure 26: Detailed view of the Tak-779 binding site within human CCR5.



Conclusions

The structure of CCR1 exhibits various interactions interhelically and with the bound ligand. The rank ordering of a series of compounds is determined correctly based on this structure. These provide validation not only of the structure but also of the MembStruk protocol in general. In addition, the homology model of CCR5 from this CCR1 structure also has binding to the Tak-779 ligand which correlates well with experimental mutation studies. This provides indirect validation of the CCR1 structure as well as providing insights into CCR5 features important for binding such quaternary amine ligands. The CCR1 structure also provides information as to the possible differences in the structure which may lead to differences in binding affinity in the mouse and human. Such 3D structures provide a wealth of information needed to explaining binding of known ligands as well as provide an avenue for designing new drugs with desired selectivities across species (mouse and human) and receptors (CCR1 vs CCR5).

References

Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W. and D.J., L. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.

Altschul, S. F.; W.;, G.; Miller, W.; Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

Archer, E.; Maigret, B.; Escrieut, C.; Pradayrol, L. and Fourmy, D. (2003). Rhodopsin crystal: new template yielding realistic models of G-protein-coupled receptors? *Trends Pharmacol. Sci.* 24, 36-40.

Dragic, T.; Trkola, A.; Thompson, D. A. D.; Cormier, E. G.; Kajumo, F. A.; Maxwell, E.; Lin, S. W.; Ying, W.; Smith, S. O.; Sakmar, T. P. and Moore, J. P. (2000). A binding pocket for a small molecule inhibitor of HIV-1 entry within the transmembane helices of CCR5. *Proc. Natl Acad. Sci. U.S.A.* 97, 5639-5644.

Farrens, D. L.; Altenbach, C.; Yang, K.; Hubbell, W. L. and Khorana, H. B. (1996). Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin. *Science* 274, 768-770.

Floriano, W. B.; Vaidehi, N.; Singer, M.; Shepherd, G. and Goddard III, W. A. (2000). Molecular mechanisms underlying differential odor responses of a mouse olfactory receptor. *Proc. Natl Acad. Sci. U.S.A.* 97, 10712-10716.

Gerard, C. R., B.J. (2001). Chemokines and disease. *Nature Immunology* 2, 108-115.

Hesselgesser, J.; Ng, H. P.; Liang, M.; Zheng, W.; May, K.; Bauman, J. G.; Monahan, S.; Islam, I.; Wei, G. P.; Ghannam, A.; Taub, D. D.; Rosser, M.; Snider, R. M.; Morrissey, M. M.; Perez, H. D. and Horuk, R. (1998). *J. Biol. Chem.* 273, 15687-15692.

Horuk, R. (2003). Development and evaluation of pharmacological agents targeting chemokine receptors. *Methods* 29, 369-375.

Houshmand, P. and Zlotnik, A. (2003). Therapeutic applications in the chemokine superfamily. *Currect Opinion in Chemical Biology* 7, 457-460.

Kazmierski, W.; Bifulco, N.; Yang, H.; Boone, L.; DeAnda, F.; Watson, C. and Kenakin, T. (2003). Recent Progress in Discovery of Small-Molecule CCR5 chemokine receptor ligands as HIV-1 inhibitors. *Bioorganic and Medicinal Chemistry* 11, 2663-2676.

Liang, M.; Mallari, C.; Rosser, M.; Ng, H. P.; May, K.; Monahan, S.; Bauman, J. G.; Islam, I.; Ghannam, A.; Buckman, B.; Shaw, K.; Wei, G. P.; Xu, W.; Zhao, Z.; Ho, E.;

Shen, J.; Oanh, H.; Subramanyan, B.; Vergona, R.; Taub, D.; Dunning, L.; Harvey, S.; Snider, R. M.; Hesselgesser, J.; Morrissey, M. M. and Perez, H. D. (2000). *J. Biol. Chem.* 275, 19000-19008.

Lim, K.-T.; Brunett, S.; Iotov, M.; McClurg, R. B.; Vaidehi, N.; Dasgupta, S.; Taylor, S. and Goddard III, W. A. (1997). Molecular dynamics for very large systems on massively parallel computers: The MPSim program. *J. Comput. Chem.* 18, 501-521.

Mackay, C. R. (2001). Chemokines: immunology's high impact factors. *Nature Immunology* 2, 95-101.

Onuffer, J. and Horuk, R. (2002). Chemokines, chemokine receptors and small-molecule antagonists: recent developments. *TRENDS in Biochemical Sciences* 23, 459-467.

Onuffer, J.; McCarrick, M. A.; Dunning, L.; Liang, M.; Rosser, M.; Wei, G.-P.; Ng, H. and Horuk, R. (2003). Structure function differences in nonpeptide CCR1 antagonists for human and mouse CCR1. *The Journal of Immunology* 170, 1910-1916.

Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C.; Motoshima, H.; Fox, B.; Trong, I.; Teller, D.; Okada, T.; Stenkamp, R.; Yamamoto, M. and Miyano, M. (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* 289, 739-745.

Tagat, J. R.; Steensma, R. W.; McCombie, S. W.; Nazareno, D. V.; Lin, S. I.; Neustadt, B. R.; Cox, K.; Xu, S.; Wojcik, L.; Murray, M. G.; Vantuno, N.; Baroudy, B. M. and Strizki, J. M. (2001). 44, 3343-3346.

Thelen, M. (2001). Dancing to the tune of chemokines. *Nature Immunology* 2, 129-134.

Thompson, J. D.; Higgins, D. G. and Gibson, T. J. (1994). Clustal-W - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673-4680.

Trabanino, R. J.; Hall, S. E.; Vaidehi, N.; Floriano, W. B.; Kam, V. W. T. and Goddard III, W. A. (2004). First principles predictions of the structure and function of G-protein-coupled receptors: validation for bovine rhodopsin. *Biophys. J.* 86, 1904-1921.

Vaidehi, N.; Floriano, W. B.; Trabanino, R.; Hall, S. E.; Freddolino, P.; Choi, E. J.; Zamanakos, G. and Goddard, W. A. (2002). Prediction of structure and function of G protein-coupled receptors. *Proc. Natl Acad. Sci. U.S.A.* 99, 12622-12627.

Appendix

Appendix: High resolution tree diagrams displaying the relationships between TM cores of human GPCRs.