Computational studies on seven-helix transmembrane receptors

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Chuong Ha Hung Nguyen, B.Sc.

aus

Lam Dong, Vietnam

Berichter:
Universitätsprofessor Dr. Paolo Carloni
Universitätsprofessor Dr. Paolo Ruggerone

Tag der mündlichen Prüfung: 26.02.2013
Diese Dissertation ist auf den Internetseiten der Hochschulbibliothek online verfügbar

Advisors:

Prof. Dr. Paolo Carloni
Prof. Dr. Alejandro Giorgetti

PhD committee:

Prof. Dr.	Carsten Honerkamp	(Chair)
Prof. Dr.	Paolo Carloni	(Reporter)
Prof. Dr.	Paolo Ruggerone	(Reporter)
Prof. Dr.	Jörg Fitter	(Examiner)

I hereby affirm that I composed this work independently and used no other than the specified sources and tools and that I marked all quotes as such.

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Aachen, den 22.10.2012

To my parents and my brother

"When the choice is to be right or to be kind,
always make the choice that brings peace"
Wayne W. Dyer

Abstract

Seven transmembrane receptors (7TMRs) represent the largest family of integral membrane proteins across the three kingdoms of life. They are involved in the vast majority (∼80%) of signal transduction pathways across the cell membrane. They respond both to the electromagnetic field and ligands. Very importantly, more than one third of all current drugs target human 7TMRs, called G-protein coupled receptors (GPCRs). In recognition of the importance of this protein family, the 2012 Nobel Prize for Chemistry has been recently awarded to Robert J. Lefkowitz and Brian K. Kobilka for "their studies of G-protein coupled receptors."

However, detailed structural data on cellular pathways of 7TMRs is still limited. Hence, computational methods play an important role for structural predictions of 7TMRs as well as their binding to ligands. This thesis is devoted to characterize structure-function relationships in 7TMRs using a variety of computational approaches. We focus not only on the 7TMRs but also on their downstream effectors.

First, we have used classical molecular dynamics (MD) to investigate different forms of the Sensory Rhodopsin II in *Halobacterium salinarium* (*Hs*SRII). This receptor functions as a light sensor mediating repellent responses to the blue-green light. The results allow us to suggest that: (i) the alteration of charge distribution and the rearrangement of water molecules in the active site along with the interhelical hydrogen bond interactions play a key role for conformational changes associated to light absorption; (ii) two extracellular loops, a cytoplasmic loop and the C-terminus may participate to the activation of *Hs*SRII.

Next, we have developed a hybrid molecular simulation method called MM/CG to study 7TMR/ligand interactions. The accuracy of the method has been established by comparison with recent atomistic MD simulations on the 7TMRs, *Hs*SRII and human β_2 Adrenergic Receptor. The MM/CG simulations turn out to reproduce the key structural features of the active site, as found also in MD simulations carried

out by us and another research group lead by Prof. Ursula Röthlisberger (EPFL). Due to its low computational cost, this method can be applied to the study of a large number of GPCR-ligand complexes.

As the final step of the thesis, we have extended the study to 7TMR's downstream effectors including Adenylyl Cyclase III (ACIII) and Cyclic Nucleotide-Gated (CNG) ion channel, two proteins involved in the olfactory signaling cascade. By using homology modeling approaches, we have built (i) the cytoplasmic domain structure of human ACIII - in complex with human Gα-subunit of olfactory system, and (ii) the P helix, S6 and the C-linker of CNG channel in the closed and open states. Specially, in the work of CNG channel, in combination with experimental studies, we have proposed the initial conformational changes of the C-linker, S6 and P-helix upon cyclic nucleotide binding.

Acknowledgments

First of all, I would like to thank my supervisor Prof. Paolo Carloni for giving me the opportunity to work on a fascinating topic, for helping me to stay on track and complete it, and specially for his patience with me.

I wish to thank my co-supervisor Prof. Alejandro Giorgetti for his constant support for my work from the very beginning, and for guiding the progress of my work with countless helpful discussions.

My special thanks to Xevi Biarnés, Paola Lupieri, Zhaleh Ghaemi Bafghi, Emiliano Ippoliti, Chao Zhang, Giulia Rossetti, Michael Leguèbe, Luciana Capece for working together with me and providing me many useful discussions. Many thanks to our experimental collaborators: Prof. Vincent Torre and his group from SISSA, Prof. Jörg Labahn from Forschungszentrum Jülich and Illy Coffee company.

My big thanks to Salvatore Bongarzone, Fahimeh Baftizadeh Baghal, Agata Kranjc, Claudia Barrera, Marcos Brown Gonçalves for being the best office colleagues I can think of. To all the members and staff of the Computational Biophysics group from GRS and of the Statistical and Biological Physics sector from SISSA, thank you for your assistance and for making my work enjoyable.

To my former supervisors, Prof. Hoang Zung and Do Hoang Son, thank you for the constant support and advice you have been giving me.

To all my friends, thank you for being there and making my time so enjoyable.

Last, but not least, I would like to thank my family; Mom, Dad and my brother. You have been supportive of me from day one of my academic career. Whether it be financial or emotional support, you have been there all the time! Without you, I wouldn't have made it to where I am today!

Contents

Glossary

β_2AR β_2 Adrenergic Receptor. 6, 86

3D three-dimensional. 15, 20, 24, 74

7TMR seven-transmembrane receptor. 1–10, 12, 42, 74, 85, 86

AC Adenylyl Cyclase. 5, 6, 13, 75–78, 86

cAMP cyclic adenosine monophosphate. 75

CG coarse-grained. 5

CNG Cyclic Nucleotide-Gated. 5, 6, 13, 76, 78–81, 83

CSM cysteine scanning mutagenesis. 76, 79, 80

FF force field. 92–94

GPCR G-protein coupled receptor. 1–5, 7, 8, 11, 86

GS ground state. 42, 44, 45, 48–54

KS K state. 42

MD molecular dynamics. 3–6, 23, 28–30, 32, 37, 56, 96, 100

MM/CG molecular mechanics/coarse-grained. 4–6, 74, 85, 86

MS M state. 42, 44, 45, 48–54

NMR nuclear magnetic resonance. 19, 20, 22, 25

OR olfactory receptor. 11–13, 75, 77

PBC periodic boundary conditions. 32, 34, 35, 37

PDB Protein Data Bank. 16

PME particle mess Ewald. 37

RMSD root-mean-square deviation. 24, 48, 69

RMSF root-mean-square fluctuation. 69

ScB Schiff-base. 42, 44–46, 48–50, 54

SR Sensory Rhodopsin. 4, 6, 10, 86

TM transmembrane. 1, 76

CHAPTER 1

Introduction and Motivation

Cells may recognize and respond to diverse stimuli from their environment using membrane receptor proteins (Lodish *et al.*, 2007). Their responses are triggered by different signaling molecules as well as the electromagnetic field in which they are immersed. Typically, these receptors comprise an extracellular domain, a transmembrane domain and an intracellular domain. Numerous receptors, which are constructed on this basic architectural principle, are present in the three domains of life including eukaryotes, archaea and bacteria (Lodish *et al.*, 2007). Among those, there are the seven-transmembrane receptors (7TMRs) which consist of seven transmembrane (TM) helices in the membrane domain, an extracellular N-terminus and an intracellular C-terminus (Shenoy, 2007) (Fig. 1.1). They connect cells with their environment, sensing light and chemicals.

In eukaryotes, 7TMRs are called G-protein coupled receptors (GPCRs). Their signaling pathways are predominantly based on the interaction with their cognate G-proteins (Rosenbaum *et al.*, 2009; Lefkowitz, 2007; Pierce *et al.*, 2002). There is also some evidence for G-protein-independent pathways downstream of 7TMRs in animals and plants (Pierce *et al.*, 2002). The structural scaffold of the 7TMRs possesses a great degree of flexibility that allows them to sense a remarkable diversity of ligands, such as odorants, in animals (Spehr & Munger, 2009; Zarzo, 2007). They constitute the largest receptor superfamily in vertebrates and other metazoans (Lu *et al.*, 2009; Römpler *et al.*, 2007). Approximately 3-4% of mammalian genes encode GPCRs (Römpler *et al.*, 2007). Human GPCRs are involved in as many

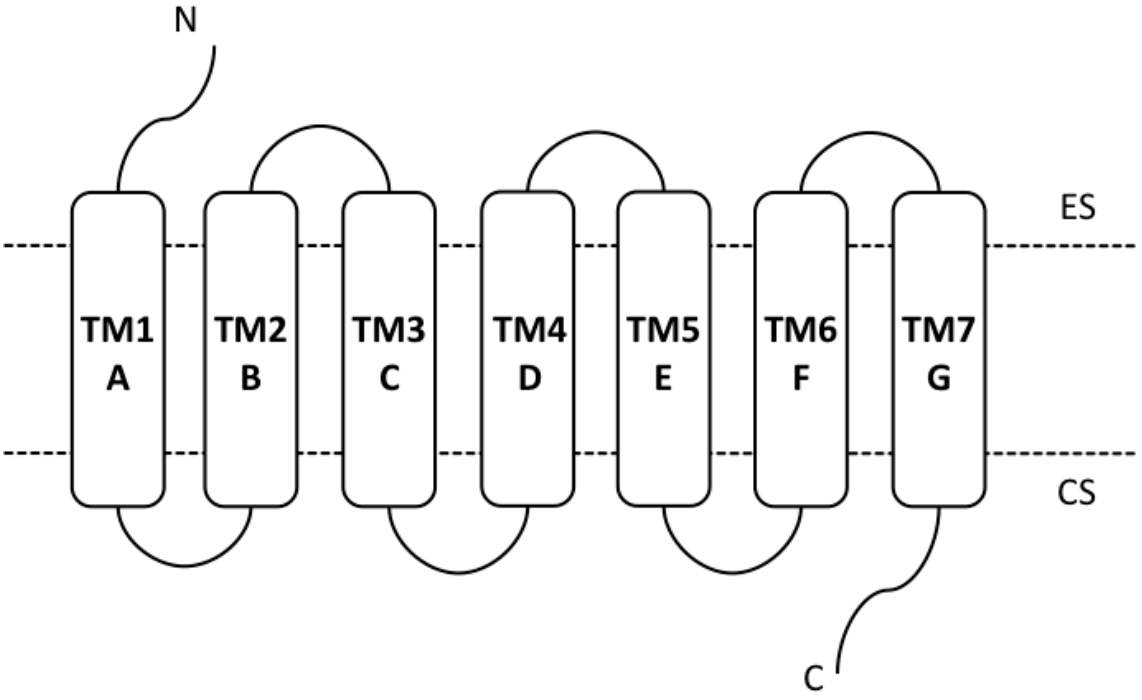


Figure 1.1 Two-dimensional topology representation of 7TMRs. Helices are usually labeled from TM1 to TM7 (in case of eukaryotic 7TMRs) or A to G (in case of bacterial and archaeal 7TMR). ES, extracellular side; CS, cytoplasmic side. The dotted black lines show the boundary of lipid bilayer.

as ∼80% of the signal transduction pathways across the cell membrane (Millar & Newton, 2010). Very importantly, more than one third of all current therapeutics target GPCRs (Overington *et al.*, 2006; Eglen & Reisine, 2009). In recognition of the importance of the GPCR protein family, the 2012 Nobel Prize for Chemistry has been recently awarded to Robert J. Lefkowitz and Brian K. Kobilka for "their studies of G-protein coupled receptors" (`http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2012/`).

7TMRs are also present in bacteria and archaea (Sharma *et al.*, 2006; Klare *et al.*, 2008). Despite the fact that the evolutionary relationship across eukaryotes and prokaryotes is unproven (Soppa, 1994), 7TMRs share structural and mechanistic similarities. For example, structural comparisons between the animal rhodopsins and the microbial rhodopsins reveal that they adopt essentially the same topology and three-dimensional fold (Spudich *et al.*, 2000). This suggests that they have most probably descended from a common ancestor despite extensive divergence of their sequence (Anantharaman & Aravind, 2003). Functions of 7TMRs in bacteria and archaea include light-driven ion transport and phototaxis signaling (Spudich *et al.*, 2000).

Understanding structure-function relationships of 7TMRs in basic physiology and pathophysiolgy is a fascinating challenge for biophysicists. In the case of human GPCRs, it may lead to the development of new therapeutic molecules (Kufareva *et al.*, 2011). Unfortunately, ascertaining the structural determinants of 7TMRs is hampered by the paucity of experimentally structural information. This is due to the difficulties in obtaining and purifying large amounts of 7TMR required for X-ray diffraction or NMR spectroscopy (Langelaan *et al.*, 2011; Chollet & Turcatti, 1999). Experimental structures are currently available for 14 members from eukaryotes and 10 from prokaryotes, respectively (see Appendix A). Thus, computational approaches play a key role in elucidating the structure-function relationships in 7TMRs as well as their binding to ligands.

Very broadly speaking, computational methods applied to biological systems use two major strategies. The first one is based on the laws of physics. In particular, molecular dynamics (MD) is a computational technique that allows a detailed description of the dynamical properties of biomolecular systems (Frenkel & Smit, 2001). In classical MD simulations, the dynamics of a system is described by Newton's second law of motion, and the interactions of all the atoms present in the system are modeled. Once the interacting forces are calculated, Newton's equations of motion are integrated. Due to the high level of detail, sometimes, this technique is considered to represent an interface between laboratory experiments and theory, and can be understood as a "virtual experiment". MD probes the relationship between molecular structure, movement and function allowing a very detailed description of biological systems at the molecular level. A key ingredient of the method is of course the force field[1], which is usually constructed so as to reproduce experimental quantities and/or quantum chemical calculations (Leach, 2001; Schwede, 2008). The quality of the force fields is continuously improving, in parallel with the availability of more powerful computational resources. This allows one nowadays to produce stable trajectories covering a time scale of hundreds of nanosecond to microsecond, or even to sub-millisecond with special-purpose machine (Raval *et al.*, 2012). Several schemes have been developed to control the temperature and the pressure of the simulation, a feature that allows one to reproduce the physiological conditions under which processes take place or experiments are performed (Berendsen *et al.*, 1984; Quigley & Probert, 2004). In this way, MD enables the exploration of the conformational energy landscape accessible to biomolecules.

[1] A force field refers here to both the functional form of the potential energy used to calculate the forces and the collection of parameters used to define the potential energy of a system of particles (typically molecules and atoms).

The second major strategy uses biological concepts, such as Darwinian evolution and knowledge based algorithms. They are, together with algorithms taken from the theory of the information, used to increase the understanding of biological processes. This strategy takes the name of Bioinformatics. Common activities in Bioinformatics include mapping and analyzing sequences of DNA and protein, aligning different DNA and protein sequences to compare them, as well as creating and analyzing 3-D models of protein structures. The latter, structural bioinformatics, allows the prediction of the proteins 3D folds, aimed at the functional assessment of the different gene products. There exists different ways for modeling structures, depending most of all, on the existence of an evolutionary related protein with known structure. Homology modeling is an approach based on the observation that two proteins diverging from a common ancestor tend to have similar 3D structures (Chothia & Lesk, 1986). Thus, once a "template", that is an homologous protein with known structure is found, a reliable model of the protein of interest can be produced by using well established programs and algorithms (Chothia & Lesk, 1986; Tramontano & Lesk, 2006).

In this thesis, we have attempted at characterizing structure-function relationships in 7TMRs through the analysis using a variety of molecular simulation and bioinformatics approaches. We focus not only on the receptors but also on their downstream effectors.

The first investigated system is the 7TMR Sensory Rhodopsin (SR) II in *Halobacterium salinarium*. The SR receptors have been studied extensively because they provide a simple model system for phototaxis, one of the fundamental functions of light sensing (Spudich & Jung, 2005). Understanding how only few switches in the binding site can induce the structural change of the receptor is one of the utmost interests for these photoreceptors. Thus we have attempted to study the conformation differences between the inactive and active receptors using MD simulations. The work has been carried out in collaboration with the experimental group lead by Prof. Jörg Labahn from Forschungszentrum Jülich.

Next, we have developed a hybrid molecular simulation method called molecular mechanics/coarse-grained (MM/CG) (Neri *et al.*, 2005; Neri *et al.*, 2008) for structural predictions of 7TMRs/ligand complexes. The motivation of this development is due to the fact that structural determinants of ligand interaction are strikingly diverse between GPCRs (Kufareva *et al.*, 2011). It is difficult (if not, even nearly impossible) to expand the atomic details of ligand binding elucidated by crystallography to cover all members of the GPCR family (Kufareva *et al.*, 2011). Although

classical MD is often the method of choice (Yarnitzky *et al.*, 2010) to identify ligand poses on 7TMR models based on bioinformatics methods, the accuracy of MD structural predictions is limited by the low sequence identity (SI) between the template and the target structures (MacCallum *et al.*, 2011; Raval *et al.*, 2012). For instance, the SI is only ∼20% across the class A GPCRs (Imai & Fujita, 2004; Michino *et al.*, 2009). In addition, such an approach can be very CPU intensive (Sgourakis & Garcia, 2010), especially, if one deals with a large number of ligand/receptor complexes. In these cases, one can sometimes tackle the problem by using reduced representations, such as coarse-grained (CG) models. CG approaches allow to study larger systems in much longer timescale (Monticelli *et al.*, 2008). The reduction of the number of degrees of freedom makes the model computationally very efficient, allowing a reduction of the simulation time by ∼2 to ∼3 orders of magnitude compared to full atom force fields (Monticelli *et al.*, 2008). However, these CG approaches cannot describe the intermolecular ligand/protein interactions. A possible strategy to address this issue is to combine atomistic with CG modeling (Kalli *et al.*, 2011; Messer *et al.*, 2010; Shi *et al.*, 2006; Villa *et al.*, 2005; Neri *et al.*, 2005; Neri *et al.*, 2008). MM/CG simulation is an approach in which different representations of the system are modeled concurrently. A coupling scheme is used to connect the boundary of models. By doing this, one can take advantages of both methods and allow extending the investigation of molecular recognition events to time-scale much longer than those accessible by MD. The developed method could be applied now to a large number of 7TMR/ligand complexes and may reveal itself very useful for computer-aided drug design.

As a final step of my thesis, we have extended the study to GPCR's downstream effectors. Understanding the molecular basis of GPCRs requires the integration of biological complexes into cellular pathways. In fact, even subtle imbalances in GPCR signaling can lead to the disorders or diseases. For example, an increase of the β adrenergic receptor kinase (βARK1) expression and activity has been found in several cardiovascular diseases (Iaccarino & Koch, 1999). Here, we have investigated Adenylyl Cyclase (AC) III and Cyclic Nucleotide-Gated (CNG) ion channel, two important proteins involved in a GPCR signaling pathway, i.e. the olfactory signaling cascade. The mutation on these proteins have been shown to have severe effects on olfactory function (Su *et al.*, 2009). Unfortunately, there is no experimental structural information at the moment on these two proteins. Thus, we have attempted to predict the structural conformation of the two proteins using homology modeling approach.

My thesis is then organized as follows:

- Chapter 2 provides a biophysical background of the 7TMR family.
- Chapter 3 discusses about the computational methods used in the thesis, including two major approaches: (i) Homology Modeling; (ii) Molecular Dynamics simulations.
- Chapter 4 shows the MD study of the 7TMR SRII present in *Halobacterium salinarium* (*Hs*SRII). Our calculations allow us to suggest some key events taking place in the binding site which might play an important role in the activation of *Hs*SRII.
- Chapter 5 presents the development of hybrid MM/CG approach to 7TMRs. We then present the performance of the MM/CG approaches in two types of 7TMRs which are SRII and β_2 Adrenergic Receptor (β_2AR).
- Chapter 6 shows the structural prediction of two proteins involving in the cascade, ACIII and CNG ion channel, using homology modeling approach.
- Chapter 7 summarizes the some of the results from the work present here, and then provides some perspectives for future work.

CHAPTER 2

Biological background

2.1 Introduction

7TMRs are the most common superfamily of transmembrane receptors across the three kingdoms of life. Members of this superfamily are characterized by seven membrane-spanning α-helices connected by intracellular and extracellular loops, along with an extracellular amino terminus and a cytoplasmic carboxyl terminus. They connect cells with their environment and are responsible for approximately 80% of signal transduction across cell membranes (Millar & Newton, 2010).

7TMRs can be divided into two main families: microbial rhodopsins (Figure 2.1a) and GPCRs expressed by eukaryotes (Figure 2.1b) (Anantharaman & Aravind, 2003). The major difference between these two families is that GPCRs associate with members of heterotrimeric G-protein family to function whereas microbial rhodopsins do not (Gether, 2000). In spite of the fact that the evolutionary relationship across them remains unproven (Soppa, 1994), microbial rhodopsins and GPCRs share a number of structural and mechanistic features (Hirai *et al.*, 2009). Structural comparisons between bacteriorhodopsin (PDB code: 1M0L) and bovine rhodopsin (PDB code: 1U19) reveal that they consist of essentially the same three-dimensional fold (Hirai *et al.*, 2009) (Figure 2.1c). In addition, spectroscopic and chemical analyses show that a retinal molecule binds covalently to a lysine residue via

a Schiff base linkage in the seventh helix in both cases (Spudich *et al.*, 2000). Moreover, conformational changes in the receptors are induced by similar photochemical reactions indicating critical features shared by both systems (Spudich *et al.*, 2000).

Microbial rhodopsins function as light-dependent ion transport or signal transduction. The ion pumps include Bacteriorhodopsin (BR), Halorhodopsin (HR), Archaerhodopsins (aR-1, aR-2), Proteorhodopsin (PR), Xanthorhodopsin (XR), Acetabularia Rhodopsins (ARI, ARII). They operate as energy converters. Other photoreceptors such as Sensory rhodopsins (SRI, SRII), Channelrhodopsins (ChR1, ChR2) operate as light sensors modulating phototaxis, i.e. movement in response to light (Klare *et al.*, 2008; Spudich *et al.*, 2000). Meanwhile, GPCRs are involved in a large number of intra- and extracellular signaling pathways, such as detection of light, sense of smell, neurotransmission, inflammation, and cardiac and smooth muscle contractility (Katritch *et al.*, 2012; Pierce *et al.*, 2002; Grote & O'Malley, 2011). Ligand (or photon) binding to GPCRs activates a cascade of events, producing an electrical signal as output. They are involved in many diseases and represent the target of approximately 30% of all marketed drugs (Landry & Gies, 2008).

The microbial rhodopsins are the best understood proteins among 7TMRs with respect to structural information from X-ray crystallography. Bacteriorhodopsin is the first solved crystallographic structure of 7TMRs (Henderson & Unwin, 1975). Since then, many high resolution structures have already been determined including BR, HR, SRI, SRII and XR (Table A.1). In addition, the X-ray structure of the complex between *Natronobacterium pharaonis* SRII (*Np*SRII) and HtrII was solved not only in ground state (Gordeliy *et al.*, 2002) but also in active states (Moukhametzianov *et al.*, 2006). Recently, the first NMR structure of a 7TMR was also reported for SRII (Gautier *et al.*, 2010).

Solving structure of GPCRs is usually more challenging due to several technical bottlenecks. They include (i) the preparation of sufficiently large amounts of functional GPCR protein since most GPCRs are expressed at low levels in native tissues (Rosenbaum *et al.*, 2009), and (ii) overcoming thermodynamic and proteolytic protein-stability problems as a result of their disordered extramembranous loop regions (Rosenbaum *et al.*, 2009). Despite that, the past couple of years have seen several novel X-ray structures of GPCRs. Crystal structures are now available for rhodopsin, adrenergic, and adenosine receptors in both inactive and activated forms, as well as for chemokine, dopamine, histamine, sphingosine-1-phosphate, muscarinic acetylcholine, and opioid receptors in inactive conformations (Table A.2). The avail-

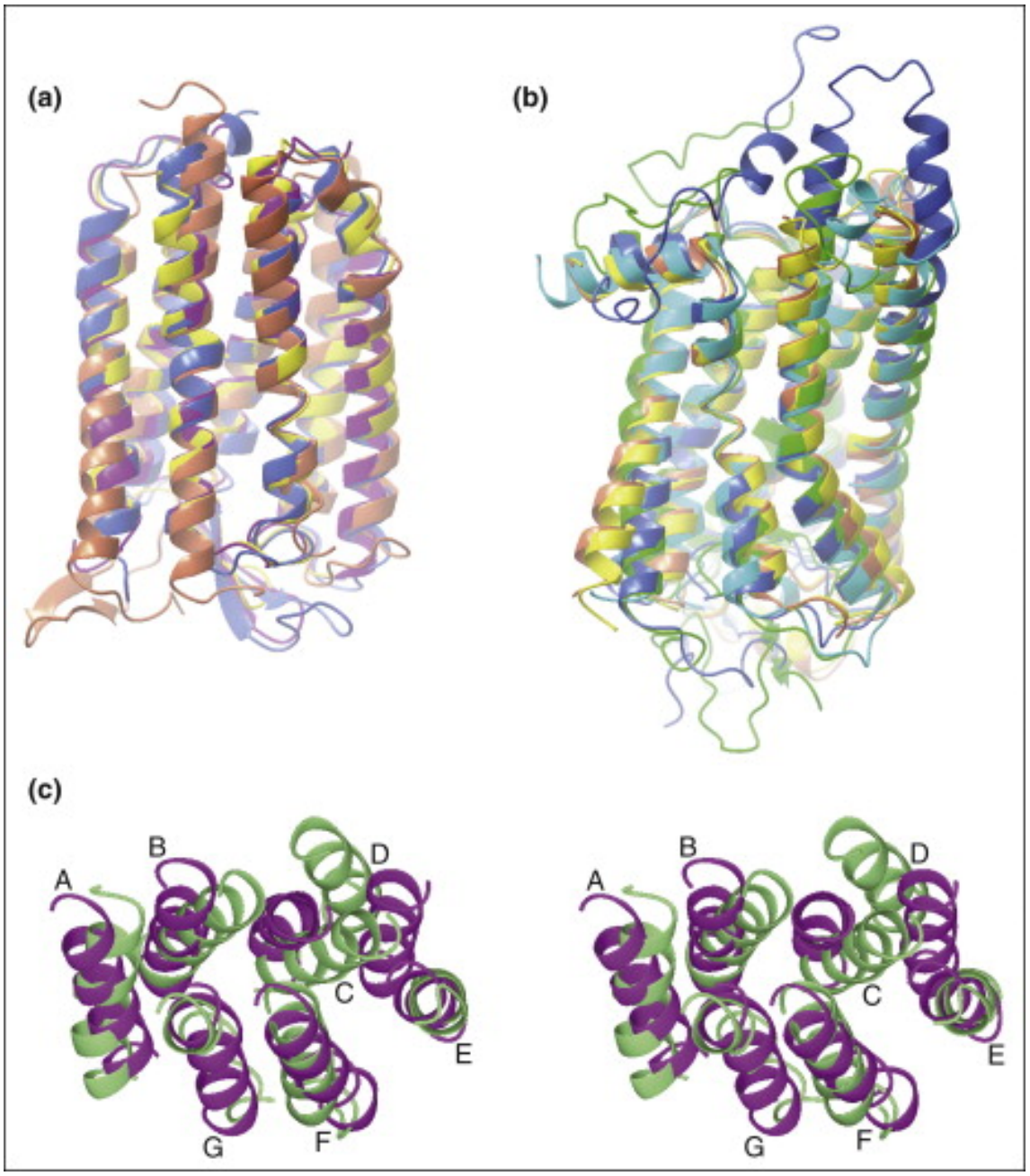


Figure 2.1 Structural determinants of selected 7TMRs. (a) Bacteriorhodopsin (PDB code; 1M0L) (Schobert *et al.*, 2002), halorhodopsin (1E12) (Kolbe *et al.*, 2000), sensory rhodopsin II (1JGJ) (Luecke *et al.*, 2001), and xanthorhodopsin (3DDL) (Luecke *et al.*, 2008) are aligned using residues 11-29, 43-59, 82-98, 110-124, 137-153, 170-189, and 205-223 of bacteriorhodopsin and equivalent residues of others and shown colored in blue, light green, yellow, and coral, respectively. (b) Bovine rhodopsin (PDB code; 1U19) (Palczewski *et al.*, 2000; Okada *et al.*, 2004), squid rhodopsin (2Z73) (Murakami & Kouyama, 2008), β2-adrenergic receptor (2RH1) (Cherezov *et al.*, 2007), β1-adrenergic receptor (2VT4) (Warne *et al.*, 2008), and A_{2A} adenosine receptor (3EML) (Jaakola *et al.*, 2008) are aligned using residues 37-62, 72-96, 109-137, 151-171, 201-223, 249-275, and 289-319 of bovine rhodopsin and equivalent residues of others and shown colored in green, light blue, yellow, coral, and cyan, respectively. (c) Structural comparisons of transmembrane regions of bacteriorhodopsin and bovine rhodopsin, as viewed from the cytoplasmic side. Residues 13-29, 43-59, 82-98, 110-124, 142-152, 170-189, and 205-222 of bacteriorhodopsin and residues of 44-60, 73-89, 117-133, 156-170, 213-223, 249-268, and 292-307 of bovine rhodopsin are aligned and shown colored in magenta and green, respectively in stereo. Figure adapted from (Hirai *et al.*, 2009) with permission from Elsevier.

able structural information of these 7TMRs promise an exciting time of studying them at molecular level.

2.2 Signal transduction by sensory rhodopsins

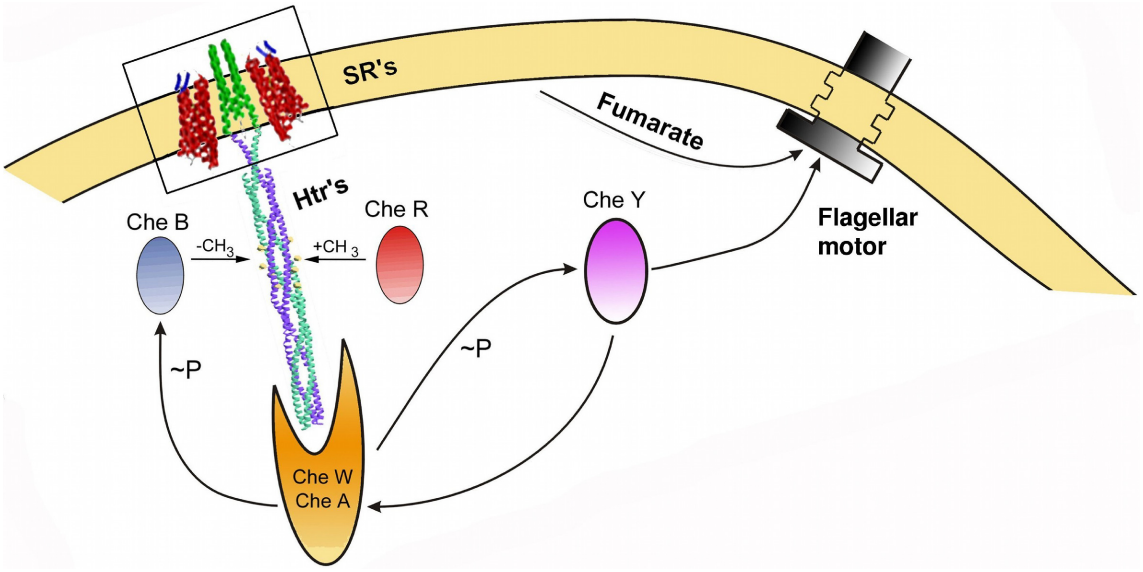


Figure 2.2 Signalling cascade for SRII response. The retinal of the receptor SRII performs upon absorption of a photon a transition from all-*trans* to 13-*cis*. This induces conformational changes within the receptor which are transferred to the transducer HtrII. Within the transducer the signal is conducted to the coupling protein CheW and the histidine kinase CheA via the coiled-coil structure formed by helices TM2 from both transducers. The next steps in the signalling cascade involve - in analogy to the bacterial sensory system - the phosphorylation of the response regulators CheY and CheB. Phosphorylated CheY functions as a switch factor of the flagellar motor. CheB, a methylesterase together with CheR, a methyltransferase are involved in the adaptation processes of the bacteria. The box highlights the receptor/transducer complex. Figure adapted from `http://www2.fz-juelich.de/isb/isb-2/members/bueldt/srII-htrII`.

One of the 7TMRs investigated in this thesis is a sensory rhodopsin present in archaea. *Halobacterium salinarium* utilizes the sensory rhodopsins (SRI and SRII) for phototaxis. SRI mediates attractant responses to the green/orange light whereas SRII mediates repellent responses to the blue-green light (Hoff *et al.*, 1997; Spudich & Luecke, 2002). Sensory rhodopsins (SRs) exist as complexes with their cognate integral membrane transducers (Htrs). SRI forms a signaling complex firmly with HtrI, while SRII forms one with HtrII. It should be noted that the HtrII in *Halobacterium salinarium* (HsHtrII) displays a dual function as a photophobic and serine receptor. The observation that HsHtrII displays a dual functionality is important

with respect to a common mechanism of signal transfer in phototaxis and chemotaxis.

The signal transduction is activated upon light absorption by a small molecule, retinal (vitamine A adehyde), which covalently bound to the sensory rhodopsins (Figure 2.2). The retinal isomerizes from all-*trans* to 13-*cis* form, causing a cyclic sequence of spectroscopically detectable intermediates. The production of these intermediates is accompanied by the conformational changes in the receptors. This conformational changes relay the signal to the cognate transducer molecules.

The photoreceptor/transducer complex interacts with a biochemical network centered around a two-component regulator consisting of two major proteins: CheA and CheY (Figure 2.2). The output of this network is the concentration of phosphorylated CheY. The flagellar motor is the target of the signaling pathway. Phosphorylated CheY functions as a switch factor of the flagellar motor. Low concentration of CheY-P prolongs the swimming period between reversals. In the same time, CheB, a methylesterase together with CheR, a methyltransferase are involved in the adaptation processes which returns system to the pre-stimulus level while the stimulus persists (Gordeliy *et al.*, 2002; Spudich, 2006).

2.3 Signal transduction by olfactory receptors

The downstream effectors investigated in this thesis is involved in the olfactory cascade. Here I briefly introduce an overview of this cascade. The olfactory receptor (OR) genes comprise the largest family of GPCRs (Buck & Axel, 1991; Kato & Touhara, 2009). There are about 900 OR genes in humans (∼3% of the total genome) and 1,500 OR genes in mice (Niimura & Nei, 2003). This clearly underlies the crucial role of the sense of smell during evolution. The olfactory system is a very efficient biological setup capable of odor information processing with neural signals. The mammalian olfactory system can recognize and discriminate a large number of different odorant molecules. It was estimated that human could detect more than 10,000 different odors (Buck, 2004). The detection of chemically distinct odorants presumably results from the association of smell molecules with specific receptors on Olfactory Sensory Neurons (OSNs). As a chemical sensor, the olfactory system detects food and influences social and sexual behavior.

To address the problem of olfactory perception at a molecular level, Nobel laureates Richard Axel (Howard Hughes Medical Institute, New York, NY) and Linda Buck (Fred Hutchinson Cancer Research , Seattle, WA) cloned and characterized

18 different members of an extremely large multigene family that encodes 7TMRs whose expression is restricted to the olfactory epithelium (Buck & Axel, 1991). The members of this novel gene family encode a diverse family of ORs.

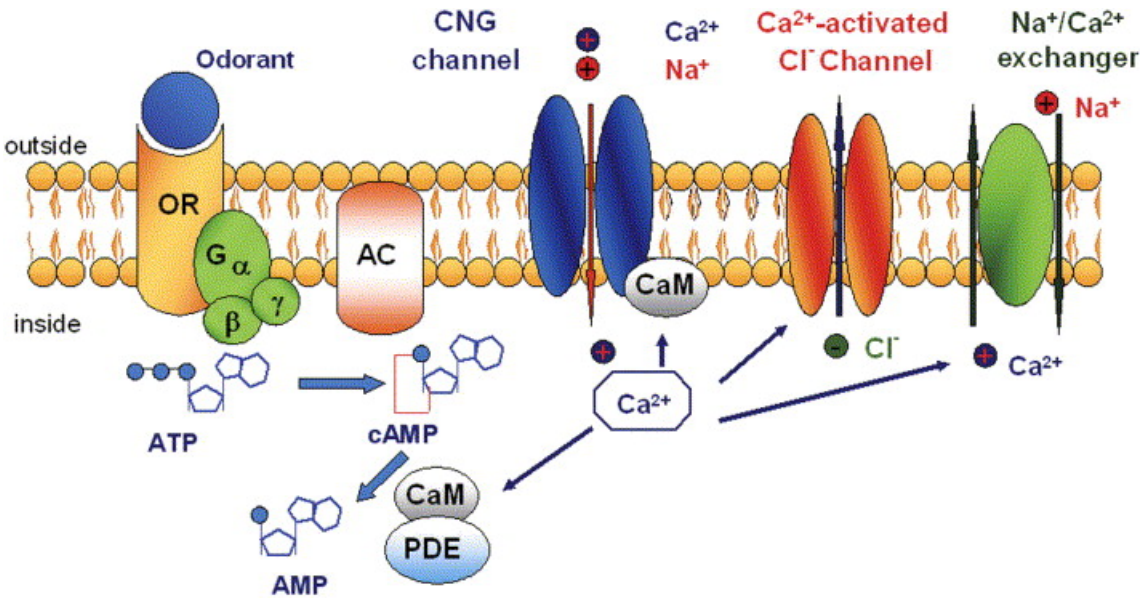


Figure 2.3 Schematic view of the odorant receptor cascade. Adapted from (Pifferi *et al.*, 2006), with permission from Elsevier.

In vertebrates, the ORs are located in the olfactory epithelium, whereas in insects they are located on the antennae. Small subsets of ORs are also expressed in non-olfactory tissues, principally the testis, taste tissues, prostate, erythroid cells and notochord. Rather than binding to specific ligands like most receptors, ORs bind to a number of similar odorant structures. In mammals, odorants enter the nasal cavity, dissolve in the mucus that covers the luminal surface of the olfactory epithelium, and bind to specific ORs on the cilia of the dendrites of OSNs (Pinto, 2011).

Before interacting with ORs, odorant molecules have to be solvated in aqueous mucus covering the olfactory epithelium (Su *et al.*, 2009). However, most airborne odorants are hydrophobic. Therefore, odorants must be transported through an aqueous fluid. This fluid is rich in so-called Odorant Binding Proteins (OBPs), which are believed to solubilize and transport odorants (Zarzo, 2007; Su *et al.*, 2009). Once an odorant has bound to an OR, the receptor undergoes structural rearrangements, with the consequence of activating its cognate G-protein (G_{α}-Olf). The activation of G-proteins, associated with a decrease of affinity for GDP and an increase of affinity for GTP, in turn, stimulates an adenylate cyclase type III (ACIII) (Bakalyar & Reed, 1990), which converts ATP into cyclic AMP (cAMP). The cAMP binds and opens

the cytoplasmatic domains of the olfactory CNG ion channels. This allows Na^+ and Ca^{2+} cations to flow along their electrochemical gradients from the extracellular to the intracellular side of the membrane (Kaupp & Seifert, 2002; Pifferi *et al.*, 2006). The increased Ca^{2+} concentration in the cilia causes the opening of Ca^{2+}-activated Cl^- channels and the subsequent Cl^- efflux, which further depolarizes the cell (Kleene, 1993; Lowe & Gold, 1993; Frings *et al.*, 2000). Thus, the chemical interactions of ORs with volatile molecules lead ultimately to the production of action potentials that will carry information about the external world to the brain (Menini, 1999; Firestein, 2001). The axons of the olfactory sensory neurons from the nasal cavity send information to second-order neurons in the olfactory bulb, which in turn project to the olfactory cortex and then to other brain areas (Pinto, 2011).

Olfactory information travels not only to the limbic system-primitive brain structures that govern emotions, behavior, and memory storage-but also to the brain's cortex, or outer layer, where conscious thought occurs (Gottfried & Wilson, 2011). In addition, it combines with taste information in the brain to create the sensation of flavor (Mombaerts, 2004). Damage to the olfactory system can occur by traumatic brain injury, cancer, inhalation of toxic fumes, or neurodegenerative diseases such as Parkinson's disease and Alzheimer's disease. These conditions can cause anosmia (Zhang & Firestein, 2002). Learning more about these links will help explain how odors affect our thoughts, emotions and behavior (Buck, 2004).

CHAPTER 3

Methods

This chapter provides an overview on the computational methods used in this thesis. The computational details on the particular methods and the simulation set up for each system can be found in their corresponding chapters.

3.1 Homology modeling

3.1.1 Principle of homology modeling

Homology modeling (in protein) is an approach to predict the three-dimensional (3D) structure of a protein from its amino acid sequence using information from one or more proteins of experimentally known structure. This method is based on the fact that proteins diverging from a common ancestor have similar 3D structures. In other words, it is based on the observation that the protein tertiary structure is better conserved during evolution than amino acid sequence. Therefore, the structure of a protein of interest (target), can be modeled based on the 3D coordinates extracted from an evolutionarily correlated protein of known structure (template) (see, e.g. (Chothia & Lesk, 1986; Tramontano & Lesk, 2006; Frishman & Valencia, 2011)).

3.1.2 Steps in model production

Homology modeling in general consists of four principal steps (Bordoli *et al.*, 2008): (see Figure 3.1).

1. *Template search.* The first step of homology modeling is the selection of evolutionarily related proteins that can be used as template(s) for modeling the target protein of interest. The identification of suitable templates is usually achieved by searching databases of experimental protein structures, such as Protein Data Bank (PDB) (Berman *et al.*, 2000), with the target sequence as the query. This step relies on the comparison between the query sequence and the sequences in the databases. The detected similarity is usually quantified in terms of sequence identity or statistical measures such as E-value or z-score, depending on the method used (Eswar *et al.*, 2007).

2. *Target-template sequence alignment.* This step is the most critical one in a modeling procedure and one of the most difficult (Tramontano & Lesk, 2006). The sequences of the target protein and the templates are aligned to identify the regions of similarity that may be a consequence of functional, structural or evolutionary relationships between the sequences. Depending on the sequence similarity between the sequences, different methods can be used to perform sequence alignment. More details of these methods will be discussed later.

3. *Model construction.* In this step, given an alignment, the coordinates from the template(s) are transferred to the target, followed by an energy optimization of the generated model based on some preselected force fields. This step involves also the modeling of conserved regions (core of the protein), modeling of the less conserved regions, in particular regions in which gaps in sequence alignment are present, and side chain modeling.

4. *Model quality assessment.* Quality assessment of the model using software validation tools, and validation with biological knowledge coming from experiments. The validation against experiments is fundamental and makes the homology modeling an iterative process. In this case, models are refined with the introduction of restraints extracted not only from the sequence alignments but also from experiments; the refined models are then used for the design of new experiments, which will be used for a subsequent validation, until arriving at convergence.

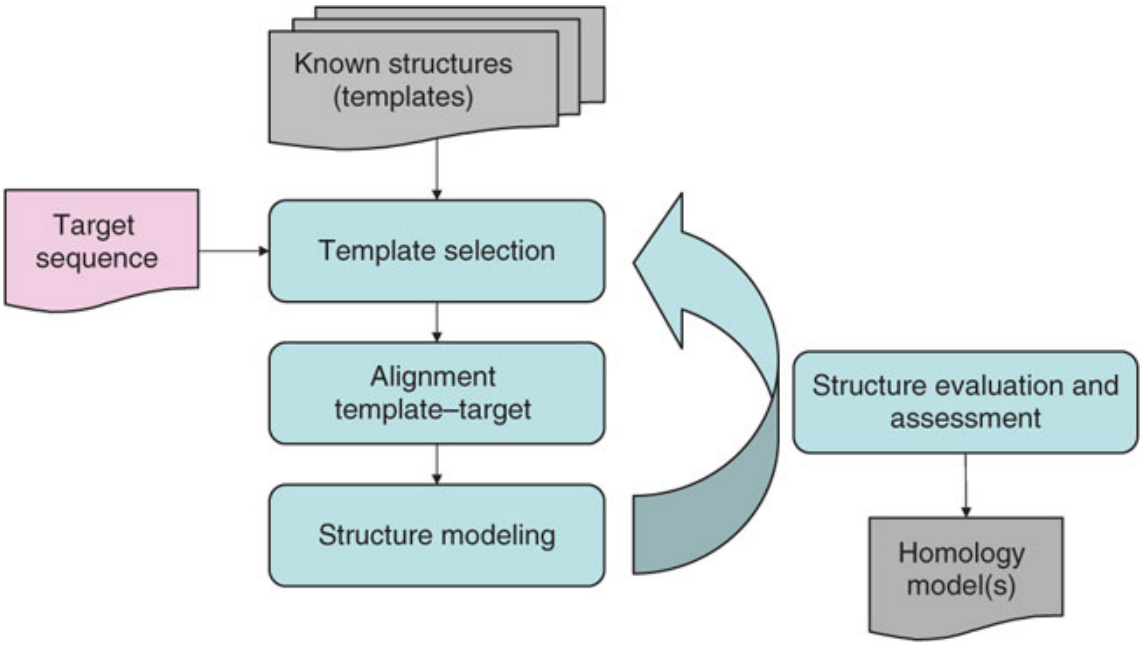


Figure 3.1 The four main steps of homology modeling: template selection, target-template alignment, model building, model quality evaluation and assessment. Figure adapted from (Bordoli *et al.*, 2008) with permission from Macmillan Publishers Ltd.

Template search and target-template sequence alignment

Although template search and target-template sequence alignment are logically two distinct steps in the process of homology modeling, in practice, these two steps are often performed together. The reason is that template identification relies on sequence alignment aided by database search techniques.

The sequence-structure relationship can be subdivided into three different regimes in the sequence similarity spectrum (Eswar *et al.*, 2007): (i) the easily detected relationships, characterized by sequence identity higher than 30%. Most protein pairs (∼90%) with more than 30% sequence identity were found to be structurally similar (Sander & Schneider, 1991); (ii) the "twilight zone" (Rost, 1999), corresponding to the 10% to 30% range of sequence identity. When the sequence identity falls in the twilight zone, the statistical measure for the evolutionary relatedness of proteins becomes uncertain (Rost, 1999); and (iii) the "midnight zone" (Rost, 1999), corresponding to statistically insignificant sequence similarity. Depending on different regimes, different sequence alignment methods can be used.

- *Pairwise sequence alignment methods.* Pairwise sequence alignment methods are used to find the best-matching piecewise (local) or global alignments of two query sequences (Eswar *et al.*, 2007). Global alignments attempt to align the

entire length of sequences while local alignments only look for highly similar regions within long sequences. Pairwise alignments can only be used between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision (such as searching a database for sequences with high similarity to a query). Two common used pairwise sequence alignment methods aided by database search techniques are FASTA (Lipman & Pearson, 1985) and BLAST (Altschul *et al.*, 1990).

- *Profile-sequence alignment methods.* Profile-sequence alignment methods, first introduced by the authors in (Gribskov *et al.*, 1987), are approaches where one compares a protein sequence with a sequence profile which typically represents a protein family. The sequence profile is derived from a multiple sequence alignment and is represented as a position-specific scoring matrix (PSSM; (Henikoff & Henikoff, 1994; Altschul *et al.*, 1997)) or as a Hidden Markov Model (HMM; (Krogh *et al.*, 1994; Eddy, 1998)). Replacing a single sequence by a frequency profile allows to discard part of the sequence information that is not conserved throughout the protein family. Therefore, the profile-sequence methods are more sensitive in detecting related structures in the "twilight zone" than the pairwise sequence-based methods. They detect approximately twice the number of homologs under 40% sequence identity (Eswar *et al.*, 2007). Some common used programs for profile-sequence alignment include PSI-BLAST (Altschul *et al.*, 1997), SAM (Karplus *et al.*, 1998), and HMMER (Eddy, 1998).

- *Profile-profile alignment methods.* As a natural extension, the profile-sequence alignment methods have led to profile-profile alignment methods. These methods compare a profile from query sequence with the profiles from the template proteins. It has been shown that profile-profile alignment methods include the most sensitive and accurate alignment protocols to date and are the method of choice for identifying and aligning templates for homology modeling (Remmert *et al.*, 2012). Profile-profile methods detect approximately 28% more relationships at the superfamily level and improve the alignment accuracy for 15% to 20%, compared to profile-sequence methods (Marti-Renom *et al.*, 2004; Zhou & Zhou, 2005; Söding, 2005). These methods are used by many of the best protein structure prediction servers in CASP[1] competition (Remmert *et al.*, 2012). For example, HHpred (Soding *et al.*, 2005), one of the best-scoring servers in

[1]CASP stands for the Critical Assessment of protein Structure Prediction, an experiment for protein structure prediction taking place every two years since 1994 (`http://predictioncenter.org/`). See section 3.1.4 for more details.

template-based structure prediction (Mariani *et al.*, 2011) in CASP9, is based on the pairwise comparison of HMM profiles.

Template selection. After a list of all related protein structures and their alignments with the target sequence have been obtained, there are several ways to select the best template(s) (Tramontano *et al.*, 2011):

- select the one with the highest sequence similarity,
- build a "theoretical template" by taking the average of the coordinates of all possible templates,
- take the conformation of different regions from different templates in such a way that the local similarity is highest in each region,
- build a model from each template and select the best model according to some criteria (see model quality assessment section below for different criteria),
- derive constraints from the templates and subsequently build a model that satisfies as many of them as possible.

It is difficult to say which way is the best in general, although the use of several templates generally is preferred. In addition, one should consider other criteria regarding the accuracy and condition of experimental structures, such as resolution of X-ray structures, number of restraints per residues for nuclear magnetic resonance (NMR) structures, holo-structures that have bound ligands of interest, states of protein (e.g. active or inactive).

Model construction

Three major classes of model generation methods have been proposed (Eswar *et al.*, 2007): (i) Modeling by assembly of rigid bodies; (ii) Modeling by segment matching or coordinate reconstruction; (iii) Modeling by satisfaction of spatial restraints. We briefly discuss only the latter as it has been used in the works presented in this thesis.

The methods in this class begin by generating many constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide. One of the most commonly used software in spatial restraint-based comparative modeling is MODELLER (`http://salilab.org/modeller`), and it was also

used in our works. The procedure is conceptually similar to that used in the determination of protein structures from NMR-derived restraints. The form of these restraints is obtained from a statistical analysis of the relationships between many pairs of homologous structures. These relationships are expressed as conditional probability density functions (*pdf's*) and can be used directly as spatial restraints. An important feature of the method is that the spatial restraints are obtained empirically, from a database of protein structure alignments. Next, the spatial restraints and CHARMM energy terms enforcing proper stereochemistry (MacKerell *et al.*, 1998) are combined into an objective function. The model is then obtained by optimizing the objective function in the Cartesian space.

Defintion of spatial restraints. A restraint is most precisely defined in terms of a probability density function, $p(f)$, for the feature f that is restrained. It can have any functional form provided that it is non-negative and that it integrates to 1 over the range of all possible values for f. The actual finite probability of an event $f_1 \leq f \leq f_2$ is obtained by integration of p:

$$p\left(f_1 \leq f \leq f_2\right) = \int_{f_1}^{f_2} p(f)\,\mathrm{d}f \tag{3.1}$$

The expression of a restraint in term of a *pdf* gives more information on the possible values of the restraint feature than the mean of measurements alone. It is also more complete than the upper and lower bounds on a certain atom-atom distance (Sali & Blundell, 1993). We now briefly review the functional form of some restraints used in our works.

Homology-derived restraints. In the first step of model building, distance and angle restraints on the target sequence are derived from its alignment with template 3D structures. The form of these restraints was obtained from a statistical analysis of the relationships between similar protein structures. For example, the distribution of the distance $(d-d')$ between two equivalent $C_\alpha - C_\alpha$ is considered as a function of four independent variables: (i) the corresponding $C_\alpha - C_\alpha$ distance in the template (i.e., d'); (ii) the fractional sequence identity between the two aligned sequences; (iii) the average solvent accessibilities of the two residues spanning the distance d' in the known structure and (iv) the average distance from a gap of the residues spanning the distance d'. It has been demonstrated that the conditional distribution of the distance differences can be approximated by a Gaussian function with a mean of zero and a standard deviation dependent on the values of the independent variables (Sali & Blundell, 1993). By considering all possible homologous protein pairs ($\sim 1,000$)

in the database of protein family alignments, it is possible to extract the conditional *pdf*, $p^d(D_{C_\alpha - C_\alpha}|a,b,c,d)$, as:

$$p^d(D_{C_\alpha - C_\alpha}|a,b,c,d) = \frac{1}{\sigma(a,b,c,d)\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{d-d'}{\sigma(a,b,c,d)}\right)^2\right] \tag{3.2}$$

where $\sigma(a,b,c,d)$ is a combination of properties (i) to (iv) mentioned above with the coefficients extracted from the database analysis. Thus, a $C_\alpha - C_\alpha$ distance restraint is defined on a target sequence once variables in (i), (ii), (iii) and (iv) are specified.

Stereochemical restraints. In the second step, the spatial restraints and the CHARMM force-field terms enforcing proper stereochemistry (MacKerell *et al.*, 1998) are combined into an objective function. All the stereochemical restraints are obtained from the amino acid sequence of the protein. The geometric restraints such as bond lengths, bond angles and dihedrals are expanded by Gaussian functions. The mean values and standard deviations are obtained from the CHARMM parameter set (MacKerell *et al.*, 1998). The Coulomb and van der Waals energy terms are the same as in the CHARMM force field (MacKerell *et al.*, 1998).

- *Bond lengths, bond angles and dihedral angles.*

  The *pdf* for the geometric feature f such as distance, angle, dihedral angle, is found to be a Gaussian *pdf*:

  $$p^f(f) = \frac{1}{\sigma_f\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{f-\bar{f}}{\sigma_f}\right)^2\right] = N(\bar{f},\sigma_f) \tag{3.3}$$

  where $\sigma_f = \sqrt{k_B T/K_f}$ with k_B is the Boltzmann constant and K_f is the force constant of the corresponding geometric feature f; $\bar{f} = f_0$ is the mean value of f. Only two parameters, the mean $\bar{f}$ and the standard deviation σ_f, are needed to describe this distribution.

  A corresponding restraint c is:

  $$c = \ln p = -\frac{1}{2}\left(\frac{f-\bar{f}}{\sigma_f}\right)^2 - \ln\frac{1}{\sigma_f\sqrt{2\pi}} \tag{3.4}$$

  where c is scaled by RT in kcal/mol with $T = 297.15K$ to allow these scores to be summed with CHARMM energies.

- *Coulomb restraint*

  $$c = \frac{1}{\epsilon_r}\frac{q_i q_j}{f} s(f, f_1, f_2) \tag{3.5}$$

$$s(f, f_1, f_2) = \begin{cases} 1; & f \leq f_1 \\ \frac{(f_2-f)^2(f_2+2f-3f_1)}{(f_2-f_1)^3}; & f_1 < f \leq f_2 \\ 0; & f > f_2 \end{cases} \quad (3.6)$$

where q_i and q_j are the atomic charges of atoms i and j, obtained from the CHARMM topology file, that are at a distance f; ϵ_r is the relative dielectric constant. Function $s(f, f_1, f_2)$ is a switching function that smoothes the potential down to zero in the interval from f_1 to f_2 ($f_2 < f_1$). The total Coulomb energy of a molecule is a sum over all pairs of atoms that are not connected by a bond or a bond angle.

- *Lennard-Jones restraint.* Usually used for non-bonded distances:

$$c = \left[\left(\frac{A}{f} \right)^{12} - \left(\frac{B}{f} \right)^{6} \right] s(f, f_1, f_2) \quad (3.7)$$

  The parameters f_1 and f_2 of the switching function can be different from those in Eq. 3.6. The parameters A and B are obtained from the CHARMM parameter file. The total Lennard-Jones energy should be evaluated over all pairs of atoms that are not connected by a bond or a bond angle.

- *α-helix restraint.* The α-helix restraint was used in our work to take into account the knowledge of secondary structure elements. It imposes restraints enforcing an α-helix for the residue segment specified (Eswar *et al.*, 2007).

Restraints derived from experimental data. Because the modeling by satisfaction of spatial restraints can use many different types of information about the target sequence, it is perhaps the most promising of all comparative modeling techniques. One of the strengths of modeling by satisfaction of spatial restraints is that restraints derived from a number of different sources can easily be added to the homology-derived restraints. For example, restraints could be provided by rules for secondary-structure packing, analyses of hydrophobicity and correlated mutations, empirical potentials of mean force, NMR experiments, cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, and site-directed mutagenesis, among other sources (Eswar *et al.*, 2007). Especially in difficult cases, a comparative model could be improved by making it consistent with available experimental data and/or with more general knowledge about protein structure (Giorgetti *et al.*, 2005b).

Optimization of the objective function. As a last step, the model is obtained by optimizing the objective function in Cartesian space. MODELLER, for example, minimizes the objective function F with respect to the Cartesian coordinates:

$$F = F(\mathbf{R}) = \sum_i c_i(\mathbf{f}_i, \mathbf{p}_i) + F_{symm} \tag{3.8}$$

where $\mathbf{R}$ represents the Cartesian coordinates of all atoms, c_i is a restraint of the feature i, $\mathbf{f}_i$ is the geometric feature of a molecule, $\mathbf{p}_i$ are parameters and F_{symm} is an optional symmetry term (symmetry restraint) defined as:

$$F_{symm} = \sum_{i<j} \omega_i \omega_j \left(d_{ij} - d'_{ij}\right)^2 \tag{3.9}$$

where the sum runs over all pairs of atoms ij, ω_i is the weight of atom i, d_{ij} is the intra-molecule distance between atoms i and j in the first segment, and d'_{ij} is the equivalent distance in the second segment. The optimization is carried out by the use of the variable target function method, employing methods of conjugate gradients and MD with simulated annealing (Eswar *et al.*, 2007)

Model quality assessment

Protein structure models generated by homology modeling may contain errors (see subsection 3.1.3). Thus, the models need to be treated with caution.

Sequence identity. The percentage sequence identity between target and template is a good predictor of the accuracy of a model. Model accuracy steadily increases with increasing sequence identity.

Stereochemistry check. The stereochemistry of the model (e.g., bond-lengths, bond-angles, backbone torsion angles, and non-bonded contacts) may be evaluated using programs such as PROCHECK (Laskowski *et al.*, 1993) and WHATCHECK (Hooft *et al.*, 1996). Although errors in stereochemistry are rare and less informative than errors detected by statistical potentials, a cluster of stereochemical errors may indicate that there are larger errors (e.g., alignment errors) in that region.

Global model quality estimation. Global indicator of the quality of a given model usually return a pseudo-energy for the entire model. Common used global indicators include DFIRE (Zhou & Zhou, 2002), an all-atom distance-dependent statistical potential; QMEAN (Benkert *et al.*, 2008) - a composite scoring function for model quality estimation; DOPE (Shen & Sali, 2006), a statistical potential optimized for model assessment. These indicators allow to distinguish “good” from “bad” models.

Local model quality estimation. The following tools are available for analyzing the local (per-residue) model reliability that can help in identifying potentially incorrect regions in the model: (i) ProQres(Wallner & Elofsson, 2006), an artificial neural network trained to predict the local model quality on the basis of the analysis of atom-atom contacts, residue-residue contacts, solvent accessibility surfaces and secondary structure propensities; (ii) ANOLEA(Melo & Feytmans, 1998), a statistical potential that can be used to analyze the packing quality of the model on the basis of nonlocal atomic interactions; (iii) the CHARMM empirical force field (MacKerell *et al.*, 1998), used to calculate the energy of each residue in the model.

Completing and refining model

Homology protein structure modeling is severely limited by errors in the alignment of a modeled sequence with related proteins of known 3D structure. Thus, once we have built our initial model, we need to "refine" it (Tramontano *et al.*, 2011; Eswar *et al.*, 2007). This can be accomplished by modeling the effect of the specific sequence changes that have occurred in our protein with respect to its template(s). One can use an iterative method that optimizes both the alignment and the model implied by it. This task can be achieved by a genetic algorithm protocol that starts with a set of initial alignments and then iterates through realignment, model building, and model assessment to optimize a model assessment score (Eswar *et al.*, 2007). During this iterative process: (i) new alignments are constructed by the application of a number of genetic algorithm operators, such as alignment mutations and crossovers; (ii) comparative models corresponding to these alignments are built by satisfaction of spatial restraints, as implemented in the program MODELLER; and (iii) the models are assessed by a composite score, partly depending on an atomic statistical potential (Eswar *et al.*, 2007). This procedure can be iterated until a satisfactory model is obtained (Figure 3.1)(Bordoli *et al.*, 2008).

3.1.3 Accuracy and errors of homology modeling

The accuracy of the structures generated by homology modeling is highly dependent on the sequence identity between target and template. As the similarity between the target and the templates decreases, the errors in the model increase above 50% sequence identity, models tend to be reliable, with only minor errors in side chain packing and rotameric state, and an overall root-mean-square deviation (RMSD) between the modeled and the experimental structures falling around 1 Å. In the

30-50% identity range, errors can be more severe and are often located in loops. Below 30% of sequence identity the structures of the template and the query will be expected to have conformational differences even though they might share the same fold.

Errors in comparative models can be divided into five categories (Figure3.2)(Eswar *et al.*, 2007) as follows:

- Errors in side-chain packing: As the sequences diverge, the packing of side chains in the protein core changes. Side-chain errors are critical if they occur in regions that are involved in protein function, such as active sites and ligand-binding sites.
- Distortions and shifts in correctly aligned regions: As a consequence of sequence divergence, the main-chain conformation changes, even if the overall fold remains the same. Therefore, it is possible that in some correctly aligned segments of a model the template is locally different (> 3 Å) from the target, resulting in errors in that region.
- Errors in regions without a template: Segments of the target sequence that have no equivalent region in the template structure (i.e., insertions or loops) are the most difficult regions to model.
- Errors due to misalignments: The largest single source of errors in comparative modeling is misalignments, especially when the target-template sequence identity decreases below 30%.
- Incorrect templates: This is a potential problem when distantly related proteins are used as templates (i.e., $< 25\%$ sequence identity). Distinguishing between a model based on an incorrect template and a model based on an incorrect alignment with a correct template is difficult. In both cases, the evaluation methods will predict an unreliable model.

3.1.4 The CASP experiments

The Critical Assessment of methods for protein Structure Prediction (CASP) is a community-wide, worldwide experiment conducted every two years that provides a mechanism for assessing how well the modeling methods perform (Moult *et al.*, 2011). Crystallographers and NMR spectroscopists who are about to solve a protein

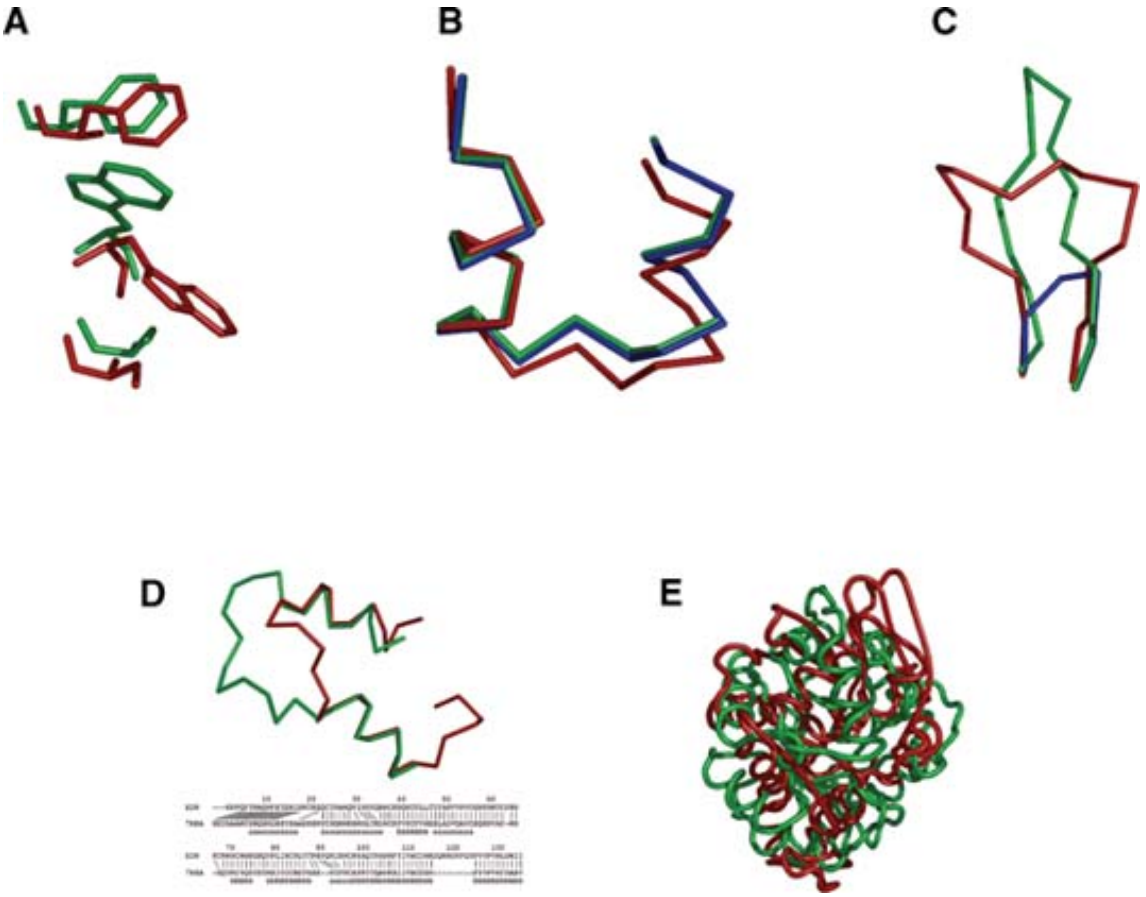


Figure 3.2 Typical errors in homology modeling. Conformations of the template are shown in red colour and that of the target are shown in green or blue colour. (A) Errors in side chain packing. (B) Distortions and shifts in correctly aligned regions. (C) Errors in regions without a template. (D) Errors due to misalignments. (E) Errors due to an incorrect template. Figure adapted from (Eswar *et al.*, 2007)

structure are asked to make the sequence of the protein available together with a tentative date for the release of the final coordinates (Tramontano *et al.*, 2011). Predictors produce and deposit models for these proteins before the structures are made available and, finally, a panel of assessors compares the models with the structures as soon as they are available and tries to evaluate the quality of the models and to draw some conclusions about the state of the art of the different methods. The results are discussed in a meeting where assessors and predictors come together and the conclusions are made available to the whole scientific community via the World Wide Web (`http://predictioncenter.org`) and the publication of a special issue of the journal Proteins: Structure, Function, and Genetics (see e.g. (Moult *et al.*, 2011)).

So far, no available method, is able to consistently produce the correct structure for regions where insertions and deletions are located or to improve the initial model and make it "better", i.e. closer to the real structure, while the accuracy of side chain modeling methods seems to be only limited by the quality of the prediction of the rest of the structure (Tramontano *et al.*, 2011). Notwithstanding the limitations of comparative modeling, this method remains the method of choice whenever possible for at least two reasons. First of all, the relative quality of a comparative model depends on the evolutionary distance between two proteins. In fact, both the probability of inferring the correct alignment between two proteins and the structural divergence between their structures are correlated with their evolutionary distance which can be estimated a priori. This implies both that it is possible to estimate the expected quality of a comparative model and its possible range of application beforehand and hence decide whether it is reasonable to embark in the task and also, perhaps most importantly, that one can attach an approximate reliability to any of the conclusions derived from the analysis of the model. The second, equally important aspect, is that the methodology will be especially effective in modeling regions of a protein that are more conserved during evolution. This implies that functionally important regions will be more correctly modeled than other, often of lower interest, regions.

Physics-based force field approaches, such as molecular dynamics simulations or Monte Carlo simulations, in principle allow the refinement of protein homology models. However, these physics-based energy function methods still have had limited success as observed in the results of CASP9 experiment (MacCallum *et al.*, 2011). Recently, MD simulations on 24 protein models, each at least 100 μs, have been performed to reveal whether the simulation time scale, force field accuracy or both is the main cause of the limit (Raval *et al.*, 2012). The authors found in most cases that simulations initiated from homology models drift away from native structure. Comparison with simulations of those proteins initiated from native structure, they suggested that force field accuracy is the primary factor limiting MD-based refinement. Development of better force fields may improve the performance of these physics-based energy function approaches on refining homology models.

3.2 Molecular dynamics simulations

3.2.1 Principle of molecular dynamics simulation

Molecular dynamics (MD) simulation is a computational technique which allows to obtain the equilibrium and transport properties of a classical many-body system (Frenkel & Smit, 2001). Given the initial positions of the atoms in the macromolecule and in the solvent, MD calculates the time evolution of the system. MD simulations can provide detailed information on the fluctuations and conformational changes of biomolecular systems such as proteins and nucleic acids. The dynamics of the system can be described by the Newton's second law:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \mathbf{F}_i,\ i = 1 \dots N. \tag{3.10}$$

where m_i and $\mathbf{r}_i$ are the mass and the coordinate of the atom i respectively; t represents the time and N is the total number of the particles in the system. The forces $\mathbf{F}_i$ acting on the atom i are the negative derivatives of the potential energy function $V(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$:

$$\mathbf{F}_i = -\frac{\partial V}{\partial \mathbf{r}_i} \tag{3.11}$$

The MD algorithm consists of five main steps:

1. Defining the initial conditions, potential interactions as a function of atoms positions; position $\mathbf{r}$ and velocities $\mathbf{v}$ of all the atoms in the system.
2. Computing the forces acting on each atom.
3. Solving numerically the Newton's equations of motion and updating the system configurations at a certain time t to the one at a time $t + \Delta t$, where Δt is the time-step.
4. Writing the new positions, velocities, energies,..., etc.
5. Back to point 2.

Thus, if one knows the initial positions of the atoms (by experiments or by computer modeling), an initial distribution of velocities consistent with the temperature simulated, and the acceleration (determined by the gradient of the potential energy function), one can provides the time-evolution of the system.

3.2.2 Calculating averages from a molecular dynamics simulation

MD simulations generate information at the microscopic level, including atomic positions and velocities or momenta. Meanwhile, an experiment is usually made on a macroscopic sample that contains an extremely large number of atoms or molecules sampling an enormous number of conformations. Therefore, macroscopic properties measured in an experiment such as pressure, energy, heat capacities, etc., are averages over billions of molecules (Lindahl, 2008). The conversion of the microscopic information from MD to macroscopic observables requires statistical mechanics. In statistical mechanics, averages corresponding to experimental observables are defined in terms of ensemble averages. An ensemble average is an average taken over a large number of replicas of the system considered simultaneously and given by:

$$\langle A \rangle_{\text{ensemble}} = \iint dp^N dr^N A\left(p^N, r^N\right) \rho\left(p^N, r^N\right) \tag{3.12}$$

where $A\left(p^N, r^N\right)$ is the observable of interest and it is expressed as a function of the momenta, p, and the positions, r, of all the particles of the system. The integration is over all variables p and r. The probability density of the ensemble, $\rho\left(p^N, r^N\right)$, is given by:

$$\rho\left(p^N, r^N\right) = \frac{1}{Q} \exp\left[-\frac{H\left(p^N, r^N\right)}{k_B T}\right] \tag{3.13}$$

where H is the Hamiltonian, T is the temperature, k_B is Boltzmann's constant and Q is the partition function:

$$Q = \iint dp^N dr^N \exp\left[-\frac{H\left(p^N, r^N\right)}{k_B T}\right] \tag{3.14}$$

The integral in equation 3.14 is generally extremely difficult to calculate because one must calculate all possible states of the system. In a MD simulation, each configuration visited by the system represents a point in the ensemble. Thus, to calculate an ensemble average by a MD simulation, the system must pass through all possible states corresponding to the particular thermodynamic condition. If this happens then we can replace the ensemble average with a time average over the configurations visited by a MD simulation:

$$\langle A \rangle_{\text{time}} = \lim_{\tau \to \infty} \int_{t=0}^{\tau} A\left(p^N(t), r^N(t)\right) dt \approx \frac{1}{M} \sum_{t=1}^{M} A\left(p^N(t), r^N(t)\right) \tag{3.15}$$

where t is the simulation time, M is the number of time steps in the simulation and $A\left(p^N(t), r^N(t)\right)$ is the instantaneous value of A.

This is the basic idea of the *ergodic hypothesis*, which states that the time average equals the ensemble average:

$$\langle A \rangle_{\text{ensemble}} = \langle A \rangle_{\text{time}} \tag{3.16}$$

However, the ergodic hypothesis is valid only in the limit $\tau \to \infty$. Replacing the time average with the finite sum in equation 3.15 could not be accurate if the number of time steps M is too small. Therefore, one goal of a MD simulation is to generate enough representative conformations such that this equality is satisfied. Because the simulations are always finite in duration, one must be certain to sample a sufficient amount of phase space.

3.2.3 Empirical force field

MD simulations numerically investigate the motions of a system of discrete particles under the influence of internal and external forces. Classical MD methods rely on empirical approximations called force field (Frenkel & Smit, 2001) to derive those forces. A force field consists of both the functional form of the potential energy used to calculate the forces and the collection of parameters used to define the potential energy. The potential energy used to describe biological systems has the following typical form:

$$V = E_{bonds} + E_{angles} + E_{dihedrals} + E_{vdw} + E_{elect}. \tag{3.17}$$

Different terms correspond to the bond distance stretching (E_{bonds}), the bond angle bending (E_{angles}), the torsional dihedral angle potential when rotating around bond, the out-of-plane "improper torsion" dihedral angle potential ($E_{dihedrals}$), van der Waals potential (E_{vdw}) and electrostatics potential (E_{elect}). The first three terms are considered to be the intramolecular bonding interactions, as they involve multiplets of atoms connected by chemical bonds. The last two terms represent the non-bonded interactions between atoms. In this work, the GROMOS[2] force field was used (Christen *et al.*, 2005). The specific form of these terms in the GROMOS force field is shown below.

[2]GROMOS is an acronym of the GROningen MOlecular Simulation computer program package (`http://www.gromos.net/`)

Bond distances stretching potential.

$$E_{bonds} = \sum_{b} \frac{1}{4} K_b \left(b^2 - b_0^2\right). \tag{3.18}$$

where b is the bond length between atom i and j. The parameters K_b and b_0 have been originally derived from experimental spectroscopic (K_b) and X-ray diffraction (b_0) data for small molecule. The functional form of equation 3.18 is anharmonic. It was chosen for computational reasons, reducing the number of square-root operations in the evaluation of the interaction energy and forces.

Bond angle bending potential.

$$E_{angles} = \sum_{\theta} \frac{1}{2} K_\theta \left(\cos\theta - \cos\theta_0\right)^2. \tag{3.19}$$

with θ being the value of the angle defined by atoms i, j, k. The parameters K_θ and θ_0 have also been derived from experimental spectroscopic and X-ray diffraction. The functional form of the bond angle bending potential has been chosen for computational reasons as well. In the GROMOS form only the value and derivative of $\cos\theta$ is needed, which saves an arccosine operation.

Torsional dihedral angle and improper dihedral angle potentials.

$$E_{dihedrals} = E_{torsional} + E_{improper}. \tag{3.20}$$

The proper torsional dihedral angle potential is treated using a trigonometric function:

$$E_{torsional} = \sum_{\phi} K_\phi \left[1 + \cos(\delta)\cos(m\phi)\right] \tag{3.21}$$

where δ is the phase shift, which is restricted to 0 or π (i.e., $\cos\delta = \pm 1.0$); m is the multiplicity of the torsional dihedral angle and ϕ is the value of the dihedral angle defined by atoms i, j, k and l. In the GROMOS force field, the parameters K_ϕ, δ and m are derived from quantum-mechanical calculations of rotational energy profiles of torsional angles.

The improper dihedral angle potential is used to keep the set of four atoms in a specific configuration. Examples are keeping four atoms in a plane, or maintaining a tetrahedral configuration:

$$E_{improper} = \sum_{\xi} \frac{1}{2} K_\psi \left(\xi - \xi_0\right)^2 \tag{3.22}$$

with ξ being the value of an improper dihedral angle defined by atoms i, j, k and l. The parameters K_ξ and ξ_0 are defined over improper dihedral angle types. In the GROMOS force field, only three types of improper dihedral angle are considered including one tetrahedral type and two planar ones with different force constants.

Van der Waals potential. The van der Waals potential is calculated using a Lennard-Jones 12/6 potential function

$$E_{vdW} = \sum_{i<j}^{atoms} \left(\frac{C12_{ij}}{r_{ij}^{12}} - \frac{C6_{ij}}{r_{ij}^{6}} \right) \tag{3.23}$$

The parameters $C12_{ij}$ and $C6_{ij}$ depend on the type of atoms involved and the character of the interaction.

Electrostatic potential.

$$E_{elect} = \sum_{i<j}^{atoms} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1} \frac{1}{r_{ij}}, \tag{3.24}$$

where q_i is the partial charge of the atom i, ϵ_0 is the dielectric permittivity of vacuum and ϵ_1 the relative permittivity of the medium in which the atoms are embedded. The electrostatic energy is evaluated by assuming the dielectric constant ϵ_1 equal to 1 (vacuum value), and using the restrained electrostatic potential model (Bayly *et al.*, 1993) to define the partial atomic point charges: in this model, the charges assigned to the atom-centered points reproduce the electrostatic potential obtained by quantum chemistry calculations of a set of small representative molecules.

Van der Waals and electrostatic potentials are calculated between atoms belonging to different molecules or for atoms in the same molecule separated by at least three bonds. In principle, the non-bonded interaction is suppose to be calculated over all pairs of atoms in the system, and they are the most expensive part of a MD calculation. In practical applications, however, the number of interactions is limited by a predefined cutoff distance, so the non-bonded interaction is calculated only between atoms separated by a distance not larger than the cutoff. For the van der Waals potential, this truncation introduce only a small error in the energy. This is not the case for the electrostatic potential because the Coulomb interaction between charges q_i and q_j decays slowly with distance. Hence it can not be truncated, but when periodic boundary conditions (PBC) are used, it can be computed with efficient schemes such as Particle Mesh Eward (Darden *et al.*, 1993) in conjunction with PBC, which approximately the exact result to an acceptable error similar to the error in the van der Waals potential (see section 3.2.6).

One should note that GROMOS is not a pure all-atom force field, since aliphatic CH_n groups are treated as united atoms. Using united atom force field can save a lot of computational time, since the number of particles involved in non-bonded interactions can be reduced significantly.

3.2.4 Integration of the equations of motion

With the knowledge of all the forces acting on the particles of a system, it is possible to calculate the dynamic behavior of the system using the Newton's equations of motion for all atoms in the system (eq. 3.10). The integration time step Δt is chosen at the beginning of the simulation and it remains unchanged during the run. The time step must be small enough to allow describing the fastest motions of the system: $\Delta t \leq 0.5$ fs is normally used when bonds involving hydrogens are allowed to stretch. Bond stretching is of little interest in most cases, therefore bonds are constrained to their equilibrium lengths using SHAKE (Ryckaert *et al.*, 1977) or LINCS (Hess *et al.*, 1997) algorithms. In this way, a time step up to 2 fs can be employed while still remaining a good accuracy.

The most used algorithms for the integration of the Newton's equations of motion are the Verlet algorithm (Frenkel & Smit, 2001) and the leap-frog algorithm (Van Gunsteren & Berendsen, 1988). In both algorithms the positions of each atom are expressed by Taylor expansions. The lack of explicit velocities in the Verlet algorithm is remedied by the leap-frog algorithm (Van Gunsteren & Berendsen, 1988). In this algorithm, positions at time $t + \Delta t$ are given by:

$$\mathbf{v}_i\left(t + \Delta t/2\right) = \mathbf{v}_i\left(t - \Delta t/2\right) + \frac{\mathbf{F}_i(t)}{m_i} \cdot \Delta t \tag{3.25}$$

$$\mathbf{r}_i\left(t + \Delta t\right) = \mathbf{r}_i\left(t\right) + \mathbf{v}_i\left(t + \Delta t/2\right) \cdot \Delta t \tag{3.26}$$

which requires the use of the positions and forces at time t and velocities at time $t + \Delta t/2$. The update of positions and of velocities leaps each other: first, the velocities are calculated at half time step, then these are used to calculate the positions at one time step. A disadvantage of this algorithm is that velocities are not calculated at the same time as positions, but this can be solved by the following approximation:

$$\mathbf{v}_i\left(t\right) = \frac{\mathbf{v}_i\left(t - \Delta t/2\right) + \mathbf{v}_i\left(t + \Delta t/2\right)}{2} \tag{3.27}$$

In this way, the kinetic and potential energies can be summed at the same time t to compute the total energy.

3.2.5 Periodic boundary conditions

Typical biomolecular simulations use periodic boundary conditions (PBC) to avoid surface artifacts and to mimic the presence of the "bulk" environment. In this approach, the system is surrounded with replicas of itself in all directions, to yield an infinite periodic lattice of identical cells. When a particle moves in the central cell, its periodic images in every other cell move accordingly (Figure 3.3). As one molecule leaves the central cell, its periodic image enters from the opposite side.

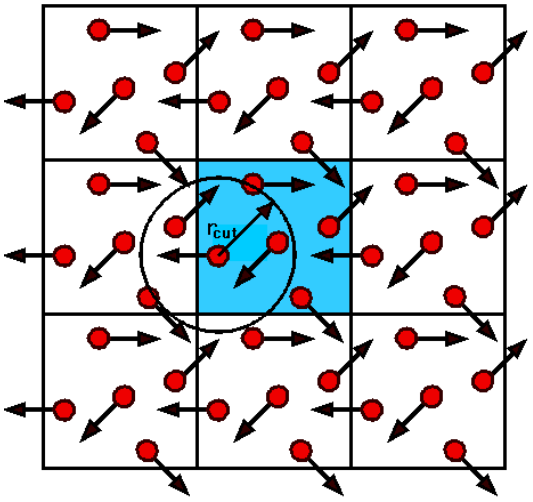


Figure 3.3 Periodic boundary conditions. As a particle moves out of the simulation box, an image particle moves in to replace it. In calculating particle interactions within the cutoff range, both real and image neighbors are included. Figure adapted from `http://www.compsoc.man.ac.uk/~lucky/Democritus/Theory/pbc-mi.html`

The PBC are taken into account only in the calculation of non-bonded interactions between atoms belonging to different molecules, and, if the potential range is not too long (the cutoff radius, r_{cut}, must not exceed half of the box side), the minimum image convention is adopted. This means that each atom interacts only with the nearest atom or image in the periodic array.

3.2.6 Long range interactions

Long range interactions are usually defined as those in which the potential decays no faster than r^{-d} where d is the dimensionality of the system ($d = 3$ in common situations). The electrostatic interaction, which decays as r^{-1}, belongs to this category.

Simple cutoffs can work for Lennard-Jones interactions that decay very rapidly, but, for Coulomb interactions, a sudden cutoff can lead to large errors. Thus, Coulomb interactions are problematic for the simulations, since their range is greater then half the box length.

There are several ways to handle this problem. One alternative is to "switch off" the interaction before the cutoff (Spoel *et al.*, 2005), but a better option for system with PBC is to use the Ewald summation(Darden *et al.*, 1993) to calculate the infinite electrostatic interactions by splitting the summation into short- and long-range parts. The Ewald summation is a technique for efficiently summing the interactions among ions and all their periodic images.

The total Coulomb energy of a system of N particles in a box of size L and their infinite replicas in PBC can be written as:

$$V = \frac{1}{2}{\sum_{\mathbf{n}}}' \left(\frac{1}{4\pi\epsilon_0} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{r_{ij,\mathbf{n}}} \right) \tag{3.28}$$

where q_i is the charge of particle i. The cell-coordinate vector is $\mathbf{n} = (n_1, n_2, n_3) = n_1 L\mathbf{x} + n_2 L\mathbf{y} + n_3 L\mathbf{z}$, where $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are the unit vectors along the cartesian axes. The first sum is primed to indicate that terms with $i = j$ are omitted when $\mathbf{n} = 0$. The distance between a particle in the origin cell and another at an image cell is $r_{ij,\mathbf{n}} = |\mathbf{r}_{j_\mathbf{n}} - \mathbf{r}_i| = |\mathbf{r}_i - \mathbf{r}_j + \mathbf{n}L|$. Due to the long-range nature of the potential, this sum is conditionally convergent, and converges very slowly. Therefore, a very large number of images is required to achieve a reliable estimate of V.

The idea behind the Ewald method is to surround every point charge by a charge distribution of equal magnitude and opposite sign ϱ_-, which spreads out radially from the charge. This distribution is conveniently taken to be Gaussian:

$$\varrho_i^G(r) = q_i \left(\frac{\alpha}{\pi}\right)^{\frac{3}{2}} \exp\left(-\alpha \left|r_i + nL\right|^2\right) \tag{3.29}$$

here α is an arbitrary parameter which does not determine the final result, but that can be adjusted to optimize the speed of convergence. In this way an efficient screening is performed, so that interactions rapidly go to 0 and direct summation is possible. This extra distribution acts like an ionic cloud, to screen the interaction between neighboring charges. The screened interactions are now short-ranged, and the total screened potential is calculated summing over all molecules. A charge distribution of the same sign as the original charge, and the same shape as the distribution $\varrho_i^G(r)$ is also added. This canceling distribution reduces the overall

potential due to the original set of charges. In order to exclude self-interactions, the contributions of these three charge densities should not be evaluated at r_i. However, it is convenient to keep self-interactions due to the canceling charge distribution ϱ_+, since ϱ_+ is in this way periodic and can be represented as a rapidly converging Fourier sum in the reciprocal space. The spurious self-interaction can be easily subtracted separately. In other words, the Fourier transforms of this distribution are added, and the result is transformed back into the real space. Thus, the total charge distribution of the system $\varrho(r)$ may be rewritten as:

$$\begin{aligned}\varrho(r) &= \sum_i q_i \delta\left(r - r_i + nL\right) \\ &= \sum_i \left[q_i \delta\left(r - r_i + nL\right) - \varrho_i^G(r)\right] + \sum_i \varrho_i^G(r)\end{aligned} \tag{3.30}$$

where the first sum corresponds to the term in the potential expression that is calculated in the direct space, while the second is calculated in the reciprocal space:

$$V = \varepsilon_{dir} - \varepsilon_{self} + \varepsilon_{rec} \tag{3.31}$$

$$\varepsilon_{dir} = \frac{1}{2} \sum_n{}' \sum_{ij}^N \frac{q_i q_j}{|r_{ij} + nL|} \mathrm{erfc}\left(\sqrt{\alpha}\,|r_{ij} + nL|\right) \tag{3.32}$$

$$\varepsilon_{self} = \sqrt{\frac{\alpha}{\pi}} \sum_{i=1}^N q_i^2 \tag{3.33}$$

$$\varepsilon_{rec} = \frac{1}{2V} \sum_{k \neq 0} \sum_{i,j=1}^N \frac{4\pi}{k^2} q_i q_j \exp\left(-i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}_j)\right) \exp\left(-\frac{k^2}{4\alpha}\right) \tag{3.34}$$

where $\mathrm{erfc}(x) \equiv 1 - (2/\sqrt{\pi}) \int_0^x \exp\left(-u^2\right) du$ is the complementary error function which falls to zero with increasing x (Frenkel & Smit, 2001); ε_{dir} is very similar to equation 3.28 although the long ranged $1/r$ function is here substituted by the short ranged $\mathrm{erfc}(r)/r$. As a result, the interaction vanishes above a cutoff roughly equal to $\alpha^{-1/2}$, and for every i and j the interaction can be approximated by only one periodic image term. Typically, α is chosen such that the truncation error is of the order of 10^{-5} or 10^{-6} of ε_{dir}; ε_{self} is the self interaction of the Gaussian charge distributions: it must be subtracted from the total, as the reciprocal space term ε_{rec} contains it, albeit it is a constant number, not depending on the atomic configuration; ε_{rec} is a sum over an infinite number of terms, but the factor $\exp\left(-\frac{k^2}{4\alpha}\right)$ ensures a fast convergence in the reciprocal space, and normally no more than 5/10 wave vectors in each direction are required. Its calculation is however the most consuming part in Ewald scheme.

In this thesis, we employ the particle mess Ewald (PME) method (Darden *et al.*, 1993) to deal with electrostatic interactions, which is a fast algorithm to perform an Ewald summation. In the PME scheme, the short-range part ε_{dir} converges rapidly in real space, and it is evaluated simply as a sum of atom-atom interactions, like in the Ewald scheme. Instead, the long-range part ε_{rec} is evaluated in a more efficient way through a fast Fourier transform, by representing the atom charges as a charge density field on a regular grid. The grid spacing is chosen compatibly with the PBC. Both the short-range and long-range parts can be safely approximated by truncating the summations to few terms. The PME algorithm is particularly suited to simulate periodic systems with a large number N of atoms, as the computational cost scales as $N \log(N)$ (Darden *et al.*, 1993).

3.2.7 Temperature and pressure control

In MD simulations it is possible to realize different types of thermodynamics ensembles which are characterized by the control of certain thermodynamic quantities. Using only the equations of motion, the system evolves in the microcanonical ensemble (NVE). However, the microcanonical ensemble does not correspond to the conditions under which most experiments are carried out. The ensembles in which most of the experiments are performed are the canonical (NVT) or the isothermal-isobaric (NPT) ensembles. To simulate the system in such ensembles, thermostat and barostat algorithms are required to control the temperature and pressure during the MD run. In this work, the stochastic velocity rescaling thermostat (Bussi *et al.*, 2007) and Parrinello-Rahman barostat (Parrinello & Rahman, 1981) were adopted, along with the Langevin thermostat (Adelman & Doll, 1976).

Stochastic velocity rescaling thermostat

Stochastic velocity rescaling thermostat is an extension of the Berendsen thermostat to which a properly constructed random force is added, so as to enforce the correct distribution for the kinetic energy (Bussi *et al.*, 2007). The Berendsen algorithm mimics weak coupling with first-order kinetics to an external heat bath with given temperature T_0. The instantaneous value of the temperature $T(t)$ of a system with

N_{df} degrees of freedom is related to the kinetic energy E_{kin} via the velocities of particles according to the equipartition theorem (Berendsen *et al.*, 1984):

$$E_{kin}(t) = \sum_{i=1}^{N} \frac{1}{2} m_i v_i^2(t) = \frac{1}{2} N_{df} k_B T(t) \tag{3.35}$$

$$T(t) = \sum_{i=1}^{N} \frac{m_i v_i^2(t)}{N_{df} k_B} \tag{3.36}$$

A simple way to alter the temperature is to scale the velocities by a factor λ:

$$\begin{aligned} \Delta T &= \sum_{i=1}^{N} \frac{m_i (\lambda v_i)^2}{N_{df} k_B} - \frac{m_i v_i^2}{N_{df} k_B} \\ &= (\lambda^2 - 1) T(t) \end{aligned} \tag{3.37}$$

with T_0 is the desired temperature and $T(t)$ is the current temperature. The time variation of temperature with the Berendsen thermostat is defined as:

$$\frac{dT(t)}{dt} = \frac{T_0 - T(t)}{\tau_T}, \qquad \Delta T = \Delta t \frac{T_0 - T(t)}{\tau_T} \tag{3.38}$$

It follows that:

$$\lambda_t = \sqrt{1 + \frac{\Delta t}{\tau_T} \left(\frac{T_0}{T(t) - 1} - 1 \right)} \tag{3.39}$$

The factor λ_t is used to scale the velocities at each time-step, in order to relax the temperature toward the desired temperature value T_0. The relaxation rate is controlled by the time coupling constant τ_T, which should be small enough to achieve the required temperature, but large enough to avoid disturbance of the physical properties of the system when it couples to the bath.

The stochastic velocity rescaling thermostat (Bussi *et al.*, 2007) is essentially a Berendsen thermostat with an additional stochastic term that ensures a correct kinetic energy distribution:

$$dK = (K_0 - K) \frac{dt}{\tau_T} + 2 \sqrt{\frac{K K_0}{N_{fd}}} \frac{dW}{\sqrt{\tau_T}} \tag{3.40}$$

where K is the kinetic energy and dW a Wiener process. There are no additional parameters, except for a random seed. This thermostat produces a correct canonical

ensemble and still has the advantage of the Berendsen thermostat: first order decay of temperature deviations and no oscillations.

Langevin thermostat

In the Langevin thermostat (Adelman & Doll, 1976), an additional random force term and a constant frictional coefficient are introduced to the equations of motion.

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i - m_i \xi_i \frac{d\mathbf{r}_i}{dt} + \eta_i(t) \tag{3.41}$$

where ξ_i is the friction constant and $\eta_i(t)$ is the white noise which is randomly extracted from a Gaussian distribution:

$$\begin{aligned} \langle \eta_i(t) \rangle &= 0 \\ \langle \eta_i(t) \cdot \eta_i(0) \rangle &= 2 k_B T_0 \beta_i \delta(t) \end{aligned} \tag{3.42}$$

with $\delta(t)$ indicating the Dirac delta. In many cases random numbers for $\eta_i(t)$ are extracted individually for each atom.

Parrinello-Rahman barostat

In the same spirit as temperature coupling, the system can also be coupled to a "pressure bath". With the Parrinello-Rahman barostat, the equations of motion for the particles are changed as following:

$$\frac{d^2\mathbf{r}_i}{dt^2} = \frac{\mathbf{F}_i}{m_i} - \mathbf{M}\frac{d\mathbf{r}_i}{dt}, \tag{3.43}$$

$$\mathbf{M} = \mathbf{b}^{-1} \left[\mathbf{b}\frac{d\mathbf{b}'}{dt} + \frac{d\mathbf{b}}{dt}\mathbf{b}' \right] \mathbf{b}'^{-1} \tag{3.44}$$

where $\mathbf{b}$ is a matrix that has box vector as column. In this scheme, $\mathbf{b}$ represents a new degree of freedom which obey the matrix equation of motion:

$$\frac{d\mathbf{b}^2}{dt^2} = V \mathbf{W}^{-1} \mathbf{b}'^{-1} (\mathbf{P} - \mathbf{P}_{ref}) \tag{3.45}$$

The volume of the box is denoted by V , and $\mathbf{W}$ is a matrix parameter that determines the strength of the coupling and how the box can be deformed. The matrices $\mathbf{P}$ and $\mathbf{P}_{ref}$ are the current and reference pressures, respectively.

CHAPTER 4

Activation mechanism of Sensory rhodopsin II in Halobacterium Salinarium investigated by molecular simulation

4.1 Introduction

The *Halobacterium salinarium* (*Hs*) and *Natronobacterium pharaonis* (*Np*) halophilic archaea avoid the damages from harmful blue-green light using a well-characterized highly efficient signaling cascade (Kandori *et al.*, 2010; Sasaki & Spudich, 2008; Klare *et al.*, 2004). The process is activated by the so-called sensory rhodopsin II (*Hs*SRII and *Np*SRII) receptors (Sasaki & Spudich, 2008). Upon light activation, they pass through a series of spectroscopically detectable intermediates denoted K, L, M and O (Tomonaga *et al.*, 2011), termed in analogy to those in bacteriorhodopsin (Jiang *et al.*, 2010). The conformational changes in these intermediate states affect the conformation of the cognate transducer proteins HtrII. This in turn initiates a cascade that regulates the cell's flagellar motor, eventually allowing the organism to move away from the blue-green light (Sasaki & Spudich, 2008).

*Hs*SRII and *Np*SRII are evolutionary related and they share a common function. Their sequence identity is $\sim 40\%$. These proteins belong to the 7TMR superfamily (Figure 4.1a-b). Crystal structures of *Np*SRII were characterized not only in ground state (GS) (Luecke *et al.*, 2001; Royant *et al.*, 2001; Gordeliy *et al.*, 2002; Gushchin *et al.*, 2011) but also in K state (KS) (Moukhametzianov *et al.*, 2006; Edman *et al.*, 2002) and M state (MS) (Moukhametzianov *et al.*, 2006; Gushchin *et al.*, 2011). Meanwhile, the structure of *Hs*SRII has been solved only in GS at low resolution of 3.6 Å (unpublished data, Labahn et. al). This is due to a small amount of *Hs*SRII in the plasma membrane and its instability in detergents and low ionic strength solutions. On the other hand, *Np*SRII exhibits higher stability: it is stable in detergent solutions, at low ionic strength (down to 50 mM) and high temperatures (range of $0 - 62\,^{\circ}\mathrm{C}$) (Moukhametzianov, 2006).

In both GS structures of *Hs*SRII and *Np*SRII, the seven helices (A-G in Figure 4.1a-b) accommodate a binding site containing the retinal in all-*trans* conformation (Figure 4.1c). The retinal is covalently bound to a lysine residue (K202/K205)[1] in helix G via a protonated Schiff-base (ScB) linkage. In this state, an aspartic acid residue (D73/D75) forms a salt bridge with the ScB. In high resolution X-ray structures of NpSRII (Royant *et al.*, 2001; Gordeliy *et al.*, 2002), an H-bond network connecting D75, D201, R72, ScB and water molecules in the active site is observed. The two structures differ in the cytoplasmic part of the helix G (Figure 4.1a-b). Specifically, the bending amplitude of this helix is higher in *Hs*SRII than in *Np*SRII. It is still unclear whether this reflects crystallization conditions or a native conformation. Helix G along with helix F is suggested to be crucial for the signaling transduction of *Np*SRII (Moukhametzianov *et al.*, 2006)(Hippler-Mreyen *et al.*, 2003).

The first step of the transduction is the chromophore isomerization from all-*trans* to 13-*cis* configuration upon light excitation (in KS) (Hirayama *et al.*, 1992). In the subsequent step, a proton transfers from ScB to the counterion residue (D73/D75, in MS). The crystal structures of *Np*SRII in MS (PDB code: 2F95 (Moukhametzianov *et al.*, 2006), 3QDC (Gushchin *et al.*, 2011)) show that (i) the H-bond network among residues and water molecules in the binding site is disrupted. In addition, the D75-ScB salt bridge is disrupted due to the proton transferred from the ScB to D75. (ii) the number of water molecules in the binding site decreases, e.g. from 5 in GS to 3 in MS (PDB code: 2F95)(Moukhametzianov *et al.*, 2006). (iii) Helix F moves towards the cytoplasm, whereas helix G moves in opposite direction towards the extracellular side (Gushchin *et al.*, 2011; Moukhametzianov *et al.*, 2006). The

[1] the first residue belongs to *Np*SRII and the second to *Hs*SRII

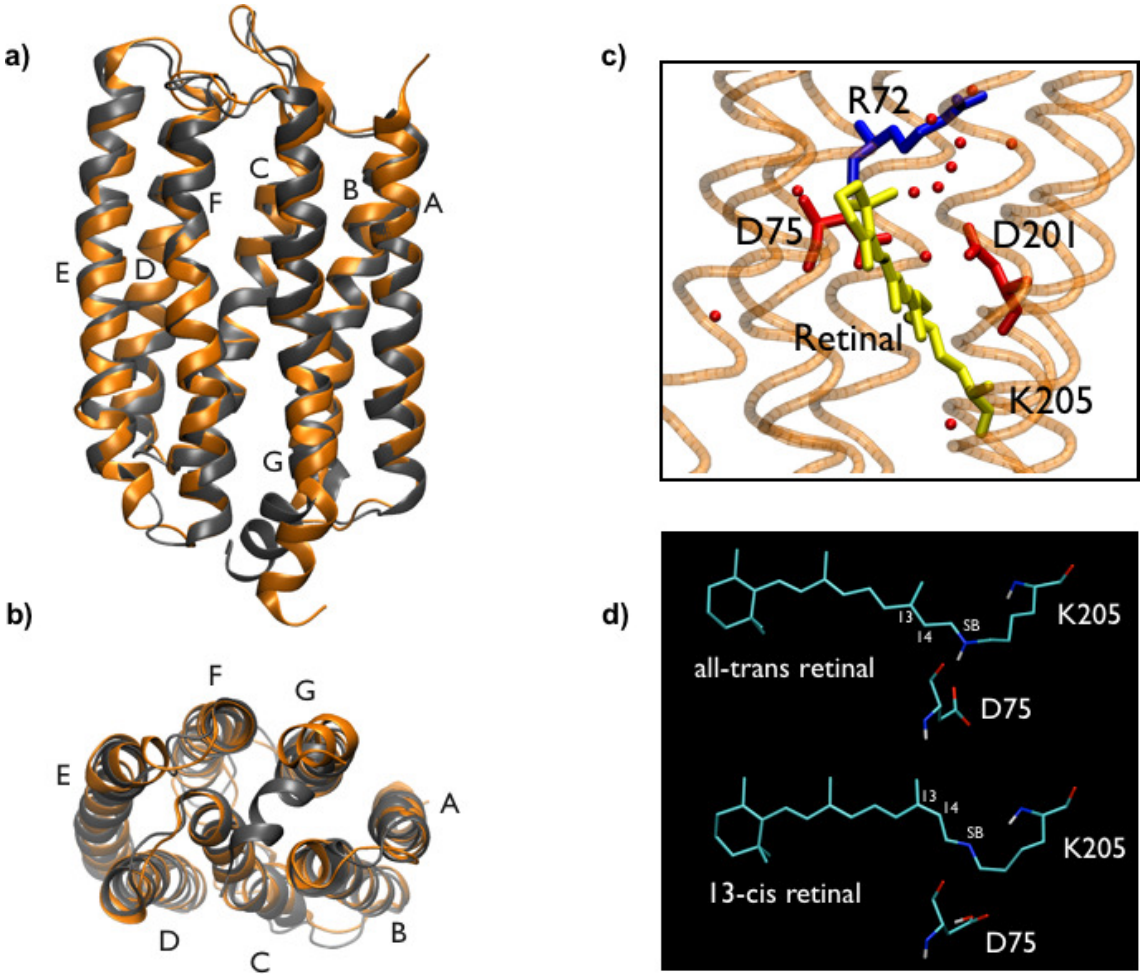


Figure 4.1 Comparison between NpSRII (Gordeliy *et al.*, 2002) (orange) and HsSRII (unpublished data) (gray) X-ray structures. a) Side view and b) view from cytoplasmic side. The helices are labeled from A to G. The only large conformational difference involves the cytoplasmic part of the Helix G (residues 215-225 for NpSRII and 213-224 for HsSRII), RMSD on backbone atoms 6.3 Å, to be compared with the RMSD of the rest including the loops, which is only 1.4 Å c) The binding pocket of NpSRII. R72, D75, D201, K205 and the retinal molecule are shown in licorice representation. Water molecules from X-ray structure are shown as red spheres. d) isomerization from all-*trans* to 13-*cis* conformation of retinal in NpSRII and subsequent proton transfer from the ScB of retinal to the counterion residue D75 upon excitation by light.

signal transfer to the transducer takes place within the interface of receptor helices F and G and the transducer helix TM2 resulting a rearrangement of the helix TM2 in *Np*HtrII (Moukhametzianov *et al.*, 2006).

A previous computational study also suggested an outward movement from the protein of the cytoplasmic half of helix F upon activation in the *Np*SRII/*Np*HtrII complex (Sato *et al.*, 2005). Additionally, in the same work, the following findings about the photocycle of *Np*SRII were suggested: (i) The H-bonds between helix F and other helices determine the direction of the movement of helix F; (ii) three amino acids (R162, T189, Y199) are essential for SRII-HtrII binding and contribute to the motion transfer from SRII to HtrII; (iii) after the isomerization of the retinal, a major conformational change of the retinal was caused by proton transfer from ScB to D75, which, in turn, triggers the steric collision of retinal with W171.

Because *Hs*SRII and *Np*SRII share a common function, the MS of both receptors is likely to be similar (Spudich & Jung, 2005), albeit the structure at atomic resolution is not yet available. In this study, we model the *Hs*SRII in MS starting from the recently solved *Hs*SRII in GS structure (unpublished data, Labahn et. al). We further attempt at gaining insights on the role of the loss of the D73-ScB salt bridge by investigating the mutant *Hs*SRII D73N in GS. Because this mutant activates the cascade without light stimuli (Sasaki *et al.*, 2007; Spudich *et al.*, 1997), the loss of the D73-ScB salt bridge might be likely to induce functionally relevant conformational changes of the receptor (Sasaki *et al.*, 2007).

This chapter presents molecular dynamics (MD) studies carried out for *Hs*SRII wild type and *Hs*SRII in MS, as well as for *Hs*SRII D73N. In the following section, we will present the materials and methods used in our studies. Analysis and results will be discussed afterward. The results allow us to suggest that the alteration of charge distribution, the rearrangement of water molecules in the active site and interhelical H-bond interactions play a key role for the movement of helices F and G in *Hs*SRII. The calculations also allow us to suggest a possible participation in the activation of *Hs*SRII of the extracellular B-C and F-G loops as well as the cytoplasmic E-F loop and the C-terminus of helix G.

4.2 Materials and methods

4.2.1 System setup

The GS of *Hs*SRII X-ray structure has been solved at 3.6 Å resolution by Labahn et al. (unpublished data). The OG1@Thr95 and CG2@Thr95 atoms, missing in the structure were added using the Modeller v.9.7 code (Eswar *et al.*, 2007). All of the residues in the refined structure are in the allowed region of Ramanchandran plot. The MS of *Hs*SRII was obtained by superimposing the backbone of refined *Hs*SRII in the GS with that of the MS of bacteriorhodopsin (PDB 1DZE) (Takeda *et al.*, 2004) using the VMD 1.9 program (Humphrey *et al.*, 1996). Then the coordinates of 13-*cis* retinal's atoms in bacteriorhodopsin were transferred to *Hs*SRII. Because proton transfers from ScB to D73 in MS, the D73 was set to be protonated. The system shows no clashes. 100% of the residues in MS model are in the allowed region of the Ramanchandran plot. D73N mutant of *Hs*SRII was built from the *Hs*SRII GS by replacing the D73 to N73. The Modeller program v9.7 (Eswar *et al.*, 2007) was used. 99,5% of D73N mutant turned out to be located in the allowed region of the Ramachandran plot. As a result, the starting configurations of D73NGS and MS are similar to that of GS except few changes in the binding site (Figure 4.2).

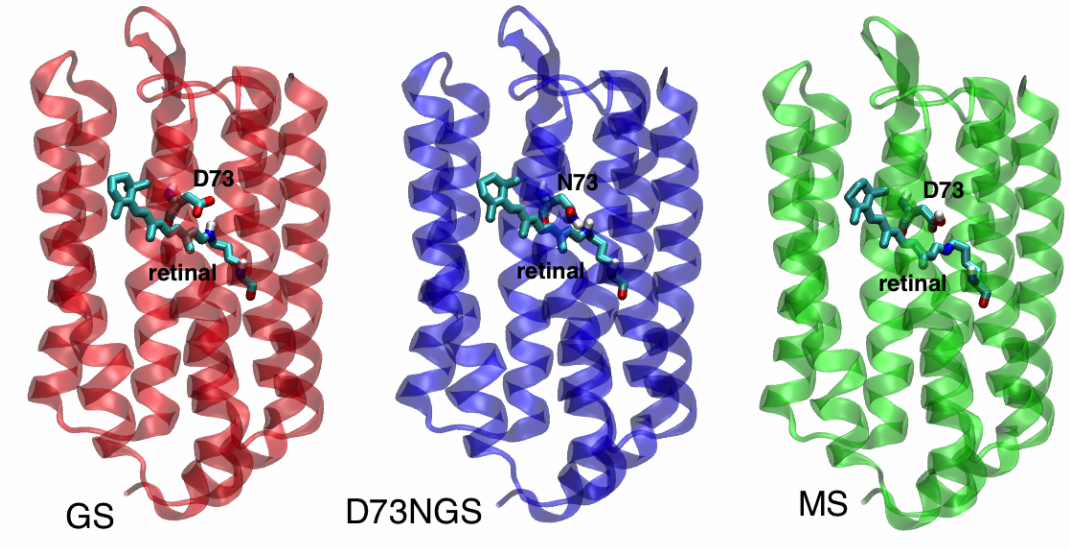


Figure 4.2 The starting configurations of three *Hs*SRII systems: GS (in red), D73NGS (in blue) and MS (in green). The retinal and the residue D73/N73 are shown in licorice representation.

The three proteins were inserted in a box of edges 8.3 nm × 8.3 nm × 9.6 nm (see Appendix B for details of inserting protein into membrane). The edges were chosen

in such a way that the minimum distance between periodic images of the protein is always larger than 30 Å throughout all simulations. The box contained 13,236 SPC water molecules (Berendsen *et al.*, 1981) and the palmitoyl oleoyl phosphatidylcholine (POPC) lipid bilayer (200 molecules). This was shown to adapt to the rhodopsin protein better than other lipids, such as dipalmitoyl phosphatidylcholine (DPPC) or dimyristoyl phosphatidylcholine (DMPC) (Cordomí & Perez, 2007). The tool g_membed from GROMACS 4.5.1 program was used (Wolf *et al.*, 2010). 39 Na+ and 44 Cl- ions were added to neutralize each system at an ionic strength of 0.1 M. The three systems consisted of approximately 51,000 atoms.

4.2.2 Simulation details

The improved united-atom 43A1-S3 GROMOS96 force field was used for both protein and lipids (Chiu *et al.*, 2009). The parameters for the force field of the moiety comprising retinal and K202 ScB in both all-*trans* and 13-*cis* conformations (Figure 4.1d) were obtained as follow: (i) The bonded and van der Waals parameters are those of the 43A1-S3 GROMOS96 force field. (ii) The partial charges were taken from (Spassov *et al.*, 2001). These authors determined the charges of retinal using density functional calculation, with a triple-zeta basis set size and frozen core orbitals.

The Coulomb interaction was calculated using the PME (particle mesh Ewald) (Darden *et al.*, 1993) with Fourier spacing of 0.12 nm and real space cutoff of 12 nm. A fourth-order of interpolation was used. For the wan der Waals interactions, a cutoff of 1.2 nm was applied. Bonds involving hydrogen atoms were constrained using the LINCS algorithm (Hess *et al.*, 1997), and the time integration step was set to 2 fs. The temperature and pressure were kept constant at 310 K and 1 bar using stochastic velocity rescaling (Bussi *et al.*, 2007) and Parrinello-Rahman semi-isotropic pressure coupling (Parrinello & Rahman, 1981), respectively. All simulations were subjected to periodic boundary condition.

The water/lipid box underwent first 20 ns of MD (see Appendix B for details). After inserting the proteins, the three systems were minimized using steepest descent algorithm (Kutzner *et al.*, 2007) in 5000 steps. Then they underwent 6 ns of position restraint runs with a force constant of 1000 kJ mol^{-1} nm^{-2} on the heavy atoms of the proteins. Finally, 250 ns of classical MD were performed. All of the calculations were carried out using the GROMACS 4.5.1 program (Kutzner *et al.*, 2007).

The following properties were calculated:

- The root mean square deviation (RMSD) was calculated using the formula:

$$\mathrm{RMSD}(t_1, t_2) = \left[\frac{1}{M} \sum_{i=1}^{N} m_i \left\| \mathbf{r}_i(t_1) - \mathbf{r}_i(t_2) \right\|^2 \right]^{1/2} \tag{4.1}$$

  where $M = \sum_{i=1}^{N} m_i$ with m_i is the mass of atom i and $\mathbf{r}_i(t)$ is the position of atom i at time t. The X-ray structure was selected as the reference conformation.

- The clustering of C_α traces during the dynamics was performed using the single linkage method (Spoel *et al.*, 2005) which adds a conformation to a cluster when its distance to any element of the cluster is less than 1 Å. The RMSD after fitting was used to define the distance between conformations. Finally, the central structure of the most populated cluster was outputted.

- The number of hydrogen bonds was determined using the geometrical criterion: $r \leq r_{HB} = 0.35\mathrm{nm}, \alpha \leq \alpha_{HB} = 30^\circ$, where r is the Donor-Acceptor radius, r_{HB} is the cutoff Donor-Acceptor radius corresponding to the first minimum of the radial distribution function of SPC water, α is the Hydrogen-Donor-Acceptor angle and α_{HB} is the cutoff for Hydrogen-Donor-Acceptor angle.

- The solvent-accessible surface area (SASA) was calculated using the double cubic lattice method (Eisenhaber *et al.*, 1995) implemented in GROMACS with a probe of 0.14 nm. The van der Waals radii used are 0.15 nm for C, 0.12 nm for F, 0.04 nm for H, 0.110 nm for N, 0.105 nm for O, and 0.16 nm for S.

- The distance between helices was computed by taking the average position of all residues belonging to each helix (residue 68 to 89 for helix C, 151 to 177 for helix F and 191 to 212 for helix G). The extracellular and cytoplasmic regions of a helix are defined by taking the average of four outmost residues in each side of the helix. For each residue, we considered the position of C_α atoms.

- The electrostatic potential of the proteins was calculated by the Poisson-Boltzmann equation using the APBS program (Baker *et al.*, 2001). The calculation parameters were 0.33 Å of grid spacing; solute surface defined by a probe sphere of radius 1.4 Å, solvent and protein dielectrics of 78.5 and 2, respectively. We first calculated the electrostatic potential of the whole protein, and then observed the potential along the plane defined by the C_α atoms of three residues R70, D73/N73 and D198.

4.3 Results and discussions

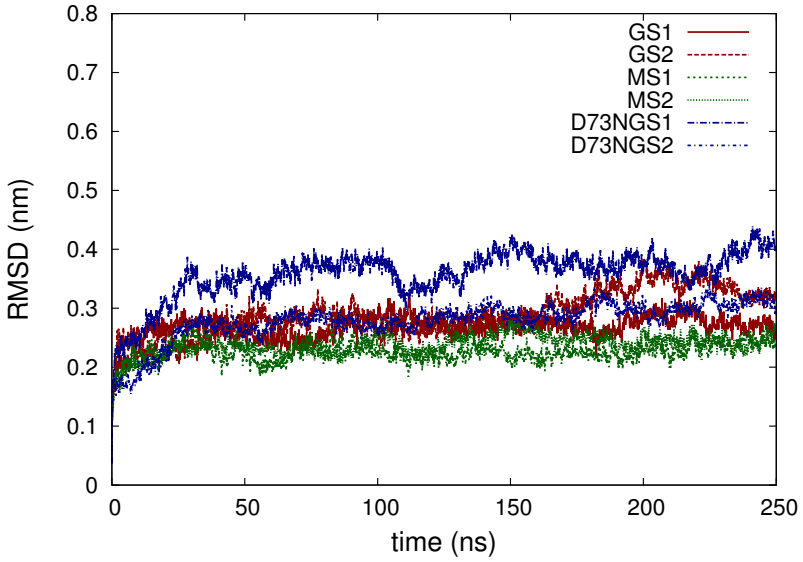


Figure 4.3 RMSD of *Hs*SRII's C_α atoms plotted as a function of time for the ground state (GS, red), the M-state (MS, green) and the mutant D73N ground state (D73NGS, blue). The indexes (1,2) indicate results for two different simulations for each state.

We have performed MD simulations of *Hs*SRII in three different states, GS, MS, and the ground state's mutant D73N (D73NGS). For each state, we carried out two independent simulations of 250 ns. Root mean square deviation (RMSD) analyses of the C_α atoms suggest that all the simulations are equilibrated within 30 ns (Figure 4.3). Hence, all our analyses have been performed for the last 220 ns. The RMSD's with respect to the crystal structure fluctuate around 2.7 Å and 3.0 Å in GS, 2.3 Å and 2.4 Å in MS, 3.7 Å and 3.0 Å in D73NGS. The higher value of RMSD in D73NGS trajectories suggests a larger rearrangement of the mutant with respect to the wild types. We next look at the MD structures in detail.

As mentioned before, the loss of the D73-ScB salt bridge in the active site might be likely to induce functionally relevant conformational changes of the receptor. Figure 4.4 shows that the D73-ScB distances in both MS simulations are larger than the corresponding ones in GS simulations, possibly due to the loss of the D73-ScB salt bridge in MS. Meanwhile, the N73-ScB distance varies largely between the two D73NGS simulations (more than 5 Å) which may be the result of mutation.

Interestingly, the distance between other two charge residues located in the binding site, R70 and D198, in MS and D73NGS is shorter than the corresponding one in GS (Figure 4.4). This may also be due to the neutralization of D73 and/or ScB.

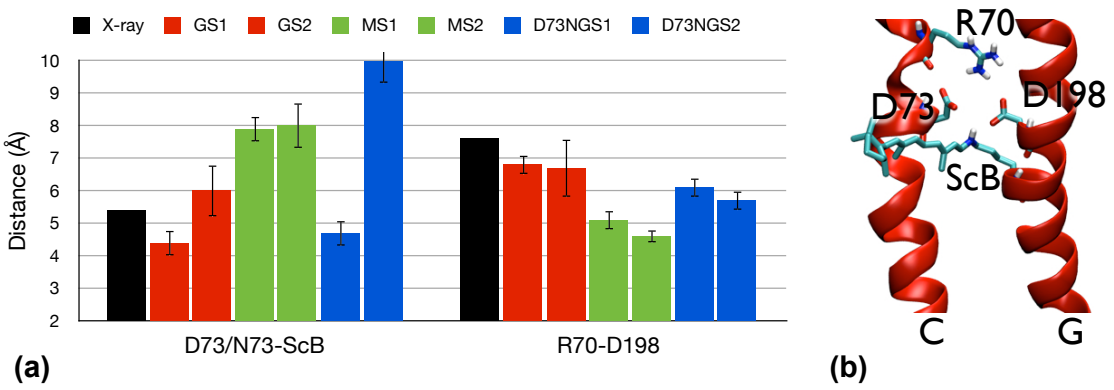


Figure 4.4 (a) Average distance between sidechain of D73/N73 and ScB as well as between two sidechains of R70 and D198 in three states: GS (red), MS (green), D73NGS (blue) and in crystal structure of *Hs*SRII (black). The indexes (1,2) indicate results for two different simulations for each state; (b) A closeup view of *Hs*SRII's binding site in GS. Only two helices C and G are shown. The ScB and three charged residues R70, D73, D198 are shown in licorice representation.

Indeed, in GS, the ScB and three charged amino acids (R70, D73, D198) form a quadrupole complex in the active site of *Hs*SRII (Rangarajan *et al.*, 2007)(Figure 4.4b). In this quadrupole, the ScB and R70 carry positive charge while D73 and D198 carry negative charge. When the residue D73 and/or ScB are neutralized, it is likely that the residues R70 and D198 attract each other more strongly. Not unexpectedly, the calculated electrostatic potentials (Figure 4.5) show a decrease (in absolute value) around ScB and D73 on passing from GS to MS. In D73NGS, the electrostatic potential is overall positive, except for a small region confined around D198.

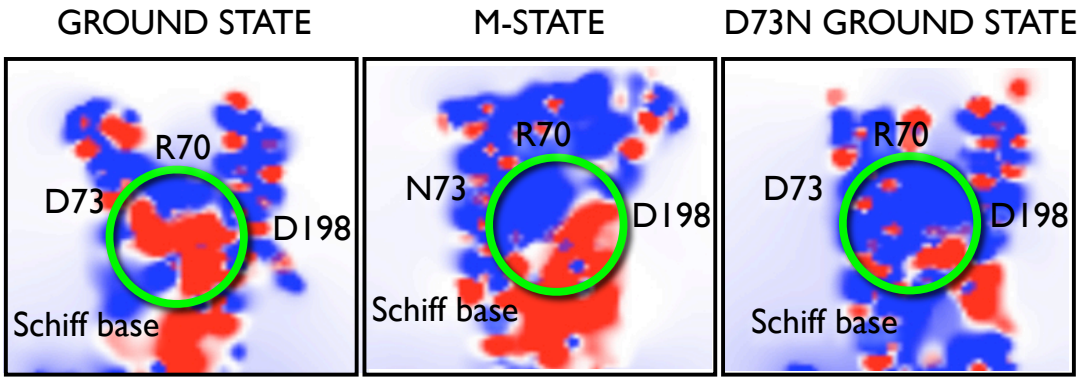


Figure 4.5 Electrostatic potential along the plane defined by C_α atoms of three residues R70, D73/N73 and D198. Positive charge is in blue color and negative charge in red color.

In addition, the number of water molecules and water-mediated H-bonds in the active site are found decreasing on passing from GS to MS and to D73NGS (Table

4.1). Consistent with these observations, in the structurally similar *Np*SRII, 5 and 3 water molecules are found in the GS and MS, respectively (Moukhametzianov *et al.*, 2006).

	GS	MS	D73NGS	X-ray
Number of water molecules in the binding site	4.5 ± 0.9	3.5 ± 0.6	1.6 ± 0.5	–
Number of H-bond between four charged groups and waters in the binding site	8.1 ± 1.5	5.3 ± 1.3	3.6 ± 1.0	–

Table 4.1 Comparison between GS, MS, D73NGS (representative configurations from this MD study) and crystal structure (unpublished data, Labahn et. al) regarding number of water molecules and number of H-bond between four charged residues and waters in binding site of *Hs*SRII.

Because R70 and D73/N73 belong to helix C while D198 and ScB belong to helix G, these changes of electrostatic environment and hydration in the active site of *Hs*SRII can affect the interactions between these two helices. In fact, the distance between the two helices in MS and D73NGS is smaller than that of GS in the extracellular region, whereas it is larger in the cytoplasmic region (Figure 4.6). Furthermore, the relative movement between the helices C and G in the cytoplasmic region of D73NGS is smaller than that of MS.

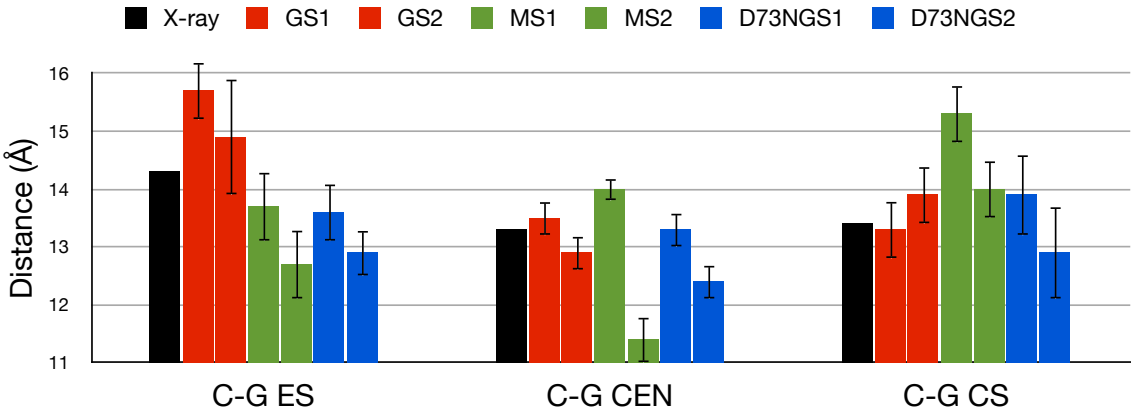


Figure 4.6 Averaged distance between two helices C and G in three different regions, extracellular (ES), transmembrane (CEN) and cytoplasmic part (CS) of the three states GS (red), MS (green), D73NGS (blue) and of the crystal structure of HsSRII (black). The indexes (1,2) indicate results for two different simulations for each state.

In *Np*SRII, FTIR studies (Ito *et al.*, 2008; Sudo *et al.*, 2003) have been shown that the H-bond between residue Y174 on helix F and T204 on helix G is weaken upon activation. This H-bond was suggested to play an important role for either

SRII activation or signal relay or both (Sudo *et al.*, 2006). However, the change of the H-bond has not been observed in the recent X-ray structure of *Np*SRII in MS (Gushchin *et al.*, 2011). Our simulations results show that the corresponding H-bond in *Hs*SRII fluctuates more significantly in MS compared to DS and D73NGS. Indeed, the H-bond between Y171 residue on helix F and S201 residue on helix G presents approximately 75% and 58% of simulation time in GS and D73NGS, respectively. Meanwhile, this H-bond presents only around 35% of simulation time in MS (Figure 4.7). This suggests a less stable H-bond between Y171 and S201 upon activation in *Hs*SRII, similar to that observed by FTIR studies for the corresponding H-bond in *Np*SRII (Ito *et al.*, 2008; Sudo *et al.*, 2003; Sudo *et al.*, 2006).

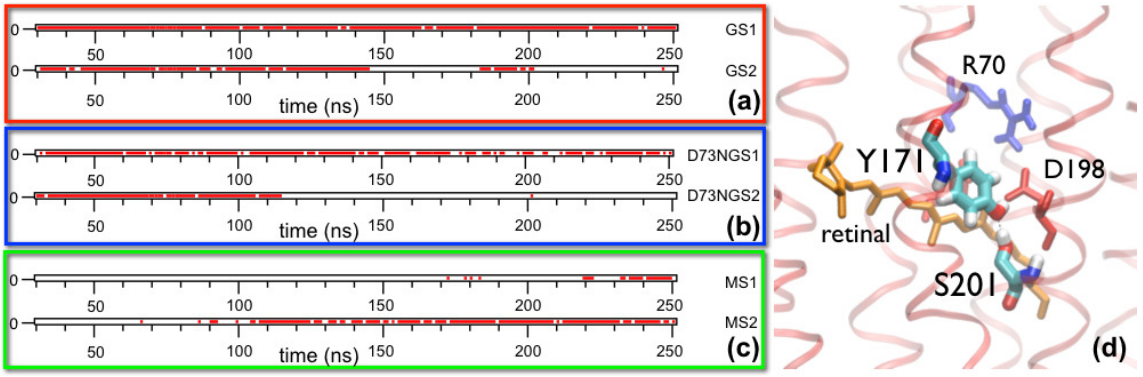


Figure 4.7 Hydrobond existence map between OH@Y171 and OG@S201 plotted as a function of time along the simulations of (a) GS, (b) D73NGS, and (c) MS. The presence of H-bond is indicated by red color; (d) closeup view of the residues Y171 and S201 in GS.

In the cytoplasmic side, we observe an increase in distance (around 1.3 Å) between helices F and G in MS compared to GS and D73NGS (Figure 4.8). On the other hand, in the extracellular and the transmembrane regions, the distance between those helices is not significantly different among three states (less than 0.8 Å). The latter indicates that the increase in distance between helices F and G happens only in the cytoplasmic moiety. Interestingly, this increase is correlated with the breaking of the Y171-S201 H-bond in one of MS simulations.[2] In another MS simulation, the increase of helix F-helix G distance also happens but around 80 ns meanwhile the Y171-S201 H-bond is already broken from 20 ns. Since the increase of helix F-helix G distance in cytoplasmic side is only observed in MS, it may be related to the difference of the retinal conformation between MS and two other states. More in detail, in MS, the retinal is in its 13-*cis* conformation while in GS and D73NGS it is in all-*trans* configuration. Therefore, we speculate that the retinal isomerization

[2]Both take place around 20 ns.

perturbs the H-bond between Y171 and S201 which then induces the displacement between the helices F and G. The displacements of these two helices may play an important role for the light-signal transduction by *Hs*SRII. Supporting to our observation, the recent FTIR study showed that the amino acid S201 regulates the photoreaction pathway of *Hs*SRII (Dai *et al.*, 2011).

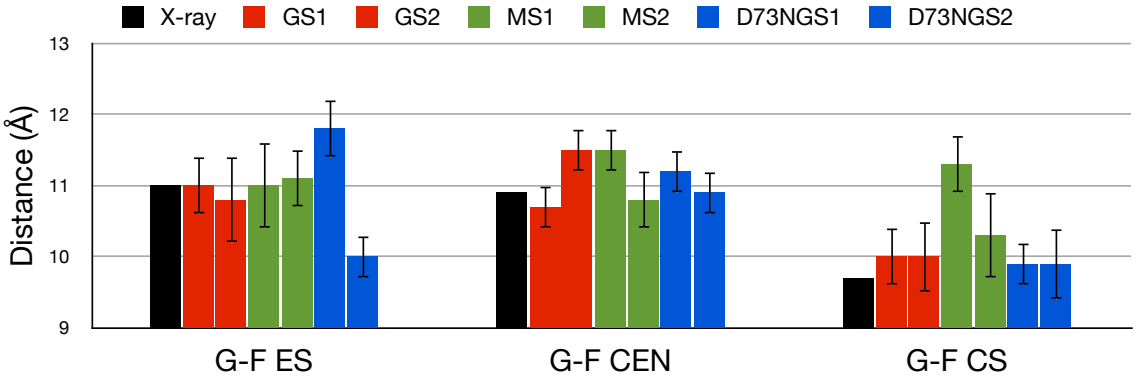


Figure 4.8 Averaged distance between two helices F and G in three different regions, extracellular (ES), transmembrane (CEN) and cytoplasmic part (CS) of the three states GS (red), MS (green), D73NGS (blue) and of the crystal structure of HsSRII (black). The indexes (1,2) indicate results for two different simulations for each state.

In the extracellular side, the F-G loop was observed to undergo conformation changes upon activation in *Np*SRII (Gushchin *et al.*, 2011). In addition, the B-C loop was shown to be important for the anion transport of halorhodopsin (Persike *et al.*, 2001), a receptor belonging to the same family of SRII. From our calculations, we observe a significant decrease of flexibility in the B-C and F-G loops passing from GS to D73NGS and MS. The average RMSF of these loops is 2.0 Å in GS meanwhile it is 1.3 Å and 1.1 Å in D73NGS and MS, respectively. We also observe in MS that the B-C and F-G loops move so as to screen the binding cavity from the extracellular side. Thus the SASA of the B-C loop is smaller in MS than that in GS and D73NGS (Figure 4.9). The reduce of flexibility upon activation in these regions may be a consequence of the displacements between helices C and G in the extracellular side. In the signaling states, MS and D73NGS, the distance between the extracellular sides of helices C and G is smaller, which may restrict the movement of the extracellular loops B-C and F-G.

In the cytoplasmic side, the loop E-F and the C-terminus of helix G are found to be less stable in MS and D73NGS. The average RMSF of these regions is 1.5 Å in GS, 1.8 Å in MS and 1.9 Å in D73NGS. In *Np*SRII, the E-F loop plays a passive role in signaling (Sasaki *et al.*, 2007) and no change of cytoplasmic half of helix G upon activation was observed in the *Np*SRII/*Np*HtrII complex (Oberbarnscheidt *et al.*, 2009). However, there is little conservation in the primary sequences between

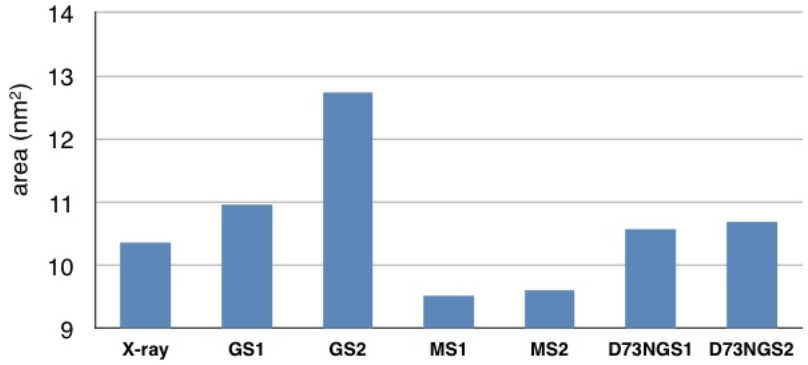


Figure 4.9 Solvent accessible surface area (SASA) of the BC loop in GS, MS and D73NGS. The indexes (1,2) indicate results for two different simulations for each state. The results show that the SASA of BC loop in MS is smaller than that of GS and D73NGS.

*Hs*SRII and *Np*SRII in the E-F loop and the C-terminus of helix G (Sasaki *et al.*, 2007). In addition, the conformation of helix G in *Hs*SRII is different from *Np*SRII in the cytoplasmic site (Figure 4.1a-b).

Based on our observations we hypothesize the following activation mechanism (Figure 4.10): the breaking of the binding site quadrupole distribution in the MS causes an increase of attractions between the two remaining charged residues. As a result, the helices C and G kink, making their extracellular regions to move closer to each other. At the same time, the perturbation of Y171-S201 interactions (in MS) causes a displacement of helix F far from helix G in the cytoplasmic part. Interestingly, this movement is found only in MS system, not in D73NGS. This may due to the fact that the two ground state systems carry the all-trans retinal while the MS contains the 13-*cis* retinal. This implicates that the outward movement of helix F from helix G in the cytoplasmic region is the consequence of the retinal isomerization. Because both helix F and G bind to the cognate transducer, we speculate that the movements of these helices are crucial in the signal transmission of SRII-HtrII complex.

The mechanism of signal transfer for SRII has been most extensively studied in *Np*SRII (reviewed in (Sasaki & Spudich, 2008; Klare *et al.*, 2008)). In the crystal structure of *Np*SRII alone (in MS) (Gushchin *et al.*, 2011), helix F was found moving towards the cytoplasm by about 0.3 Å, whereas helix G moves in opposite direction towards the extracellular side by 0.7 Å. EPR measurements shows an increase in the mobilization of a spin-labelled side chain between helices F and G upon light activation for NpSRII alone and in complex (Bordignon *et al.*, 2007), indicating an increase in the distance of helices G and F. The outward movement of helix F then ejects TM2 of the transducer in the way of a clockwise rotation (Klare *et al.*, 2008). In agreement with these experiments, we also observed the displacement between

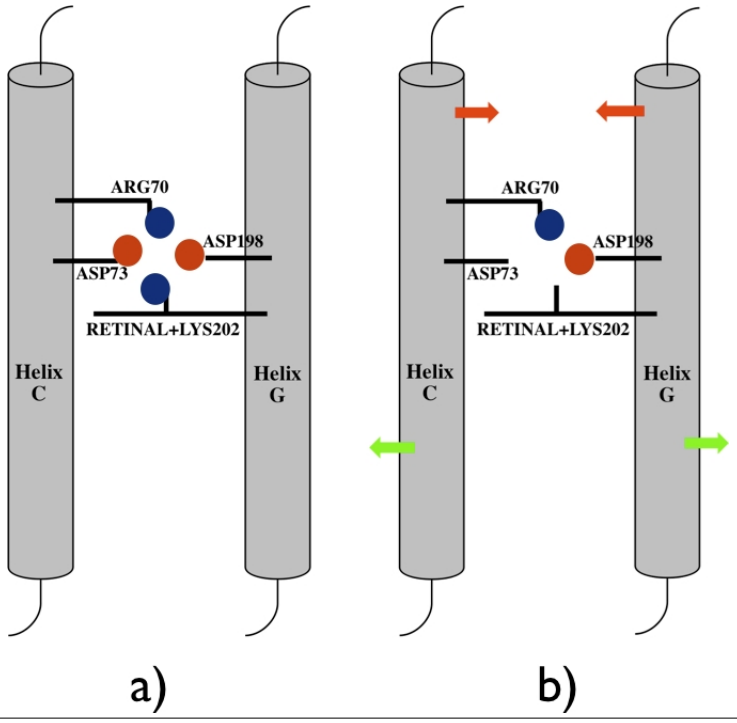


Figure 4.10 Schematic representation of two helices C and G in GS (a) and MS (b). Only the charged residues around retinal were shown. Positive charged residues are represented in blue color and negative charged in red color. The relative movement between two helices upon activation is shown in the MS.

helix F and G in the cytoplasmic side of *Hs*SRII in MS. The disruption of the salt-bridge between D73 and ScB as well as the change in the H-bond between Y171 and S20 are suggested to play important roles in inducing the conformational changes of helix F and G. In addition, the conformational changes of the B-C and F-G loops in the extracellular side as well as of the E-F loop and the C-terminus of helix G in the cytoplasmic side may involve in the activation of *Hs*SRII.

CHAPTER 5

Hybrid molecular mechanics/coarse-grained approach for 7TMRs

5.1 Introduction

Seven-transmembrane-domain receptors (7TMRs) are involved in an enormous number of biochemical processes at the cell membrane across the three kingdoms of life (Römpler *et al.*, 2007; Sharma *et al.*, 2006). 7TMRs comprise the largest membrane protein superfamily across eukaryotes (Römpler *et al.*, 2007). They are also present in prokaryotes that utilize light to harvest energy or promote phototaxis (Klare *et al.*, 2008). Human 7TMRs, called GPCR's, are among the most important targets of pharmaceutical intervention, constituting the target for ∼30% of clinically used drugs (Overington *et al.*, 2006). Thus, methods for investigating how ligands bind to 7TMRs are crucial not only for characterizing processes in cells but also for drug development.

Experimental structures of 7TMRs are available for only 14 members from eukaryotes and 10 from prokaryotes (see Table A.2, A.1, Appendix A). Thus, detailed structural and biophysical data on their interactions with their cognate ligands is currently limited (Langelaan *et al.*, 2011). Molecular Dynamics simulations, in many cases

based on structures predicted by bioinformatics methods, are often used to identify ligand poses on 7TMRs for which experimental structural information is not available (Yarnitzky *et al.*, 2010). This approach can be very CPU intensive (Sgourakis & Garcia, 2010), specially in order to characterize a large number of ligand/receptor complexes. On the other hand, coarse-grained (CG) approaches allow the study of large systems on longer timescales than those usually explored with MD (Ayton *et al.*, 2007; Rompler *et al.*, 2011). Coarse graining in MD consists of replacing an atomistic description of a biological molecule with a lower-resolution coarse-grained model that averages or smooths away fine details. As a result, the reduction of the number of degrees of freedom allows a reduction of the simulation time by $\sim$2 to $\sim$3 orders of magnitude compared to full atom force fields (Monticelli *et al.*, 2008). However, these approaches cannot describe the intermolecular ligand/protein interactions at atomic detail as required in ligand pose predictions.

A possible strategy to address this issue is to combine atomistic with CG modeling (Kalli *et al.*, 2011; Messer *et al.*, 2010; Shi *et al.*, 2006; Villa *et al.*, 2005; Neri *et al.*, 2005; Neri *et al.*, 2008). One can use CG simulation results as reference configurations which are afterwards converted to atomistic resolution. The systems are then further refined and characterized in detail by all-atom simulations (Kamerlin *et al.*, 2011; Stansfeld & Sansom, 2011a; Kalli *et al.*, 2011). This is so-called the *serial* multiscale approach (Ayton *et al.*, 2007). This strategy has been proven to be very successful in a number of applications (Kamerlin *et al.*, 2011; Stansfeld & Sansom, 2011a; Kalli *et al.*, 2011).

Molecular Mechanics/Coarse-Grained (MM/CG) simulation, on the other hand, is an *parallel* approach in which different representations of the system are modeled concurrently. A coupling scheme is used to connect the boundary of models. This approach has been developed for proteins by several groups, including ours (Shi *et al.*, 2006; Villa *et al.*, 2005; Neri *et al.*, 2005; Neri *et al.*, 2008). In our scheme (Neri *et al.*, 2005; Neri *et al.*, 2008), a region of interest (i.e. the active site of an enzyme, MM region) is treated at a molecular level using an atomistic force field. The one used here is the GROMOS96 43a1 force field (Christen *et al.*, 2005). Hydration is accounted for including a droplet of water molecules around the MM region. The protein frame is described at a CG level using a Go-like model (Go & Abe, 1981). Such a model includes only the backbone C_α atoms. An interface (I region) is defined between the MM and CG regions to bridge the two different resolution models.

Our MM/CG approach was originally developed for enzymes where the MM region is exposed to the water solvent (Neri *et al.*, 2008). In 7TMs, the binding sites are

located in the transmembrane region (Deupi & Standfuss, 2011). Thus, the presence of the lipid bilayer should be taken into account. In addition, the model requires modification to avoid the waters diffusing into the hydrophobic regions of the lipid bilayer.

Hence, we have modified our MM/CG scheme and tested the accuracy of the method by comparison with available atomistic Molecular Dynamics data of two 7TMRs:

- The *Halobacterium salinarium* sensory rhodopsin II (*Hs*SRII) which covalently binds to the retinal molecule. This 7TMR allows the *Halobacterium salinarium* archaea to avoid the damages from harmful blue-green light. The Molecular Dynamics data is taken from the chapter 4.

- The human β2 adrenergic receptor (hβ2-AR) which binds endogenous adrenaline and noradrenaline. It is involved in cardiac functions (Taylor, 2007; Hoffmann *et al.*, 2004). The trajectories of the all-atom MD simulations (Vanni *et al.*, 2011) were kindly provided by Ursula Rothlisberger.

5.2 Principle of MM/CG approach

Within our MM/CG approach, the small biologically relevant region of the protein is treated at the level of detail allowed by classical Molecular Dynamics while the rest of the protein is treated at the CG level, by only considering C_α centroids. An interface region is located between the two MM and CG regions, bridging the large discontinuity between full-atom and CG descriptions (Figure 5.1).

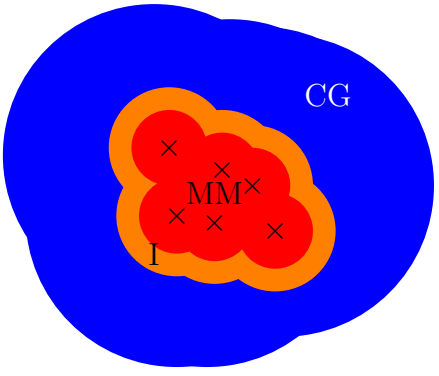


Figure 5.1 Schematic representation of the regions described by the MM/CG model. The MM region is represent in red color, the interface (I) region in orange color and the CG region is in blue color.

The total potential energy of the system in the MM/CG approach is split into terms corresponding to different sets of atoms, belonging to the MM, CG and CG/I regions:

$$V = E_{\mathrm{MM}} + E_{\mathrm{I}} + E_{\mathrm{CG}} + E_{\mathrm{MM/I}} + E_{\mathrm{CG/I}} + E_{\mathrm{SD}}. \tag{5.1}$$

Here, the first term E_{MM} denotes the GROMOS96 43a1 force field (Christen *et al.*, 2005). Atoms in the I region are represented at the atomistic level, thus the terms E_{I} and $E_{\mathrm{MM/I}}$ have the same form as E_{MM}. E_{CG} is given by a Go-like model (Go & Abe, 1981):

$$E_{\mathrm{CG}} = \frac{1}{4}\sum_{i} K_b \left(|\mathbf{R}_i - \mathbf{R}_{i+1}|^2 - b_{ii+1}^2\right)^2 + \sum_{i>j} V_0 \left\{1 - \exp\left[-B_{ij}\left(|\mathbf{R}_i - \mathbf{R}_j| - b_{ij}\right)\right]\right\}^2 \tag{5.2}$$

where (i) the first term describes the interaction between consecutive CG beads (the C_α atoms); K_b is the force constant; b_{ij} is the equilibrium distance, corresponding to the native distance between CG atoms; and (ii) nonbonded interactions are taken into account in the second (Morse-potential-type) term, in which $V_0 = 5.3$ (kJ mol^{-1}) is the well depth, B_{ij} is the modulating coefficient which measures the curvature of the potential and is a function of b_{ij}: $B_{ij} = 6/b_{ij}(\mathrm{nm}^{-1})$. The smaller the value of B_{ij}, the softer is the potential. The values of the two parameters, V_0 and B_{ij}, are chosen in a way that each contact in the native conformation is stabilized at the minimum of the potential. They have been already used for both soluble and membrane enzymes (Neri *et al.*, 2005; Neri *et al.*, 2008).

The interaction between the CG and I regions, described by the term $E_{\mathrm{CG/I}}$, is treated in the same form as E_{CG}. The bonded term is considered between CG atoms and consecutive C_α atoms of I region and has a similar form to the first term in Equation 5.2:

$$E_{\mathrm{CG/I}}^{bonded} = \frac{1}{4}\sum_{i,j} K_b \left(\left|\mathbf{r}_{i,C_\alpha}^I - \mathbf{R}_j\right|^2 - b_{ij}^2\right)^2 \tag{5.3}$$

where $\mathbf{r}_{i,C_\alpha}^I$ represents the coordinates of the ith C_α atoms in I region, $\mathbf{R}_j$ represents the coordinate of the jth CG atoms in CG region. To this, the non-bonded term between CG atoms and both C_α and C_β atoms is added. It reads:

$$E_{\mathrm{CG/I}}^{non-bonded} = \frac{1}{2}\sum_{i\in[C_\alpha,C_\beta],j} V_0 \left\{1 - \exp\left[-B_{ij}\left(\left|\mathbf{r}_i^I - \mathbf{R}_j\right| - b_{ij}\right)\right]\right\}^2 \tag{5.4}$$

where the interface ith atom is either a C_α or a C_β atom and the factor 1/2 stands for the interaction energy equally distributed between the two types of atoms. All

the coefficients are chosen similarly to those in 5.2. This interface potential ensures the integrity of the backbone.

The E_{SD} term mimics the thermal and viscous solvent effects acting on the system (Nadler *et al.*, 1987). If the I and MM regions are solvent exposed, the solvent is treated in an explicit way using the SPC water model (Berweger *et al.*, 1995). In the framework of the MM/CG approach: a drop of water molecules is centered around the MM and I regions and if a molecule exits from the water shell, its velocity is reflected towards the inside. Within this approach, water properties are very similar to those of the bulk water in proximity of the all-atom region, but approaching the drop border located approximately at the interface region, the water density lowers, providing a rough approximation of the bulk behavior (Neri *et al.*, 2005).

5.3 Development of MM/CG approach for 7TMRs

The MM/CG approach described above has been developed for enzymes where the binding site is exposed to the water solvent (Neri *et al.*, 2005; Neri *et al.*, 2008). In this study, we extend our approach to 7TMRs where the binding site is located in the transmembrane region. To take into account the presence of the lipid bilayer in the membrane, we have defined surfaces around the protein and modified the potential energy function as explained below.

The main goal of our set-up is to prevent water molecules from diffusing (i) away from the protein and (ii) into the hydrophobic region of the membrane. In oder to do that, we introduced five walls around the protein (Figure 5.2). These are described by five functions $\varphi_i (i = 1\text{–}5)$ using a level-set approach (Osher & Sethian, 1988). The region where the set of $\{\varphi_i\}$ is positive characterizes the protein site. The wall itself is the set of points for which the set $\{\varphi_i\}$ vanishes (Fig. 5.2). Two planar walls $(\varphi_i, i = 1, 2)$ coincide with the height of the membrane lipid's head. Two hemispheric walls ('outer walls', $\varphi_i, i = 3, 4$) capping the extracellular and cytoplasmic ends of the 7TMR (Fig. 5.2) are defined as follow:

$$\varphi_i(\mathbf{r}) = r_i - \|\mathbf{r} - \mathbf{c}_{hi}\|, \qquad i = 3, 4 \tag{5.5}$$

where their locations are defined only outside the membrane region. The center $\mathbf{c}_{hi}$ of each hemisphere is located at the height of the membrane lipid heads, above/under the center of mass of the protein. The radius r_i of each hemisphere is such that the minimum distance between any protein atoms and the wall is 15 Å. This value can

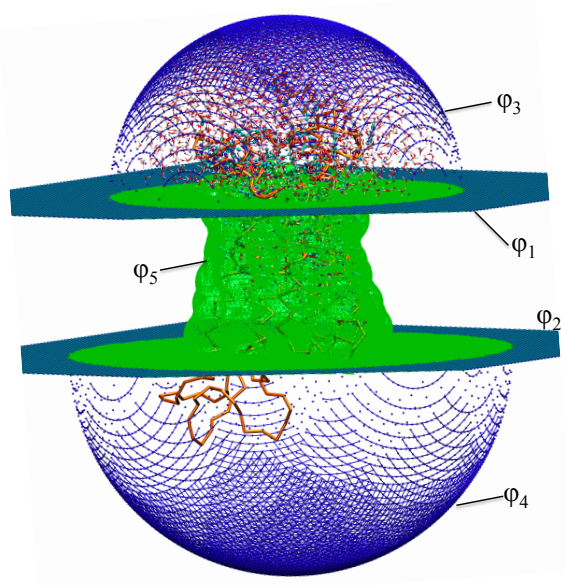


Figure 5.2 Five walls around the GPCR are used to mimic the presence of lipid bilayer: the planar walls ($\varphi_{1,2}$) are the two sheets located at the height of the membrane lipids head, the outer walls ($\varphi_{3,4}$) are the blue hemispheres, the membrane wall (φ_5) is the surface in green.

be adjusted depending on the number of water molecules. This creates a droplet of waters around the MM region.

The last wall representing the membrane ('membrane wall', φ_5) approximately follows the initial shape of the interface between protein and membrane (Fig. 5.3a). This wall is defined as:

$$\varphi_5(\mathbf{r}) = r_p - \min \|\mathbf{r} - \mathbf{c}_j\| \tag{5.6}$$

where the distance between the point $\mathbf{r}$ and the closest initial position of C_α atoms $\mathbf{c}_j$ is computed, and r_p is a distance parameter. Choosing a relatively small value of r_p, i.e. less than 1.5 Å, the whole protein may be overly constrained. A large value of r_p, i.e. larger than 3 Å, allows water molecules to enter the region compressed between the wall and the protein. This may cause a distortion of the global fold of the protein. By testing different values in the range from 1.5 Å to 3.5 Å, we have found the optimum value is between 2.0 Å and 2.5 Å. The surface is therefore constituted by pieces of spheres around the protein (Fig. 5.3a).

In addition, to avoid large discontinuities of the surface φ_5, a larger radius, $3r_p$, was used to construct the surface around the C_α atoms with the center-point shifted by $2r_p$ inwards (Fig. 5.3a). In this way, the surface is smoothed out and the minimum distance r_p from the surface φ_5 to the C_α atoms is still ensured.

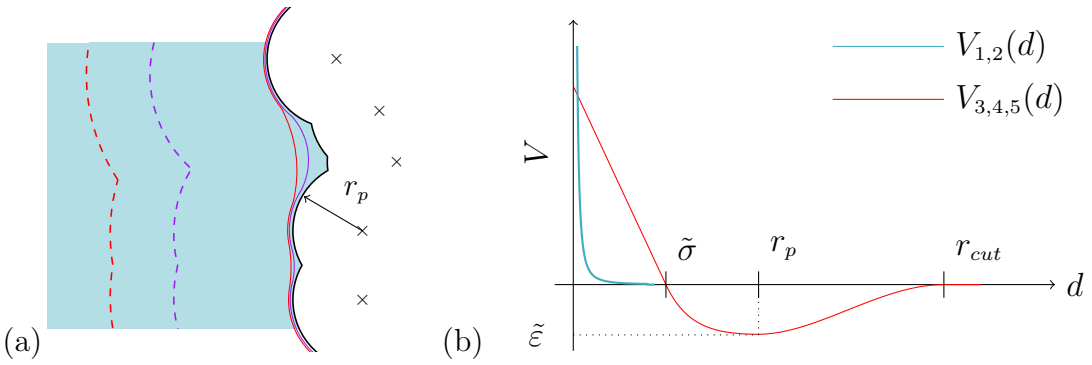


Figure 5.3 (a) The membrane wall ϕ_5. To avoid large discontinuities of the surface φ_5, a larger radius, $3r_p$ (red line), was used to construct the surface around the C_α atoms with the center-point shifted by $2r_p$ inwards. (b) Schematic of the wall potentials $V_i(d)(i = 1-5)$ plotted as a function of distance to the walls d.

Furthermore, the polar-aromatic residues (Tyr and Trp) have a specific affinity for a region near the lipid carbonyl in the membrane (Killian & von Heijne, 2000). The presence of these residues has been shown to anchor the membrane protein in a lipid bilayer (Killian & von Heijne, 2000). This effect has been taken into account by including not only the C_α atoms but also all atoms of Tyr and Trp residues in the set of initial positions c_j in Eq. 5.6, and by setting $r_p = 1.4$ for the side chain atoms of these residues. Hence the location of potential well, defined by r_p, is closer to the wall, and the derived force is strengthened as well.

Boundary potentials are added to the MM/CG potential energy function. They are defined as functions of the distance $V_i(d)(i = 1-5)$ with $d = \min \varphi_i(i = 1-5)$ from the corresponding walls, as followed (Figure 5.3b):

$$V_i(d) = 1/d, \qquad i = 1, 2 \tag{5.7}$$

$$V_i(d) = 4\tilde{\epsilon}\left[\left(\frac{\tilde{\sigma}}{d}\right)^2 - \left(\frac{\tilde{\sigma}}{d}\right)\right], \qquad i = 3, 4, 5 \tag{5.8}$$

The index i of boundary potential corresponds to the index of the surfaces φ_i. The potential applied to an atom is the one corresponding to the closest wall from that atom $d = \min \varphi_i(i = 1, 5)$. The potential $V_i(i = 1, 2)$ is purely repulsive and $V_i(i = 3, 4, 5)$ is a softened Lennard-Jones-type potential; $\tilde{\epsilon}$ is the depth of the potential well; $\tilde{\sigma}$ is the finite distance at which the potential $V_i(i = 3, 4, 5)$ is zero. The minimum of the potential is at $d = r_p$ (Fig. 5.3b). Water molecules, C_α atoms of both MM and CG regions, and all atoms of Tyr and Trp residues are influenced by

these potentials. This model neither includes electrostatics nor allows distinguishing between different types of bilayers.

MM/CG method has been successfully applied for soluble proteins (Neri *et al.*, 2005) and membrane proteins (Neri *et al.*, 2008) of which the binding site locates outside of the membrane region. All the parameters from this model are kept in the new version except the modulating coefficient in the Morse-potential-type B_{ij} (Equation 5.2). By increasing the value of this parameter, the attraction among CG beads increases more rapidly when atoms move away from their equilibrium distance. We find that increasing the value of B_{ij} from $6/b_{ij}$ (nm^{-1}) to $5 + 6/b_{ij}$ (nm^{-1}) ensures the stability of the protein inside the transmembrane site.

The force $\mathbf{F}(\mathbf{r})$ due to the presence of the wall is derived from the potentials V_i:

$$\mathbf{F}(\mathbf{r}) = -\frac{\partial V}{\partial d}\nabla d(\mathbf{r}) \tag{5.9}$$

where the $\nabla d(\mathbf{r})$ is normal to the walls. The force is applied within a given cutoff from the surface. The cutoff distance of the force is chosen to be 7 Å for the repelling walls $V_i(i = 1, 2)$, and to be $1.5r_p$ for the outer wall and membrane wall $V_i(i = 3, 4, 5)$. The first value is chosen so as a water molecule is not able to pass through this distance during one time step, while the second value guarantees that the force does not affect directly the MM region. The force is shift to be continuous at the cutoff distance to avoid a sharp disruption. In addition, it is set to a finite value (1000 kJ mol^{-1} nm^{-1}) near the wall to prevent too large of a force being applied to the system.

5.4 Testing of MM/CG approach with HsSRII

5.4.1 System setup and simulation details

The MM/CG calculations have been carried out based on the X-ray structure of the *Hs*SRII, solved for the ground state at resolution of 3.6 Å (unpublished data, Labahn et. al). The binding site contains the all-*trans* retinal covalently bound to the residue K202. The starting configuration of the receptor was obtained following the procedure in section 4.2. Similarly to previous works (Neri *et al.*, 2005; Neri *et al.*, 2008), the MM region included all the residues at less than 5 Å from the retinal molecule (37 amino acids) and the retinal itself . A cut-off of 6 Å (measured from the MM boundary) was considered in order to calculate the residues included

in the interface between the MM and the CG subsystems. 142 water molecules in the radius of 15 Å from the MM region were added using SPC model (Berendsen *et al.*, 1981). The overall system was made of 2030 atoms (Figure 5.4B). The number of atoms of MM/CG setup is 25 times smaller than the MD setup for the same system.

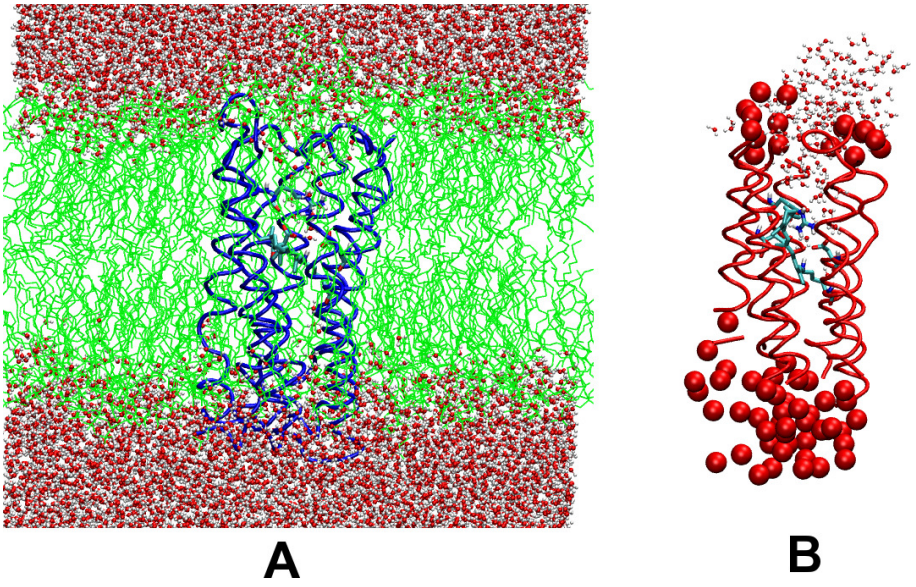


Figure 5.4 (A) Atomistic model of the HsSRII receptor (in blue color) inserted in a box of palmitoyl oleoyl phosphatidylcholine (POPC) lipid bilayer (in green color) and waters (in red CPK representation). (B) MM/CG model of HsSRII; MM and I regions are represent by red tube, CG region is represent by red spheres, waters are shown in red licorice representation.

The protein complexes were encapsulated in a 31 Å thick implicit membrane, with the transmembrane wall at $r_p = 2.5$ from the C_α atoms. Cutoffs of 14 Å, 16 Å and 16 Å were used for the electrostatic, van der Waals and Go-like nonbonded interactions, respectively. The SHAKE algorithm (Ryckaert *et al.*, 1977) was used to keep fixed the distance of bonds containing hydrogen(s). The time-step was set at 2 fs. Temperature was set at 310K, using stochastic dynamics, controlled with an inverse friction constant at a value of 0.4 ps. Periodic boundary conditions were used. RESP charges (Bayly *et al.*, 1993; Wang *et al.*, 2000) for the retinal were taken from (Spassov *et al.*, 2001). Our MM/CG simulations were carried out up to 1 μs using our MM/CG implementation in Gromacs 4.5.1 (Kutzner *et al.*, 2007).

CG simulation was carried out up to 1μs, using 2 fs time-step. Within this Go-model, each residue is treated as one bead at position of C_α atom. Thus, the atom number of this system is 222, as the number of residues present in X-ray structure. The temperature in CG simulation was also kept at 310 K using stochastic dynamics (Nadler *et al.*, 1987). The periodic boundary condition was used. The CG simulations were performed in order to test whether the MM/CG simulations were

able to reproduce the local structural features, which can be studied by Molecular Dynamics, but not by standard CG model.

The following properties were calculated:

- root mean square deviation (RMSD) of the receptor,
- root mean square fluctuation (RMSF) of the receptor,
- the number of H-bond formed by the charged residues and water molecules in the binding site,
- the side-chain torsional angle χ_1 of the charged residues in the binding site,
- the distances among the charged residues in the binding site.

5.4.2 Results

First, the stability of the receptor with respect to the crystal structure has been shown by the low RMSD value of the C_α atoms during the MM/CG trajectory (Figure 5.5a). Next, the flexibility of the receptor was compared among the three approaches. The RMSF calculated using the MM/CG approach follows the trend of that obtained by MD (Figure 5.5b), with the difference less than 0.5 Å for the MM regions. The results show that the global fluctuations of the MM regions in MM/CG simulation are similar to those in the atomistic MD simulation. Mean while, in the CG regions, the fluctuations are much smaller possibly due to the higher rigidity of the CG Go-model force field in comparison with the atomistic force field. Indeed, this is observed in the CG simulation (green line in Figure 5.5b).

As observed in chapter 4, the four charged groups in the binding site of *Hs*SRII (R70, D73, D198 and the retinal covalently bound to K202) play an important role in the receptor's function. Therefore, the following properties of these four residues were examined in detail (i) the side-chain torsional angle χ_1 (ii) the distances among the residues (iii) the number of H-bond formed by the residues and water molecules.

Firstly, we compare the side-chain torsional angle distribution χ_1 between MM/CG and MD simulations for the four charged residues. The torsion angle χ_1 provides the information on the internal dynamics of protein side-chains. It is defined by four atoms $N - C_\alpha - C_\beta - C_\gamma$ (Fig. 5.6e). The preferred conformations of χ_1 observed by experimental data are generally in three regions denoted g^+, t and g^- corresponding to rotation of 60°, 180° and −60° (Janin *et al.*, 1978; Kominos *et al.*,

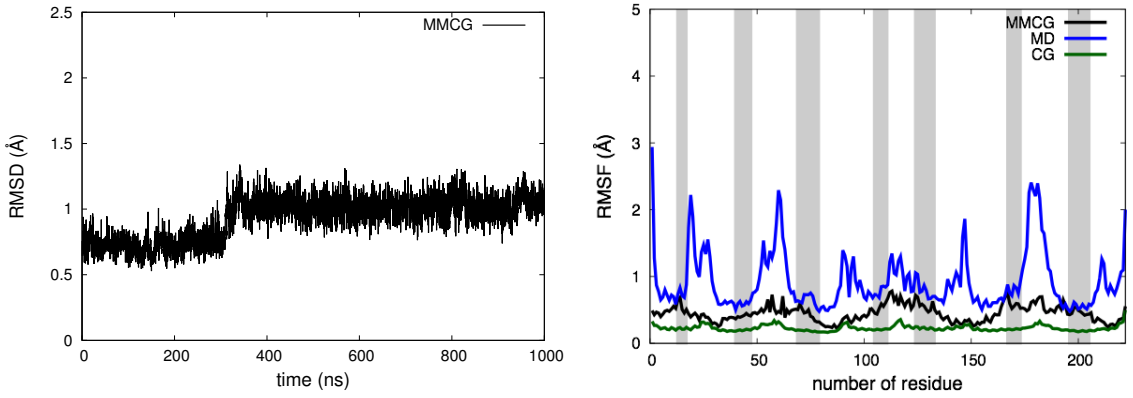


Figure 5.5 (a) Root-mean-square-deviation of *Hs*SRII's C_α atoms in the MM/CG simulation, in relative to the initial X-ray structure, plotted as a function of time. b) Root-mean-square-fluctuations of *Hs*SRII's C_α calculated based on atomistic simulation (blue), MM/CG simulation (black) and CG simulation (green). Residues included in the MM regions are highlighted with grey bars on the plots.

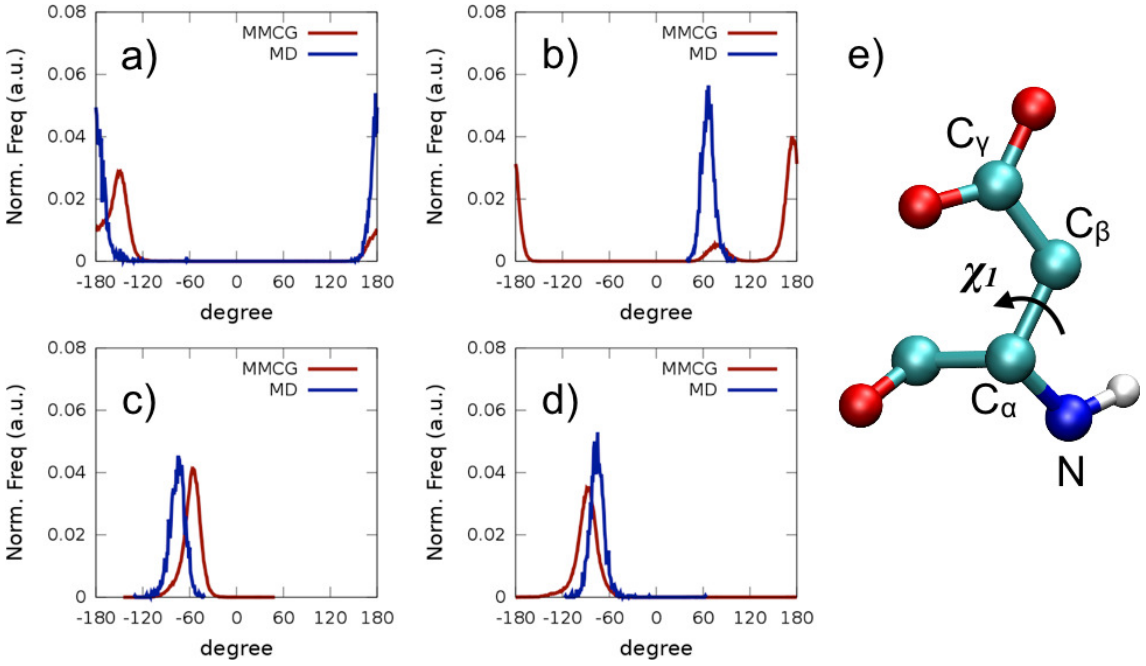


Figure 5.6 Distribution over 200 ns of the side-chain torsional angle χ_1 for four residues (a) R70, (b) D73, (c) D198, and (d) K202. The result of MD and MM/CG simulations are shown in blue and red, respectively; (e) Representation of the side-chain torsional angle χ_1 in an Asp residue. It is defined by four atoms $N - C_\alpha - C_\beta - C_\gamma$.

1990; Fadouloglou *et al.*, 2001). The distribution of χ_1 of four residues (R70, D73, D198 and K202) in both MD and MM/CG simulations show a good agreement with this observation (Figure 5.6a-d).

In addition, the results show a good agreement between two methods with a big overlap of the histograms in the cases of R70, D198 and K202. In the case of D73, the distribution of χ_1 in MM/CG simulation shows two peaks corresponding to two conformations, g^+ rotamer and t rotamer. Meanwhile, only the g^+ rotamer is observed in MD simulation of *Hs*SRII. This indicates that the side-chain of D73 fluctuate between two rotamers during MM/CG simulation.

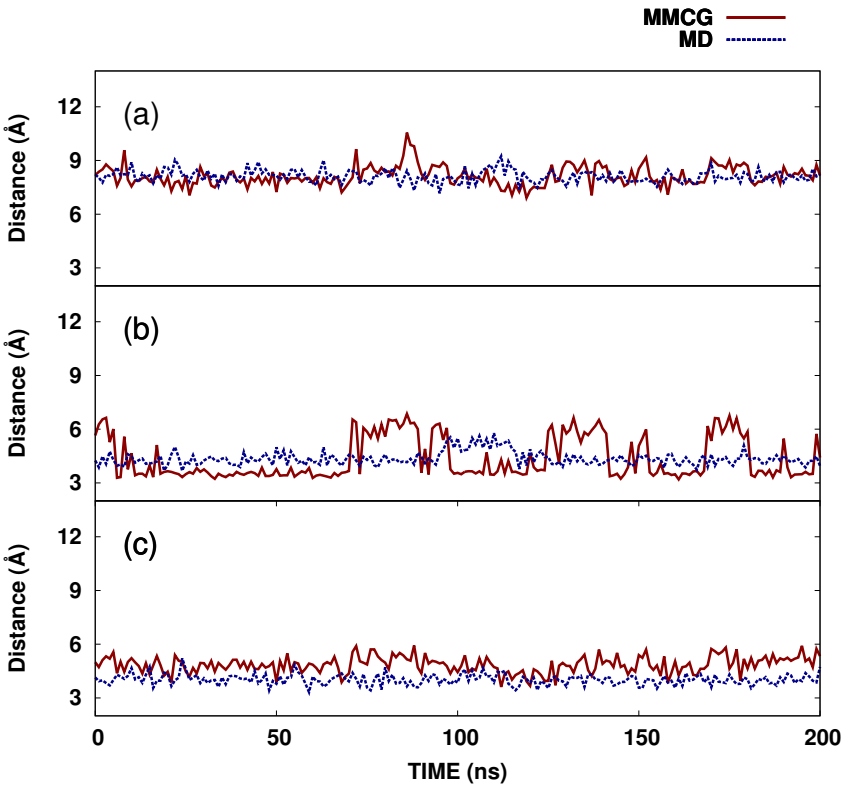


Figure 5.7 Distances as functions of time between Shiff-base and the center of mass of the three residues (a) R70 (b) D73 (c) D198. Data were derived from MD (blue) and MM/CG (red) simulations.

Secondly, the distances among side-chain's center of mass of charged residues were calculated. We observed that the results from MM/CG, in average, do not differ from those calculated from MD by more than 1 Å (Figure 5.7). Indeed, along MM/CG simulation, the distances between Schiff-base and the charged residues, R70, D73 and D198, are 8.1 ± 0.6 Å, 4.4 ± 1.2 Å, and 4.9 ± 0.5 Å, respectively. While those calculated from MD are 8.1 ± 0.3 Å, 4.4 ± 0.4 Å, and 4.0 ± 0.3 Å. In the case of D73, the deviation of the distance between this residue and the Schiff-base is relatively large (around 1.2 Å). This can be observed from the Figure 5.7b as this distance changes between two values, around 3Å and around 6 Å. This is due to the side-chain of D73 fluctuating between two conformations, g^+ rotamer and t rotamer.

Thirdly, the average number of H-bonds between waters and charged residues in the binding site were calculated. The analysis showed that the average of the number of H-bonds keeps fluctuating around 8 in MM/CG and MD simulations. Comparing this result of H-bonds with our MM/CG simulation further established the accuracy of our method (Fig. 5.8).

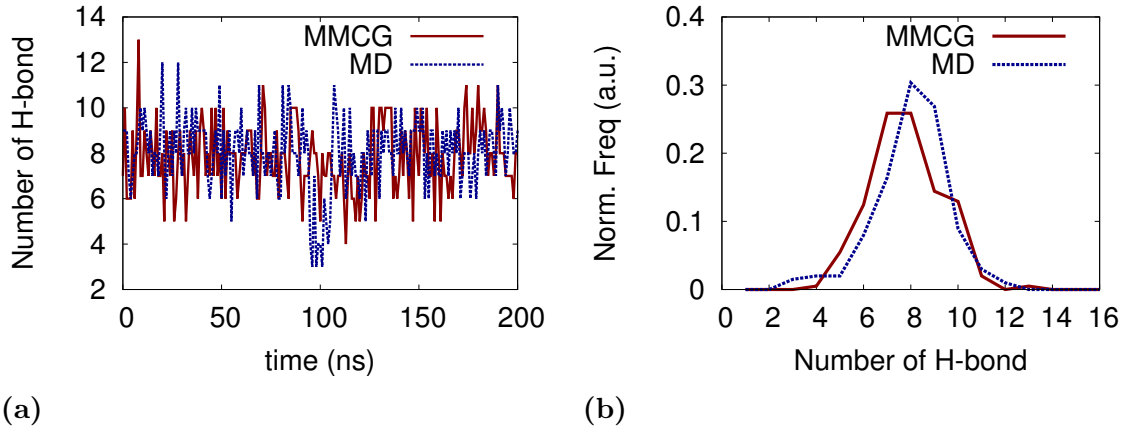


Figure 5.8 The H-bond numbers formed between water molecules and key residues in the binding site including R70, D73, D198 and the retinal (a) The numbers of H-bond as a function of time (b) Distribution of H-bond numbers. Data were derived from MD (blue) and MM/CG (red) simulations.

In summary, our MM/CG simulations are able to reproduce both (i) the residue root mean square fluctuations (RMSF) in the MM regions of *Hs*SRII and (ii) the local microscopic details including the side-chain torsional angle of key residues in the active site, the distances between them and the H-bond patterns of *Hs*SRII. These results suggest that our method can be conveniently applied to 7TMR's systems for which either the size or the necessity of long-time sampling prevents the application of standard MD techniques.

5.5 MM/CG simulations of human β_2 adrenergic receptor

Our MM/CG setup has been shown applicable for *Hs*SRII in section 5.4. *Hs*SRII is a 7TMR with covalently bound ligand. Meanwhile, most of 7TMRs interact non-covalently with their endogenous ligands or drug molecules. Here we validate our MM/CG approach to 7TMRs in complex with non-covalently bound ligands. Hence, we have tested the accuracy of this method on the human β_2 adrenergic

receptor (hβ_2-AR),a member of the GPCRs superfamily involved in cardiac functions (Taylor, 2007; Hoffmann *et al.*, 2004), in complex with S-carazolol (S-Car) or R-isoprenaline (R-Iso) molecules. The results were compared with recent extensive atomistic molecular dynamics simulations on the same system (Vanni *et al.*, 2011).

5.5.1 System setup and simulation details

The calculations are based on the X-ray structure of the hβ2AR in complex with S-Carazolol (PDB code 2RH1) (Cherezov *et al.*, 2007). The third intracellular loop (residues 231-262) is not present in the structure, it was predicted using the Modeller9v3 program (Sali & Blundell, 1993). The structure of the complex between the receptor and the agonist R-Iso was obtained following the procedure of (Vanni *et al.*, 2011). 975 water molecules using SPC model (Berendsen *et al.*, 1981) were added. This constitutes a layer of approximately 15 Å around the MM region. Similarly to our previous works (Neri *et al.*, 2005; Neri *et al.*, 2008), the MM region included all the residues at less than 5 Å from the bound ligand (43 amino acids) which consisted in residues 79-82, 86, 109 to 118, 164-165, 193-195, 199-208, 282, 286, 289-290, 293, 308, 311-316, the corresponding ligands R-Iso or S-Car, and the water molecules. A cut-off of 6 Å (measured from the MM boundary) was considered in order to calculate the residues included in the interface between the MM and the CG subsystems. The resulting total number of atoms is 4597 and 4587 for the hβ2-AR-S-Car and hβ2-AR-R-Iso complexes, respectively.

Starting from this structure, 800 ns MM/CG simulations were performed using a 2 fs time step. Similar to the MM/CG setup of *Hs*SRII, the hβ2-AR complexes were encapsulated in a 31 Å thick implicit membrane, with the transmembrane wall 2.0 Åfrom the C_α atoms. Cutoffs of 16 Å were used for the electrostatic, van der Waals and Go-like interactions. The SHAKE algorithm (Ryckaert *et al.*, 1977) was used to fix the distance in bonds containing hydrogen(s). The temperature was set to 300 K using stochastic dynamics, controlled by an inverse friction constant with a value of 0.4 ps. Periodic boundary conditions were used. RESP charges (Bayly *et al.*, 1993; Wang *et al.*, 2000) for the ligands were taken from ref. (Vanni *et al.*, 2011). All simulations have been performed using our MM/CG implementation in Gromacs 4.5.1 (Kutzner *et al.*, 2007). CG simulations were also carried out for up to 1 μs using a 2 fs time-step. The atom number of this CG system is 314, which is the number of residues present in X-ray structure.

5.5.2 Results

Since each system was only made of approximately 4600 atoms, this allows us to run more than 80 ns/day on 16 CPUs (2.93 GHz Intel Xeon), a 15-fold speedup compared to MD simulations of the same system (Ursula Roethlisberger, private communication). With respect to this speed-up, we notice that a significant speed-up could also have been obtained using an all-atom representation for the whole protein while adding waters on both the intracellular and the extracellular region. However, the combined MM/CG representation allows a further reduction of the system size, since (i) waters are only needed in the extracellular site of the system; and (ii) the use of the CG representation for residues far away from the binding cavity reduces the system size.

Similarly to the case of *Hs*SRII, along the MM/CG trajectories, both complex structures remain stable, as observed in the RMSD calculation (Figure 5.9a). The residue root-mean-square fluctuation (RMSF) calculated using the MM/CG approach follows the trend obtained with all-atom MD (Figure 5.9c,d). The results in panels c) and d) show that globally the fluctuations of the MM and I regions are similar to the fluctuations observed in the all-atom MD simulations. In the CG regions, the fluctuations are much lower, possibly due to the higher rigidity of the CG Go-model force field in comparison to the all-atom force field (Neri *et al.*, 2005). This is observed in the full CG simulations (green line in Figure 5.9c,d).

The main discrepancies between all atom MD and MM/CG simulations are observed in the intracellular loop 3 (ICL3, ranging from residues 233 to 253 approximately), which exhibits larger flexibility in the all-atom simulations. As shown in panel b), the ICL3, which is not present in the crystal structure (Cherezov *et al.*, 2007), is located in the cytoplasmic region far away from the binding cavity. Thus, it was included in the CG region, and it is not expected to directly affect the properties of the binding site. Small differences in the fluctuations (smaller than 1Å) observed in the MM region with respect to the all-atom simulation can be expected due to the different force fields applied in both simulations (Amber99 (Cornell *et al.*, 1995) and Gromos43a1 (Christen *et al.*, 2005), respectively). The differences observed here are of the same order of magnitude as those observed in other comparisons between different force fields (Rueda *et al.*, 2007). Indeed, these differences were also observed in two other independent test simulations of the hβ_2-AR/S-Car complex (Figure S5). The velocity autocorrelation function and the radial distribution function for the water molecules (Figure S2, Table S2) do not differ significantly from the data obtained for bulk water (Mark & Nilsson, 2001; Paesani *et al.*, 2006). This suggests

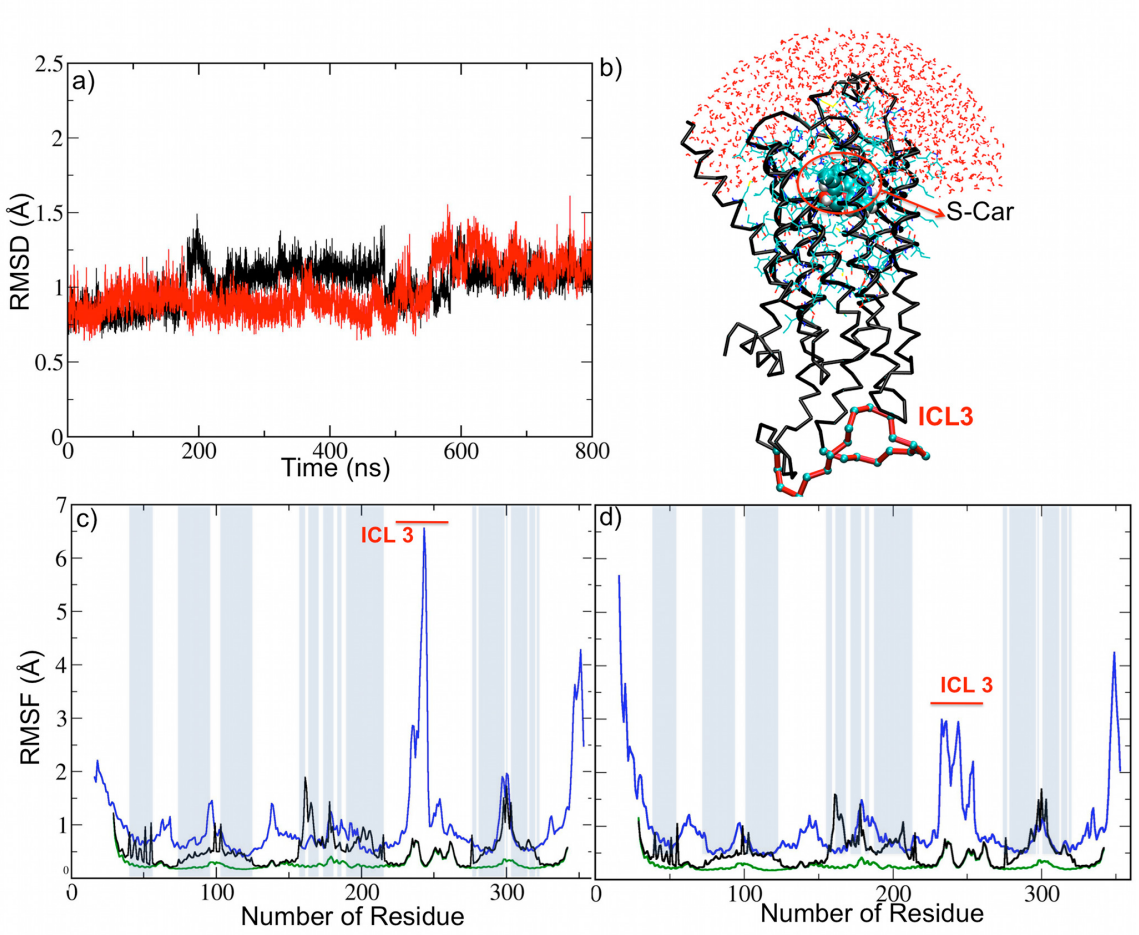


Figure 5.9 a) Root-mean-square-deviation of hβ_2-AR's backbone atoms in the MM/CG simulation of hβ_2-AR.S-Car (black lines) and hβ_2-AR.R-Iso (red lines), relatively to the initial X-ray structure, plotted as a function of time. b) MM/CG representation of the hβ_2-AR.S-Car complex. The MM and I regions, together with the water molecules, are shown in lines representation, the ligand S-Car is shown in spheres and the ICL3 is highlighted in red. c) and d) Root Mean-Square Fluctuations of hβ_2-AR's backbone atoms calculated based on MD simulations (Vanni *et al.*, 2011). Results for all-atom simulations, MM/CG simulations and CG simulations are shown in blue, black and green lines, respectively. Results for hβ2-AR.S-Car and hβ_2-AR.R-Iso complexes are shown in panels c) and d), respectively. Residues included in the MM and I regions (which feature all-atom representation) are highlighted with grey bars on the plots.

that the MM/CG scheme does not significantly constraint either the structure or the mobility of water molecules.

The structural determinants at the active site were well maintained (Figure 5.10). The interactions between the ligands and the protein matrix evolved in good agreement with the all-atom simulations. These included H-bond interactions between the ligands and residues Asn7.39, Asp3.32, Ser5.42, Ser5.43 and Asn6.55 (the last two residues, only in the case of the hβ_2-AR/R-Iso complex). However, small differences have been observed in some specific interactions. The H-bonds between Asn7.39 carbonyl group (OCO) and the NH2+ group of the agonist (R-Iso) or the reverse agonist (S-car) are only partially formed in the MM/CG simulation. (Figure 5.10 panel b.iii, d.ii). The H-bond between NH2+ group of Asn6.55 and R-Iso.OH is longer in the MM/CG simulations than in the all-atom simulations (Figure 5.10, panel d.v). Finally, both R-Iso.OH groups form H-bonds to Ser5.42 and Ser5.43 in the MM/CG simulation, while only the R-Iso.O(2)H group forms an H-bond to Ser5.42 in the all-atom simulation (Figure 5.10, panel d.iv)(Vanni *et al.*, 2011). The fact that the simulations performed in this work show a high level of agreement with the all-atom simulations allows us to suggest that the results do not critically depend on the choice of the force field.

To investigate the predictive power of the method at the structural level, we ran a simulation in which we located the ligand S-Car in a position different from the crystallographic pose (Figure 5.11). In this new position, none of the interactions with the residues found in the X-ray structure of hβ_2-AR/S-Car complex (Cherezov *et al.*, 2007) are present. In these new simulations, the ligand migrates to the correct pose between 150 and 200 ns, forming the key interactions with Ser5.42, Ser5.43, Asp 3.32 and Asn7.39 (Figure 5.11). Hence, our method is not only able to conserve the ligand-receptor structure but also able to predict the correct pose of the ligand in the binding site.

All together, the MM/CG simulations reproduce the key structural features of the active site found in the MD simulations. The introduction of the potential wall to represent the membrane leads to a large reduction in the computational cost of the simulation, conserving the stability of the protein structure. Moreover, the ligand remains in a stable position inside the binding cavity throughout the long-scale simulation, conserving the key interactions with the protein matrix at the binding site.

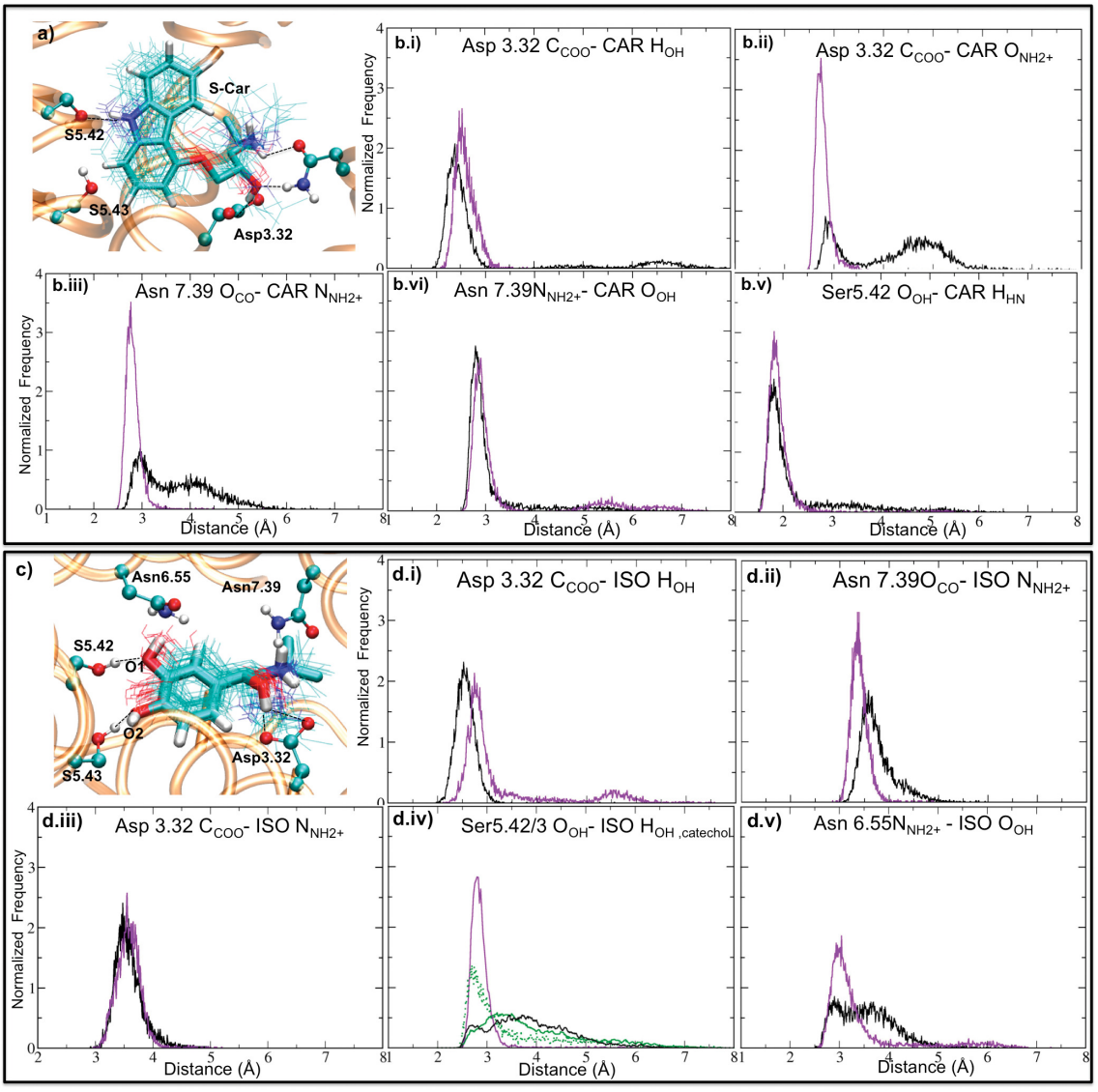


Figure 5.10 MM/CG and MD simulations of hβ_2-AR/S-Car and hβ_2-AR/R-Iso complexes. H-bond interactions between S-Car and hβ_2-AR, reported in ref. (Vanni *et al.*, 2011). Panels a) and c) display snapshots of the binding site, obtained from the MM/CG trajectory of the S-Car and R-Iso complexes respectively. Superimposed positions of the agonist and reverse agonist along the trajectories are shown in lines representation (snapshots were taken every 40 ns). The distribution of all H-bonds and salt involved in S-Car and R-Iso binding are shown in Panels b.i) to b.v) and d.i) to d.v). Results of the MM/CG and the all-atom MD simulations (Vanni *et al.*, 2011) are shown in black and violet lines respectively. In d.iv), black and violet lines correspond to the distance between the O1 (labeled in panel a) and the OH group of Ser5.42 in the MM/CG simulations, and the O2 and the OH group of Ser5.42 in the all-atom simulations. The distance between O1 and Ser5.43-OH and between O2 and Ser5.43-OH in the MM/CG simulation are shown in full and dotted green lines respectively.

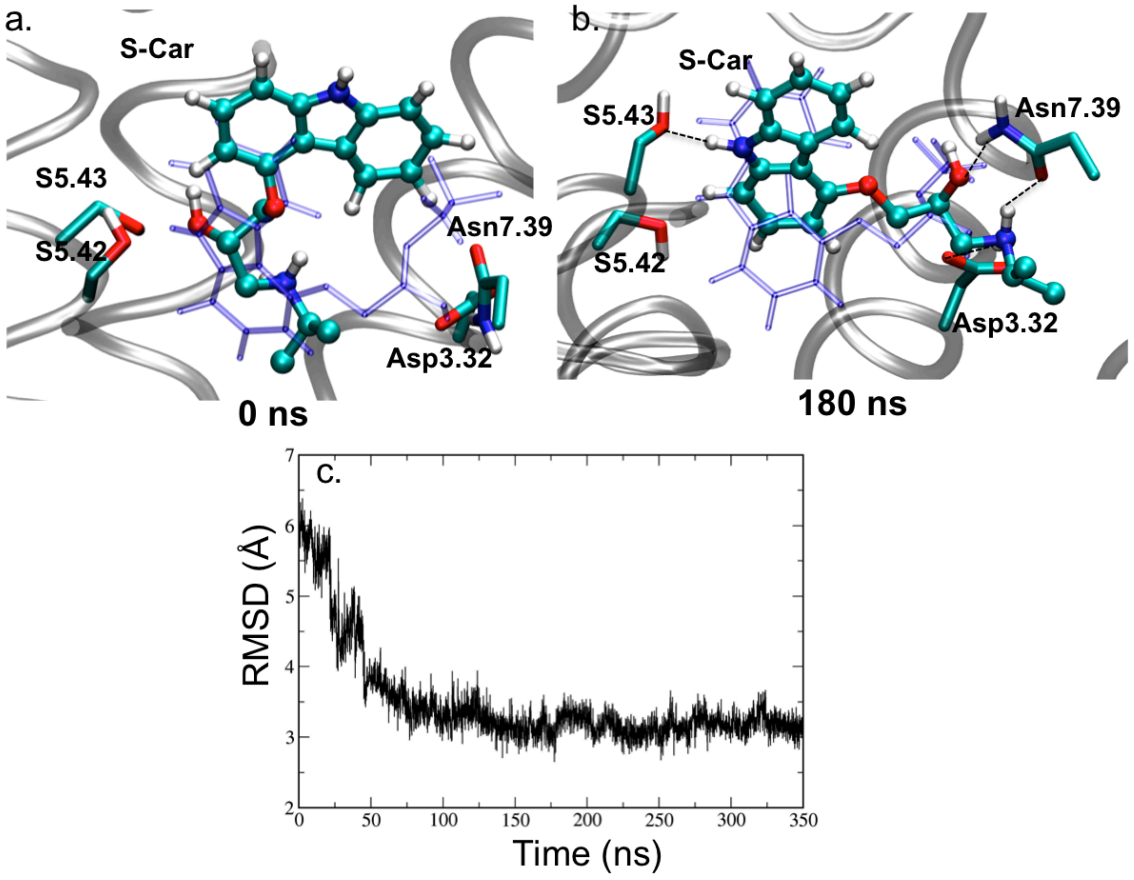


Figure 5.11 MM/CG simulation of hβ_2-AR/S-Car complex. Here the S-Car ligand is originally located at a position different from the crystallographic pose. Panels a) and b) show snapshots taken at 0 ns and 180 ns of the simulation. Panel c) shows the RMSD of the S-Car ligand with respect to the crystallographic position.

5.6 Discussions and conclusions

The results obtained indicate that the MM/CG simulations reproduce the key structural features of the active site found in the MD simulations. Particularly, for the case of *Hs*SRII where the ligand binds covalently to the protein, the distances between key residues and the ligand are similar to those of MD simulations. Moreover, the distribution of side-chain torsional angles and the H-bond numbers between the key residues and water molecules in the binding site are also preserved. In the case of hβ2-AR, the ligands remain in a stable position inside the binding cavity along the long-scale simulation, conserving the key interactions with the protein matrix in the binding site. In addition, in both testing cases, the stability of the protein with respect to the X-ray structure has been shown by the low value of the RMSD of the C_α atoms during the dynamics.

The introduction of the potential wall to represent the membrane leads to a large reduction in the computational cost of the simulation. The overall system of *Hs*SRII was made of 2030 atoms. This allowed us running up more than 110 ns/day on 8 CPUs (2.93 GHz Intel Xeon Cluster). This speeded up ∼25 times comparing to the time needed to perform atomistic MD simulations of the same system. Mean while, the overall system were made of 4597 and 4587 atoms for the hβ2-AR.S-Car and hβ2-AR.R-Iso complexes respectively. This allows us to run more than 80 ns/day on 16 CPUs (2.93 GHz Intel Xeon). This is a 15-fold speedup compared to MD simulations of the same system (Ursula Roethlisberger, private communication).

Due to this low computational cost, it is very powerful to combine MM/CG approach with homology modeling and molecular docking methods, as the data are often complementary one to the other (Schwede, 2008). Homology modeling approach is a very useful tool to predict the 3D structures of protein of which experimentally structural information is not available (Yarnitzky *et al.*, 2010). Molecular docking is then often used to predict poses of ligands onto their targets (Kitchen *et al.*, 2004; Biarnés *et al.*, 2010). Because of its computational efficiency, molecular docking is routinely used in drug screening from large databases (Moitessier *et al.*, 2008) as well as in the design of new ligands (Kitchen *et al.*, 2004). Unfortunately, standard docking procedures on homology modeling suffer from severe limitations which greatly limit the predictive power of these methods. These limitations include the difficulty in predicting correctly side chains orientations in the binding site (Eswar *et al.*, 2007) and neglecting the presence of explicit solvent (Camacho, 2005). This is particularly important for 7TMRs, as water molecules can be found in the binding site of these receptors and they may be crucial to stabilize the ligand (Angel *et al.*, 2009; Nygaard *et al.*, 2010). The MM/CG approach can be used to perform relatively long simulations, up to μs time scale, needed to explore the conformational space of side-chain orientations and water molecules positions in the binding site of the predicted models. Thus, the MM/CG method may allow us to predict the ligand poses on the 7TMR models and hence provide insights into the molecular basis of ligand selectivity in 7TMRs.

In conclusion, we have presented a hybrid MM/CG method to investigate hβ2-AR and *Hs*SRII, two members of 7TMRs. The method could be extended to a large number of 7TMR/ligand complexes and may reveal itself very useful for computer-aided drug design.

CHAPTER 6

Structural predictions of downstream effectors in the olfactory signaling cascade

6.1 Overview

As described in the chapter 2, the interaction of an OR with an odorant leads to the activation of the G-protein (i.e., G-olf) which in turn triggers the enzyme AC resulting in an increase of cyclic adenosine monophosphate (cAMP) in the olfactory cells. This sequence of events is referred to as a multi-step "cascade". The direct activation of a cation permeable channel by cAMP is the final step in producing the odor induced ionic current. In the presence of normal physiological extracellular Ca^{2+}, the second messengers (i.e., cAMP) elicits the opening of the channel allowing Ca^{2+} to flow in, and the increase in intracellular Ca^{2+} concentration appears to activate a chloride current that helps depolarize the olfactory cell. Thus, the chemical interactions of ORs with volatile molecules lead ultimately to the production of action potentials that will carry information about the external world to the brain (Figure 2.3).

When I started my thesis work, different components involved in the cascade had been studied in our lab, including ORs homology modeling (Khafizov *et al.*, 2007),

G-protein modeling (Khafizov *et al.*, 2009), CNG channel homology modeling (Giorgetti *et al.*, 2005b) and chloride channel modeling (Kranjc *et al.*, 2009). In an effort to complete the molecular description of the olfactory receptor cascade, two studies have been performed in this work including (i) homology modeling of the ACIII in complex with G-olf, and (ii) refining and extending the homology models of CNG ion channel using recent cysteine scanning mutagenesis (CSM) data (Nair *et al.*, 2009a). Almost all responses to odorants in olfactory sensory neurons (OSNs) are facilitated by AC enzymes and CNG channels (Su *et al.*, 2009). Thus, the molecular description of their structures will allow a better understanding of their functions as well as the olfactory signal transduction at the molecular level.

6.2 Structural prediction of the Adenylyl cyclase III

6.2.1 Introduction

Adenylyl cyclases (ACs) are the most abundant enzymes catalyzing the synthesis of the universal second messenger cAMP from ATP. They can be activated or inhibited by binding with GTP-bound α-subunit of specific G-protein. Following activation of AC, the resulting cAMP activates target proteins such as protein kinases, ion channels and transcription factors, finally resulting in a cellular response to the primary stimulus. Cloning of mammalian ACs have been identified nine isoforms of ACs (Cooper, 2003), in which the isoform ACsIII has been known to be involved in olfactory transduction (Bakalyar & Reed, 1990). Beside G-protein, ACIII can be also stimulated by Ca^{2+}/calmodulin, a calcium-binding messenger protein expressed in all eukaryotic cells (Ronnett & Moon, 2002).

These proteins contain two transmembrane (TM) domains (M1 and M2), each crossing the membrane six times. The main functional parts are located in the cytoplasm and can be subdivided into the N-terminus, C1a, C1b, C2a and C2b. The C1 region exists between TM helices six and seven and the C2 region follows TM helix 12. The C1a and C2a domains form a catalytic dimer where ATP binds and is converted to cAMP, a molecule that plays a fundamental role in olfaction as a second messenger and mediates signal transduction for a wide variety of odorants (Takeuchi *et al.*, 2003).

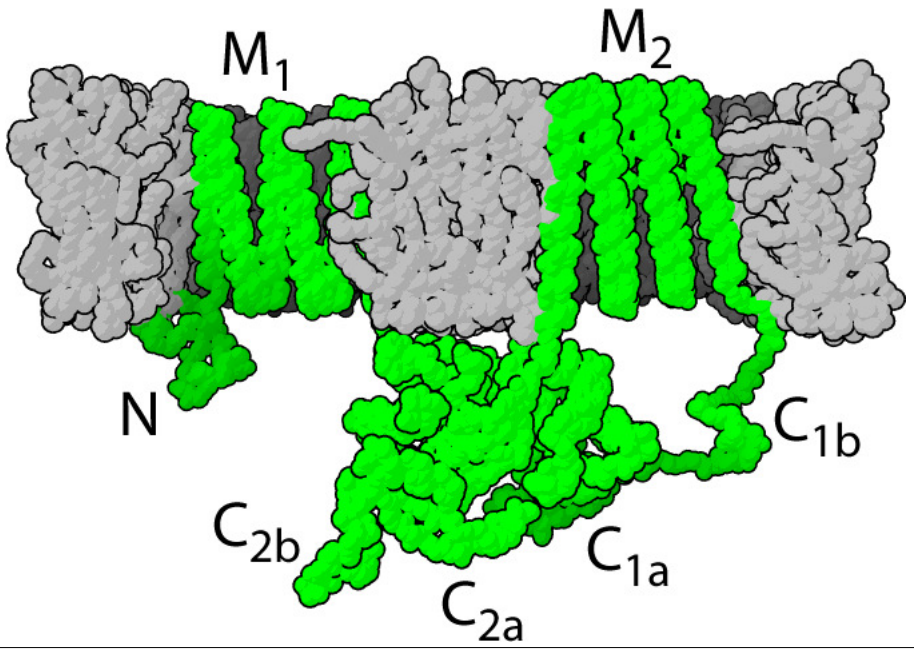
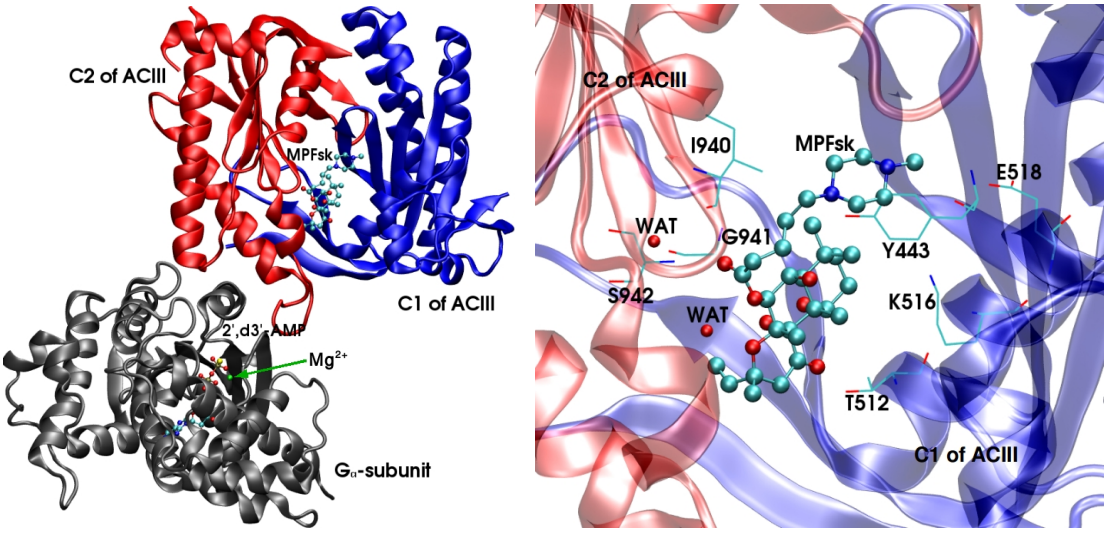


Figure 6.1 Structure of adenylate cyclase. Adapted from the Wikimedia Commons file `http://upload.wikimedia.org/wikipedia/commons/c/c2/Adenylyl_cyclase.png`

6.2.2 Material and Methods

Using homology modeling, we have here built the cytoplasmic domain structure of human AC III - the enzyme involved in ORs signaling - in complex with human G_α-subunit of olfactory system, known as G_α-olf (Uniprot accession code: O60266 and P38405 respectively) (Figure 6.2a). The template for the C1 and G_α-subunit was the X-ray structure of bovine AC in complex with $G_{s\alpha}$-GTP$_\gamma$S (Tesmer *et al.*, 1997). The template used as scaffold for the modeling of the C1, C2 heterodimer (PDB code 1AZS) is the catalytically active form of adenylyl cyclase in a complex with its stimulatory heterotrimeric G protein α-subunit ($G_{s\alpha}$), forskolin and a Mg^{2+} ion in the G-protein subunit, and with an adenosine analog, known also as P-site inhibitor (2'-deoxy-3'-adenosine monophosphate, 2',d3'-AMP) of AC's. The sequence identities of C1 domain and G_α- subunit are relatively large, 58% and 79%, respectively. The sequence identity of C2 domain is low (23%) across the AC family, because of the structural diversity. Hence, C2 was constructed by multiple templates modeling technique (Biegert *et al.*, 2006), using the X-ray structure of rat ACII in complex with forskolin (Zhang *et al.*, 1997) and the crystal structure of the mycobacterial ACIV (Ketkar *et al.*, 2006) as templates. The templates were aligned to the target structures using the HHsearch and HHalign programs (Biegert *et al.*, 2006). The 3D structure was built using the MODELLER 9v3 program (Eswar *et al.*, 2007).

6.2.3 Results and Discussions



(a) (b)

Figure 6.2 (a) Structural model of the cytoplasmic domain of human AC type III in complex with its cognate Gα-subunit (G_α-olf). The C1 domain, C2 domain (of ACIII), and the G-olf are showed in blue, red, and black, respectively; (b) forskolin (MPFsk) binding site in C1-C2 heterodimer of ACIII is showed in the close-up view on the right.

Figure 6.2 shows the structural model of the cytoplasmic domain of human AC type III in complex with its cognate Gα-subunit (G_α-olf). This model suggests that the active site residues binding to the ligand MPFsk (Y443, T512, K516, E518, I940, G941, and S942) are the same as in the template (Tesmer *et al.*, 1997). The presence of a ligand that is specific for the template structure and the amino-acid conservation in the putative binding site may allow us to make hypothesis about the mechanism underlying the signal transduction and the active site localization. However, docking studies should have to be performed in order to have a more complete picture of the model. In addition, MD simulations could be performed to relax the structure and investigate the hydration of the active site cavity.

6.3 Structural prediction of the CNGA1 channel

6.3.1 Introduction

Cyclic nucleotide-gated (CNG) ion channels are gated by cGMP and cAMP second messengers. CNG channels are nonselective cation channels that are found in the

membranes of various tissue and cell types, and are significant in sensory transduction as well as cellular development. They produce the electrical signal in response not only to odor stimulation but also to light in the vision process, as well as in other key cellular functions such as hormone release and chemotaxis.

CNG channels belong to the superfamily of tetrameric voltage-gated ion channels (Biel & Michalakis, 2007; Anselmi *et al.*, 2007). CNG channels have a very complex structure with various subunits and domains that play a critical role in their function (Fig. 6.3). They consist of two domains: (i) a transmembrane domain formed by six transmembrane helices (S1-S6) and a pore helix (P-helix); (ii) a cytoplasmic domain formed by the cyclic nucleotide binding domain (CNBD), which is linked to the transmembrane domain through the so called C-linker region.

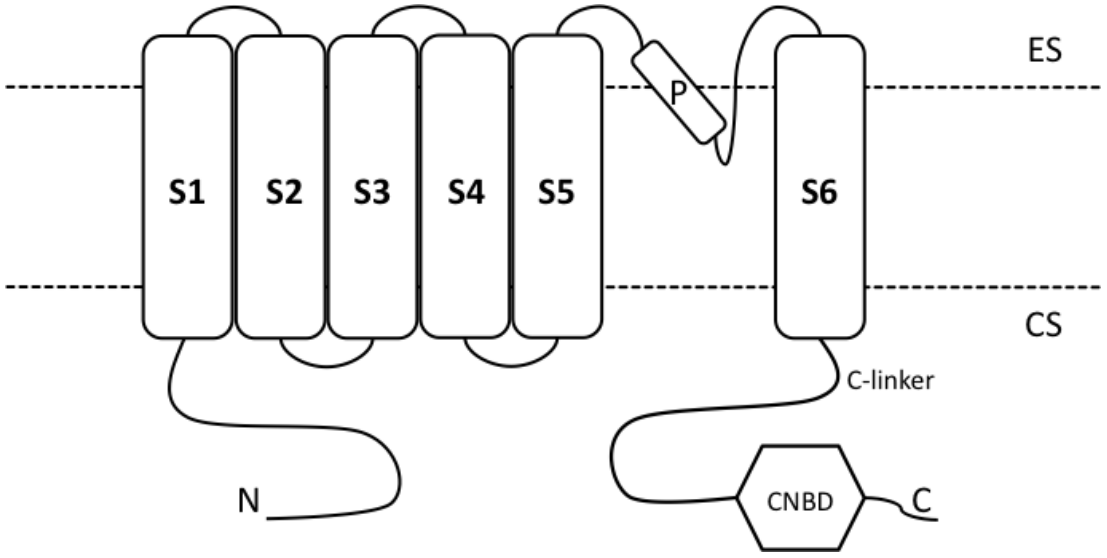


Figure 6.3 Topological representation of a CNG channel A subunit, consisting of six transmembrane helices (S1-S6), a pore helix between S5 and S6 (P-helix), a cyclic nucleotide binding domain (CNBD) at the C terminus which is linked to the transmembrane domain through the C-linker region.

CNG channels have been found less voltage-dependent than the voltage-gated K+ (Kv) channels (Anselmi *et al.*, 2007). Namely, the charge of their voltage sensor, the S4 helix, is lower than that of Kv channels and moreover a conserved proline in the S4-S5 linker region of all CNGA channels is present. This is likely to reduce the mechanical coupling between S4, S5 and S6 helices. The alignment of the sequences across K+ channels, along with experimental constraints, has then been used to provide a structural basis of CNG channels (Giorgetti *et al.*, 2005a).

The analysis of residue accessibility in the pore of CNG channels, based on the cysteine CSM method (Akabas *et al.*, 1992; Kürz *et al.*, 1995; Bénitah *et al.*, 1997),

has shown that CNG and Kv channels share the same gross topology (Becchetti & Gamel, 1999). A similar analysis performed in the S6 domain of CNGA1 channels from V384 to S399 (Flynn & Zagotta, 2001; Flynn & Zagotta, 2003) suggested that also in CNG channels this domain has a helical configuration, possibly crossing at a hypothetical constriction located between residue V391 and S399. On the basis of their results (Flynn & Zagotta, 2003), the authors proposed that the closed and open conformations of the CNGA1 channels are similar to the 3D structure of KcsA and MthK, respectively. Another work based on CSM method, using the oxidizing agent CuP (Hua & Gordon, 2005), concluded that residues from Q417 to V424 in neighboring subunits become closer in the open state and are far apart in the closed state. Therefore, during channel gating, residues in homologous subunits from V384 and S399 move away from each other in a different way from the residues from Q417 to V424.

The purpose of the present study is to verify and extend previous investigations (Flynn & Zagotta, 2001; Flynn & Zagotta, 2003; Hua & Gordon, 2005; Giorgetti *et al.*, 2005a; Mazzolini *et al.*, 2008; Nair *et al.*, 2009b) on the spatial rearrangement of amino acids during gating in the S6 domain and in the initial portion of the C-linker, focusing in particular on understanding the relative motion of residues from S399 to Q417. Specific residues between F375 and V424 were mutated to a cysteine in the CNGA1 and $CNGA1_{cys\text{-}free}$ background; modifications induced by intracellular Cd^{2+} (Mazzolini *et al.*, 2002; Rothberg *et al.*, 2003) were analyzed in these mutant channels in the open and closed state, and the effect of other sulfhydryl reagents, such as 2-(trimethylammonium) ethyl methanethiosulfonate bromide (MTSET) and methanethiosulfonate (MTS) cross-linkers (Loo & Clarke, 2001), was investigated on selected mutants.

6.3.2 Material and Methods

Estimation of the distance between C_α of coordinating cysteines

The experimental constraints were obtained by CSM of residues present principally along the channel axis. Mutated channels were then studied by measuring the differences of current blockage upon the introduction of metals, such as Cd^{2+}, and agents capable of interacting with cysteines in the solution (Nair *et al.*, 2006). The distance between the C_α of cysteines coordinating Cd^{2+} ions or cross-linkers such as 1,2-ethanediyl bismethanethiosulfonate (M-2-M) and 1,4-butanediyl bismethanethiosulfonate (M-4-M) was estimated (Mazzolini *et al.*, 2008). We calculated that, if Cd^{2+}

inhibits the channel by coordinating to two cysteines, the distance between the C_α of coordinating cysteines is $\leq$ 10.5 Å. If channel inhibition is observed with M-2-M but not with Cd^{2+} the distance between the C_α of coordinating cysteines is between 10.5 and 12.3 Å. Similarly, if the inhibition is caused by M-4-M and not by Cd^{2+} and M-2-M, the distance between the C_α of coordinating cysteines is between 12.3 and 14.7 Å (Mazzolini *et al.*, 2008; Nair *et al.*, 2009b).

Homology modeling

All of the steps have been performed following the ref. (Giorgetti *et al.*, 2005b). The residues belonging to the template for the pore region, the pore helices and the S6 helices were aligned with residues from 39 to 124 of the X-ray structures of the KcsA K+ channel from Streptommyces lividans (Doyle *et al.*, 1998) and with residues from 67 to 98 of the K+ MthK channel from M. Thermautotrophicum (Jiang *et al.*, 2002). The sequence of the template for the cytoplasmatic portion of S6 helices (from 402 to 422) was aligned with the residues from 443 to 460 of the X-ray structure mHCN2 channel (Zagotta *et al.*, 2003). The alignment was performed using the program Clustal W (J D Thompson, 1994). The 3D structural models of the P helix, S6 and the C-linker of CNG channel in the closed and open states were produced using Modeller 6v2 (Eswar *et al.*, 2007). This region was restrained to satisfy distance constraints between filter residues and between the different subunits of C-linker (Nair *et al.*, 2009a).

6.3.3 Results and Discussions

The modifications caused by intracellular Cd^{2+} and other sulfhydryl reagents, such as MTSET and MTS cross-linkers (Loo & Clarke, 2001), were investigated in presence (open state) and absence (closed state) of 1 mM cGMP when residues in the S6 domain and C-linker were mutated to cysteines in CNGA1 and in $CNGA1_{cys\text{-}free}$ background. These regions can be divided into three segments with different properties (Figure 6.4): (1) the first segment, from F375 to S399, where the effect of Cd^{2+} ions in mutant channels constructed in the CNGA1 is very similar in the closed and open state; (2) the second segment, from N400 to Q409, in which Cd^{2+} ions inhibited in the closed state many mutant channels constructed in the CNGA1 background but not in $CNGA1_{cys\text{-}free}$ background; and (3) the third segment, from A410 to V424, where the effect of Cd^{2+} ions was similar for some, but not all, of the mutant channels constructed in the CNGA1 and $CNGA1_{cys\text{-}free}$ background.

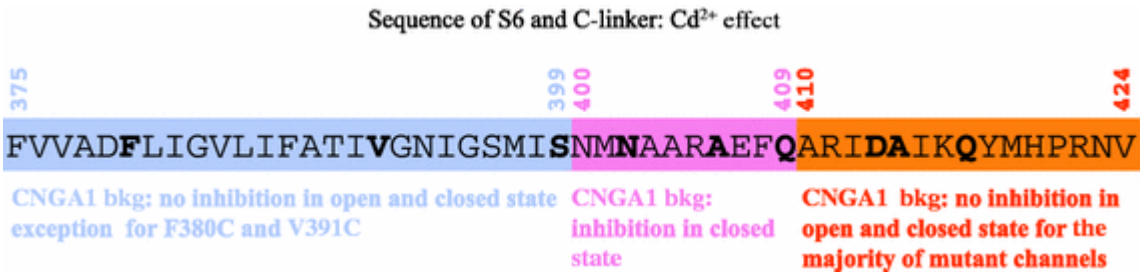


Figure 6.4 Sequence of the S6 and C-linker regions. Schematic representation of Cd^{2+} effect and relative subdivision into three segments: from 375 to 399 in violet, from 400 to 409 in pink, and from 410 to 424 in orange. *Bkg*: background.

These experimental observations can be rationalized in the molecular model presented in Figure 6.5. This model assumes that, in the closed state, residues from F375 to H420 have an alpha-helix conformation. Mutant channels V391C are inhibited by Cd^{2+} ions in both the closed and open state, and a spontaneous rundown of the cGMP-activated current of mutant channels S399C is observed in both the closed and open state. These results suggest that residues from F375 to S399 in homologous subunits are near each other in the closed state and remain at a close distance also during channel gating.

Residues in homologous subunits from N400 to Q409 are near each other in the closed state (Figure 6.5a) but at a relative greater distance than V391 because of a reduced efficiency of Cd^{2+} inhibition. Mutant channels A406C and Q409C are not blocked by Cd^{2+} ions in the open state, suggesting that residues from A400 to Q409 move apart in the open state (Figure 6.5b). In contrast, mutant channels D413C and Y418C were inhibited by Cd^{2+} ions in the open state but not in the closed state, suggesting that residues from D413 to T418, move differently: these residues are nearer each other in the open state (Figure 6.5b).

These conclusions are in agreement with the model proposed by Giorgetti et al. (Giorgetti *et al.*, 2005a) for residues from F375 to approximately Q409 and with the model proposed by Hua and Gordon (Hua & Gordon, 2005) for residues from Q417 to V424. Conformational changes of these domains during channel gating have also a rotational component, as previously suggested (Johnson & Zagotta, 2001; Nair *et al.*, 2006).

On the basis of the combination of experimental and theoretical studies, we have proposed that a rotational movement begins in the C-linker region (Figure 6.6). This rotational movement is then transmitted upwards, making the upper part of S6 rotate anticlockwise. Due to the direct interaction of S6 with the P-helix, this

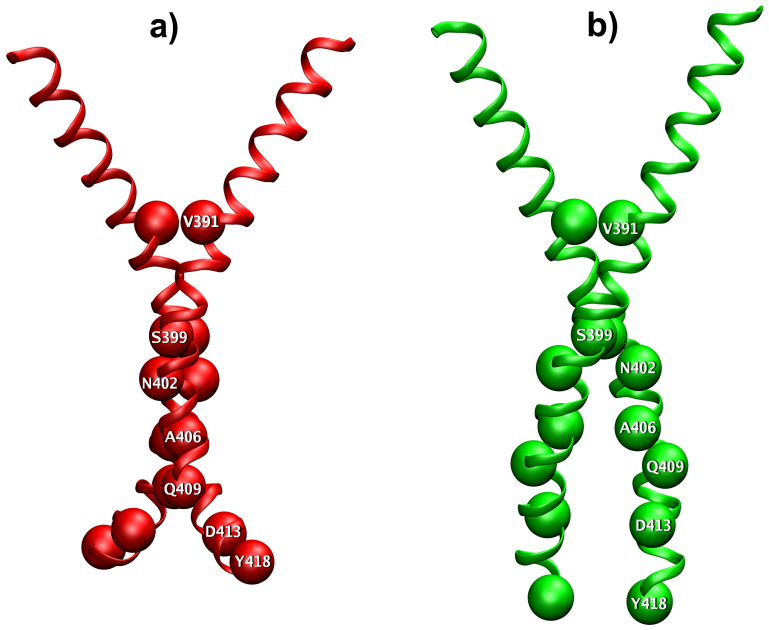


Figure 6.5 Model of possible movement of the S6-C-linker transmembrane helices of CNG channel. In red, front view of closed state (a) and in green, front view of open state (b). The residues from N400 to Q409 are near in the closed state (red) and residues from D413 to Y418 are near in the open state (green). For clarity only two subunits are shown.

motion is transmitted to the latter, which rearranges itself so as its terminal T360 residues and, therefore, the lower part of the pore wall, lead to the opening of the pore lumen. Thus, the initial event of cyclic nucleotide binding is transmitted to the pore walls by a remarkable and sophisticated coupling of conformational changes spanning throughout the entire cytoplasmic and transmembrane domains of the channel.

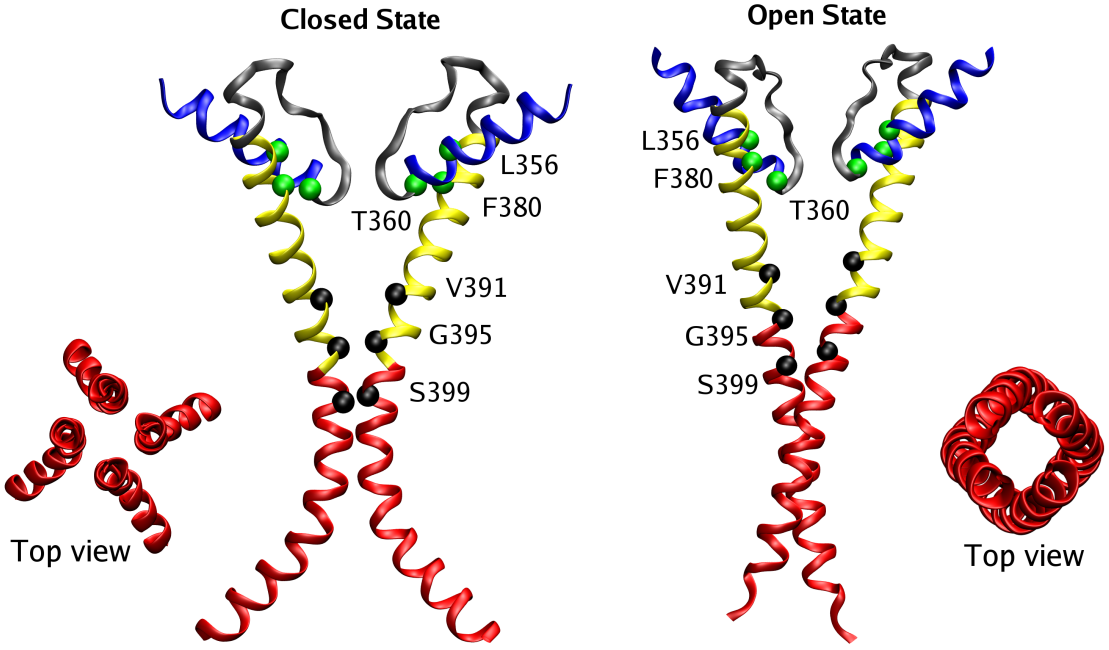


Figure 6.6 Structural models of the P-helix (in blue), S6 (in yellow), and the C-linker (in red) of CNG channel in the closed state and open state. For clarity, only two subunits are shown. Insets: top view of the N-term@C-linker (in red).

CHAPTER 7

Conclusions and Perspectives

In this thesis, I have used different computational approaches to study different components involving in the 7TMRs signaling. These include the Sensory Rhodopsin II in phototaxis signaling of *Halobacterium salinarium*, Cyclic nucleotide-gated ion channel and Adenylyl cyclase III in olfactory signaling of human. Another important contribution of this thesis is the development of the hybrid MM/CG approach to 7TMRs.

First, we have used molecular dynamics simulations to investigate different forms of *Hs*SRII (the ground state, the so-called intermediate M-state, and the D73N mutant) based on recent structural information from Prof. Labahn's lab in Forschungszentrum Jülich. Our calculations allow us to suggest that the reorganization of charged residues in the binding site may produce the displacements of helix G respect to helix C in both M-state and D73N mutant. In addition, we observed the correlation between the Y171-S201 H-bond with the displacement between helix F and helix G in M-state. The rearrangements of these two helices are expected to trigger the signal cascade of *Hs*SRII, as already observed in the functional homologue of SRII from *Natronobacterium pharaonis*. Finally, our calculations suggested that the conformational changes of the B-C and F-G loops in the extracellular side as well as of the E-F loop and the C-terminal of helix G in the cytoplasmic side may also be involved in the activation of *Hs*SRII.

Next, we have developed a hybrid molecular mechanics/coarse-grained (MM/CG) approach to 7TMR/ligand complexes. The accuracy of the method was established

by comparison with recent atomistic molecular dynamics simulations on the *Halobacterium salinarium* SRII and human β_2AR (Vanni *et al.*, 2011), two members of 7TMRs superfamily. The MM/CG simulations reproduce the key structural features of the active site found in the MD simulations. Due to its low computational cost, this method can be applied to the study of a large number of GPCRs-ligand complexes.

Currently, our MM/CG method is being applied in our group to study ligands binding to several GPCRs including the human bitter taste receptor hTAS2R38, the mouse odorant receptor MOR174-9, and the human chemokine receptor CCR5.

Finally, by using homology modeling approaches, we have here built (i) the cytoplasmic domain structure of human ACIII - the enzyme involved in ORs signaling - in complex with human Gα-subunit of olfactory system, and (ii) the P helix, S6 and the C-linker of CNG channel in the closed and open states. Specially, in the work of CNG channel, in combination with experimental studies, we have proposed that a rotational movement begins in the C-linker region. This rotational movement is then transmitted upwards, making the upper part of S6 rotate anticlockwise. Due to the direct interaction of S6 with the P-helix, this motion is transmitted to the latter which leads to the opening of the pore lumen.

As future perspectives of this work, our lab has been attempting to determine the impact of GTP-bound G-proteins on the catalytic activity of the AC enzymes. The structural model of the cytoplasmic domain of human AC type III in complex with its cognate G_α-subunit (G_α-olf) predicted from this thesis will be used as the starting point for the bioinformatics part of this study. Our lab plans to model the complexes between GTP-bound $G_{\alpha s}$ and all AC isoforms, as well as the ones of GTP-bound $G_{\alpha i}$ with all AC isoforms. By performing molecular simulations, we plan to investigate the enzymatic reaction of AC in the presence and in the absence of the GTP-bound G-proteins.

APPENDIX A

7TMR structures in PDB

No.	Protein name	Species	PDB code	Resolution (Å)
1	Bacteriorhodopsin (BR)	Halobacterium salinarium	2BRD	3.5
2	Bacteriorhodopsin (BR)	Halobacterium salinarium	1AT9	3.0
3	Bacteriorhodopsin (BR)	Halobacterium salinarium	1AP9	2.35
4	Bacteriorhodopsin (BR)	Halobacterium salinarium	1BRR	2.90
5	Bacteriorhodopsin (BR)	Halobacterium salinarium	1BRX	2.30
6	Bacteriorhodopsin (BR)	Halobacterium salinarium	1QKP	2.10
7	Bacteriorhodopsin (BR)	Halobacterium salinarium	1QKO	2.10
8	Bacteriorhodopsin (BR)	Halobacterium salinarium	1M0K	1.43
9	Bacteriorhodopsin (BR)	Halobacterium salinarium	1M0L	1.43
10	Bacteriorhodopsin (BR)	Halobacterium salinarium	1QHJ	1.90
11	Bacteriorhodopsin (BR)	Halobacterium salinarium	1C3W	1.55
12	Bacteriorhodopsin (BR)	Halobacterium salinarium	1C8R	1.80
13	Bacteriorhodopsin (BR)	Halobacterium salinarium	1C8S	1.80
14	Bacteriorhodopsin	Halobacterium salinarium	3NS0	1.78
15	Bacteriorhodopsin	Halobacterium salinarium	3NSB	1.78
16	Bacteriorhodopsin	Halobacterium sp. nrc-1	3T45	3.01
17	Halorhodopsin (HR)	Halobacterium salinarium	1E12	1.8
18	Halorhodopsin (HR)	Natronomonas pharaonis	3A7K	2.0
19	Sensory Rhodopsin I (SRI)	Anabaena (Nostoc) sp. PCC7120	1XIO	2.0
20	Sensory Rhodopsin II (SRII)	Natronomonas pharaonis	1JGJ	2.40
21	Sensory Rhodopsin II (SRII)	Natronomonas pharaonis	1H68	2.10
22	Sensory Rhodopsin II (SRII)	Natronomonas pharaonis	1H2S	1.93
23	Sensory Rhodopsin II (SRII)	Natronomonas pharaonis	2KSY	NMR structure
24	Sensory Rhodopsin II (SRII)	Natronomonas pharaonis	3QDC	2.50
25	Sensory Rhodopsin II (SRII)	Natronomonas pharaonis	3QAP	1.90
26	Archaerhodopsin-1 (aR-1)	Halorubrum sp. aus-1	1UAZ	3.4
27	Archaerhodopsin-2 (aR-2)	Haloroubrum sp. aus-2	1VGO	2.5
28	Archaerhodopsin-2 (aR-2)	Haloroubrum sp. aus-2	2EI4	2.10
29	Archaerhodopsin-2 (aR-2)	Haloroubrum sp. aus-2	2Z55	2.50
30	Xanthorhodopsin	Salinibacter ruber	3DDL	1.9
31	Acetabularia Rhodopsin II (ARII)	Acetabularia acetabulum	3AM6	3.20
32	Channelrhodopsin (ChR)	Chlamydomonas reinhardtii	3UG9	2.30

Table A.1 Available prokaryote 7TMR structures in Protein Data Bank (PDB), adapted from `http://blanco.biomol.uci.edu/mpstruc` on May 13, 2013

No.	Protein name	Species	PDB code	Resolution (Å)
1	Rhodopsin	Bos taurus	1F88	2.80
2	Rhodopsin	Bos taurus	1L9H	2.60
3	Rhodopsin	Bos taurus	1GZM	2.65
4	Rhodopsin	Bos taurus	1U19	2.20
5	Rhodopsin	Bos taurus	2J4Y	3.40
6	Rhodopsin	Bos taurus	2I35	3.80
7	Rhodopsin	Bos taurus	2I36	4.10
8	Rhodopsin	Bos taurus	2I37	4.15
9	Rhodopsin	Bos taurus	3CAP	2.90
10	Rhodopsin	Bos taurus	3PXO	3.00
11	Rhodopsin	Bos taurus	3PQR	2.85
12	Rhodopsin	Bos taurus	3DQB	3.20
13	Rhodopsin	Bos taurus	4A4M	3.30
14	Rhodopsin	Todarodes pacificus	2Z73	2.50
15	Rhodopsin	Todarodes pacificus	2ZIY	3.70
16	Rhodopsin	Todarodes pacificus	3AYN	2.70
17	Rhodopsin	Todarodes pacificus	3AYM	2.80
18	beta-1 adrenergic receptor	Meleagris gallopavo	2VT4	2.70
19	beta-1 adrenergic receptor	Meleagris gallopavo	2Y00	2.50
20	beta-1 adrenergic receptor	Meleagris gallopavo	2Y01	2.65
21	beta-1 adrenergic receptor	Meleagris gallopavo	2Y02	2.65
22	beta-1 adrenergic receptor	Meleagris gallopavo	2Y03	2.85
23	beta-1 adrenergic receptor	Meleagris gallopavo	2Y04	3.05
24	beta-1 adrenergic receptor	Meleagris gallopavo	2YCW	3.00
25	beta-1 adrenergic receptor	Meleagris gallopavo	2YCX	3.25
26	beta-1 adrenergic receptor	Meleagris gallopavo	2YCY	3.15
27	beta-1 adrenergic receptor	Meleagris gallopavo	2YCZ	3.65
28	beta-1 adrenergic receptor	Meleagris gallopavo	4AMJ	2.30
29	beta-1 adrenergic receptor	Meleagris gallopavo	4AMI	3.20
30	beta-2 adrenergic receptor	Homo sapiens	2R4R	3.4/3.7
31	beta-2 adrenergic receptor	Homo sapiens	2R4S	3.4/3.7
32	beta-2 adrenergic receptor	Homo sapiens	3KJ6	3.40
33	beta-2 adrenergic receptor	Homo sapiens	2RH1	2.40
34	beta-2 adrenergic receptor	Homo sapiens	3D4S	2.80
35	beta-2 adrenergic receptor	Homo sapiens	3P0G	3.50
36	beta-2 adrenergic receptor	Homo sapiens	3PDS	3.50
37	beta-2 adrenergic receptor	Homo sapiens	3SN6	3.20
38	A2A adenosine receptor	Homo sapiens	3EML	2.60
39	A2A adenosine receptor	Homo sapiens	3QAK	2.71
40	A2A adenosine receptor	Homo sapiens	2YDO	3.00
41	A2A adenosine receptor	Homo sapiens	2YDV	2.60
42	A2A adenosine receptor	Homo sapiens	3RFM	3.60
43	A2A adenosine receptor	Homo sapiens	3PWH	3.30
44	A2A adenosine receptor	Homo sapiens	3REY	3.30
45	A2A adenosine receptor	Homo sapiens	3VG9	2.70
46	A2A adenosine receptor	Homo sapiens	3VGA	3.10
47	CXCR4 Chemokine Receptor	Homo sapiens	3ODU	2.50
48	CXCR4 Chemokine Receptor	Homo sapiens	3OE8	3.10
49	CXCR4 Chemokine Receptor	Homo sapiens	3OE9	3.10
50	CXCR4 Chemokine Receptor	Homo sapiens	3OE6	3.20
51	CXCR4 Chemokine Receptor	Homo sapiens	3OE0	2.90
52	Dopamine D3 Receptor	Homo sapiens	3PBL	2.89
53	Histamine H1 receptor	Homo sapiens	3RZE	3.10
54	Sphingosine 1-phosphate receptor	Homo sapiens	3V2W	3.35
55	Sphingosine 1-phosphate receptor	Homo sapiens	3V2Y	2.80
56	M2 human muscarinic acetylcholine receptor	Homo sapiens	3UON	3.00
57	kappa-opioid receptor	Homo sapiens	4DJH	2.90
58	mu-opioid receptor	Mus musculus	4DKL	2.80
59	delta-opioid receptor	Mus musculus	4EJ4	3.40
60	Nociceptin/orphanin FQ (N/OFQ) receptor	Homo sapiens	4EA3	3.01

Table A.2 Available GPCR's structures in Protein Data Bank (PDB), adapted from `http://blanco.biomol.uci.edu/mpstruc` on May 13, 2013

APPENDIX B

Setting up MD simulation of membrane protein

The field of membrane protein simulation has matured during recent years, and worldwide there are now many groups performing simulations. One reason for the recent increase in the number of research groups is that the computational facilities required to perform membrane protein simulations are now accessible to more people. In this appendix, we discuss about membrane protein simulation, focusing on how to set up and run MD simulation of a membrane protein embedded within a lipid bilayer. This procedure was applied to set up the *Hs*SRII systems in Chapter 4.

Until recently, the setup required a large amount of interactive input from the researcher. Now, principally because of increases in computational power, the setup and running of such simulations is much simpler. The process of setting up simulation of membrane protein can be divided into four distinct steps:

- the preparation of the protein itself,
- the preparation of the lipid,
- the actual insertion and establishing of a stable system,
- running the simulation.

Of these steps, it is perhaps the preparation of the protein itself that requires the most care and interactive input from the researcher (Biggin & Bond, 2008). Before discussing these steps in more detail, it's worth mentioning on the available force fields applied for membrane protein simulations.

Force fields

While force-field (FF) parameters for proteins and water are well established and published with the major simulation software packages, this is not always the case for parameters for the lipids in the membrane (Domanski *et al.*, 2010). It has proven to be very difficult to obtain accurate force fields (FFs) for lipid bilayers and often specialized FFs have to be used; e.g., CHARMM uses specialized parameters for lipids that differ from the parameters used for proteins to describe the same type of atoms. One possible explanation for this is the anisotropic nature of these systems. Lipid bilayers are very peculiar systems with a polar head group in contact with the aqueous environment and a strong attraction of the tails driven by the hydrophobic effect and van der Waals (vdW) forces. A complexity like this poses a challenge for the modeling community. Especially, the fact that in order to study complexes including both lipids and proteins we need FFs that work well together implies further complications. The interactions between amino acids and lipid molecules need to be carefully evaluated before embarking on larger simulations (Jämbeck & Lyubartsev, 2012).

Recent developments have led to improved FF parameters for lipids, which now cover a wider range of lipids and better reflect experimental measurements (Stansfeld & Sansom, 2011b). Many of these parameters are now deposited in Lipidbook (Domanski *et al.*, 2010)[1]. Two major types of FFs are available for lipid bilayers: united atom (UA) and all-atomistic (AA). In UA FFs all nonpolar hydrogens have been included in the heavier atoms, creating pseudo-atoms. The popular Berger lipids (Berger *et al.*, 1997), the G53A6L FF(Poger *et al.*, 2010), the 43A1-S3 FF (Chiu *et al.*, 2009) and the potential developed by Ulmschneider and Ulmschneider(Ulmschneider & Ulmschneider, 2009) and Kukol(Kukol, 2009) are examples of these FFs. The AA FFs have explicit hydrogens included and the most common ones for lipid simulations are CHARMM (Lim *et al.*, 2012) and the general AMBER FF (GAFF) (Wang *et al.*, 2004). Another AA FF has just been developed by Jämbeck and Lyubartsev (Jämbeck & Lyubartsev, 2012). Other available FFs are

[1] see http://lipidbook.bioch.ox.ac.uk/

nonadditive (polarizable)(Harder *et al.*, 2009)(Davis *et al.*, 2009) or coarse-grained. (Marrink *et al.*, 2007; Monticelli *et al.*, 2008; Wang & Deserno, 2010)(Orsi & Essex, 2011)(De Nicola *et al.*, 2011)

Different force fields exist that are fundamentally similar, but have their own strengths and weaknesses. Because the parameters are empirical, there is no unique solution for an optimal set of parameters. Hence there is a certain degree of experience and judgement involved in picking force fields and interpreting results from simulations (Tieleman, 2010).

In this thesis, the GROMOS 43A1-S3 FF was chosen to study *Hs*SRII system due to the following reasons (i) the GROMOS FF has been applied to bacteriorhodopsin (Kandt *et al.*, 2004), a structurally and functionally closely related to *Hs*SRII, as well as various membrane proteins (Tieleman, 2010); (ii) the GROMOS FFs offer an advantage over other FFs (CHARMM, AMBER) in that they are united atom FFs which decreases the number of atoms in the system and speeding up the simulations.

Preparation of the Protein

Typically, the starting point for the protein will be a structure deposited in the protein data bank (PDB; `www.rcsb.org`). Often, however, these structures will need a certain amount of preparation before production level MD can be run. The preparations are subject to be similar to soluble proteins, but membrane proteins appear to be somewhat harder. Below are some issues that one should consider while setting up membrane proteins:

- *Missing atoms.* Many membrane protein structures are solved at low resolution. Often, part of the structures are missing, which can range from a couple of side chain atoms to entire loops. Dealing with this problem depends on the question one is trying to address with the simulation. For the case in which only a few atoms are missing from a small number of side chains, one can manually build in the missing atoms using an interactive modeling program such as PyMOL (Schrödinger, LLC, 2010) or What-If (Vriend, 1990). For the more complicated case in which whole loops are missing, typically, one has to resort to programs that can build random structures that are geometrically correct, such as Modeller (Eswar *et al.*, 2007). Indeed, in some cases, it may be that construction of an entire homology model is required.

- *Termini in the structure.* Frequently, the structure is not the whole sequence of the protein, and, therefore, charged termini may not be appropriate. One common procedure has been to build on capping groups that help to best mimic the continuing protein chain. A simpler approach involves simply protonating the C terminus and deprotonating the N terminus.

- *Add hydrogen atoms.* In all but the very high-resolution structures, one will still have to add hydrogen atoms because these will not be present in the PDB file. Although this is a very simple process, there are decisions to be made even for this process: (i) the choice of FF - all atom versus a united-atom model in which only polar hydrogens are explicit, and (ii) the protonation states of ionizable side chains. United-atom force fields will give the benefit of reduced computational effort because of reduction in the number of particles, but all-atom models might be preferred in some cases in which greater accuracy is required. It may be the case that the protonation state is not important, in which case, default ionization states at pH 7.0 are assumed. However, there are examples in which the protonation state may be critical, as exemplified by the protonation state of Glu71 in KcsA or Asp85 in bacteriorhodopsin (Tieleman, 2010). In these cases, the position of the hydrogens on histidine residues should also be considered carefully.

- *Crystallographic waters.* Although solvation of the system is generally automated, oxygen atoms from water molecules are often included in protein crystal structures. These reflect low-energy minima for a water molecule and, thus, it is usual to include these before "bulk solvation". Deciding whether a water molecule belongs to the subunit of interest from the PDB file is a problem and usually one simply chooses an arbitrary cutoff within which to include these crystallographic waters in the simulation. A recommend cutoff which has been used by different groups is 4 Å (Tieleman, 2010).

Preparation of the Lipid

In general, the term lipid is used to denote a large number of different molecules that are not soluble in water. Here, we will restrict the term to refer only to amphiphilic lipids as they occur in biological membranes. These molecules have an elongated, hydrophobic moiety (the tail of the lipid) and a more globular, hydrophilic head. The head and tail moieties can be of very different chemical nature. The lipids

most abundant in the membranes of animals, plants and bacteria are the phospholipids (Figure B.1), in particular phosphatidylcholines (lecithins, abbr. PCs), phosphatidylethanolamines (PEs), and sphingolipids.

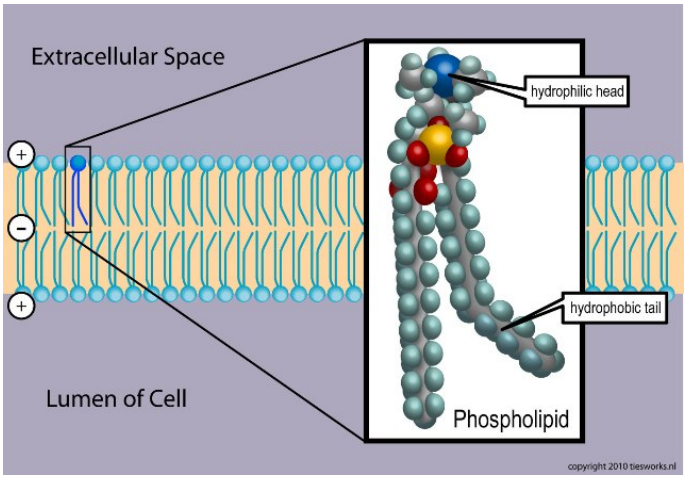


Figure B.1 Phospholipid. Adapted from the Wikimedia Commons file `http://upload.wikimedia.org/wikipedia/commons/3/39/Phospholipid_TvanBrussel.jpg`

Equilibrated conformations of many lipid bilayer systems are now deposited in Lipidbook database (Domanski *et al.*, 2010), which provide a good starting point for the main procedure. However, sometimes it will be necessary to generate a new lipid bilayer system from scratch. Bilayers consisting of a single type of lipid appear relatively straightforward and can be made in a number of ways, including a rigid-body packing procedure implemented in CHARMM (Brooks *et al.*, 1983), random placement on a grid (Kandt *et al.*, 2007) and self-assembly from a random mixture in solution (Marrink *et al.*, 2001). Water can be added by geometric criteria (avoiding overlap with lipid atoms), although this will often place water in the centre of the bilayer where there is a significant amount of free volume. Such water molecules can simply be deleted, as they may take a long simulation time to move.

Not all of the methods of insertion protein into lipid bilayer rely on a preformed lipid bilayer. However, one invariably will need a lipid-only system for control purposes, therefore, simulation of the pure system should be done at some point. One then needs a measure of how stable or good that pure system is before proceeding to insert a protein into it. The most commonly used measure of equilibration or stability is to analyze the mean area per lipid; a quantity for which there frequently exists experimental data with which to make a direct comparison. Furthermore, if this is incorrect, then it is likely that most other properties will also be inaccurate (Kandt *et al.*, 2007).

Figure B.2 shows the mean area per lipid (APL) of 200 palmitoyl oleoyl phosphatidylcholine (POPC) during 20 ns of MD simulations. The simulation was performed under temperature of 310 K using stochastic velocity rescaling (Bussi *et al.*, 2007) and pressure of 1 bar using Parrinello-Rahman semi-isotropic pressure coupling (Parrinello & Rahman, 1981). The average value of APL calculated from this simulation (0.65 ± 0.01 nm^2) is consistent with the experimental value (0.66 nm^2) measured at 310 K for POPC bilayer (Hyslop *et al.*, 1990).

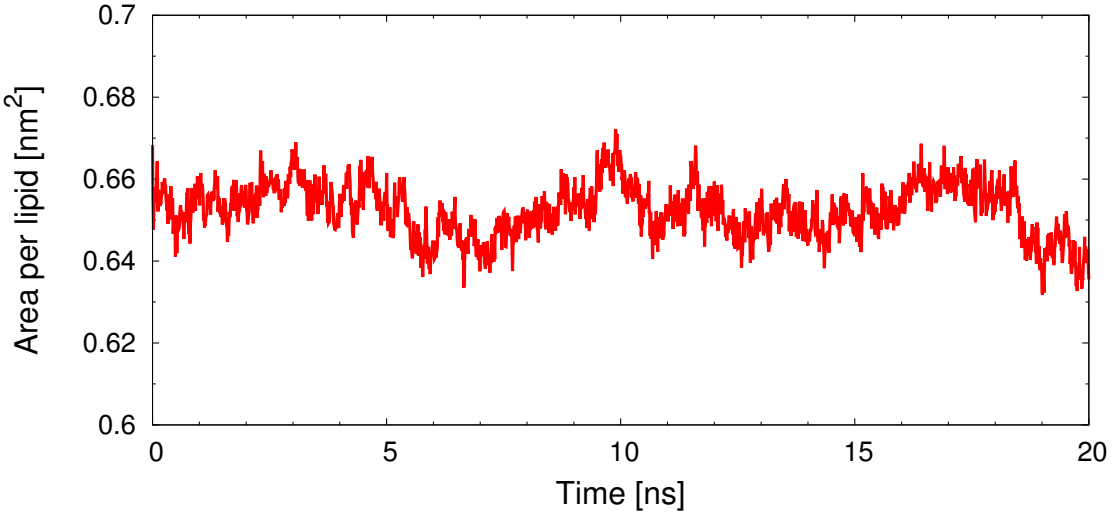


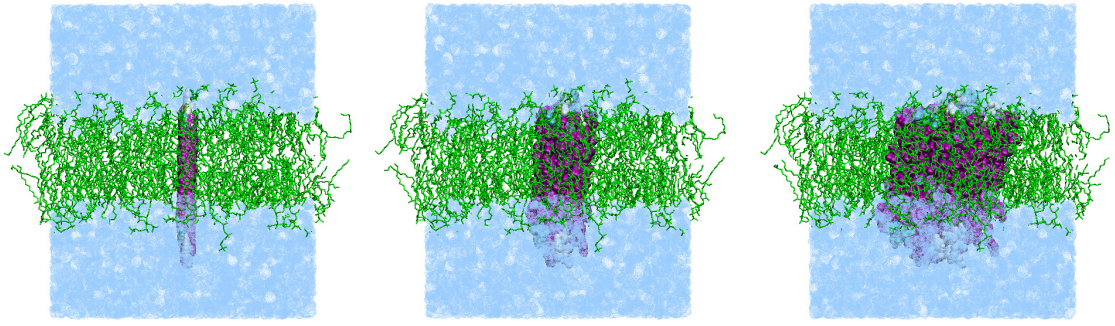
Figure B.2 The mean area per lipid (APL) of 200 palmitoyl oleoyl phosphatidylcholine (POPC) molecules calculated as a function of time.

Insertion of the Protein into the Membrane

Orienting the protein. The protein is needed to be oriented in such a way that its hydrophobic belt is aligned with the non-polar lipid tails. The term hydrophobic belt refers to a common feature of membrane proteins regarding the distribution of charged residues on the molecule's surface: the area exposed to the hydrophobic component of a bilayer usually contains no charged residues (Tieleman, 2010). Recently, a number of computational methods have emerged that are able to predict the bilayer-spanning region of a membrane protein structure (Lomize *et al.*, 2006). The orientation of a significant number of membrane proteins from the PDB are publically accessible (e.g., OPM; `http://opm.phar.umich.edu/`).

Embedding protein in bilayer. The method used in this thesis is developed by Wolf et al. (Wolf *et al.*, 2010) and has been implemented in GROMACS as the tool

g_membed. The input consists of an equilibrated membrane system, either flat or curved, and a protein structure in the right position and orientation with respect to the lipid bilayer. g_membed first decreases the width of the protein in the xy-plane and removes all molecules (generally lipids and waters) that overlap with the narrowed protein. Then the protein is grown back to its full size in a short molecular dynamics simulation (typically 1000 steps), thereby pushing the lipids away to optimally accommodate the protein in the membrane (Figure B.3).



(a) **(b)** **(c)**

Figure B.3 The embedding process. The protein *Hs*SRII grown within the membrane (POPC lipid bilayer and waters) using g_membed are shown in side view. The protein is displayed as Van der Waals spheres, lipid molecules as licorice in green color, and water as ice blue color. Three snapshots were taken at step 1 (a), step 500 (b), and step 1000 (c), respectively.

In details, g_membed performs the following steps to embed the protein (or any other desired group of atoms defined as a Gromacs index group) within a lipid bilayer.

1. Protein narrowing - The coordinates of the atoms in the protein are scaled with respect to the geometrical center of the transmembrane part of the protein by a user specified scaling factor of the original (input) coordinates in the xy-plane and, if applicable, in the z-direction. Normally the protein should not be scaled in the z-direction. However in special cases, such as a protein that has the same height as the bilayer, increasing the size of the protein in the z-direction prevents lipids to envelop the protein during the growth phase.

2. Remove overlapping molecules - Every molecule not part of the protein for which at least one atom is within a user defined radius of a protein atom will be removed. If the difference between the number of lipids removed from the lower (n_{lower}) and the upper (n_{upper}) membrane leaflet is not equal to a user defined number (n_{diff}), i.e., $n_{\text{lower}} - n_{\text{upper}} = n_{\text{diff}}$, additional lipids will be

removed, such that this equality will be obtained. This option is useful when inserting an asymetrically shaped protein.

3. Growth phase - Iterate steps 3a and 3b $nxy + nz$ times to grow the narrowed protein to its original size.

   a. md step - Do a normal md step.

   b. Protein resizing - Change the atom coordinates of the protein by linear interpolation between the coordinates of the narrowed protein (step 1) and those of the input configuration by

$$\mathbf{r}_i = \mathbf{r}_{geom} + \mathbf{s}_i \cdot (\mathbf{r}_0 - \mathbf{r}_{geom}) \tag{B.1}$$

   with $\mathbf{r}_i$ the protein atom coordinates at step i, $\mathbf{r}_0$ the input atom coordinates of the protein, $\mathbf{r}_{geom}$ the coordinates of the geometrical center of the transmembrane region, $\mathbf{s}_i$ the scaling factor at step i.

After embedding the protein in the membrane, both the lipid properties and the hydration layer are still close to equilibrium. Thus, only a short equilibration run (less then 1 ns in the cases tested) is required to re-equilibrate the membrane. Its simplicity makes g_membed very practical for use in scripting and high-throughput molecular dynamics simulations.

Simulation parameters

Among the numerous run parameters the most crucial ones for membrane protein simulations include the type of pressure coupling and the actual length of the simulation.

Pressure coupling

Roughly, there are three different methods of pressure coupling: Isotropic, semi-isotropic and anisotropic (Figure B.4). With isotropic pressure coupling the single contributions in x, y, and z direction are coupled, thus only a proportional scaling of the system is possible (Figure B.4a). Generally this leads to very small changes in box size due to the incompressibility of water. These changes are often negligible. In regard to membrane protein simulations it is important to keep in mind that

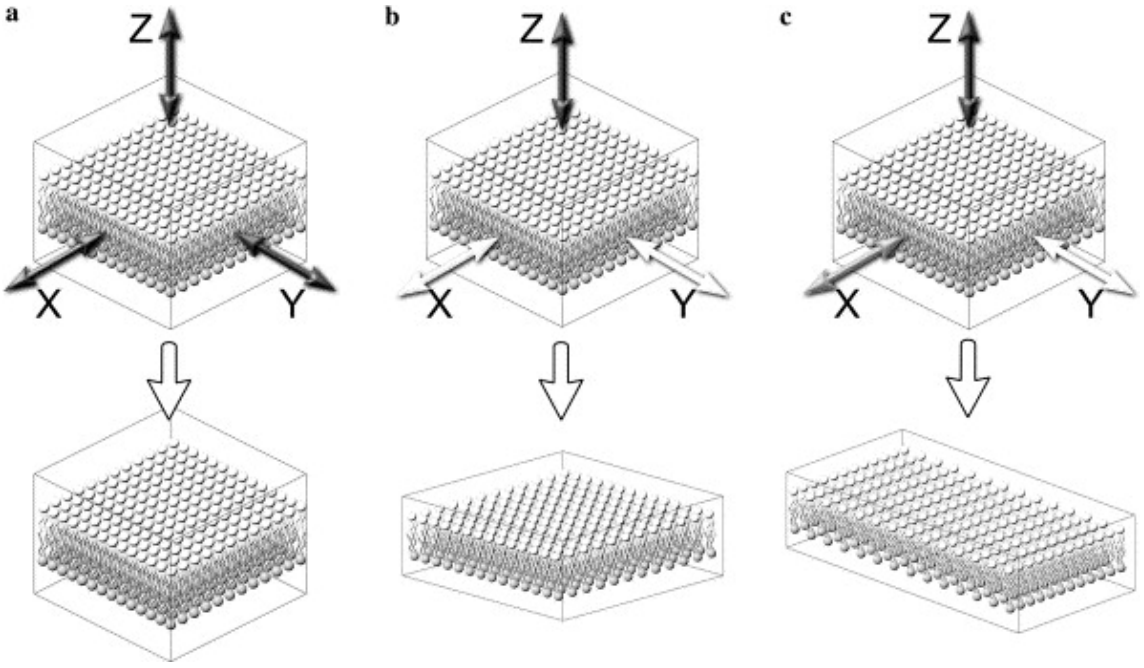


Figure B.4 There are three types of pressure coupling: (a) isotropic pressure coupling has all three pressure contributions (in x, y, and z, assuming the box has to remain rectangular) coupled, permitting only a proportional scaling of the simulation system. Only small fluctuations can occur this way, fluctuations in surface area are not possible; (b) semi-isotropic pressure coupling does allow area fluctuations as here only the pressure contributions in x and y direction are coupled together but the one in z is not; (c) anisotropic pressure coupling too enables fluctuations of the membrane area but can also lead to large deformations of the simulation system as there is no coupling between any of the directions of pressure contributions. Figure reprinted from (Kandt *et al.*, 2007), with permission from Elsevier.

isotropic pressure coupling does not allow fluctuations in the surface area, which is a key feature of lipid bilayers, and does not specify a surface tension. As such it is inappropriate for membrane simulations. One possible method to deal with this is to impose a constant surface area, calculated from the area per lipid value for the particular lipid species of interest, or a constant volume. Unfortunately knowledge of the right area per lipid is very hard to obtain and actually not possible to get for arbitrary systems. Furthermore, this approach will not allow reproducing effects such as temperature dependent swelling of the bilayer.

Semi-isotropic pressure coupling does allow area fluctuations. In this case, only the pressure contributions in x and y direction are (isotropically) coupled, but the one in z is not (Figure B.4b). It is important to stress here that having the x and y pressure contributions coupled does not imply a constant surface area: if the size of the simulation box needs to be adjusted this will affect both dimensions equally to allow area fluctuations without changing the aspect-ratio of the box. Semi-isotropic

is the recommended type of pressure coupling for membrane protein simulations. With the united-atom parameters in GROMACS a reference pressure of 1 bar in both xy and z is appropriate, and corresponds to a realistic pressure (Tieleman, 2010).

Although anisotropic coupling also enables fluctuations of the membrane area it should be used carefully: as there is no coupling between any of the directions of pressure contributions, this can result in large deformations of the whole simulation system (Figure B.4).

Equilibration length

The simulation time required to equilibrate a membrane protein simulation system depends on many factors. Two major ones are the actual system size (the bigger the system, the slower equilibration) and the starting structure used. For example, how big is the degree of disturbance caused by the insertion of the protein? How good is the quality of the protein model used? Is the system solvated sufficiently? Regarding the protein, this concerns both external and internal hydration.

It is generally useful to keep in mind that 10-20 ns are required to equilibrate lipids if the starting structure is reasonably close to the equilibrium structure (Kandt *et al.*, 2007). A time scale of at least nanoseconds is a good practical starting time for a typical membrane equilibration run, but will often be too short. Protein coordinates should be kept restrained at this stage. Progress in equilibration can be judged by monitoring the dimensions of the simulation box, lipid properties, and energies. In the case of our MD simulations with *Hs*SRII, 30 ns is long enough for the protein getting equilibrated. Unfortunately it is difficult to suggest exact criteria, but one of the most common problems in the literature on membrane protein simulations is lack of equilibration and invalid conclusions drawn from such simulations. Lastly, one should keep in mind that sampling is often by far the greatest source of errors in MD simulations of lipids and typically much more limiting than relatively modest differences caused by force field choice or choice of simulation algorithms (Kandt *et al.*, 2007).

References

Adelman, S A, & Doll, J D. 1976. Generalized Langevin equation approach for atom/solid-surface scattering: General formulation for classical scattering off harmonic solids. *The journal of chemical physics*, **64**(6), 2375.

Akabas, M H, Stauffer, D A, Xu, M, & Karlin, A. 1992. Acetylcholine receptor channel structure probed in cysteine-substitution mutants. *Science*, **258**(5080), 307–310.

Altschul, S F, Gish, W, Miller, W, Myers, E W, & Lipman, D J. 1990. Basic local alignment search tool. *J mol biol*, **215**(3), 403–410.

Altschul, S F, Madden, T L, Schäffer, A A, Zhang, J, Zhang, Z, Miller, W, & Lipman, D J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids res*, **25**(17), 3389–3402.

Anantharaman, Vivek, & Aravind, L. 2003. Application of comparative genomics in the identification and analysis of novel families of membrane-associated receptors in bacteria. *Bmc genomics*, **4**(1), 34.

Angel, Thomas E, Chance, Mark R, & Palczewski, Krzysztof. 2009. Conserved waters mediate structural and functional activation of family A (rhodopsin-like) G protein-coupled receptors. *Proc natl acad sci usa*, **106**(21), 8555–8560.

Anselmi, Claudio, Carloni, Paolo, & Torre, Vincent. 2007. Origin of functional diversity among tetrameric voltage-gated channels. *Proteins*, **66**(1), 136–146.

Ayton, Gary S., Noid, Will G., & Voth, Gregory A. 2007. Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr opin struct biol*, **17**(2), 192–198.

Bakalyar, H, & Reed, R. 1990. Identification of a specialized adenylyl cyclase that may mediate odorant detection. *Science*, **250**(4986), 1403–1406.

Baker, N A, Sept, D, Joseph, S, Holst, M J, & McCammon, J A. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc natl acad sci usa*, **98**(18), 10037–10041.

Bayly, Christopher I, Cieplak, Piotr, Cornell, Wendy, & Kollman, Peter A. 1993. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. phys. chem.*, **97**(40), 10269–10280.

Becchetti, A, & Gamel, K. 1999. The properties of cysteine mutants in the pore region of cyclic-nucleotide-gated channels. *Pflï¿½gers archiv european journal of physiology*, **438**(5), 587–596.

Bénitah, J P, Ranjan, R, Yamagishi, T, Janecki, M, Tomaselli, G F, & Marban, E. 1997. Molecular motions within the pore of voltage-dependent sodium channels. *Biophys j*, **73**(2), 603–613.

Benkert, Pascal, Tosatto, Silvio C E, & Schomburg, Dietmar. 2008. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins*, **71**(1), 261–277.

Berendsen, Herman J. C., Postma, James P. M., van Gunsteren, Wilfred F, & Hermans, Jan. 1981. Interaction models for water in relation to protein hydration. *Pages 331–342 of:* Pullman, B (ed), *Intermolecular forces*. D. Reidel Publishing Company.

Berendsen, HJC, Postma, JPM, Vangunsteren, W F, DiNola, A, & Haak, JR. 1984. Molecular-Dynamics with Coupling to an External Bath. *J chem phys*, **81**(8), 3684–3690.

Berger, O, Edholm, O, & Jahnig, F. 1997. Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidylcholine at full hydration, constant pressure, and constant temperature. *Biophys j*, **72**(5), 2002–2013.

Berman, H M, Westbrook, J, Feng, Z, Gilliland, G, Bhat, T N, Weissig, H, Shindyalov, I N, & Bourne, P E. 2000. The Protein Data Bank. *Nucleic acids res*, **28**(1), 235–242.

Berweger, Christian D, van Gunsteren, Wilfred F, & Müller-Plathe, Florian. 1995. Force field parametrization by weak coupling. Re-engineering SPC water. *Chemical physics letters*, **232**(5-6), 429–436.

Biarnés, Xevi, Marchiori, Alessandro, Giorgetti, Alejandro, Lanzara, Carmela, Gasparini, Paolo, Carloni, Paolo, Born, Stephan, Brockhoff, Anne, Behrens, Maik, & Meyerhof, Wolfgang. 2010. Insights into the binding of Phenyltiocarbamide (PTC) agonist to its target human TAS2R38 bitter receptor. *Plos one*, **5**(8), e12394.

Biegert, Andreas, Mayer, Christian, Remmert, Michael, Söding, Johannes, & Lupas, Andrei N. 2006. The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic acids res*, **34**(Web Server issue), W335–9.

Biel, Martin, & Michalakis, Stylianos. 2007. Function and dysfunction of CNG channels: insights from channelopathies and mouse models. *Mol. neurobiol.*, **35**(3), 266–277.

Biggin, Philip C, & Bond, Peter J. 2008. *Methods in Molecular Biology*. John M.WalkerMethods in Molecular Biology1064-37451940-6029, vol. 443. Totowa, NJ: Humana Press.

Bordignon, Enrica, Klare, Johann P, Holterhues, Julia, Martell, Swetlana, Krasnaberski, Aliaksei, Engelhard, Martin, & Steinhoff, Heinz-Jürgen. 2007. Analysis of light-induced conformational changes of Natronomonas pharaonis sensory rhodopsin II by time resolved electron paramagnetic resonance spectroscopy. *Photochem photobiol*, **83**(2), 263–272.

Bordoli, Lorenza, Kiefer, Florian, Arnold, Konstantin, Benkert, Pascal, Battey, James, & Schwede, Torsten. 2008. Protein structure homology modeling using SWISS-MODEL workspace. *Nat protoc*, **4**(1), 1–13.

Brooks, BR, Bruccoleri, RE, Olafson, BD, States, DJ, Swaminathan, S, & Karplus, M. 1983. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J comput chem*, **4**(2), 187–217.

Buck, Linda, & Axel, Richard. 1991. A Novel Multigene Family May Encode Odorant Receptors - a Molecular-Basis for Odor Recognition. *Cell*, **65**(1), 175–187.

Buck, Linda B. 2004. The search for odorant receptors. *Cell*, **116**(2 Suppl), S117–9– 1 p following S119.

Bussi, Giovanni, Donadio, Davide, & Parrinello, Michele. 2007. Canonical sampling through velocity rescaling. *The journal of chemical physics*, **126**(1), 014101.

Camacho, Carlos J. 2005. Modeling side-chains using molecular dynamics improve recognition of binding region in CAPRI targets. *Proteins*, **60**(2), 245–251.

Cherezov, V, Rosenbaum, D M, Hanson, M A, Rasmussen, S G F, Thian, F S, Kobilka, T S, Choi, H J, Kuhn, P, Weis, W I, Kobilka, B K, & Stevens, R C. 2007. High-Resolution Crystal Structure of an Engineered Human 2-Adrenergic G Protein Coupled Receptor. *Science*, **318**(5854), 1258–1265.

Chiu, See-Wing, Pandit, Sagar A, Scott, H L, & Jakobsson, Eric. 2009. An improved united atom force field for simulation of mixed lipid bilayers. *J phys chem b*, **113**(9), 2748–2763.

Chollet, A, & Turcatti, G. 1999. Biophysical approaches to G protein-coupled receptors: structure, function and dynamics. *J comput aided mol des*, **13**(3), 209–219.

Chothia, C, & Lesk, A M. 1986. The relation between the divergence of sequence and structure in proteins. *Embo j*, **5**(4), 823–826.

Christen, Markus, Hünenberger, Philippe H, Bakowies, Dirk, Baron, Riccardo, Bürgi, Roland, Geerke, Daan P, Heinz, Tim N, Kastenholz, Mika A, Kräutler, Vincent, Oostenbrink, Chris, Peter, Christine, Trzesniak, Daniel, & van Gunsteren, Wilfred F. 2005. The GROMOS software for biomolecular simulation: GROMOS05. *J comput chem*, **26**(16), 1719–1751.

Cooper, Dermot M F. 2003. Regulation and organization of adenylyl cyclases and cAMP. *Biochem. j.*, **375**(Pt 3), 517–529.

Cordomí, Arnau, & Perez, Juan J. 2007. Molecular dynamics simulations of rhodopsin in different one-component lipid bilayers. *J phys chem b*, **111**(25), 7052–7063.

Cornell, Wendy D, Cieplak, Piotr, Bayly, Christopher I, Gould, Ian R, Merz, Kenneth M, Ferguson, David M, Spellmeyer, David C, Fox, Thomas, Caldwell, James W, & Kollman, Peter A. 1995. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J am chem soc*, **117**(19), 5179–5197.

Dai, Gang, Zhang, Yu, Tamogami, Jun, Demura, Makoto, Kamo, Naoki, Kandori, Hideki, & Iwasa, Tatsuo. 2011. An amino acid residue (S201) in the retinal binding pocket regulates the photoreaction pathway of phoborhodopsin. *Biochemistry*, **50**(33), 7177–7183.

Darden, Tom, York, Darrin, & Pedersen, Lee. 1993. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *The journal of chemical physics*, **98**(12), 10089.

Davis, Joseph E., Raharnan, Obaidur, & Patel, Sandeep. 2009. Molecular Dynamics Simulations of a DMPC Bilayer Using Nonadditive Interaction Models. *Biophys j*, **96**(2), 385–402.

De Nicola, Antonio, Zhao, Ying, Kawakatsu, Toshihiro, Roccatano, Danilo, & Milano, Giuseppe. 2011. Hybrid Particle-Field Coarse-Grained Models for Biological Phospholipids. *J chem theory comput*, **7**(9), 2947–2962.

Deupi, Xavier, & Standfuss, Joerg. 2011. Structural insights into agonist-induced activation of G-protein-coupled receptors. *Curr opin struct biol*, **21**(4), 541–551.

Domanski, Jan, Stansfeld, Phillip J., Sansom, Mark S P, & Beckstein, Oliver. 2010. Lipidbook: A Public Repository for Force-Field Parameters Used in Membrane Simulations. *J membrane biol*, **236**(3), 255–258.

Doyle, D A, Morais Cabral, J, Pfuetzner, R A, Kuo, A, Gulbis, J M, Cohen, S L, Chait, B T, & MacKinnon, R. 1998. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science*, **280**(5360), 69–77.

Eddy, S R. 1998. Profile hidden Markov models. *Bioinformatics*, **14**(9), 755–763.

Edman, Karl, Royant, Antoine, Nollert, Peter, Maxwell, Carrie A, Pebay-Peyroula, Eva, Navarro, Javier, Neutze, Richard, & Landau, Ehud M. 2002. Early structural rearrangements in the photocycle of an integral membrane sensory receptor. *Structure*, **10**(4), 473–482.

Eglen, Richard M, & Reisine, Terry. 2009. New insights into GPCR function: implications for HTS. *Methods mol biol*, **552**, 1–13.

Eisenhaber, Frank, Lijnzaad, Philip, Argos, Patrick, Sander, Chris, & Scharf, Michael. 1995. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *J comput chem*, **16**(3), 273–284.

Eswar, Narayanan, Webb, Ben, Marti-Renom, Marc A, Madhusudhan, M S, Eramian, David, Shen, Min-Yi, Pieper, Ursula, & Sali, Andrej. 2007. Comparative protein structure modeling using MODELLER. *Curr protoc protein sci*, **Chapter 2**(Nov.), Unit 2.9.

Fadouloglou, V E, Glykos, N M, & Kokkinidis, M. 2001. Side-chain conformations in 4- -helical bundles. *Protein engineering design and selection*, **14**(5), 321–328.

Firestein, S. 2001. How the olfactory system makes sense of scents. *Nature*, **413**(6852), 211–218.

Flynn, Galen E, & Zagotta, William N. 2001. Conformational Changes in S6 Coupled to the Opening of Cyclic Nucleotide-Gated Channels. *Neuron*, **30**(3), 689–698.

Flynn, Galen E, & Zagotta, William N. 2003. A cysteine scan of the inner vestibule of cyclic nucleotide-gated channels reveals architecture and rearrangement of the pore. *J. gen. physiol.*, **121**(6), 563–582.

Frenkel, Daan, & Smit, Berend. 2001. *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications (Computational Science)*. 2 edn. Academic Press.

Frings, S, Hackos, D H, Dzeja, C, Ohyama, T, Hagen, V, Kaupp, U B, & Korenbrot, J I. 2000. Determination of fractional calcium ion current in cyclic nucleotide-gated channels. *Meth. enzymol.*, **315**, 797–817.

Frishman, D, & Valencia, Alfonso (eds). 2011. *Modern Genome Annotation: The Biosapiens Network*. Softcover reprint of hardcover 1st ed. 2008 edn. Springer.

Gautier, Antoine, Mott, Helen R, Bostock, Mark J, Kirkpatrick, John P, & Nietlispach, Daniel. 2010. Structure determination of the seven-helix transmembrane receptor sensory rhodopsin II by solution NMR spectroscopy. *Nat struct mol biol*, **17**(6), 768–774.

Gether, U. 2000. Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. *Endocr. rev.*, **21**(1), 90–113.

Giorgetti, A, Carloni, P, Mistrik, P, & Torre, V. 2005a. A homology model of the pore region of HCN channels. *Biophys j*, **89**(2), 932–944.

Giorgetti, Alejandro, Nair, Anil V, Codega, Paolo, Torre, Vincent, & Carloni, Paolo. 2005b. Structural basis of gating of CNG channels. *Febs lett*, **579**(9), 1968–1972.

Go, N, & Abe, H. 1981. Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers*, **20**(5), 991–1011.

Gordeliy, Valentin I, Labahn, Jörg, Moukhametzianov, Rouslan, Efremov, Rouslan, Granzin, Joachim, Schlesinger, Ramona, Büldt, Georg, Savopol, Tudor, Scheidig, Axel J, Klare, Johann P, & Engelhard, Martin. 2002. Molecular basis of transmembrane signalling by sensory rhodopsin II-transducer complex. *Nature*, **419**(6906), 484–487.

Gottfried, Jay A, & Wilson, Donald A. 2011. *Smell.* Frontiers in Neuroscience. Boca Raton (FL): CRC Press.

Gribskov, M, McLachlan, A D, & Eisenberg, D. 1987. Profile analysis: detection of distantly related proteins. *Proc natl acad sci usa*, **84**(13), 4355–4358.

Grote, Mathias, & O'Malley, Maureen A. 2011. Enlightening the life sciences: the history of halobacterial and microbial rhodopsin research. *Fems microbiol rev*, May.

Gushchin, Ivan, Reshetnyak, Anastasia, Borshchevskiy, Valentin, Ishchenko, Andrii, Round, Ekaterina, Grudinin, Sergei, Engelhard, Martin, Büldt, Georg, & Gordeliy, Valentin. 2011. Active state of sensory rhodopsin II: structural determinants for signal transfer and proton pumping. *J mol biol*, **412**(4), 591–600.

Harder, Edward, Mackerell, Alexander D, & Roux, Benoit. 2009. Many-body polarization effects and the membrane dipole potential. *J am chem soc*, **131**(8), 2760–2761.

Henderson, R, & Unwin, P N. 1975. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature*, **257**(5521), 28–32.

Henikoff, S, & Henikoff, J G. 1994. Position-based sequence weights. *J mol biol*, **243**(4), 574–578.

Hess, B, Bekker, H, Berendsen, HJC, & Fraaije, JGEM. 1997. LINCS: A linear constraint solver for molecular simulations. *J comput chem*, **18**(12), 1463–1472.

Hippler-Mreyen, Silke, Klare, Johann P, Wegener, Ansgar A, Seidel, Ralf, Herrmann, Christian, Schmies, Georg, Nagel, Georg, Bamberg, Ernst, & Engelhard, M. 2003. Probing the sensory rhodopsin II binding domain of its cognate transducer by calorimetry and electrophysiology. *J mol biol*, **330**(5), 1203–1213.

Hirai, Teruhisa, Subramaniam, Sriram, & Lanyi, Janos K. 2009. Structural snapshots of conformational changes in a seven-helix membrane protein: lessons from bacteriorhodopsin. *Curr opin struct biol*, **19**(4), 433–439.

Hirayama, Junichi, Imamoto, Yasushi, Shichida, Yoshinori, Kamo, Naoki, Tomioka, Hiroaki, & Yoshizawa, Toru. 1992. Photocycle of phoborhodopsin from haloalkaliphilic bacterium (Natronobacterium pharaonis) studied by low-temperature spectrophotometry. *Biochemistry*, **31**(7), 2093–2098.

Hoff, WD, Jung, KH, & Spudich, JL. 1997. Molecular mechanism of photosignaling by archaeal sensory rhodopsins. *Annu rev bioph biom*, **26**, 223–258.

Hoffmann, C, Leitz, M R, Oberdorf-Maass, S, Lohse, M J, & Klotz, K-N. 2004. Comparative pharmacology of human beta-adrenergic receptor subtypes–characterization of stably transfected receptors in CHO cells. *Naunyn schmiedebergs arch. pharmacol.*, **369**(2), 151–159.

Hooft, R W, Vriend, G, Sander, C, & Abola, E E. 1996. Errors in protein structures. *Nature*, **381**(6580), 272.

Hua, Li, & Gordon, Sharona E. 2005. Functional interactions between A' helices in the C-linker of open CNG channels. *J. gen. physiol.*, **125**(3), 335–344.

Humphrey, W, Dalke, A, & Schulten, K. 1996. VMD: visual molecular dynamics. *J mol graph*, **14**(1), 33–8, 27–8.

Hyslop, PA, Morel, B, & Sauerheber, RD. 1990. Organization and Interaction of Cholesterol and Phosphatidylcholine in Model Bilayer-Membranes. *Biochemistry*, **29**(4), 1025–1038.

Iaccarino, G, & Koch, W J. 1999. Therapeutic potential of G-protein coupled receptor kinases in the heart. *Expert opin investig drugs*, **8**(5), 545–554.

Imai, Takashi, & Fujita, Norihisa. 2004. Statistical sequence analyses of G-protein-coupled receptors: structural and functional characteristics viewed with periodicities of entropy, hydrophobicity, and volume. *Proteins*, **56**(4), 650–660.

Ito, Motohiro, Sudo, Yuki, Furutani, Yuji, Okitsu, Takashi, Wada, Akimori, Homma, Michio, Spudich, John L, & Kandori, Hideki. 2008. Steric constraint in the primary photoproduct of sensory rhodopsin II is a prerequisite for light-signal transfer to HtrII. *Biochemistry*, **47**(23), 6208–6215.

J D Thompson, D G Higgins T J Gibson. 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**(22), 4673.

Jaakola, Veli-Pekka, Griffith, Mark T, Hanson, Michael A, Cherezov, Vadim, Chien, Ellen Y T, Lane, J Robert, Ijzerman, Adriaan P, & Stevens, Raymond C. 2008. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science*, **322**(5905), 1211–1217.

Jämbeck, Joakim P M, & Lyubartsev, Alexander P. 2012. Derivation and systematic validation of a refined all-atom force field for phosphatidylcholine lipids. *J phys chem b*, **116**(10), 3164–3179.

Janin, Joël, Wodak, Shoshanna, Levitt, Michael, & Maigret, Bernard. 1978. Conformation of amino acid side-chains in proteins. *J mol biol*, **125**(3), 357–386.

Jiang, Xiue, Engelhard, Martin, Ataka, Kenichi, & Heberle, Joachim. 2010. Molecular impact of the membrane potential on the regulatory mechanism of proton transfer in sensory rhodopsin II. *J am chem soc*, **132**(31), 10808–10815.

Jiang, Youxing, Lee, Alice, Chen, Jiayun, Cadene, Martine, Chait, Brian T, & MacKinnon, Roderick. 2002. The open pore conformation of potassium channels. *Nature*, **417**(6888), 523–526.

Johnson, J P, & Zagotta, W N. 2001. Rotational movement during cyclic nucleotide-gated channel opening. *Nature*, **412**(6850), 917–921.

Kalli, Antreas C, Campbell, Iain D, & Sansom, Mark S P. 2011. Multiscale simulations suggest a mechanism for integrin inside-out activation. *Proc natl acad sci usa*, **108**(29), 11890–11895.

Kamerlin, Shina C L, Vicatos, Spyridon, Dryga, Anatoly, & Warshel, Arieh. 2011. Coarse-grained (multiscale) simulations in studies of biophysical and chemical systems. *Annu rev phys chem*, **62**(May), 41–64.

Kandori, Hideki, Sudo, Yuki, & Furutani, Yuji. 2010. Protein-protein interaction changes in an archaeal light-signal transduction. *J biomed biotechnol*, **2010**(Jan.), 424760.

Kandt, Christian, Schlitter, Jürgen, & Gerwert, Klaus. 2004. Dynamics of Water Molecules in the Bacteriorhodopsin Trimer in Explicit Lipid/Water Environment. *Biophys j*, **86**(2), 705–717.

Kandt, Christian, Ash, Walter L, & Peter Tieleman, D. 2007. Setting up and running molecular dynamics simulations of membrane proteins. *Methods*, **41**(4), 475–488.

Karplus, K, Barrett, C, & Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**(10), 846–856.

Kato, Aya, & Touhara, Kazushige. 2009. Mammalian olfactory receptors: pharmacology, G protein coupling and desensitization. *Cell mol life sci*, **66**(23), 3743–3753.

Katritch, Vsevolod, Cherezov, Vadim, & Stevens, Raymond C. 2012. Diversity and modularity of G protein-coupled receptor structures. *Trends pharmacol sci*, **33**(1), 17–27.

Kaupp, U Benjamin, & Seifert, Reinhard. 2002. Cyclic nucleotide-gated ion channels. *Physiological reviews*, **82**(3), 769–824.

Ketkar, Amit D, Shenoy, Avinash R, Ramagopal, Udupi A, Visweswariah, Sandhya S, & Suguna, Kaza. 2006. A structural basis for the role of nucleotide specifying residues in regulating the oligomerization of the Rv1625c adenylyl cyclase from M. tuberculosis. *J mol biol*, **356**(4), 904–916.

Khafizov, Kamil, Anselmi, Claudio, Menini, Anna, & Carloni, Paolo. 2007. Ligand specificity of odorant receptors. *J mol model*, **13**(3), 401–409.

Khafizov, Kamil, Lattanzi, Gianluca, & Carloni, Paolo. 2009. G protein inactive and active forms investigated by simulation methods. *Proteins*, **75**(4), 919–930.

Killian, J A, & von Heijne, G. 2000. How proteins adapt to a membrane-water interface. *Trends biochem. sci.*, **25**(9), 429–434.

Kitchen, DB, Decornez, H, Furr, JR, & Bajorath, J. 2004. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat rev drug discov*, **3**(11), 935–949.

Klare, Johann P, Gordeliy, Valentin I, Labahn, Jörg, Büldt, Georg, Steinhoff, Heinz-Jürgen, & Engelhard, Martin. 2004. The archaeal sensory rhodopsin II/transducer complex: a model for transmembrane signal transfer. *Febs lett*, **564**(3), 219–224.

Klare, Johann P, Chizhov, Igor, & Engelhard, Martin. 2008. Microbial rhodopsins: scaffolds for ion pumps, channels, and sensors. *Results probl cell differ*, **45**, 73–122.

Kleene, S J. 1993. The cyclic nucleotide-activated conductance in olfactory cilia: effects of cytoplasmic Mg2+ and Ca2+. *J. membr. biol.*, **131**(3), 237–243.

Kolbe, M, Besir, H, Essen, L O, & Oesterhelt, D. 2000. Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution. *Science*, **288**(5470), 1390–1396.

Kominos, D, Bassolino, D A, Levy, R M, & Pardi, A. 1990. Analysis of side-chain conformational distributions in neutrophil peptide-5 NMR structures. *Biopolymers*, **29**(14), 1807–1822.

Kranjc, Agata, Grillo, Federico W, Rievaj, Juraj, Boccaccio, Anna, Pietrucci, Fabio, Menini, Anna, Carloni, Paolo, & Anselmi, Claudio. 2009. Regulation of Bestrophins by Ca2+: A Theoretical and Experimental Study. *Plos one*, **4**(3), e4672.

Krogh, A, Brown, M, Mian, I S, Sjölander, K, & Haussler, D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J mol biol*, **235**(5), 1501–1531.

Kufareva, Irina, Rueda, Manuel, Katritch, Vsevolod, Stevens, Raymond C, Abagyan, Ruben, & GPCR Dock 2010 participants. 2011. Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment. *Structure*, **19**(8), 1108–1126.

Kukol, Andreas. 2009. Lipid Models for United-Atom Molecular Dynamics Simulations of Proteins. *J chem theory comput*, **5**(3), 615–626.

Kürz, L L, Zühlke, R D, Zhang, H J, & Joho, R H. 1995. Side-chain accessibilities in the pore of a K+ channel probed by sulfhydryl-specific reagents after cysteine-scanning mutagenesis. *Biophys j*, **68**(3), 900–905.

Kutzner, Carsten, van der Spoel, David, Fechner, Martin, Lindahl, Erik, Schmitt, Udo W, de Groot, Bert L, & Grubmüller, Helmut. 2007. Speeding up parallel GROMACS on high-latency networks. *J comput chem*, **28**(12), 2075–2084.

Landry, Yves, & Gies, Jean-Pierre. 2008. Drugs and their molecular targets: an updated overview. *Fundam clin pharmacol*, **22**(1), 1–18.

Langelaan, David N, Ngweniform, Pascaline, & Rainey, Jan K. 2011. Biophysical characterization of G-protein coupled receptor-peptide ligand binding. *Biochem. cell biol.*, **89**(2), 98–105.

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 crystallogr*, **26**(2), 283–291.

Leach, Andrew. 2001. *Molecular Modelling: Principles and Applications (2nd Edition)*. 2 edn. Prentice Hall.

Lefkowitz, R J. 2007. Seven transmembrane receptors: something old, something new. *Acta physiol (oxf)*, **190**(1), 9–19.

Lim, Joseph B., Rogaski, Brent, & Klauda, Jeffery B. 2012. Update of the Cholesterol Force Field Parameters in CHARMM. *J phys chem b*, **116**(1), 203–210.

Lindahl, Erik R. 2008. Molecular dynamics simulations. *Methods mol biol*, **443**, 3–23.

Lipman, D J, & Pearson, W R. 1985. Rapid and sensitive protein similarity searches. *Science*, **227**(4693), 1435–1441.

Lodish, Harvey, Berk, Arnold, Kaiser, Chris A, Krieger, Monty, Scott, Matthew P, Bretscher, Anthony, Ploegh, Hidde, & Matsudaira, Paul. 2007. *Molecular Cell Biology (Lodish, Molecular Cell Biology)*. 6th edn. W. H. Freeman.

Lomize, M A, Lomize, A L, Pogozheva, I D, & Mosberg, H I. 2006. OPM: Orientations of Proteins in Membranes database. *Bioinformatics*, **22**(5), 623–625.

Loo, T W, & Clarke, D M. 2001. Determining the dimensions of the drug-binding domain of human P-glycoprotein using thiol cross-linking compounds as molecular rulers. *J biol chem*, **276**(40), 36877–36880.

Lowe, G, & Gold, G H. 1993. Nonlinear amplification by calcium-dependent chloride channels in olfactory receptor cells. *Nature*, **366**(6452), 283–286.

Lu, Guoqing, Wang, Zhifang, Jones, Alan M, & Moriyama, Etsuko N. 2009. 7TMRmine: a Web server for hierarchical mining of 7TMR proteins. *Bmc genomics*, **10**(Jan.), 275.

Luecke, H, Schobert, B, Lanyi, J K, Spudich, E N, & Spudich, J L. 2001. Crystal structure of sensory rhodopsin II at 2.4 angstroms: insights into color tuning and transducer interaction. *Science*, **293**(5534), 1499–1503.

Luecke, Hartmut, Schobert, Brigitte, Stagno, Jason, Imasheva, Eleonora S, Wang, Jennifer M, Balashov, Sergei P, & Lanyi, Janos K. 2008. Crystallographic structure of xanthorhodopsin, the light-driven proton pump with a dual chromophore. *Proc natl acad sci usa*, **105**(43), 16561–16565.

MacCallum, Justin L, Pérez, Alberto, Schnieders, Michael J, Hua, Lan, Jacobson, Matthew P, & Dill, Ken A. 2011. Assessment of protein structure refinement in CASP9. *Proteins*, **79**(S10), 74–90.

MacKerell, A D, Bashford, D, Bellott, 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, Wiórkiewicz-Kuczera, J, Yin, D, & Karplus, M. 1998. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J phys chem b*, **102**(18), 3586–3616.

Mariani, Valerio, Kiefer, Florian, Schmidt, Tobias, Haas, Juergen, & Schwede, Torsten. 2011. Assessment of template based protein structure predictions in CASP9. *Proteins*, **79 Suppl 10**, 37–58.

Mark, P, & Nilsson, L. 2001. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *J phys chem a*, **105**(43), 9954–9960.

Marrink, S J, Lindahl, E, Edholm, O, & Mark, A E. 2001. Simulation of the spontaneous aggregation of phospholipids into bilayers. *J am chem soc*, **123**(35), 8638–8639.

Marrink, Siewert J, Risselada, H Jelger, Yefimov, Serge, Tieleman, D Peter, & de Vries, Alex H. 2007. The MARTINI force field: coarse grained model for biomolecular simulations. *J phys chem b*, **111**(27), 7812–7824.

Marti-Renom, Marc A, Madhusudhan, M S, & Sali, Andrej. 2004. Alignment of protein sequences by their profiles. *Protein sci*, **13**(4), 1071–1087.

Mazzolini, Monica, Punta, Marco, & Torre, Vincent. 2002. Movement of the C-helix during the gating of cyclic nucleotide-gated channels. *Biophys j*, **83**(6), 3283–3295.

Mazzolini, Monica, Nair, Anil V, & Torre, Vincent. 2008. A comparison of electrophysiological properties of the CNGA1, CNGA1tandem and CNGA1cys-free channels. *Eur biophys j*, **37**(6), 947–959.

Melo, F, & Feytmans, E. 1998. Assessing protein structures with a non-local atomic interaction energy. *J mol biol*, **277**(5), 1141–1152.

Menini, A. 1999. Calcium signalling and regulation in olfactory neurons. *Curr. opin. neurobiol.*, **9**(4), 419–426.

Messer, Benjamin M, Roca, Maite, Chu, Zhen T, Vicatos, Spyridon, Kilshtain, Alexandra Vardi, & Warshel, Arieh. 2010. Multiscale simulations of protein landscapes: using coarse-grained models as reference potentials to full explicit models. *Proteins*, **78**(5), 1212–1227.

Michino, Mayako, Abola, Enrique, GPCR Dock 2008 participants, Brooks, Charles L, Dixon, J Scott, Moult, John, & Stevens, Raymond C. 2009. Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nat rev drug discov*, **8**(6), 455–463.

Millar, Robert P, & Newton, Claire L. 2010. The year in G protein-coupled receptor research. *Mol endocrinol*, **24**(1), 261–274.

Moitessier, N, Englebienne, P, Lee, D, Lawandi, J, & Corbeil, C R. 2008. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br j pharmacol*, **153 Suppl 1**(Mar.), S7–26.

Mombaerts, Peter. 2004. Genes and ligands for odorant, vomeronasal and taste receptors. *Nat rev neurosci*, **5**(4), 263–278.

Monticelli, Luca, Kandasamy, Senthil K., Periole, Xavier, Larson, Ronald G., Tieleman, D Peter, & Marrink, Siewert-Jan. 2008. The MARTINI coarse-grained force field: Extension to proteins. *J chem theory comput*, **4**(5), 819–834.

Moukhametzianov, Rouslan. 2006 (Apr.). *X-ray Crystallographic Study on the Mechanisms of Bacteriorhodopsin and the Sensory Rhodopsin/Transducer Complex*. Ph.D. thesis, docserv.uni-duesseldorf.de.

Moukhametzianov, Rouslan, Klare, Johann P, Efremov, Rouslan, Baeken, Christian, Göppner, Annika, Labahn, Jörg, Engelhard, Martin, Büldt, Georg, & Gordeliy, Valentin I. 2006. Development of the signal in sensory rhodopsin and its transfer to the cognate transducer. *Nature*, **440**(7080), 115–119.

Moult, John, Fidelis, Krzysztof, Kryshtafovych, Andriy, & Tramontano, Anna. 2011. Critical assessment of methods of protein structure prediction (CASP)-round IX. *Proteins*, **79**(S10), 1–5.

Murakami, Midori, & Kouyama, Tsutomu. 2008. Crystal structure of squid rhodopsin. *Nature*, **453**(7193), 363–367.

Nadler, W, Brünger, A T, Schulten, K, & Karplus, M. 1987. Molecular and stochastic dynamics of proteins. *Proc natl acad sci usa*, **84**(22), 7933–7937.

Nair, Anil V, Mazzolini, Monica, Codega, Paolo, Giorgetti, Alejandro, & Torre, Vincent. 2006. Locking CNGA1 channels in the open and closed state. *Biophys j*, **90**(10), 3599–3607.

Nair, Anil V, Nguyen, Chuong H H, & Mazzolini, Monica. 2009a. Conformational rearrangements in the S6 domain and C-linker during gating in CNGA1 channels. *Eur biophys j*, **38**(7), 993–1002.

Nair, Anil V, Anselmi, Claudio, & Mazzolini, Monica. 2009b. Movements of native C505 during channel gating in CNGA1 channels. *Eur biophys j*, **38**(4), 465–478.

Neri, Marilisa, Anselmi, Claudio, Cascella, Michele, Maritan, Amos, & Carloni, Paolo. 2005. Coarse-Grained Model of Proteins Incorporating Atomistic Detail of the Active Site. *Phys rev lett*, **95**(21), 218102.

Neri, Marilisa, Baaden, Marc, Carnevale, Vincenzo, Anselmi, Claudio, Maritan, Amos, & Carloni, Paolo. 2008. Microseconds dynamics simulations of the outer-membrane protease T. *Biophys j*, **94**(1), 71–78.

Niimura, Y, & Nei, M. 2003. Evolution of olfactory receptor genes in the human genome. *Proc natl acad sci usa*, **100**(21), 12235–12240.

Nygaard, Rie, Valentin-Hansen, Louise, Mokrosinski, Jacek, Frimurer, Thomas M, & Schwartz, Thue W. 2010. Conserved Water-mediated Hydrogen Bond Network between TM-I, -II, -VI, and -VII in 7TM Receptor Activation. *Journal of biological chemistry*, **285**(25), 19625–19636.

Oberbarnscheidt, Leoni, Janissen, Richard, Martell, Swetlana, Engelhard, Martin, & Oesterhelt, Filipp. 2009. Single-molecule force spectroscopy measures structural changes induced by light activation and transducer binding in sensory rhodopsin II. *J mol biol*, **394**(3), 383–390.

Okada, Tetsuj, Sugihara, Minor, Bondar, Ana-Nicolet, Elstner, Marcu, Entel, Pete, & Buss, Volke. 2004. The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure. *Journal of molecular biology*, **342**(2), 0.

Orsi, Mario, & Essex, Jonathan W. 2011. The ELBA Force Field for Coarse-Grain Modeling of Lipid Membranes. *Plos one*, **6**(12), –.

Osher, S, & Sethian, JA. 1988. Fronts Propagating with Curvature-Dependent Speed - Algorithms Based on Hamilton-Jacobi Formulations. *J comput phys*, **79**(1), 12–49.

Overington, John P, Al-Lazikani, Bissan, & Hopkins, Andrew L. 2006. How many drug targets are there? *Nat rev drug discov*, **5**(12), 993–996.

Paesani, Francesco, Zhang, Wei, Case, David A, Cheatham, Thomas E, & Voth, Gregory A. 2006. An accurate and simple quantum model for liquid water. *The journal of chemical physics*, **125**(18), 184507.

Palczewski, K, Kumasaka, T, Hori, T, Behnke, C A, Motoshima, H, Fox, B A, Le Trong, I, Teller, D C, Okada, T, Stenkamp, R E, Yamamoto, M, & Miyano, M. 2000. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*, **289**(5480), 739–745.

Parrinello, M, & Rahman. 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of applied physics*, **52**(12), 7182–7190.

Persike, N, Pfeiffer, M, Guckenberger, R, Radmacher, M, & Fritz, M. 2001. Direct observation of different surface structures on high-resolution images of native halorhodopsin. *J mol biol*, **310**(4), 773–780.

Pierce, Kristen L, Premont, Richard T, & Lefkowitz, Robert J. 2002. Seven-transmembrane receptors. *Nat. rev. mol. cell biol.*, **3**(9), 639–650.

Pifferi, Simone, Boccaccio, Anna, & Menini, Anna. 2006. Cyclic nucleotide-gated ion channels in sensory transduction. *Febs lett*, **580**(12), 2853–2859.

Pinto, Jayant M. 2011. Olfaction. *Proc am thorac soc*, **8**(1), 46–52.

Poger, David, van Gunsteren, Wilfred F, & Mark, Alan E. 2010. A new force field for simulating phosphatidylcholine bilayers. *J comput chem*, **31**(6), 1117–1125.

Quigley, D, & Probert, M I J. 2004. Langevin dynamics in constant pressure extended systems. *The journal of chemical physics*, **120**(24), 11432–11441.

Rangarajan, Rekha, Galan, Jhenny F, Whited, Gregg, & Birge, Robert R. 2007. Mechanism of spectral tuning in green-absorbing proteorhodopsin. *Biochemistry*, **46**(44), 12679–12686.

Raval, Alpan, Piana, Stefano, Eastwood, Michael P, Dror, Ron O, & Shaw, David E. 2012. Refinement of protein structure homology models via long, all-atommolecular dynamicssimulations. *Proteins*, Apr.

Remmert, Michael, Biegert, Andreas, Hauser, Andreas, & Söding, Johannes. 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat meth*, **9**(2), 173–175.

Rompler, H, Staubert, C, Thor, D, Schulz, A, Hofreiter, M, & Schoneberg, T. 2011. Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat commun*, **2**(1), 17–25.

Römpler, Holger, Stäubert, Claudia, Thor, Doreen, Schulz, Angela, Hofreiter, Michael, & Schöneberg, Torsten. 2007. G protein-coupled time travel: evolutionary aspects of GPCR research. *Mol interv*, **7**(1), 17–25.

Ronnett, Gabriele V, & Moon, Cheil. 2002. G proteins and olfactory signal transduction. *Annu rev physiol*, **64**, 189–222.

Rosenbaum, Daniel M, Rasmussen, Søren G F, & Kobilka, Brian K. 2009. The structure and function of G-protein-coupled receptors. *Nature*, **459**(7245), 356–363.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein engineering design and selection*, **12**(2), 85–94.

Rothberg, Brad S, Shin, Ki Soon, & Yellen, Gary. 2003. Movements near the gate of a hyperpolarization-activated cation channel. *J. gen. physiol.*, **122**(5), 501–510.

Royant, A, Nollert, P, Edman, K, Neutze, R, Landau, E M, Pebay-Peyroula, E, & Navarro, J. 2001. X-ray structure of sensory rhodopsin II at 2.1-A resolution. *Proc natl acad sci usa*, **98**(18), 10131–10136.

Rueda, Manuel, Ferrer-Costa, Carles, Meyer, Tim, Pérez, Alberto, Camps, Jordi, Hospital, Adam, Gelpí, Josep Lluis, & Orozco, Modesto. 2007. A consensus view of protein dynamics. *Proc natl acad sci usa*, **104**(3), 796–801.

Ryckaert, Jean-Paul, Ciccotti, Giovanni, & Berendsen, Herman J. C. 1977. Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J comput phys*, **23**(3), 327–341.

Sali, Andrej, & Blundell, Tom L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J mol biol*, **234**(3), 779–815.

Sander, C, & Schneider, R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**(1), 56–68.

Sasaki, Jun, & Spudich, John L. 2008. Signal transfer in haloarchaeal sensory rhodopsin- transducer complexes. *Photochem photobiol*, **84**(4), 863–868.

Sasaki, Jun, Nara, Toshifumi, Spudich, Elena N, & Spudich, John L. 2007. Constitutive activity in chimeras and deletions localize sensory rhodopsin II/HtrII signal relay to the membrane-inserted domain. *Mol. microbiol.*, **66**(6), 1321–1330.

Sato, Y, Hata, M, Neya, S, & Hoshino, T. 2005. Computational analysis of the transient movement of helices in sensory rhodopsin II. *Protein sci*, **14**(1), 183–192.

Schobert, Brigitte, Cupp-Vickery, Jill, Hornak, Viktor, Smith, Steven, & Lanyi, Janos. 2002. Crystallographic structure of the K intermediate of bacteriorhodopsin: conservation of free energy after photoisomerization of the retinal. *J mol biol*, **321**(4), 715–726.

Schrödinger, LLC. 2010 (Aug.). *The PyMOL Molecular Graphics System, Version 1.3r1.*

Schwede, Torsten. 2008. *Computational Structural Biology: Methods and Applications.* 1 edn. World Scientific Publishing Company.

Sgourakis, Nikolaos G, & Garcia, Angel E. 2010. The membrane complex between transducin and dark-state rhodopsin exhibits large-amplitude interface dynamics on the sub-microsecond timescale: insights from all-atom MD simulations. *J mol biol*, **398**(1), 161–173.

Sharma, Adrian K, Spudich, John L, & Doolittle, W Ford. 2006. Microbial rhodopsins: functional versatility and genetic mobility. *Trends microbiol.*, **14**(11), 463–469.

Shen, Min-Yi, & Sali, Andrej. 2006. Statistical potential for assessment and prediction of protein structures. *Protein sci*, **15**(11), 2507–2524.

Shenoy, S K. 2007. Seven-Transmembrane Receptors and Ubiquitination. *Circulation research*, **100**(8), 1142–1154.

Shi, Qiang, Izvekov, Sergei, & Voth, Gregory A. 2006. Mixed atomistic and coarse-grained molecular dynamics: Simulation of a membrane-bound ion channel. *J phys chem b*, **110**(31), 15045–15048.

Soding, J, Biegert, A, & Lupas, A N. 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids res*, **33**(Web Server issue), W244–W248.

Söding, Johannes. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**(7), 951–960.

Soppa, J. 1994. Two hypotheses–one answer. Sequence comparison does not support an evolutionary link between halobacterial retinal proteins including bacteriorhodopsin and eukaryotic G-protein-coupled receptors. *Febs lett*, **342**(1), 7–11.

Spassov, V Z, Luecke, H, Gerwert, K, & Bashford, D. 2001. pK(a) Calculations suggest storage of an excess proton in a hydrogen-bonded water network in bacteriorhodopsin. *J mol biol*, **312**(1), 203–219.

Spehr, Marc, & Munger, Steven D. 2009. Olfactory receptors: G protein-coupled receptors and beyond. *J neurochem*, **109**(6), 1570–1583.

Spoel, David van der, Lindahl, Erik, Hess, Berk, Kutzner, Carsten, Buuren, Aldert R van, Apol, Emile, Meulenhoff, Pieter J, Tieleman, D Peter, Sijbers, Alfons LTM, Feenstra, K Anton, Drunen, Rudi van, & Berendsen, Herman JC. 2005. GROMACS USER MANUAL Version 4.0. *www.gromacs.org*.

Spudich, E N, Zhang, W, Alam, M, & Spudich, J L. 1997. Constitutive signaling by the phototaxis receptor sensory rhodopsin II from disruption of its protonated Schiff base-Asp-73 interhelical salt bridge. *Proc natl acad sci usa*, **94**(10), 4960–4965.

Spudich, J L, Yang, C S, Jung, K H, & Spudich, E N. 2000. Retinylidene proteins: structures and functions from archaea to humans. *Annu rev cell dev biol*, **16**(Jan.), 365–392.

Spudich, John L. 2006. The multitalented microbial sensory rhodopsins. *Trends microbiol.*, **14**(11), 480–487.

Spudich, John L, & Jung, Kwang-Hwan. 2005. *Handbook of Photosensory Receptors*. BRIGGS:PHOTORECEPTORS O-BK. Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA.

Spudich, John L, & Luecke, Hartmut. 2002. Sensory rhodopsin II: functional insights from structure. *Curr opin struct biol*, **12**(4), 540–546.

Stansfeld, Phillip J., & Sansom, Mark S P. 2011a. From Coarse Grained to Atomistic: A Serial Multiscale Approach to Membrane Protein Simulations. *J chem theory comput*, **7**(4), 1157–1166.

Stansfeld, Phillip J., & Sansom, Mark S P. 2011b. Molecular simulation approaches to membrane proteins. *Structure*, **19**(11), 1562–1572.

Su, Chih-Ying, Menuz, Karen, & Carlson, John R. 2009. Olfactory perception: receptors, cells, and circuits. *Cell*, **139**(1), 45–59.

Sudo, Yuki, Furutani, Yuji, Shimono, Kazumi, Kamo, Naoki, & Kandori, Hideki. 2003. Hydrogen bonding alteration of Thr-204 in the complex between pharaonis phoborhodopsin and its transducer protein. *Biochemistry*, **42**(48), 14166–14172.

Sudo, Yuki, Furutani, Yuji, Kandori, Hideki, & Spudich, John L. 2006. Functional importance of the interhelical hydrogen bond between Thr204 and Tyr174 of sensory rhodopsin II and its alteration during the signaling process. *J biol chem*, **281**(45), 34239–34245.

Takeda, K, Matsui, Y, Kamiya, N, Adachi, S, Okumura, H, & Kouyama, T. 2004. Crystal structure of the M intermediate of bacteriorhodopsin: Allosteric structural changes mediated by sliding movement of a transmembrane helix. *J mol biol*, **341**(4), 1023–1037.

Takeuchi, Hiroko, Imanaka, Yukie, Hirono, Junzo, & Kurahashi, Takashi. 2003. Cross-adaptation between olfactory responses induced by two subgroups of odorant molecules. *J. gen. physiol.*, **122**(3), 255–264.

Taylor, M. R. G. 2007. Pharmacogenetics of the human beta-adrenergic receptors. *Pharmacogenomics j*, **7**(1), 29–37.

Tesmer, JJG, Sunahara, R K, Gilman, A G, & Sprang, S R. 1997. Crystal structure of the catalytic domains of adenylyl cyclase in a complex with G(s alpha).GTP gamma S. *Science*, **278**(5345), 1907–1916.

Tieleman, D Peter. 2010. Methods and Parameters for Membrane Simulations. *Page 330 of:* Sansom, Mark S P, & Biggin, Philip C (eds), *Molecular simulations and biomembranes: From biophysics to function (rsc biomolecular sciences)*. Royal Society of Chemistry.

Tomonaga, Yuya, Hidaka, Tetsurou, Kawamura, Izuru, Nishio, Takudo, Ohsawa, Kazuhiro, Okitsu, Takashi, Wada, Akimori, Sudo, Yuki, Kamo, Naoki, Ramamoorthy, Ayyalusamy, & Naito, Akira. 2011. An active photoreceptor intermediate revealed by in situ photoirradiated solid-state NMR spectroscopy. *Biophys j*, **101**(10), L50–2.

Tramontano, Anna, & Lesk, Arthur M. 2006. *Protein Structure Prediction: Concepts and Applications*. 1 edn. Wiley-VCH.

Tramontano, Anna, Jones, David, Rychlewski, Leszek, Casadio, Rita, Martelli, Pierluigi, Raimondo, Domenico, & Giorgetti, Alejandro. 2011. Structure prediction of globular proteins. *In: Modern genome annotation: The biosapiens network*. Springer.

Ulmschneider, Jakob P., & Ulmschneider, Martin B. 2009. United Atom Lipid Parameters for Combination with the Optimized Potentials for Liquid Simulations All-Atom Force Field. *J chem theory comput*, **5**(7), 1803–1813.

Van Gunsteren, W F, & Berendsen, H J C. 1988. A Leap-Frog Algorithm for Stochastic Dynamics. *Mol simulat*, **1**(3), 173–185.

Vanni, Stefano, Neri, Marilisa, Tavernelli, Ivano, & Rothlisberger, Ursula. 2011. Predicting novel binding modes of agonists to beta adrenergic receptors using all-atom molecular dynamics simulations. *Plos comput biol*, **7**(1), e1001053.

Villa, Elizabeth, Balaeff, Alexander, & Schulten, Klaus. 2005. Structural dynamics of the lac repressor-DNA complex revealed by a multiscale simulation. *Proc natl acad sci usa*, **102**(19), 6783–6788.

Vriend, G. 1990. What if - a Molecular Modeling and Drug Design Program. *J mol graph*, **8**(1), 52–&.

Wallner, Björn, & Elofsson, Arne. 2006. Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein sci*, **15**(4), 900–913.

Wang, JM, Wolf, RM, Caldwell, JW, Kollman, PA, & Case, DA. 2004. Development and testing of a general amber force field. *J comput chem*, **25**(9), 1157–1174.

Wang, Junmei, Cieplak, Piotr, & Kollman, Peter A. 2000. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? . *J comput chem*, **21**(July), 1049–1074.

Wang, Zun-Jing, & Deserno, Markus. 2010. A Systematically Coarse-Grained Solvent-Free Model for Quantitative Phospholipid Bilayer Simulations. *J phys chem b*, **114**(34), 11207–11220.

Warne, Ton, Serrano-Vega, Maria, Baker, Jillian, Moukhametzianov, Rousla, Edwards, Patricia, Henderson, Richar, Leslie, Andrew G, Tate, Christopher, & Schertler, Gebhard F. 2008. Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature*, **454**(7203), 7.

Wolf, Maarten G, Hoefling, Martin, Aponte-Santamaría, Camilo, Grubmüller, Helmut, & Groenhof, Gerrit. 2010. g_membed: Efficient insertion of a membrane protein into an equilibrated lipid bilayer with minimal perturbation. *J comput chem*, **31**(11), 2169–2174.

Yarnitzky, Talia, Levit, Anat, & Niv, Masha Y. 2010. Homology modeling of G-protein-coupled receptors with X-ray structures on the rise. *Curr opin drug discov devel*, **13**(3), 317–325.

Zagotta, William N, Olivier, Nelson B, Black, Kevin D, Young, Edgar C, Olson, Rich, & Gouaux, Eric. 2003. Structural basis for modulation and agonist specificity of HCN pacemaker channels. *Nature*, **425**(6954), 200–205.

Zarzo, Manuel. 2007. The sense of smell: molecular basis of odorant recognition. *Biol rev camb philos soc*, **82**(3), 455–479.

Zhang, G, Liu, Y, Ruoho, A E, & Hurley, J H. 1997. Structure of the adenylyl cyclase catalytic core. *Nature*, **386**(6622), 247–253.

Zhang, Xinmin, & Firestein, Stuart. 2002. The olfactory receptor gene superfamily of the mouse. *Nat neurosci*, **5**(2), 124–133.

Zhou, Hongyi, & Zhou, Yaoqi. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein sci*, **11**(11), 2714–2726.

Zhou, Hongyi, & Zhou, Yaoqi. 2005. SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics*, **21**(18), 3615–3621.

List of Figures

List of Tables

Curriculum Vitae

Personal Details

Date of birth: November 09th, 1984
Place of birth: Lam Dong, Vietnam
Sex: Male
Nationality: Vietnamese
Work address: German Research School for Simulation Sciences GmbH,
Wilhelm-Johnen-Strasse, 52428 Juelich, Germany
Phone: +49 2461 61 8916
Email: n.h.chuong@grs-sim.de

Qualifications

12/2009 – 02/2013: Doctoral candidate in Computational Biophysics at German Research School for Simulation Sciences GmbH, Juelich, Germany
Thesis: "Computational studies on seven-helix transmembrane receptors" (defense on 26/02/2013).
Supervisor: Prof. Dr. Paolo Carloni.

09/2002 – 10/2006: B.Sc. of Physics at Vietnam National University Ho Chi Minh City (VNU), University of Science, Faculty of Physics
Thesis: "Study of vortex glass transition in disordered type II superconductor by simulating linear resistivity using three dimensional lattice XY model."
Supervisor: Prof. Dr. Hoang Zung