Chapter 1

Introduction

Cells and organelles are bounded by membranes, which are composed of lipids and proteins. The lipids form a bilayered structure that is hydrophilic on its two outer surfaces and hydrophobic in between, and proteins are embedded in this layer. These membrane proteins can be classified into two broad categories—integral and peripheral—based on the protein-membrane interactions[1]. Most integral membrane proteins span the entire membrane (i.e., transmembrane protein). The regions of the protein that are actually crossing the bilayer are in most cases α helices, but are in some cases mutiple β strands as in porins. Although some proteins only pass through the membrane once as an α helix, others may be multipass, having several transmembrane α helices connected by hydrophilic loops. Some of integral proteins are anchored to the membrane by one α helix parallel to the plane of the membrane. Peripheral membrane proteins are usually bound to the membrane indirectly by non-covalent interactions with integral membrane proteins or directly by interactions with lipid polar head groups.

The transmembrane proteins play a role as active mediators between the cell and its environment or the interior of an organelle and the cytosol. They catalyze specific transport of ions across the membrane barriers (e.g., ion channels). They convert the energy of sunlight into chemical and electrical energy (e.g., photosynthetic reaction centers). They serve as signal receptors, for example, the G protein-coupled receptors (GPCRs) that are the main subject in this thesis, and transduce signals across the membrane. The signals can be neurotransmitters, growth factors, hormones, light or chemotactic stimuli.

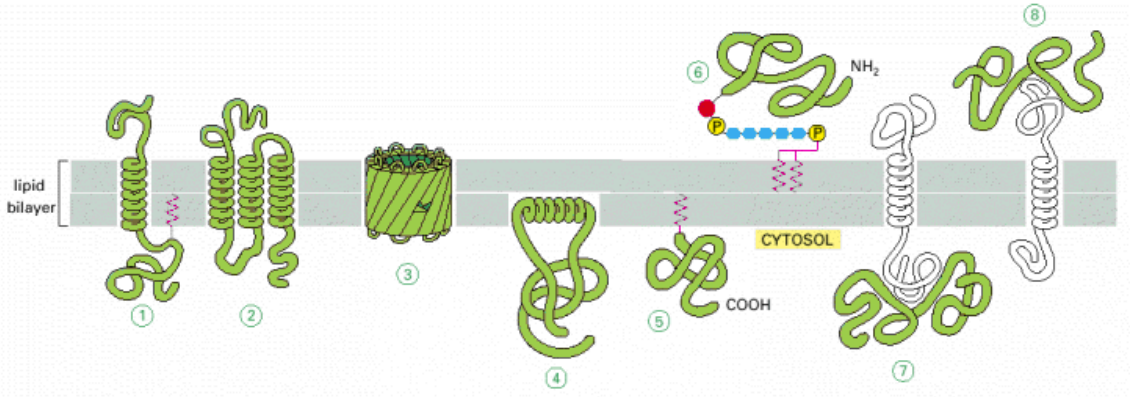


Figure 1.1 Various ways in which membrane proteins associate with the lipid bilayer. Most transmembrane proteins are thought to extend across the bilayer (1) as a single α helix, (2) as multiple α helices, or (3) as a rolled-up β sheet (a β barrel). Some of these "single-pass" and "multipass" proteins have a covalently attached fatty acid chain inserted in the cytosolic lipid monolayer (1). Other membrane proteins are exposed at only one side of the membrane. (4) Some of these are anchored to the cytosolic surface by an amphipathic a helix that partitions into the cytosolic monolayer of the lipid bilayer through the hydrophobic face of the helix. (5) Others are attached to the bilayer solely by a covalently attached lipid chain – either a fatty acid chain or a prenyl group in the cytosolic monolayer or, (6) via an oligosaccharide linker, to phosphatidylinositol in the noncytosolic monolayer. (7, 8) Finally, many proteins are attached to the membrane only by noncovalent interactions with other membrane proteins[1].

In this chapter we outline GPCRs, one of important transmembrane receptor families, on the structural and functional aspects, and discuss orphan GPCRs and an effort to identify their endogenous ligands and physiological functions (deorphanization). Lastly, the principles of molecular modeling are explained, focusing on the techniques used in our studies for the structural and functional prediction of GPCRs.

1.1 G protein-coupled receptors

GPCRs comprise a large and diverse family of proteins whose primary function is to induce extracellular stimuli into intracellular signals. These stimuli include light, neurotransmitters, odorants, biogenic amines, lipids, proteins, amino acids, hormones, nucleotides, and chemokines. They are among the largest and most diverse protein families in mammalian genomes[2]. The common structural feature is that they have seven transmembrane-spanning α-

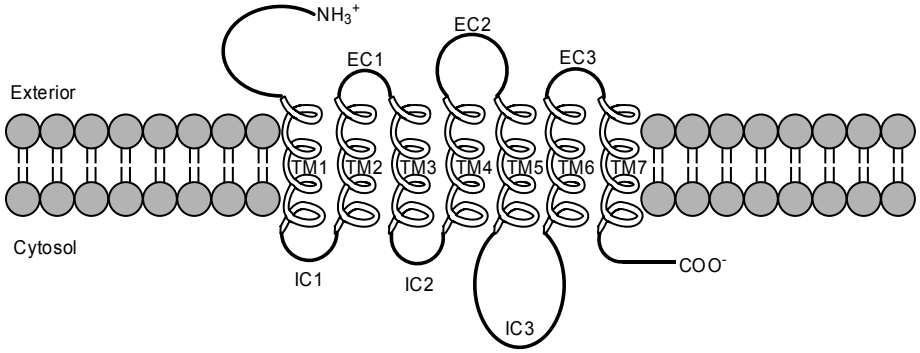


Figure 1.2 Schematic diagram of the general structure of G protein-coupled receptors. All receptors of this type contain seven transmembrane α-helical regions. The loop between α helices 5 and 6, and in some cases the loop between helices 3 and 4, which face the cytosol, are important for interactions with the coupled G protein. TM1–TM7 = transmembrane domains; EC1–EC3 = extracellular loops; IC1–IC3 = intracellular loops.

helical segments connected by alternating intracellular and extracellular loops, with the amino terminus located on the extracellular side and the carboxyl terminus on the intracellular side (fig. 1.2). GPCRs can be divided into three major subfamilies; rhodopsin-like family (family A), glucagon receptor-like family (family B) and metabotropic neurotransmitter/calcium receptors (family C)[3]. The family A has the largest number of receptors including biogenic amine receptors (adrenergic, serotonin, dopamine, muscarinic, histamine), neurotensin receptors, chemokine receptors, opioid receptors, and olfactory receptors. In a recent analysis of the GPCRs in the human genome more than 800 human GPCRs were listed[4]. Among them a total of 701 receptors belong to the rhodopsin-like family and, of these, 241 are non-olfactory.

GPCRs have been named based on their ability to recruit and regulate the activity of intracellular heterotrimeric G proteins (α, β and γ subunits)[3]. The extracellular signaling (ligand binding) is followed by a change in the conformation of the receptor. This activated receptor induces a conformational change in the associated G protein α subunit, leading to release of a guanosine diphosphate (GDP) followed by binding of a guanosine triphosphate (GTP). Subsequently, the GTP-bound form of the α subunit dissociates from the receptor as well as from

the stable βγ-dimer. Both the GTP-bound α subunit and the free βγ-dimer modulate several intracellular signaling pathways. These include stimulation or inhibition of adenylate cyclase and activation of phospholipases, in addition to regulation of potassium and calcium channel activity[5]. This variety of intracellular signaling pathways is dictated by the different G protein types in α, β and γ subunits and multiplicity in G protein coupling, that is, the simultaneous functional coupling of GPCRs with distinct unrelated G proteins[6]. There are at least 18 different human Gα proteins, at least 5 types of Gβ subunits and at least 11 types for Gγ subunits.

Signaling is then attenuated (desensitized) by GPCR internalization, which is facilitated by arrestin binding[7]. Arrestins bind specifically to GPCRs phosphorylated by G protein-coupled receptor kinases (GRKs) and lead to an interaction which participates in the desensitization of the receptor by disturbing their coupling to G proteins. Arrestins also target the receptors for internalization by means of their ability to interact with clathrin. Thus signaling, desensitization and eventual resenstization are regulated by complex interactions of various intracellular domains of the GPCRs with numerous intracellular proteins.

1.2 Orphan GPCRs and deorphanization

Although the biology of GPCRs is certainly intriguing, their ultimate importance is underscored by the fact that approximately 25% of the top 200 best-selling drugs target GPCRs (http://www.mindbranch.com/products/R359-0071.html) although only 10% of non-sensory GPCRs are known drug targets, emphasizing the potential of the remaining 90% of the GPCR superfamily for the treatment of human disease[8]. Among the non-sensory approximately 360 GPCR genes, the endogenous ligands have been identified for around 210 receptors leaving ~150 receptors for which the ligands remain unknown (“orphan receptors”)[9]. These orphan receptors may play important, albeit unknown, functions in various cells, so that some of them may be potential candidates for new drug targets.

Discovery of the endogenous ligand for an orphan receptor is the preferred strategy in deorphanization process since it provides additional biological information derived from the ligand that might give initial clues to the utility of receptor in disease and address pharmacological anomalies. The orphan receptor strategy has been developed with the aim of discovering novel natural ligands[10]. In this strategy, the cloned orphan GPCR is transfected in cells, which are then exposed to a tissue extract. Activation of the orphan GPCR is monitored by second messenger response. The tissue extract is fractionated and isolated to determine the chemical structure of the active compound. Melanin concentrating hormone (MCH), urotensin II and neuromedin U are example peptide ligands paired with orphan GPCRs through this strategy.

In the reverse pharmacology strategy, orphan GPCRs are screened using mixtures of synthetic ligands (naturally occurring). This approach can be extended with use of small-molecule focused libraries designed using known GPCR modulators (agonists or antagonists) as templates.

The widely used cell-based screening assays are based on calcium ion mobilization or modulation of intracellular cyclic adenosine monophosphate (cAMP) level. The calcium ion is naturally produced in cells upon activation of GPCRs coupled to α subunits belonging to G_q family (fig. 1.3)[11]. The release α subunit couples to phosphoinositidases of the phospholipase β class (PLCβ). Activation of PLCβ induces the formation of inositol-triphosphate and diacylglycerol from phosphatidylinositol diphosphate. Inositol-triphosphate in turn stimulates the release of intracellular calcium from endoplasmic reticulum. The heterologous expression of a member of the $G\alpha_q$ family, $G\alpha_{15}$ or $G\alpha_{16}$, can allow coupling of a wide range of GPCRs to PLCβ activity through an alternative pathway. Therefore it is possible to force a receptor to response to an agonist via PLCβ activation, thus considerably broadening the range of receptors that will give a measurable calcium mobilization response.

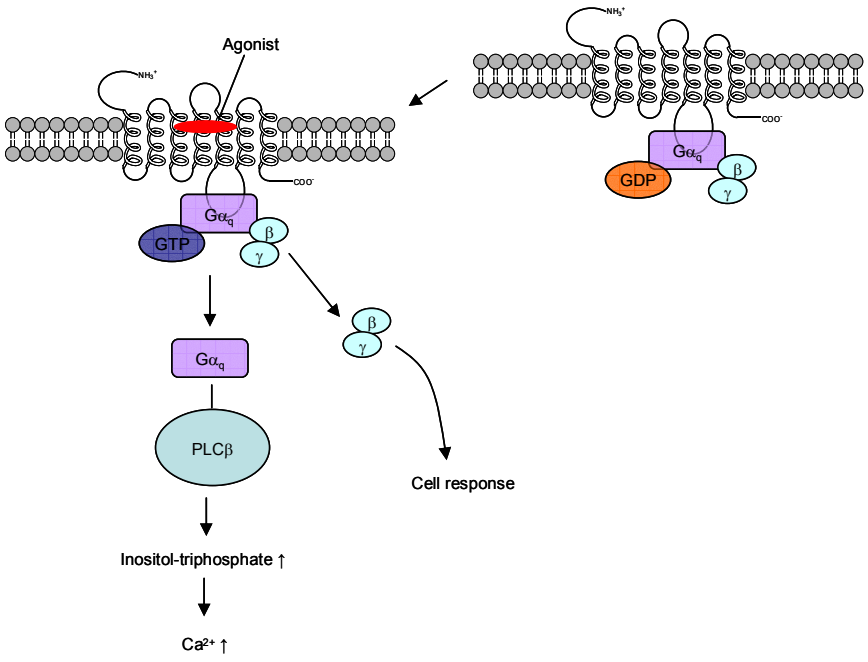


Figure 1.3 Classical examples of GPCR signalling. After agonist binding, a transient high-affinity complex of agonist, activated receptor and G protein is formed. GDP is released from the G protein and is replaced by GTP. This leads to dissociation of the G-protein complexes into a subunits and bg dimers, which both activate several effectors. Gaq, for instance, couples to phosphoinositidases of the phospholipase beta class (PLCb), which leads to an increase in inositol-triphosphate. Inositol-triphosphate in turn stimulates the release of intracellular calcium.

Recently Dong et al.[12] and Lembo et al.[13] have identified a novel family of GPCRs called the Mas-related gene (Mrg) receptor for mouse or the sensory neuron specific receptor (SNSR) in mice and human. A subset of these receptors including mouse MrgA1 (mMrgA1) and mouse MrgC11 (mMrgC11) is distributed mainly to isolectin $B4^{+}$, small diameter nociceptors in the dorsal root ganglia (DRG), which are suggested to be involved in pain sensation or modulation. Mrg receptors have been paired with structurally diverse transmitter peptides and provide a daunting case for deorphanization[14]. Although these receptors remain orphans, and their precise physiological function remains unknown, distinct and selective peptides activating some of these receptors have been identified:

- BAM22 derived from preproenkephalin A, one of endogenous opioid peptides activates SNSR3 (EC_{50} ~ 13 nM) or SNSR4 (EC_{50} ~ 16 nM)[13].

- The neuropeptide RF amides are potent for mouse Mrg receptors, for example, NPFF for MrgA1 (EC_{50} ~ 200 nM) and MrgC11 (EC_{50} ~ 54 nM) and NPAF for MrgA4 (EC_{50} ~ 60 nM)[12, 15].

- In addition, adenine shows high affinity (K_i ~ 18 nM) and potency for rat MrgA receptor[16].

- Cortistatin has been identified to activate potently human MrgX2 (EC_{50} ~ 25 nM)[17].

- More recently Grazzini *et al*. have observed that γ2-MSH is highly potent in rat MrgC receptor and the active moiety recognized by rat MrgC receptor is the C-terminal RF-amide motif of γ2-MSH[18].

- Recent studies also show that MrgD receptors specifically respond to β-alanine with micromolar concentration[19].

Our studies aimed to contribute to deorphanization of Mrg receptors, especially focusing on mMrgC11, mMrgA1 and rat MrgA, by characterizing the active site and screening the chemical libraries to search for the potential agonist or antagonists.

1.3 The 3D structure of GPCR and molecular modeling

Clearly it would be most useful to have the three-dimensional (3D) structure of the receptor to help select the most promising new ligands for experimental assays. Moreover the structural information is essential in designing receptor subtype-specific drugs. However, GPCRs, like other membrane proteins, are difficult to crystallize. Membrane proteins, which have both hydrophobic and hydrophilic regions on their surfaces, are not soluble in aqueous buffer and denature in organic solvent. In addition, because membrane proteins are typically produced in a

heterogeneous manner by cells with substantial variability in glycosylation, obtaining high-quantity and high-purity GPCR proteins is very challenging[20]. All GPCRs are known to have a common motif of seven transmembrane helical structures, but the only GPCR crystal structure published at atomic resolution is of inactive conformation of rhodopsin[21]. Here comes the demand for prediction of the 3D structures of GPCRs. The low (<25 %) sequence homology with rhodopsin sheds some uncertainties on the accuracy of a 3D structure constructed by using the comparative homology modeling method. Clearly, then it is necessary to devise a general method that predicts more reliable structures.

Recently MembStruk computational method to predict the 3D structure of GPCRs has been developed in Goddard's group[22]. It includes prediction of transmembrane (TM) α helices using hydrophobicity profile with a set of homologous sequences, subsequent optimization in relative orientations of helices and then conformational optimization of the entire receptor structure using molecular mechanics (MM) and molecular dynamics (MD). The binding site of the GPCR is further predicted using the HierDock method to validate the predicted protein structure, and the binding modes of the ligand are suggested. In our study of Mrg receptors, we also applied the Membstruk method in prediction of their 3D structures and the HireDock method in characterization of the binding site. Chapter 2 describes the details in each step of the procedure.

In the following sections, the basic principles of molecular modeling are explained with specific technique used in prediction of the 3D GPCR structure and the binding site.

1.3.1 Hydrophobicity scale: TM prediction from the primary sequence

The membrane helices are embedded in a hydrophobic environment and are built up from continuous regions of predominantly hydrophobic amino acids. Thus from the amino acid sequences, the regions that comprise the TM helices can be predicted with reasonable confidence. In order to determine whether the segment of amino acid sequences is likely to be a TM helix, we

Table 1.1 Eisenberg hydrophobicity scale

Amino acid	I	F	V	L	W	M	A	G	C	Y
Hydrophobicity	0.73	0.61	0.54	0.53	0.37	0.26	0.25	0.16	0.04	0.02
Amino acid	P	T	S	H	E	N	Q	D	K	R
Hydrophobicity	-0.07	-0.18	-0.26	-0.40	-0.62	-0.64	-0.69	-0.72	-1.1	-1.8

need to measure the amount of hydrophobicity. The numerical hydrophobicity scales of each amino acid have been derived in several groups on the basis of solubility measurements of the amino acids in different solvents, vapor pressure of side-chain analogs, analysis of side-chain distributions within soluble proteins, and theoretical energy calculations. These values generally correspond to the free energy of transfer of the side chain of the amino acid from water to a nonpolar environment. In our study, we used the “consensus” hydrophobicity scale that Eisenberg *et al*. introduced by averaging the normalized hydrophobicities for each residue over the five known scales[23]. The hydrophobicity values of 20 amino acids in the Eisenberg scale are shown in table 1.1.

With the given hydrophobicity scale, the hydropathy index, the mean value of the hydrophobicity of the amino acids within a window (12 to 20 residues long in MembStruk), is calculated for each position in the sequence. In MembStuck, the hydropathy plot, the curve of the hydropathy indices against residue numbers is evaluated from the multiple sequence alignment of the set of homologous sequences with a target protein sequence[22]. First, the hydrophobicity at each residue position is averaged over all the sequences in the multiple sequence alignment. Then we calculate the mean hydrophobicity over a window size of residues around every residue position. Figure 1.4 shows one example of a hydropathy plot obtained from MembStruk.

1.3.2 Force field

The molecular state can be accurately described by solving the Schrödinger equation:

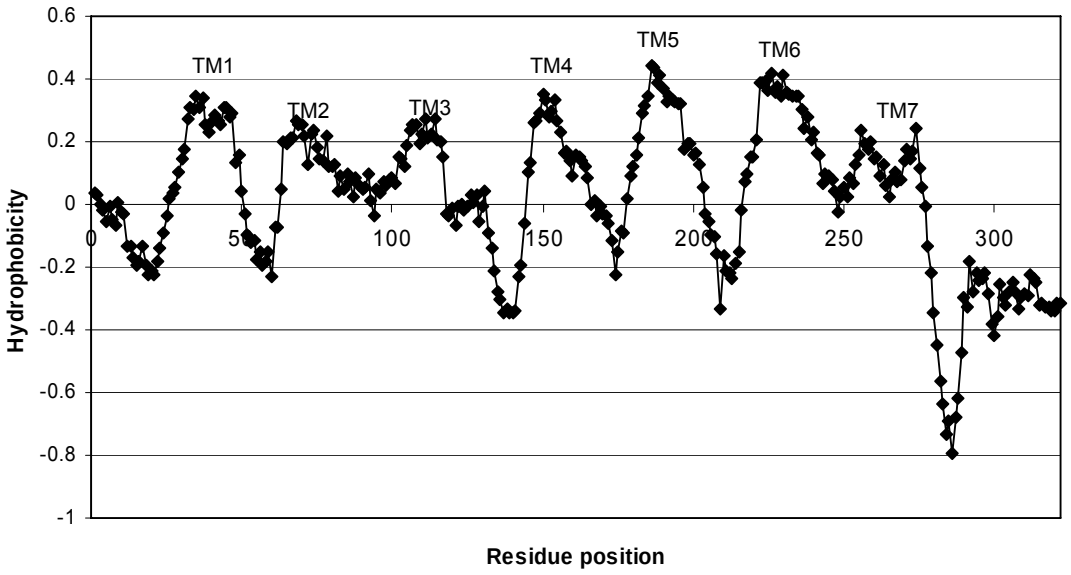


Figure 1.4 Hydrophobicity profile for mouse MrgC11 sequence set (window size = 12)

$$H\Psi(R,r) = E(R,r)\Psi(R,r)\,, \tag{1.1}$$

where H is the Hamiltonian for the system, Ψ is the wavefunction, and E is the energy. In general, Ψ is a function of the coordinates of the nuclei (R) and of the electrons (r). Although this equation is quite general, it is too complex for any practical use, so approximations are made. Based on the Born-Oppenheimer approximation that the electrons are several thousands of times lighter than the nuclei and therefore move much faster, the motion of the electrons can be decoupled from that of the nuclei, giving two separate equations. The first equation describes the electronic motion:

$$(H_{el} + V_{NN})\psi_{el}(r;R) = U(R)\psi_{el}(r;R)\,, \tag{1.2}$$

where the purely electronic Hamiltonian H_{el} includes nuclear repulsion V_{NN}. It depends only parametrically on the positions of the nuclei. This equation defines the energy, $U(R)$, which is a function of only the coordinates of the nuclei. This energy is usually called the *potential energy surface*.

The second equation then describes the motion of the nuclei on this potential energy surface $U(R)$:

$$H_N \Phi_N(R) = E\Phi_N(R). \tag{1.3}$$

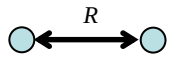
In principle, (1.2) could be solved for the potential energy U, and then (1.3) could be solved. However, the effort required to solve (1.2) is extremely large, so usually an empirical fit to the potential energy surface, commonly called the forcefield (V), is used. Since the nuclei are relatively heavy objects, quantum mechanical effects are often insignificant, in which case (1.3) can be replaced by Newton's equation of motion:

$$-\frac{dV}{dR} = m\frac{d^2R}{dt^2}. \tag{1.4}$$

The solution of (1.4) using an empirical fit to the potential energy surface $U(R)$ is called "molecular dynamics". Molecular mechanics ignores the time evolution of the system and instead focuses on finding particular geometries and their associated energies or other static properties.

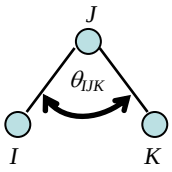
The potential energy is expressed as a sum of valence interaction, nonbonded interaction and additional terms such as constraints. The valence interactions consist of bond stretching (E_{bond}, two-body), bond angle bending (E_{angle}, three-body), dihedral angle torsion ($E_{torsion}$, four-body) and inversion ($E_{inversion}$, four-body), that are in nearly all force fields of covalent systems plus cross-terms that are included in more sophisticated force fields developed to produce accurate vibrational frequencies. The nonbonded interactions are composed of van der Waals or dispersion (E_{vdw}), electrostatic ($E_{coulomb}$) and explicit hydrogen bonds (E_{hbond}) terms. Figure 1.5 shows the schematic representation of these valence and nonbonded interactions with the functional forms of potentials used in DREIDING force field[24].

1.3.3 Molecular mechanics



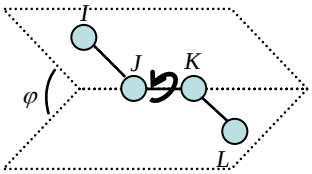
$$E_{bond} = \frac{1}{2} k_e (R - R_e)^2$$

k_e = force constant and R_e = equilibrium bond distance



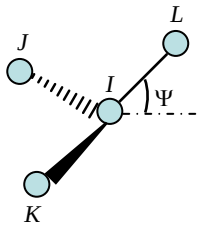
$$E_{angle} = \frac{1}{2} C_{IJK} [\cos\theta_{IJK} - \cos\theta_J^0]^2$$

θ = angle between bonds IJ and JK



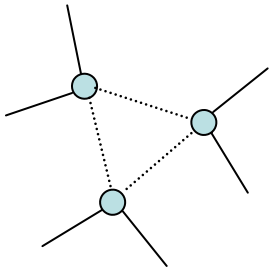
$$E_{torsion} = \frac{1}{2} V_{JK} \{1 - \cos[n_{JK} (\varphi - \varphi_{JK}^0)]\}$$

φ = angle between the IJK and JKL planes, n_{JK} = periodicity and V_{JK} = rotation barrier



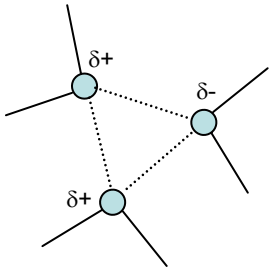
$$E_{inversion} = \frac{1}{2} C_I (\cos\Psi - \cos\Psi_I^0)^2$$

Ψ = angle between the IL bond and the JIK plane



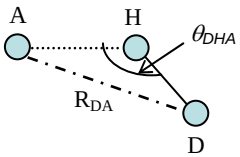
$$E_{vdw} = D_0 [\rho^{-12} - 2\rho^{-6}]$$

ρ = scaled distance and D_0 = well depth



$$E_{Coulomb} = (322.0637) Q_i Q_j / \varepsilon R_{ij}$$

Q_i, Q_j = charges, R_{ij} = distance and ε = dielectric constant



$$E_{hbond} = D_{hb} [5(R_{hb} / R_{DA})^{12} - 6(R_{hb} / R_{DA})^{10}] \cos^4(\theta_{DHA})$$

θ_{DHA} = angle between the hydrogen donor (D), the hydrogen (H) and the hydrogen acceptor (A), R_{DA} = distance between D and A

Figure 1.5 Schematic representation of the six key contributions of molecular mechanics force field; bond stretching, angle bending, inversion, non-bonded (van der Waals and Coulomb) and hydrogen bond interactions.

The potential energy of a system of N particles, $U=U(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)$, is minimized with the respect to their positions r_i (and, possibly, some other internal coordinates). After an initial configuration has been specified, the positions of particles are adjusted using an iterative computational method until the minimum energy configuration is attained. It should be emphasized that U, which is a function of $3N$ variables, may possess a number of minima. No method guarantees that the lowest energy minimum will be found.

All minimization methods pursue the following algorithm: if in the mth iteration the system of particles is described by position vectors $r_i^{(m)}$ then in the $(m + 1)$th interation the position vectors are

$$r_i^{(m+1)} = r_i^{(m)} + \Delta r_i^{(m)}, \tag{1.5}$$

where $\Delta \boldsymbol{r}_i^{(m)}$ is determined so as to decrease the potential energy and approach, eventually, a minimum of $U(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)$. Different molecular mechanics (MM) methods of relaxation differ in the way $\Delta \boldsymbol{r}_i^{(m)}$ is determined. There are three commonly used methods for finding minima: steepest descent, Newton's method and conjugate gradient. Here the conjugate gradient method that we used is explained briefly. The conjugate gradient method is based on the idea that the convergence to the energy minimum could be accelerated if we minimize a function (here U) over the hyperplane that contains all previous search directions. In steepest descent, the position vectors $\boldsymbol{r}_i$ are being adjusted in proportion to the negative gradient of U, that is, the force $\boldsymbol{F}_i$ at any given iteration. However, in the conjugate gradient the directions of the displacements of the $(m + 1)$th iteration are not determined only on the basis of the forces calculated in the mth iteration but also using values of the forces found in previous iterations. This is carried out as follows:

The increment of the $3N$ dimensional vector $\mathbf{R} = \{r_i^{\alpha}\}$ is

$$\Delta R = \sum_{k=1}^{3N} \lambda_k \Phi_k , \tag{1.6}$$

where $\mathbf{\Phi}_k$ are $3N$ vectors in the $3N$-dimensional space that have been gradually constructed in the previous $3N$ iterations. In the first iteration $\mathbf{\Phi}_1 = \boldsymbol{F}^{(1)}$, where $\boldsymbol{F}^{(1)}$ is the $3N$-dimensional vector of forces evaluated in the first iteration, and all other vectors $\mathbf{\Phi}_k$ for $k > 1$ are set to zero. In the second iteration the vector $\mathbf{\Phi}_2$ is constructed as $\mathbf{\Phi}_2 = \boldsymbol{F}^{(2)}$, similarly as in the first iteration; all other vectors $\mathbf{\Phi}_k$ for $k > 2$ are set to zero. In the following iterations the recursive formula

$$\Phi_m = F^{(m)} + \frac{F^{T(m-1)}F^{(m-1)}}{F^{T(m-2)}F^{(m-2)}}\Phi_{m-1} \tag{1.7}$$

is used to construct gradually additional vectors $\mathbf{\Phi}_k$; T denotes the transpose of the corresponding vector. Thus in every iteration, m, a new vector $\mathbf{\Phi}_k$ is added until $3N$ vectors have been constructed in the first $3N$ iterations. At this point these $3N$ vectors are used to determine $\Delta\boldsymbol{R}^{(3N+1)}$ in the $3N$+1 iteration according to (1.7). When the number of iterations, M, is larger than $3N$, then $3N$ vectors constructed in the previous $3N$ iterations are used in determining $\Delta\boldsymbol{R}^{(M+1)}$ in the M+1 iteration.

1.3.4 Molecular dynamics

In molecular dynamics (MD) that investigates the motion of atoms in time as discussed in section 1.3.2, successive configurations of a system are generated by integrating Newton's law of motion, hence resulting in a trajectory that specifies the positions and velocities of the atoms as function of time. In (1.4), the accelerations of atoms are determined from the gradient of the potential energy and therefore their velocities can be derived, resulting in new positions of the atoms.

The approach taken by MD is to solve the equations of motion numerically on a computer. The most widely used algorithm of integrating the equations of motion is Verlet algorithm. It uses the positions and acceleration (= $\boldsymbol{F}_i/m_i$) at time t and the positions from the previous step, $\boldsymbol{r}_i(t\text{-}\Delta t)$,

to calculate the new positions at $t+\Delta t$, $\boldsymbol{r}_i(t+\Delta t)$. Using the central difference method for numerical evaluation of the second derivative, the equation of motion for $\boldsymbol{r}_i$ can be written as

$$\frac{d^2 r_i(t)}{dt^2} = \frac{1}{(\Delta t)^2}[r_i(t+\Delta t) - 2r_i(t) + r_i(t-\Delta t)] = \frac{1}{m_i}F_i(t)\,, \tag{1.8}$$

and therefore

$$r_i(t+\Delta t) = 2r_i(t) - r_i(t-\Delta t) + \frac{(\Delta t)^2}{m_i}F_i(t)\,. \tag{1.9}$$

The basic recurrent formula for the MD simulation proceeds as follows:

The forces $\boldsymbol{F}_i(J\Delta t)$ are first evaluated at the time step J.

Positions $\boldsymbol{r}_i((J+1)\Delta t)$ at the time step $J+1$ are calculated using (1.9)

Velocities $\boldsymbol{v}_i(J\Delta t)$ at the time step $J+1$ may be calculated as

$$v_i(t) = \frac{r_i(t+\Delta t) - r_i(t-\Delta t)}{2\Delta t}\,. \tag{1.10}$$

Implementation of the Verlet algorithm is straightforward and the storage requirements are modest, comprising two sets of positions and the force. One of its drawbacks is that the positions $\boldsymbol{r}_i(t+\Delta t)$ are obtained by adding a small term $(\Delta t)^2\boldsymbol{F}_i/m_i$ to the difference of two much larger terms, $2\boldsymbol{r}_i(t)$ and $\boldsymbol{r}_i(t-\Delta t)$. This may lead to a loss of precision. Some other disadvantages are that it does not have an explicit velocity term in the equation and indeed velocities are not available until the positions have been computed at the next step. Moreover it is not a self-starting algorithm; the new positions are calculated from the current positions $\boldsymbol{r}_i(t)$ and the previous time step, $\boldsymbol{r}_i(t-\Delta t)$.

The velocity Verlet method is one of the variations on the Verlet algorithm. It gives positions, velocities and forces at the same time and does not compromise precision. The MD simulation then proceeds as follows:

The forces $\boldsymbol{F}_i(J\Delta t)$ are first evaluated at the time step J.

Positions $\boldsymbol{r}_i((J+1)\Delta t)$ and velocities at the time step $J+1$ are evaluated as

$$r_i((J+1)\Delta t) = r_i(J\Delta t) + \Delta t v_i(J\Delta t) + \frac{(\Delta t)^2}{2m_i} F_i(J\Delta t) \tag{1.11}$$

$$v_i((J+1)\Delta t) = v_i(J\Delta t) + \frac{(\Delta t)}{2m_i}(F_i((J+1)\Delta t) + F_i(J\Delta t)) . \tag{1.12}$$

In the above formalism, the coupling of the system with a heat bath is not considered yet. Actually in the ensemble such as the canonical ensemble or the isobaric-isothermal ensemble where the temperature, T is kept constant, that is, the kinetic energy of the system should be constant, the scaling of the velocity is necessary during MD simulation. The simplest approach is to first compute the instantaneous kinetic energy $\frac{1}{2}\sum_{i=1}^{N} m_i v_i^2$ from the velocities obtained from (1.10) or (1.12) and then scale velocities by a factor λ chosen such as to preserve the temperature T

$$\lambda = \left[\frac{3Nk_B T}{\sum_{i=1}^{N} m_i v_i^2} \right]^{1/2} . \tag{1.13}$$

The more sophisticated schemes are Anderson thermostat and Nose-Hoover thermostat in which the exchange of heat with a bath is explicitly included.

1.3.5 Molecular docking

In molecular docking, we attempt to predict the structure of the intermolecular complex formed between two molecules. Most docking cases target at the identification of the low-energy binding modes of a small molecule (a ligand) within the active site of a macromolecule such as a protein receptor, whose structure is known. Therefore solving a docking problem computationally

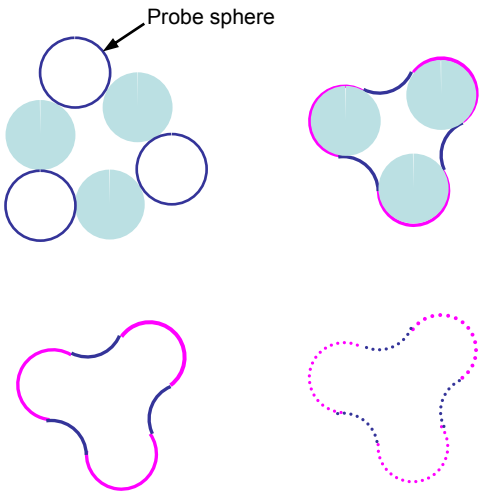


Figure 1.6 Construction of molecular surface in 2D. The filled circles (cyan) correspond to the van der Waals spheres of the atoms. The molecular surface is obtained with a spherical probe and the contact surface is in magenta and the reentrant surface is in blue. Actually the molecular surface is a collection of points and vectors normal to the surface at each point.

requires an accurate description of the molecular energetics (scoring function) as well as an efficient algorithm to search for the potential binding modes.

The docking problem involves many degrees of freedom; three translational and three rotational freedom of one molecule relative to the other as well as the conformational degrees of freedom for each molecule. In reality, it is almost impossible to consider all possible degrees of freedom since one of the molecules in the docking problems is a macromolecule. Therefore the simplest algorithms treat the two molecules as rigid bodies and explore the six degrees of translational and rotational freedom. A well-known example is the DOCK program of Kuntz and co-workers[25]. DOCK is based on the shape complementarity between a ligand and the pocket in a receptor that forms the binding site. To describe the shape of the binding site in a receptor, the molecular surface is calculated first. The molecular surface is divided into two classes; the contact surface and the reentrant surface. The contact surface is the part of the van der Waals surface that can be touched by a probe sphere. The reentrant surface consists of the inward-facing

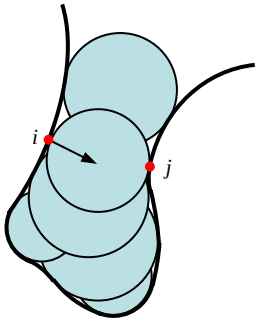


Figure 1.7 A binding site represented as a collection of overlapping spheres. Each sphere touches the molecular surface at two points.

part of the probe sphere when it is in contact with more than one atom. The surface can only be defined completely with reference to a probe object of some form, and indeed depend on the probe size (the probe radius for a spherical probe). A spherical probe of radius 1.4Å to approximate a water molecule is most commonly used. In the diagram of figure 1.6 where the molecular surface is obtained with a spherical probe the contact surface is in magenta and the reentrant surface is in blue.

Next a collection of overlapping spheres of varying radii filling the binding pocket is generated. Each sphere touches the molecular surface at two points (*i*, *j*) and has its center on the surface normal from point *i* and lies on the outside of the receptor surface ("negative image"). Ligand atom are matched to the sphere centers to find matching sets in which all the distances between the ligand atoms in the set are equal to the corresponding sphere center-sphere center distances within some tolerance (1 to 2Å). Actually matching four pairs is sufficient to determine the rigid docking. Then the ligand is positioned within the site by performing the least square fits of the atoms to the sphere centers, as shown in figure 1.8. The orientation may be checked to make sure that there is no unacceptable steric interaction between the ligand and the receptor. If the ligand orientation is acceptable, the interaction energy is calculated to give the "score" for that binding mode. The DOCK uses the grid-based energy evaluation in which the receptor-dependent terms in the potential function are pre-calculated at points on a 3D grid in order to minimize the

overall computational costs of evaluation[26]. Grid-based scoring can be accomplished when the ligand and receptor terms in the evaluation function are separable. It could be achieved in the following ways. The energy scores are calculated as a sum of van der Waals and electrostatic components:

$$E = \sum_{i=1}^{lig} \sum_{j=1}^{rec} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + 332.0 \frac{q_i q_j}{D r_{ij}} \right], \tag{1.14}$$

where each term is a double sum over ligand atoms i and receptor atoms j, A_{ij} and B_{ij} are van der Waals repulsion and attraction parameters, r_{ij} is the distance between atoms i and j, q_i and q_j are the point charges on atoms i and j, D is the dielectric constant and 332.0 is a factor that converts the electrostatic energy into kcal/mol. By using a geometric mean approximation, the van der Waals parameters A_{ij} and B_{ij} can be expressed with the single-atom-type parameters as follows:

$$A_{ij} = \sqrt{A_{ii}} \sqrt{A_{jj}} \quad \text{and} \quad B_{ij} = \sqrt{B_{ii}} \sqrt{B_{jj}} \ . \tag{1.15}$$

Therefore Eq. 1.14 can be rewritten as:

$$E = \sum_{i=1}^{lig} \left[\sqrt{A_{ii}} \sum_{j=1}^{rec} \frac{\sqrt{A_{jj}}}{r_{ij}^{12}} - \sqrt{B_{ii}} \sum_{j=1}^{rec} \frac{\sqrt{B_{jj}}}{r_{ij}^{6}} + q_i \sum_{j=1}^{rec} \frac{332.0 q_j}{D r_{ij}} \right]. \tag{1.16}$$

Three values are stored for every grid point k, each a sum over receptor atoms that are within a user-defined distance of the point:

$$aval = \sum_{j=1}^{rec} \frac{\sqrt{A_{jj}}}{r_{jk}^{12}} \quad bval = \sum_{j=1}^{rec} \frac{\sqrt{B_{jj}}}{r_{jk}^{6}} \quad esval = \sum_{j=1}^{rec} \frac{332.0 q_j}{D r_{jk}} . \tag{1.17}$$

The final scoring function can be expressed in the multiplication of these values (which is may be values at the nearest point from the corresponding ligand atom or the results of trilinearly interpolating the values for the eight surrounding points) by the appropriate ligand values:

$$E = \sum_{i=1}^{lig} \left[\sqrt{A_{ii}} (aval) - \sqrt{B_{ii}} (bval) + q_i (esval) \right]. \tag{1.18}$$

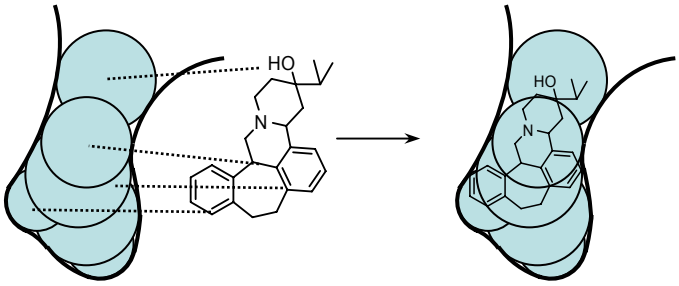


Figure 1.8 Matching algorithm in DOCK. Atoms are matched to spheres centers and then molecule is placed in the binding pocket (Reproduced from [27]).

New orientations are generated by matching different sets of ligand atoms and sphere centers and then scored. The top-scoring orientations are retained for subsequent analysis.

To perform the flexible docking, the conformational degrees of freedoms should be considered. Most of the methods including DOCK take into account only the conformational space of the ligand and assume that the receptor is fixed. In DOCK, the rotatable bonds are defined with the possible discrete torsion angles based on the hybridizations of two atoms in the bond. The conformations of a ligand are searched or relaxed by modifying only the torsion angles with the bond lengths or angles fixed. DOCK uses two search strategies: incremental construction and random conformation search.

To briefly explain, in the incremental construction (anchor and grow) technique a rigid portion of the ligand, the anchor, is first identified and docked using a geometrical matching procedure[28]. To select the anchor, all rotatable bonds in the ligand are identified and the ligand molecule is divided into rigid, overlapping segments, then the anchor segment is selected (fig. 1.8). Usually the largest overlapping segment is chosen as the anchor. In the next step, the molecular atoms of the ligand organized into non-overlapping segments arranged concentrically around anchor. In the conformation search step, the remaining molecular segments are added to the docked anchor starting from the inner layer. On each cycle, a molecular segment is added to the current set of partial binding configurations and sampling the appropriate torsion positions of

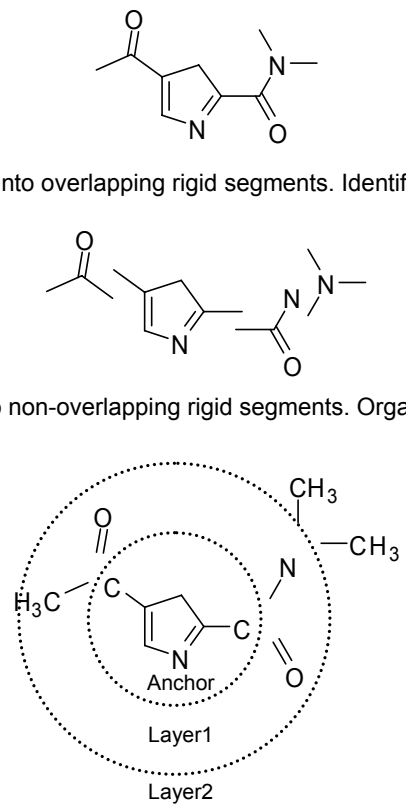


Figure 1.9 Atom pre-organization and anchor selection[27].

the intervening rotatable bond. The set of partial binding configurations are pruned based on score and positional diversity to avoid the exponential growth of a systematic conformation search.

When the conformational freedom is given to flexible ligand molecules during construction, the intramolecular energy term of the ligand should be considered in scoring. In addition to prevention of internal clash, the van der Waals and coulombic energies are computed for interaction between atoms in different rigid segments in DOCK. Atoms within a rigid segment are excluded because their contribution is a constant. The overall scoring includes both the intramolecular energy and the intermolecular energy between the ligand and the receptor discussed earlier.

The HierDock protocol[29] used in our study applies more sophisticated scoring method to the set of configurations generated from the DOCK run in order to complement the crude scoring function in DOCK. The selection of the top configurations proceeds in the hierarchical way; along with scoring steps the number of selected configurations decreases, and on the other hand the more degrees of freedom are taken into account in the energy scoring. Moreover the recent development of MSCDock (a new version of HierDock) incorporates the diversity and enrichment scheme into DOCK 4.0 to enhance the completeness in the conformation search. All the details are described in the next chapter.

1.4 Outline of Thesis

The following part of the thesis is composed of four chapters:

- In chapter 2, we predict the 3D structure of the mMrgC11 and mMrgA1 receptors using the MembStruk computational method. We also predict the binding sites of the di- and tetra-peptide ligands containing the RF amide motif that have been identified as agonists for these receptors. The subsequent mutagenesis experiments validate our prediction of the binding site in the mMrgC11 receptor.
- Chapter 3 describes the all-atom MD simulation of mMrgC11/F-(D)M-R-F-NH_2 complex in the explicit lipid and water environment.
- In chapter 4, the virtual ligand screening for the predicted binding site of mMrgC11 receptor is carried out as an effort to identify novel non-peptide ligands.
- In chapter 5, the 3D structure and the binding site of rat MrgA receptor are predicted using the homology modeling and docking method.

In appendix A, the quantum mechanics and molecular dynamics study of the 5-formyluracil, which was my earlier PhD subject, is discussed.

References

1. Alberts, B., et al., *Molecular biology of the cell*. 4th ed. 2002, New York: Garland Science.
2. Kroeze, W.K., D.J. Sheffler, and B.L. Roth, *G-protein-coupled receptors at a glance.* Journal of Cell Science, 2003. **116**: p. 4867-4869.
3. Gether, U., *Uncovering molecular mechanisms involved in activation of G protein-coupled receptors.* Endocrine Reviews, 2000. **21**(1): p. 90-113.
4. Fredriksson, R., et al., *The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints.* Molecular Pharmacology, 2003. **63**(6): p. 1256-1272.
5. Hamm, H.E., *The many faces of G protein signaling.* Journal of Biological Chemistry, 1998. **273**(2): p. 669-672.
6. Hermans, E., *Biochemical and pharmacological control of the multiplicity of coupling at G-protein-coupled receptors.* Pharmacology & Therapeutics, 2003. **99**(1): p. 25-44.
7. Bockaert, J. and J.P. Pin, *Molecular tinkering of G protein-coupled receptors: an evolutionary success.* Embo Journal, 1999. **18**(7): p. 1723-1729.
8. Vassilatis, D.K., et al., *The G protein-coupled receptor repertoires of human and mouse.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(8): p. 4903-4908.
9. Wise, A., S.C. Jupe, and S. Rees, *The identification of ligands at orphan G-protein coupled receptors.* Annual Review of Pharmacology and Toxicology, 2004. **44**: p. 43-66.
10. Civelli, O., et al., *Novel neurotransmitters as natural ligands of orphan G-protein-coupled receptors.* Trends in Neurosciences, 2001. **24**(4): p. 230-237.
11. Robas, N., et al., *Maximizing serendipity: strategies for identifying ligands for orphan G-protein-coupled receptors.* Current Opinion in Pharmacology, 2003. **3**(2): p. 121-126.

12. Dong, X.Z., et al., *A diverse family of GPCRs expressed in specific subsets of nociceptive sensory neurons.* Cell, 2001. **106**(5): p. 619-632.

13. Lembo, P.M.C., et al., *Proenkephalin A gene products activate a new family of sensory neuron-specific GPCRs.* Nature Neuroscience, 2002. **5**(3): p. 201-209.

14. Civelli, O., *GPCR deorphanizations: the novel, the known and the unexpected transmitters.* Trends in Pharmacological Sciences, 2005. **26**(1): p. 15-19.

15. Han, S.K., et al., *Orphan G protein-coupled receptors MrgA1 and MrgC11 are distinctively activated by RF-amide-related peptides through the G alpha(q/11) pathway*. Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(23): p. 14740-14745.

16. Bender, E., et al., *Characterization of an orphan G protein-coupled receptor localized in the dorsal root ganglia reveals adenine as a signaling molecule.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(13): p. 8573-8578.

17. Robas, N., E. Mead, and M. Fidock, *MrgX2 is a high potency cortistatin receptor expressed in dorsal root ganglion.* Journal of Biological Chemistry, 2003. **278**(45): p. 44400-44404.

18. Grazzini, E., et al., *Sensory central neuron-specific receptor activation elicits and peripheral nociceptive effects in rats.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(18): p. 7175-7180.

19. Shinohara, T., et al., *Identification of a G protein-coupled receptor specifically responsive to beta-alanine.* Journal of Biological Chemistry, 2004. **279**(22): p. 23559-23564.

20. Filmore, D., *It's a GPCR world*, in *Modern Drug Discovery*. 2004. p. 24-28.

21. Palczewski, K., et al., *Crystal structure of rhodopsin: A G protein-coupled receptor.* Science, 2000. **289**(5480): p. 739-745.

22. Trabanino, R.J., et al., *First principles predictions of the structure and function of G-protein-coupled receptors: Validation for bovine rhodopsin.* Biophysical Journal, 2004. **86**(4): p. 1904-1921.

23. Eisenberg, D., et al., *Hydrophobic Moments and Protein-Structure.* Faraday Symposia of the Chemical Society, 1982(17): p. 109-120.

24. Mayo, S.L., B.D. Olafson, and W.A. Goddard, *Dreiding - a Generic Force-Field for Molecular Simulations.* Journal of Physical Chemistry, 1990. **94**(26): p. 8897-8909.

25. Kuntz, I.D., et al., *A Geometric Approach to Macromolecule-Ligand Interactions.* Journal of Molecular Biology, 1982. **161**(2): p. 269-288.

26. Meng, E.C., B.K. Shoichet, and I.D. Kuntz, *Automated Docking with Grid-Based Energy Evaluation.* Journal of Computational Chemistry, 1992. **13**(4): p. 505-524.

27. Leach, A.R., *Molecular Modelling; principles and applications.* 2nd ed. 2001: Prentice Hall.

28. Ewing, T.J.A., et al., *DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases.* Journal of Computer-Aided Molecular Design, 2001. **15**(5): p. 411-428.

29. Vaidehi, N., et al., *Prediction of structure and function of G protein-coupled receptors.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(20): p. 12622-12627.