COMPUTATIONAL STUDIES OF ORPHAN G PROTEIN-COUPLED RECEPTORS

Thesis by

Jiyoung Heo

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2007

(Defended October 30, 2006)

Acknowledgments

I am very fortunate to have enjoyed my graduate studies under the wonderful research environment at Caltech. Working with many intelligent and knowledgeable people here has been another great honor to me. First, I would like to thank my advisor Bill Goddard for his support during my PhD studies. He was always willing to discuss my research work with me and to guide my research direction. His long-time passion for science was inspiring. I also thank Mel Simon for suggesting this great project, for helpful discussion and for letting me use his facilities. I thank my committee members, Professor Jack Beauchamp, Doug Rees and Judy Campbell who came across campus on candidacy and proposition exams, discussing my research work and providing valuable feedback.

I am indebted to many people in the Goddard group. I especially thank Nagarajan Vaidehi for managing the project until she left the group. Bill gave me the big picture and insights, and she helped me a lot to start my research project and figure out practical problems. I thank Wely, Rene and Spencer for their past contribution to the methods that I used for my research. I thank John (Wendel), Victor and Pete for their effort in developing methods, Scott at UNC for collaboration on screening, Ravi for helping me to start NAMD runs, and other former and present Biogroup people for valuable discussions.

Three terrific Korean post-docs in the Simon group were essential in my experimental work. I especially thank Sang-kyou for training me to perform biology experiments. I learned most experimental techniques from him. I also thank Jong-Ik and Keum-Joo for additional help and comments in experiments.

I thank my 056C officemates Julius, Sam, Victor, Santiago, Candy and John (Keith) for support and friendship. I also thank WAG Koreans, Seung-Soon, Yunhee, Hyon-Jee and Eun-Jung (a pseudo-WAG Korean). I will not forget the time we chatted in Korean about science and

everything. Thanks to my friends that I met outside the lab, In-Jung and Junghi, I could feel the Korean culture (called “Jung”) that I have missed during my stay in the U.S.

I am so grateful to my parents, my sisters, and my brother who have been always supportive through my whole life. I would like to thank my parents-in-law for taking care of their grandson in Korea. Without their sacrifice, I could not be devoted in my research to finish my PhD studies.

Finally, my deepest gratitude goes to Nam-Joon, my husband, for his steadfast love and support.

Abstract

G protein-coupled receptors (GPCRs) play an essential role in cell communications and sensory functions. Consequently, they are involved in wide variety of diseases and are targets for many drug therapies. Particularly important is the large number of orphan GPCRs, which may play important, albeit unknown, functions in various cells. To understand their respective physiological roles, it is important to identify their endogenous ligands, and to find small molecule ligands that would serve as selective agonists or antagonists. The *mas-related gene G protein-coupled receptors* (Mrg receptors) belong to the orphan GPCR family, which is expressed in a specific subset of sensory neurons known to detect painful stimuli, suggesting that they could be involved in pain sensation or modulation.

The primary focus of this thesis is to predict the 3D structure and binding site of Mrg receptors and to identify novel ligands that would be potential agonists or antagonists. We predict the 3D structure for the mouse MrgC11 (mMrgC11) and the binding site for five chiral FMRF-NH_2 ligands. We correctly predict the relative binding observed for these five ligands. We find that Tyr110 (TM3), Asp161 (TM4), and Asp179 (TM5) are particularly important to binding the ligands. Subsequently, we carry out mutagenesis experiments followed by intracellular calcium release assays that demonstrate the dramatic decrease in activity for the Y110A, D161A, and D179A mutants predicted by our model.

The all-atom molecular dynamics simulation of the mMrgC11/F-(D)M-R-F-NH_2 complex structure in explicit water and infinite lipid membrane system shows that some conformational fluctuations are present, but no significant instability is detected, thus validating our structure prediction method.

The virtual screening with the combination of QSPR and docking methods is carried out for the predicted mMrgC11 receptor. The compounds showing the antagonistic effect are

identified by competitive functional assays. These hit compounds are certainly good staring points in designing better agonists or antagonists.

The binding site of rat MrgA receptor that shows differential binding between adenine and guanine is also predicted. The predicted binding affinity correlates with the availability of the hydrogen bonds to two Asn residues, which would be primary mutation candidates to validate the structure.

Table of Contents

Figures and Tables

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.

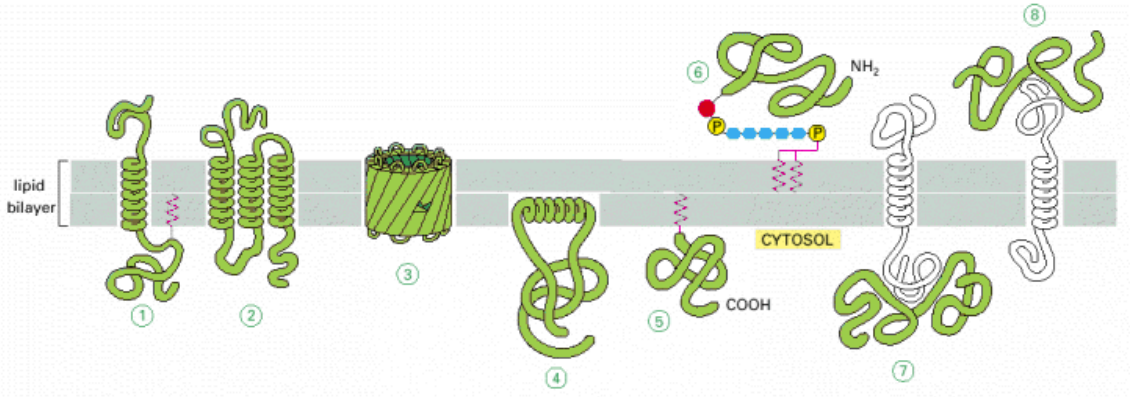


Figure 1.1 Various ways in which membrane proteins associate with the lipid bilayer. Most transmembrane proteins are thought to extend across the bilayer (1) as a single α helix, (2) as multiple α helices, or (3) as a rolled-up β sheet (a β barrel). Some of these "single-pass" and "multipass" proteins 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-

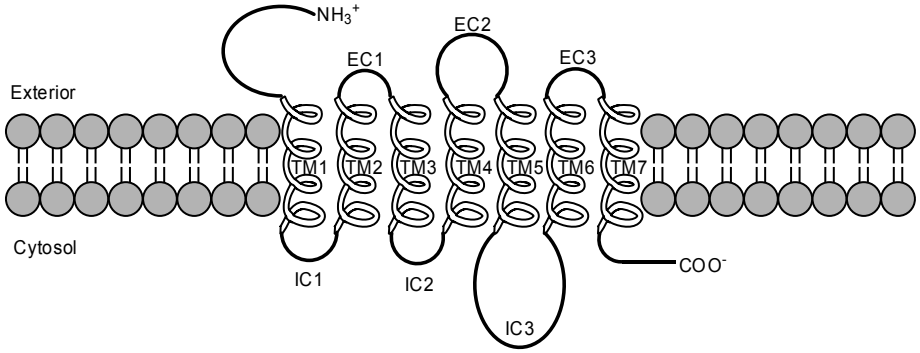


Figure 1.2 Schematic diagram of the general structure of G protein-coupled receptors. All receptors of this type contain seven transmembrane α-helical regions. The loop between α helices 5 and 6, and in some cases the loop between helices 3 and 4, which face the cytosol, are important for interactions 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.

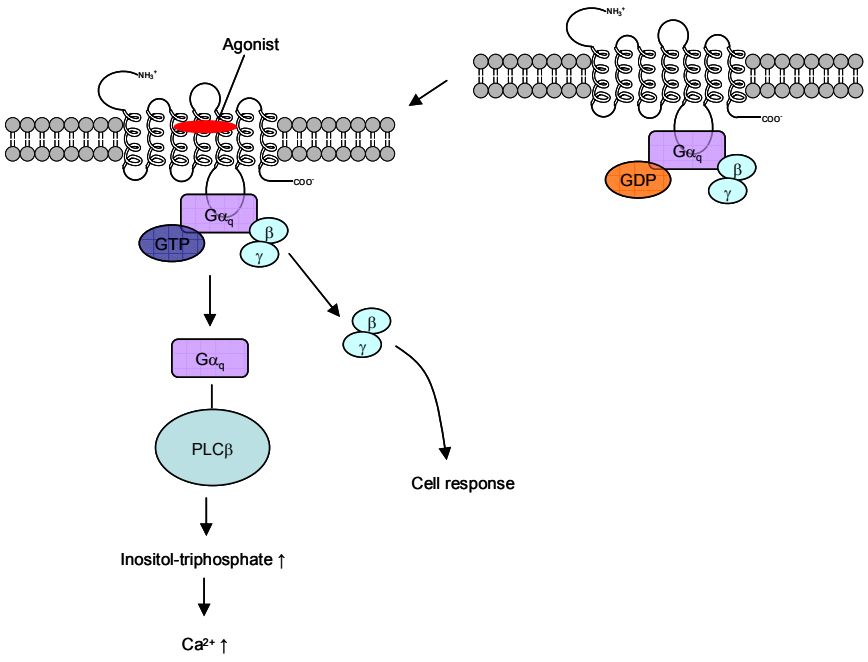


Figure 1.3 Classical examples of GPCR signalling. After agonist binding, a transient high-affinity complex of agonist, activated receptor and G protein is formed. GDP is released from the G protein and is replaced by GTP. This leads to dissociation of the G-protein complexes into a subunits and bg dimers, 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:

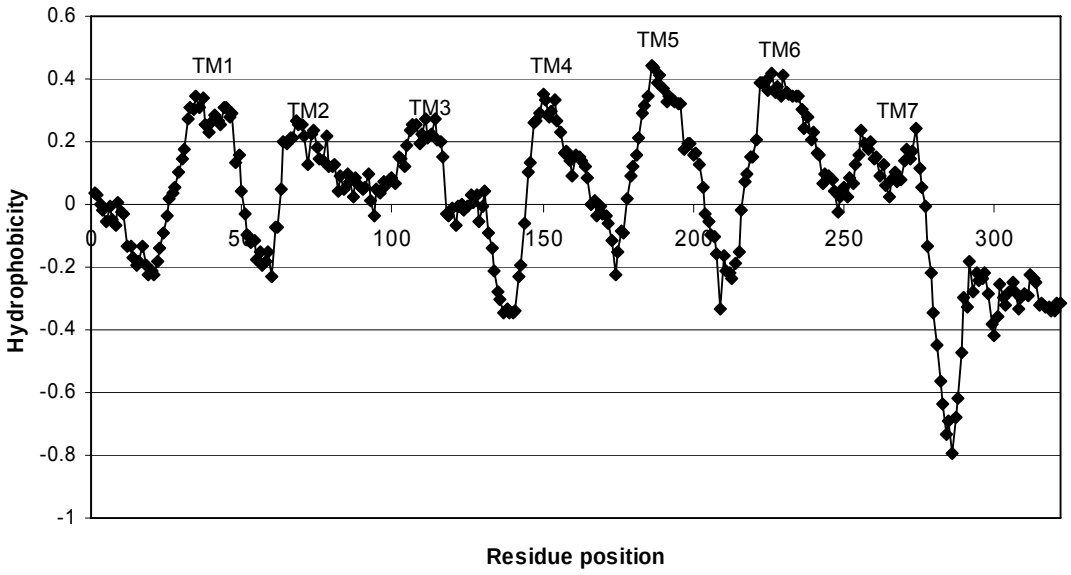


Figure 1.4 Hydrophobicity profile for mouse MrgC11 sequence set (window size = 12)

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

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 $U(R)$:

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

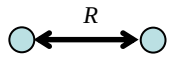
In principle, (1.2) could be solved for the potential energy U, and then (1.3) could be solved. However, the effort required to solve (1.2) is extremely large, so usually an empirical fit to the potential energy surface, commonly called the forcefield (V), is used. Since the nuclei are relatively heavy 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.

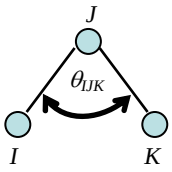
The potential energy is expressed as a sum of valence interaction, nonbonded interaction and additional terms such as constraints. The valence interactions consist of bond stretching (E_{bond}, two-body), bond angle bending (E_{angle}, three-body), dihedral angle torsion ($E_{torsion}$, four-body) and 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



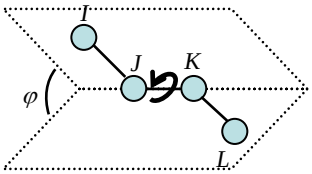
$$E_{bond} = \frac{1}{2} k_e (R - R_e)^2$$

k_e = force constant and R_e = equilibrium bond distance



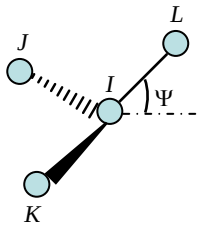
$$E_{angle} = \frac{1}{2} C_{IJK} [\cos\theta_{IJK} - \cos\theta_J^0]^2$$

θ = angle between bonds IJ and JK



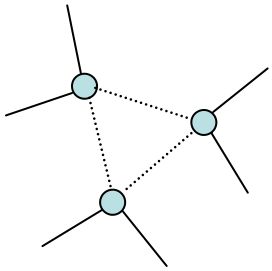
$$E_{torsion} = \frac{1}{2} V_{JK} \{1 - \cos[n_{JK} (\varphi - \varphi_{JK}^0)]\}$$

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



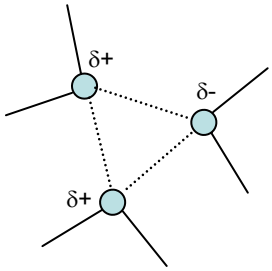
$$E_{inversion} = \frac{1}{2} C_I (\cos\Psi - \cos\Psi_I^0)^2$$

Ψ = angle between the IL bond and the JIK plane



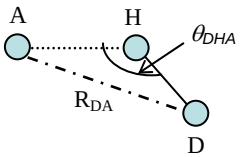
$$E_{vdw} = D_0 [\rho^{-12} - 2\rho^{-6}]$$

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

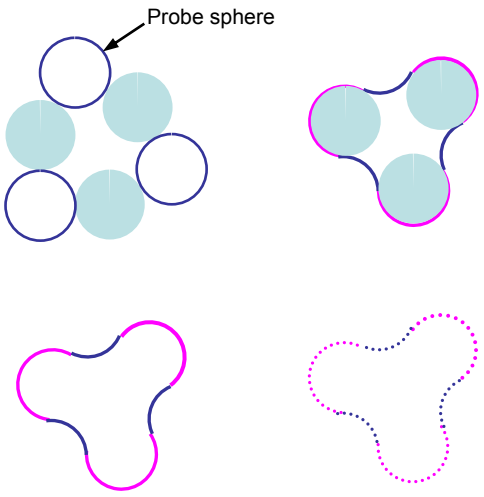


Figure 1.6 Construction of molecular surface in 2D. The filled circles (cyan) correspond to the van der Waals spheres of the atoms. The molecular surface is obtained with a spherical probe and the contact surface is in magenta and the reentrant surface is in blue. Actually the molecular surface is a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

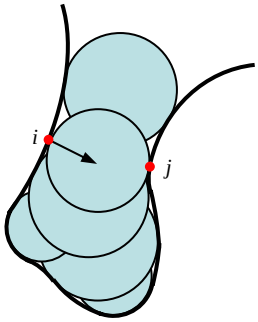


Figure 1.7 A binding site represented as a collection of overlapping spheres. Each sphere touches the molecular surface at two points.

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:

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

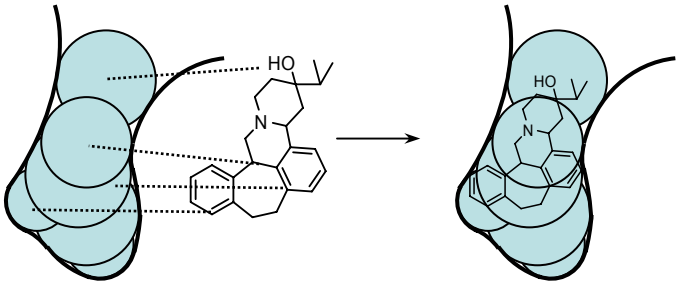


Figure 1.8 Matching algorithm in DOCK. Atoms are matched to spheres centers and then molecule is placed in the binding pocket (Reproduced from [27]).

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

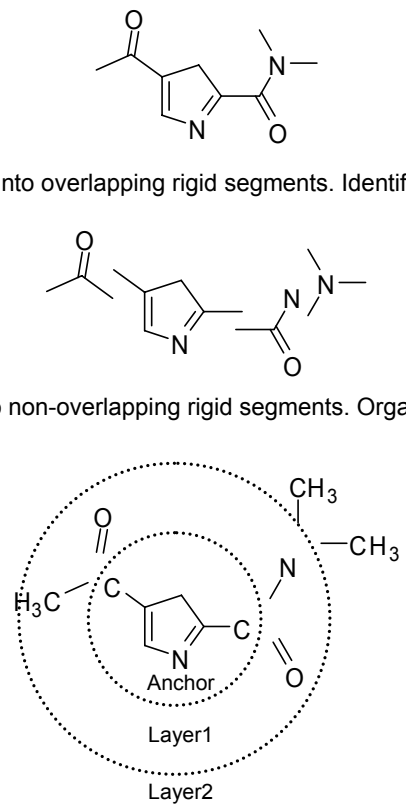


Figure 1.9 Atom pre-organization and anchor selection[27].

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.

Chapter 2

Prediction of the 3D Structure for FMRF-amide Peptides Bound to Mouse MrgC11 Receptor with Subsequent Experimental Verification[1]

2.1 Introduction

G protein-coupled receptors (GPCRs) play an essential role in cell communications and sensory functions as mentioned in chapter 1. Consequently they are involved in wide variety of diseases and are targets for many drug therapies. Particularly important is the large number of orphan GPCRs (for which the native ligands remain unknown), which may play important, albeit unknown, functions in various cells. To understand their respective physiological roles, it is important to identify their endogenous ligands, and to find small molecule ligands that would serve as selective agonists or antagonists. One example here is the family of GPCRs called the Mas-related gene (Mrg) receptor for mouse or the sensory neuron specific receptor (SNSR) in mouse and human[1, 2]. A subset of these receptors including mMrgC11 and mMrgA1 is localized mainly to isolectin $B4^{+}$, the small diameter nociceptors in the dorsal root ganglia (DRG). Dong et al. showed that some of these receptors were activated by RFamide neuropeptides such as NPFF and NPAF and suggested them to be involved in pain sensation or modulation[1]. These Mrg receptors have been paired with structurally diverse transmitter peptides[3].

[1] Portions of this chapter have been submitted to the *Journal of Medical Chemistry* for publication.

Clearly deorphanization would be greatly aided by having three-dimensional (3D) structures of the orphan receptors to help select the most promising new ligands for experimental assays, but it is not yet possible to obtain experimental 3D structures for human GPCRs. Consequently our group developed the MembStruk computational method[4, 5] to predict such structures and we demonstrate in this study that the predicted structures are sufficiently accurate to predict binding sites and relative binding energies. Previously MembStruk was applied to several GPCRs, obtaining ligand binding sites in excellent agreement with experiments. However in these studies the structural data were known prior to our calculations. Although any experimental data were not utilized in making our predictions, such validations are not completely convincing. We undertook this study on mMrgC11 and mMrgA1 receptors for the specific purpose of validating the MembStruk method. Thus prior to our calculations there were no data on how mutations affect binding. In addition the experiments had shown that the F-M-R-F-NH_2, (D)F-M-R-F-NH_2 and F-(D)M-R-F-NH_2 tetrapeptides activate mMrgC11 receptor at ~100 nM concentration, while F-M-(D)R-F-NH_2 and F-M-R-(D)F-NH_2 are inactive (>10 μM). We assumed that explaining such an effect of chirality on binding should provide a strong test of the predicted structures.

2.2 Computational methods

All energy and force calculations were done using DREIDING force field (FF)[6] with the charges from CHARMM22[7] FF and were executed in the molecular dynamics program, MPSIM[8]. The cell multipole method[9] was used for the calculation of nonbond interaction. Unless otherwise specified all simulations were performed in gas phase with the dielectric constant of 2.5.

2.2.1 Structure predictions of the Mrg receptor

The 3D structure of the mMrgC11 and mMrgA1 receptors were predicted independently using MembStruk (version 4.05)[10]. The details of the MembStruk (version 3.5) were described

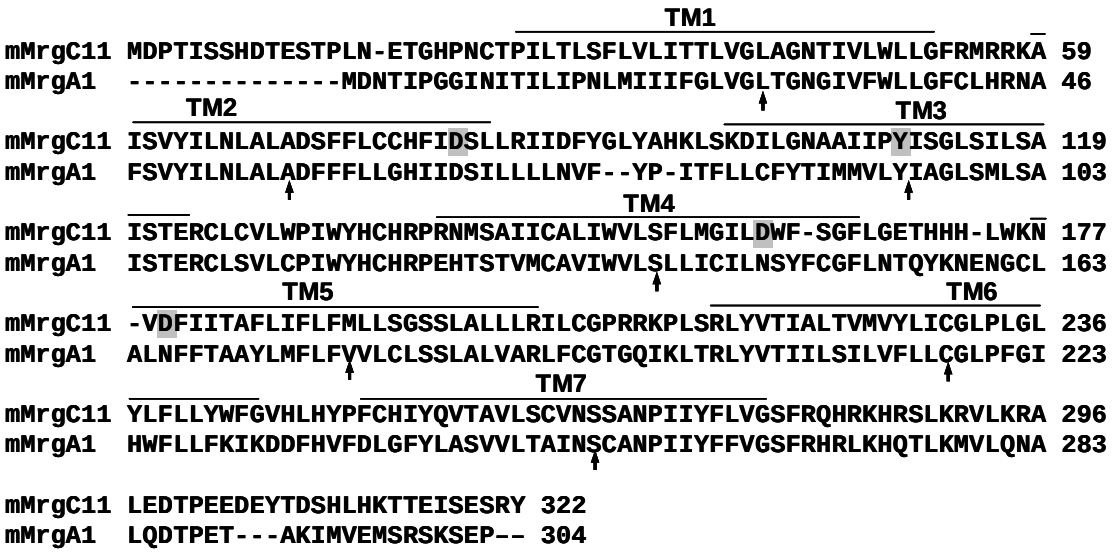


Figure 2.1 Predicted transmembrane (TM) regions. The sequence alignment of mMrgC11 and mMrgA1 is based on the alignment with the entire set of sequences obtained by BLAST search with the mMrgC11 sequence (see Fig. S2.1). The hydrophobic center of each TM is indicated with an arrow. The residues involved in the mutagenesis experiment are highlighted.

in reference 10. Here we outline the procedure, highlighting aspects also relevant to Mrg receptors or that were improved in version 4.05.

Prediction of transmembrane regions

The transmembrane (TM) regions and the hydrophobic maximum for each TM helix were predicted using the TM2ndS method[5]. We used NCBI BLAST[11] to search the non redundant protein database to find sequences homologous to the mMrgC11 receptor with bit scores greater than 200. These 27 sequence hits had sequence identities to mMrgC11 ranging from 41% to 88%. This set of sequences included the mMrgA1 receptor whose sequence identity to mMrgC11 is 44%. Twenty-two of these 27 sequences belong to Mrg receptor family, with remaining 5 corresponding to unnamed GPCRs. We then carried out a multiple sequence alignment with these 27 sequences using ClustalW[12]. These results (Fig. S2.1) were used as input to TM2ndS. The hydrophobicity profile (Fig. S2.2) resulting from TM2ndS had no clear separation between TM2 and TM3, leading to uncertainty in the boundaries between TM2 and TM3. A similar ambiguity

was observed between TM6 and TM7. To eliminate such problems, the MembStruk 4.05 procedure calculates the hydrophobicity profile from a second round of seven TM predictions in which each sequence of the core of the seven TM (15 amino acids around the hydrophobic center) was used as a template. This second set of independent BLAST searches was executed under high gap penalty with each TM core. Here we selected GPCR sequences with sequence identities of >50% (the identity with the entire sequence of mMrgC11 was as low as 23%), see table S1. Then a second round of TM predictions was performed using the multiple sequence alignment of these 7 sets of sequences, see Table S1. The final refined TM region and its hydrophobic center for each of the 7 TM domains were determined from this second round of prediction. For mMrgA1 we used the same TM regions as assigned from alignment with mMrgC11.

Assembly of TM helical bundle

For each TM domain we built canonical α-helices with fully extended conformation of side chains. These were assembled such that the 7 predicted hydrophobic centers are all in the xy plane with the x and y coordinates adapted from the 7.5 Å electron density map of frog rhodopsin[13]. Each helix oriented about its axis so that its hydrophobic moment pointed away from the center of the seven helices (toward the membrane). The tilt of each helix with respect to the z axis and its azimuthal angle were adapted from the 7.5 Å electron density map of frog rhodopsin[13].

Then we carried out 200 ps of molecular dynamics (MD) at 300 K without solvent or lipid, but with charged side chains neutralized by adding Na^+ or Cl^- ions. This allows the conformation of each individual helix to bend or kink as appropriate. We then selected the snapshot with the lowest potential energy from the last 100 ps of the MD trajectory and the net hydrophobic moment was calculated for the middle 15 residues around the hydrophobic center for each helix using this conformation. Each helix was rotated again so that its hydrophobic moment faces toward the membrane. This hydrophobicity-based rotation works well for the six TM helices with

extensive contacts to lipid bilayers. Moreover, since the optimal orientation of a helix depends on the relative orientations of the neighboring helices, we often carry out a combinatorial rotation of the 7 helices. However we found that the structure predicted by the above process placed the highly conserved Asn44-Asp71 pair between TM1 and TM2 and the Asn66-Trp151 pair between TM2 and TM4 close enough to form hydrogen bonds (based only on the coarse hydrophobicity-based rotation step). Therefore we carried out extensive 360º rotational orientation optimization only for TM3, TM5 and TM6. Here the rotational angle of TM3 was scanned for 360 º (in 30 º increments) because TM3 has the least surface area exposed to lipid, but TM5 and 6 were rotated only over the range of -60º to 60º since the orientation had already been optimized roughly using the hydrophobic moment. For every rotation we reassigned the side chain conformation using SCWRL3.0[14] before energy-minimization. The orientation with the best energy was then selected. The results of these scans are shown in Table S2. The rotational orientation of TM7 was scanned over 360º in 5º increments, where for each angle all atoms were optimized. In fact the initial orientation showed the best energy.

Rigid body dynamics in lipid bilayers and addition of loops

Next we added two layers of explicit lipid molecules (52 molecules of dilauroylphosphatidyl choline (DPC) lipid) surrounding the TM bundle. The initial structures for the lipid DPC these were based on the crystal structure in Cambridge Structural Database (ID: LAPETM10). To achieve proper packing of the TM helices, the 7-helix-lipid complex was optimized using rigid body MD for 50 ps where each helix and lipid molecule was treated as a rigid body, with just 6 degrees of freedom (translation and rotation).

The conformation of each TM helix was further optimized in the lipid environment with full atom Cartesian MD simulation for 50 ps while the coordinates of lipid molecules were kept fixed. Then we carried out an additional equilibration of the whole system for 40 ps and selected the structure with the lowest potential energy. For this structure each side chain conformation was

re-assigned using SCWRL and the bundle (helices plus lipid) was minimized to an RMS force of 0.5 (kcal/mol)/Å using conjugate gradients.

The loops were added to the helices using MODELLER6v2[15]. The side chains were re-assigned using SCWRL and subsequently a full atom conjugate gradient minimization of the receptor was performed.

In many GPCRs (including bovine rhodopsin and the catechol amine receptors, such as dopamine and adrenergic receptors) there are conserved cysteines near the top of TM3 and in the second extracellular loop (EC2) that are expected to form a disulfide bond leading to a closed loop. However the mMrgC11 and mMrgA1 receptors do *not* contain such cysteines so the loops were allowed to remain in an open conformation. From five loop structures generated with MODELLER6v2 we selected the one with the lowest internal strain and then optimized the coordinates using annealing MD while keeping the coordinates of TM helices fixed. In this process the system was heated from 50 K to 600 K and cooled down back to 50 K in 50 K steps, with 1 ps of equilibration between the temperature jumps. At the end of the annealing cycle the structure was fully optimized using the conjugate gradients. This final structure shown in Figure 2.3 (top and side views) was used for all docking studies.

2.2.2 Docking predictions with peptide ligands

Using the 3D structure of the mMrgC11 structure we used a refined version (MSCDock) of the docking procedure described in Cho et al.[16]. Since peptide ligands are highly flexible we modified the step in HierDock2.0 (described in Vaidehi et al.[4]), involving scan of the entire receptor with RFa to locate the binding site. This hierarchical docking protocol to predict the binding sites for various ligands as used in this study is described below.

Scanning of the binding sites

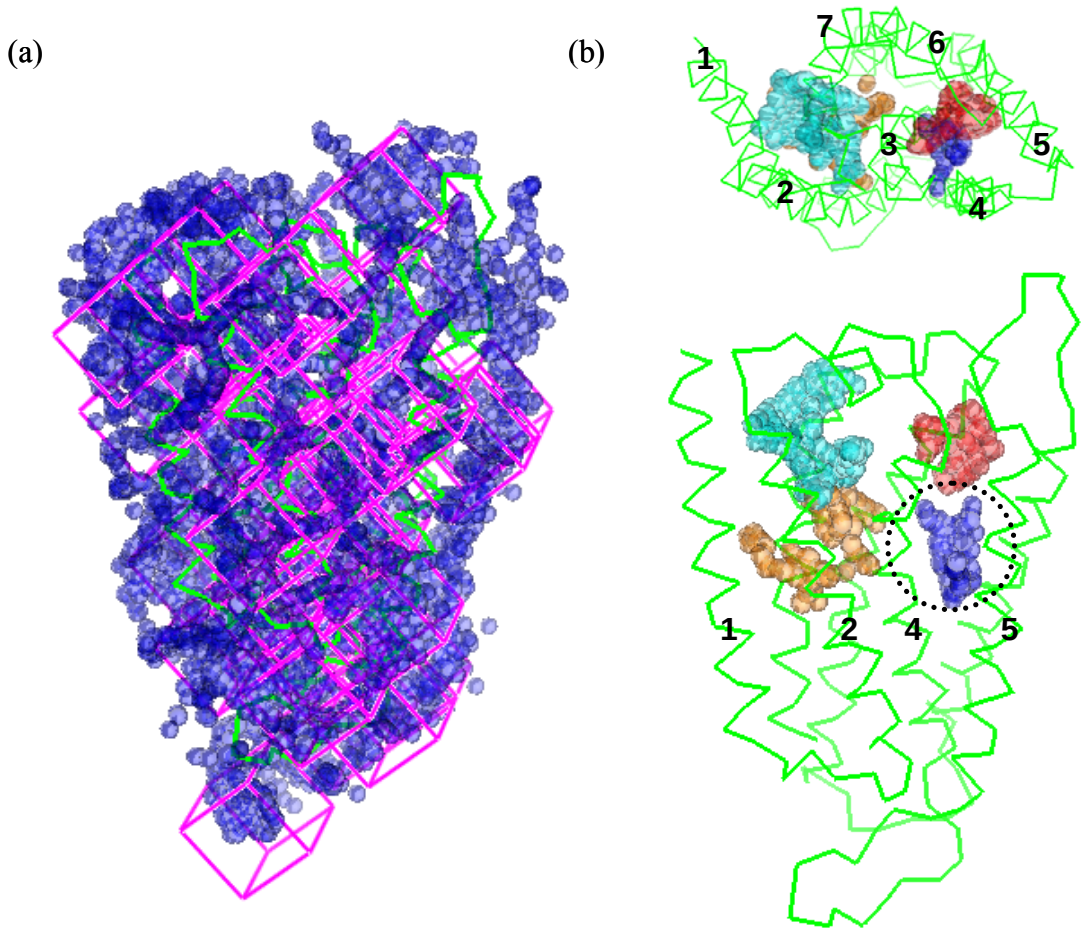


Figure 2.2 The scanning regions used to determine the binding sites for the mMrgC11 receptor. (a) The 9186 Spheres generated with SPHGEN to fill the void spaces of the receptor. The 40 cubic boxes used for docking are shown. (b) The four regions pre-selected for the docking studies. The region enclosed by the dotted circle was identified as the best site.

The entire receptor structure was scanned with the Arg-Phe-NH_2 (RFa) dipeptide known to agonize the receptor (EC50 = 460 nM) to locate the putative binding site. First, the molecular surface was created using the autoMS utility in DOCK4.0[17] with the default values for surface density (3.0dots/$Å^2$) and probe radius (1.4Å). Then we generated spheres from each that filled the void space in the receptor. To do this we used SPHGEN in DOCK4.0. We then constructed a total of 40 cubic boxes (sides of 10Å) and spaced by 8Å that covered this set of spheres. Assuming that the ligand binds inside the TM bundle from the extracellular region, we analyzed these

spheres using a buried surface criterion to pre-select four non-peripheral regions, located on the upper half of the receptor as shown in Figure 2.2. The spheres inside each box were used to define the docking region as input to DOCK4.0.

The RFa ligand was docked independently into each of the four regions as follows. Since the peptide ligands have a significant number of independent dihedral angles (the smallest dipeptide, RFa, contains 10 torsions), we wanted to ensure that this extensive conformation space is sampled in the docking. Thus for each peptide ligand we used the Metropolis Monte Carlo (MC) Method in Cerius2[18] (with a MC temperature of 5000 K in vacuum) to generate a set of 1,000 low energy conformations having a diversity of CRMS = 1.0 Å.

Level0: Then for each of the 1,000 conformers we used DOCK4.0 to generate a set of 3,000 configurations within each of the four binding regions of the receptor. From these we selected the 100 best configurations for each of the 1,000 conformers based on the DOCK score. This led to a total of 100*1000 = 100,000 configurations which were combined together and saved for the next scoring step. In these configurational searches, the rigid ligand and torsion drive options in DOCK4.0 were used. The bump filter option was turned on (maximum bump = 10) and the reduced (to 75%) van der Waals radius was used.

Level1: The configurations from level0 with a ligand-buried surface area below 65% were discarded and the remaining configurations were ranked by the number of hydrogen bonds between receptor and ligand, then by the percentage of buried surface area, and then by DOCK4.0 energy score. This ordered list was trimmed using a diversity criterion of CRMS = 0.6 Å and the top 100 configurations selected. Each of these was minimized in MPSIM using 100 steps of conjugate gradient method, while the receptor coordinates were fixed.

Level2: The 10 best configurations by energy were selected from level1 and the full ligand-protein complex was minimized in MPSIM with 100 steps. The side chain rotamers for all the residues within 5 Å of the ligand were reassigned using the SCREAM side chain replacement

program [Kam, Vaidehi and Goddard unpublished], which uses a side chain rotamer library of 1,478 rotamers with a diversity of 1.0 Å in coordinates.

Level3: The binding energies were then calculated for these 10 optimized ligand-receptor complex configurations. The calculated binding energy (BE) is defined by

BE = E (ligand in fixed protein) – E (ligand in water)

where the E (ligand in fixed protein is the potential energy of the ligand calculated in the ligand-receptor complex with the coordinates of the receptor fixed. This potential energy includes the internal energy of the ligand and the interaction energy of the ligand with the receptor. E (ligand in water) is the potential energy of the free ligand in its docked conformation (snap bind energy) and its solvation energy calculated using the analytical volume generalized born (AVGB) continuum solvation method[19]. In these calculations the dielectric constant was set to 78.2 for the exterior region and to 1.3 for the interior region. The final best ligand-receptor structure was selected as the one with the most negative binding energy.

Among four regions we found that the RFa ligand had the best binding energy in the region involving TM3, 4, 5 and 6 (blue in Fig. 2.2(b)), which we call the putative binding site. The best structure of the RFa-receptor complex was further refined using one cycle of annealing MD heating from 50 K to 600 K and cooling down back to 50 K in 50 K steps, with 1 ps of equilibration between the temperature jumps. Here only the ligand and the residues within 10 Å of the binding pocket (including backbone atoms) were allowed to move during the annealing cycle. At the end of the annealing cycle, the system was minimized to an RMS force of 0.3 (kcal/mol)/Å and the side chains of the residues in the receptor within 4 Å from the ligand was reassigned again with SCREAM. The spheres for docking other peptide ligands were defined with this final optimized RFa-receptor complex.

Docking of other peptide ligands

For the five F-M-R-F tetra-peptide stereoisomers, we first docked the R-F amide part of the C-terminal. This motif is common to most peptide agonists of mMrgC11[20]. Indeed the efficacy results for the five chirally modified F-M-R-F-amides (see Fig. 2.11) show that the chirality of the R-F part dramatically affects activation. Therefore we first docked the three dipeptides: acetylated R-F-NH_2, (D)R-F-NH_2 and R-(D)F-NH_2. Then we used these as an anchor in building the remaining F-M amino acids to construct the docked tetrapeptide.

For docking the three dipeptides, we used the Dock-Diversity Completeness protocol (DDCP) described in Cho et al.[16] to generate a set of diverse configurations and improve completeness in searching the configurations in DOCK (Level0). Briefly, DDCP attempts to generate a complete set of ligand configurations families with a fixed coordinate diversity (1.0 Å). Completeness is defined as the point where the fraction of new configuration that belong to previously generated families to the fraction that leads to a new family is 2.2 (but restricted the list to 5000 families). Then we selected the 50 families with the best energies (by DOCK4.0 energy score) and continued generating configurations while keeping only those that belonged to one of these 50 families until there was an average of six members in each family. Then 50 family heads (best energy in each family) were conjugate gradient minimized (100 steps or 0.1 kcal/mol/Å of RMS force) with the ligand atoms movable and the receptor atoms fixed. Then the 10 best scoring ligands (one from each family by binding energy) were selected for further side chain optimization. Here the binding energy was calculated as the difference between the energy of the ligand in the fixed receptor and the energy of the ligand in solution. The energy of the free ligand was calculated for the docked conformation and its solvation energy was calculated using surface generalized Born model (SGB)[21]. The side chain rotamers of the residues in the receptor within 5 Å of the bound ligand were reassigned by using the SCREAM side chain replacement program. After side chain optimization, the final 10 complex structures were minimized (100 steps or 0.1 kcal/mol/Å of RMS force) with all atoms movable.

The above docking procedure was applied to each of the 1,000 conformers of each peptide ligand generated using the MC Method in Cerius2 (with a MC temperature of 5,000 K in vacuum) using diversity of CRMS = 2.0 Å. Prior to docking the structures of the 1,000 conformers were minimized in gas phase and ordered by energy. Then they were re-clustered with the diversity of 2.0 Å and the conformer of each family head (the best energy among the family) was chosen for docking. This led at least 10 family heads for the R-F dipeptides and over 20 family heads for acetylated R-F dipeptides.

For each such structure the docking process ends up with 10 structures for the ligand/protein complex. Thus we obtained ~100 structures for the dipeptide and ~200 for acetylated peptides. The number of hydrogen bonds (intermolecular between receptor and ligand and intramolecular for a ligand) was calculated for each structure of each ligand/protein complex. This was combined with the binding energy and the number of hydrogen bonds to select the final best structure.

The final structure of ligand-receptor complex obtained from the hierarchical docking procedure was further refined by annealing MD as described in section 2.1.3. Here only the ligand and the side chains of residues within 3.5 Å of the binding pocket were allowed to move. At the end of the annealing cycle, the system was minimized to an RMS force of 0.1 (kcal/mol)/Å.

Building the terminal F-M residues from the bound acetylated R-F-NH_2

The conformations of the terminal F-M residues were sampled using moleculeGL, a recursive, Metropolis Monte Carlo-based rotamer design technique [Kekenes-Huskey, Vaidehi and Goddard in preparation] from the R-F-NH_2 dipeptide docked in mMrgC11 receptor where the extracellular loops were removed. Either the psi angle of Met or the phi angle of Arg is defined as an anchor. We used moleculeGL to generate 1000 structures for the terminal FM, using a diversity of 1.0. Then we selected the lowest energy conformation and minimized the ligand structure (0.3 kcal/mol/Å of RMS force) with the coordinates of receptor fixed. Then the side

chain rotamers of residues within 5 Å of the ligand were assigned using SCREAM and the structure of the whole complex was minimized. The final best structure was refined by annealing as described in previous section.

2.3 Experimental procedures

2.3.1 *In vitro* mutagenesis

The point mutation was incorporated into mMrgC11-GFP coding sequence in pcDNA3.1/Zeo (+) plasmid (Invitrogen) using the QuickChange site-directed mutagenesis kit (Stratagene, La Jolla, CA). The mutagenic oligonucleotide primers were synthesized and purified in the oligonucleotide synthesis center of Caltech. All mutant constructs were verified by DNA sequencing. Later the wild type and mutant gene in pcDNA3.1/Zeo (+) were sub-cloned into pcDNA5/FRT expression vector (Invitrogen) for stably expressing cell lines.

2.3.2 Cell culture and transfection

Flp-In™-293 cells (Invitrogen) were co-transfected with mMrgC11-GFP gene in pcDNA5/FRT vector and pOG44 plasmid (Invitrogen) using FuGENE-6 reagent (Roche Applied Science) according to the manufacturer's instructions. The cells were maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), penicillin/streptomycin and L-glutamine. The cells were split into fresh medium 48 h after transfection and then selected with 400 μg/ml of hygromycin. After two weeks of selection period the hygromycin-resistant clones were picked and then maintained in the selective medium with 200 μg/ml of hygromycin.

2.3.3 Biotinylation and immunoprecipitation

Flp-In™-293 cells stably expressing wild type and mutant receptors were placed into 10 cm culture dish coated with poly-L-lysine and cultured for 24h. The cells were washed twice with ice-cold PBS and incubated with 3 mL of 0.5 mg/mL Sulfo-NHS-LC-Biotin (Pierce

Biotechnology) in PBS supplemented with 0.1 mM HEPES, pH 7.5 at room temperature for 30 min. The biotinylation reaction was quenched by washing cells three times with Tris-buffered saline (10 mM Tris pH 7.5, 154 mM NaCl). The washed cells were incubated with 5 mL of cold lysis buffer (10 mM HEPES, pH 7.4, 1 mM EGTA) supplemented with 100 μM 4-(2-aminoethyl)-benzene sulfonyl fluoride hydrochloride at 4 °C for 15 min. Cells were scraped from the dish and homogenized with a Dounce homogenizer (20-25 strokes with a tight pestle). The cell lysate was centrifuged at 750x g for 10min at 4 °C to remove the nuclei and cell debris. The resulting supernatant was centrifuged at 75,000x g for 30 min at 4 °C. The membrane pellet was solubilized in 500 μL of ice-cold TX/G buffer (300 mM NaCl, 1% TX-100, 10% Glycerol, 1.5 mM $MgCl_2$, 1 mM $CaCl_2$, 50 nM Tris pH 7.4, 0.5 mM PMSF, protease inhibitor cocktail) and incubated with gentle mixing at 4 °C for 1h. Insoluble material was removed by centrifugation at 10,000x g for 15 min at 4 °C. The protein concentration was estimated using the *DC* Protein Assay Kit (Bio-Rad).

The solubilized protein was incubated with 50 μL of streptavidin-agarose (Pierce Biotechnology) overnight at 4 °C on an inversion wheel. The streptavidin-agarose was washed four times with ice-cold TX/G buffer in absence of protease inhibitor and then twice with ice-cold PBS. The precipitates were resuspended with protein sample buffer and then boiled for 15 min. The protein sample was analyzed by SDS-PAGE and transferred to nitrocellulose membrane. The membrane was blocked in Tris-buffered saline with 0.1% Tween 20 containing 5% non-fat milk for 1h. GFP-tagged mMrgC11 receptors were detected by blotting with anti-GFP polyclonal primary antibody (Molecular Probes) in blocking solution followed by anti-rabbit horseradish peroxidase-conjugated secondary antibody and an ECL detection kit (Amersham Biosciences).

2.3.4 Intracellular calcium assay

The cells were placed into 96-well cell culture plate coated with MATRIGEL matrix (BD Biosciences). After 16-24 h, the cells were washed twice with Hank's balanced salt solution

supplemented with 10 mM D-glucose, 20 mM HEPES and 1.6 mM NaOH (assay buffer) and loaded with 2 μM fura-2/AM (Molecular Probes) in assay buffer at room temperature for 20min. Then the cells were washed four times with assay buffer to get rid of the residual fura-2/AM present outsides cell membranes. The fluorometric imaging plate reader (FLIPR) assay was carried out at various concentrations of peptide ligands (1 nM to 10 μM) with the FlexStation II system (Molecular Devices). The fluorescence emitted from the excitation at 340 nm and 380 nm was measured respectively along the time and the ratio of emission at two excitation wavelengths was evaluated together. The difference between maximum and minimum value of the ratio was plotted along with the logarithm of the ligand concentration. The curve was fitted with ORIGIN6.0 software to compute EC_{50} value.

2.4 Results and discussion

2.4.1 Characteristics of the predicted mMrgC11 receptor structure

The predicted TM regions for mMrgC11 are given in Figure 2.1 and the predicted 3D structure of the mMrgC11 receptor is shown in Figure 2.3. TM6 is bent by 28º at Pro233 and TM7 is bent by 15º at Pro271. These two prolines are highly conserved over all family A GPCRs including rhosopsin (in rhodopsin TM 6 and TM7 are bent by 24º and 33º, respectively). Moreover, Pro109 in the middle of TM3 leads to bending of 23º (in rhodopsin TM3 is bent by 13º). We find that these distortions lead to a cavity lined by TM3, TM5, and TM6 that provides the space required for binding our tetrapeptides. The remaining four TMs have relatively straight α-helical conformations.

The predicted 3-D structure of mMrgC11 receptor is superimposed with the 2.2 Å X-ray crystal structure of bovine rhodopsin[22] in Figure 2.3. Here each TM between mMrgC11 and rhodopsin was aligned separately with Clustal-W, imposing a high gap penalty and only the TM

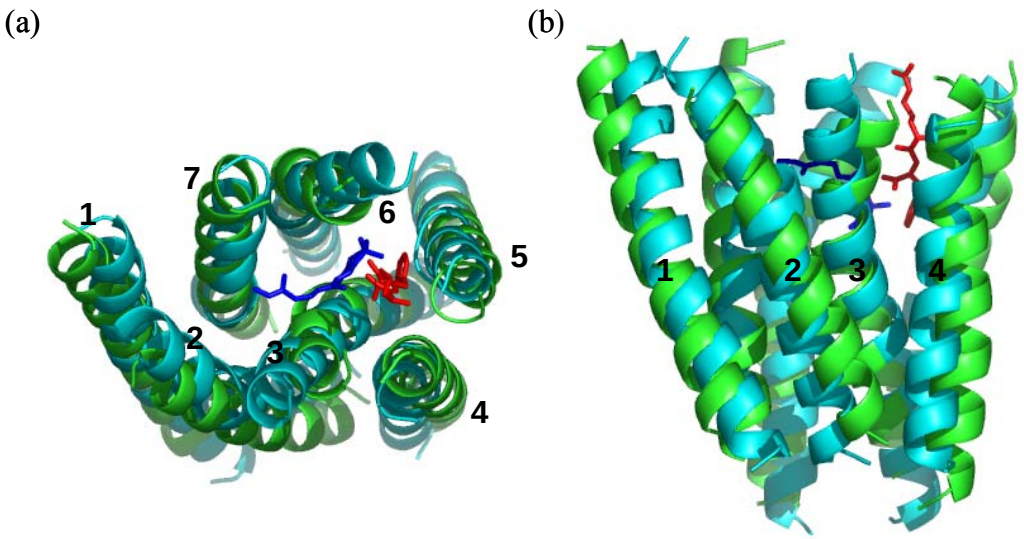


Figure 2.3 Comparison of the predicted 3D structure for the RFa/mMrgC11 complex (green) with the X-ray crystal structure of retinal/rhodopsin (PDB code: 1U19, 2.2 Å resolution). The RFa dipeptide is colored red while the retinal is blue. (a) top view (from extracellular region); (b) side view (with EC at the top). As expected by the low sequence identity (22% for the TM regions) there are significant differences. The CRMS difference in the Cα atoms is 3.75 Å.

regions were fitted with each other for superposition. The sequence identity between the TM regions is ~22%, averaged over the seven TM region sequences. The RMSD in coordinates (CRMSD) of the Cα atoms in the TM regions between bovine rhodopsin and mMrgC11 is 3.75 Å. As expected from the low sequence identity the structures are rather different, but they share such structural features as the kink in the TM6 and TM7 helices. Indeed TM3 of rhodopsin has a slight kink at the two consecutive glycines present at the same position as the proline in mMrgC11.

Several conserved residues participate in the inter-helical hydrogen bonds that maintain the stability of the mMrgC11 receptor structure just as in the rhodopsin crystal structure. Thus Asn44 (TM1) (highly conserved in the family A GPCRs) forms a hydrogen bond with the Ser268 carbonyl group of the backbone in TM7 as shown in Figure 2.4. Asp71 (TM2) forms an interhelical hydrogen bond with this Asn in rhodopsin is in the proximity, but is not in hydrogen bond contact in the mMrgC11 receptor. Such differences are plausible since Miura and Karnik reported TM2 movement from activation in angiotensin II type 1 receptor (using substituted

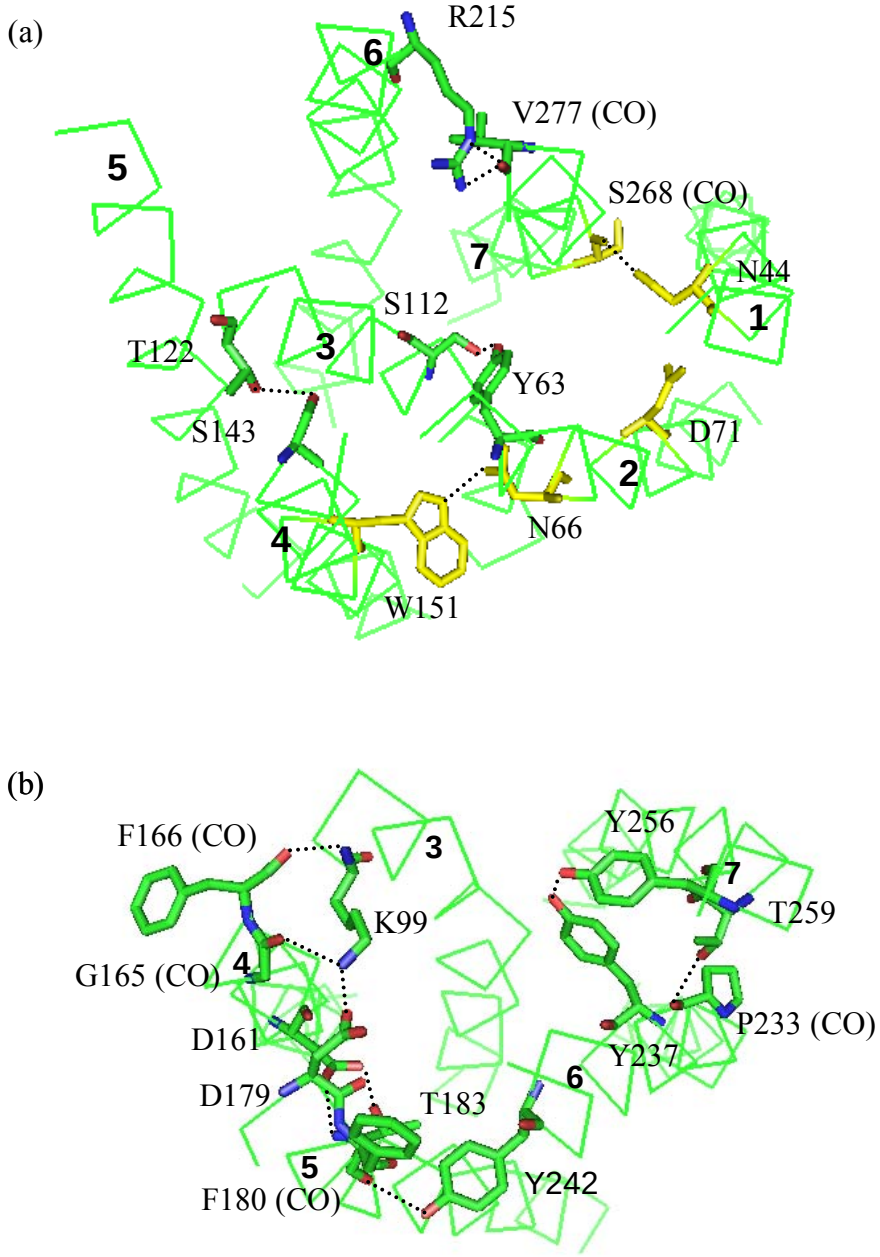


Figure 2.4 Interhelical hydrogen bond networks in the mMrgC11 receptor. The interhelical hydrogen bonds (dashed lines) are specified with residues participating in hydrogen bonds. The highly conserved residues in the family A of GPCRs that form interhelical hydrogen bonds in rhodopsin are colored by yellow. (a) Viewed from the intracellular region. (b) Viewed from the extracellular region. The HBPLUS[23] program was used to calculate hydrogen bonds (maximum D-A distance = 3.9 Å, minimum D-H-A angle = 90.0°).

cystein accessibility mapping)[24]. Thus Asp in TM2 might interact differently, compared to one in the inactive rhodopsin structure. The Asn66 (TM2)–Trp151 (TM4) pair does form a hydrogen bond just like the analogous pair in rhodopsin.

Important points to note in the structure are:

Tyr63 (TM2) (one of residues conserved in the Mrg receptor family (with 39 sequences available on Swiss-Prot and TrEMBL)) participates in hydrogen bonding with Ser112 (TM3) as shown in Figure 2.4.

Another conserved residue, Ser143 (TM4) forms a hydrogen bond with the hydroxyl group in Thr122 (TM3) as shown in Figure 2.4.

Arg215 (TM6) contacts with the backbone carbonyl group of Val277 in TM7, as shown in Figure 2.4.

Asp179 (TM5), which is identified as a key residue for the ligand binding in this study is in contact with Lys99 (TM3) in the apo protein. Asp161 (TM4) also interacts with Thr183 (TM5) in the absence of a ligand.

Several other inter-helical hydrogen bonds are formed with non-conserved hydrophilic residues. Most of these are found in the regions of the TM regions near the intracellular loop. These regions pack more compactly than the near-extracellular regions as appropriate for ligand binding.

No direct contact between TM3 and TM6 or between TM3 and TM7 is found in the TM regions. However, these TM helices interact with each other through well-stacked aromatic rings as shown in Figure 2.5. Tyr110 (TM3), one of the aromatic residues participating in these interactions is conserved through the Mrg receptors (5 of 39 have Phe at this position instead of Tyr). Also Trp265 in TM6 known to be responsible in activating rhodopsin is replaced with Gly

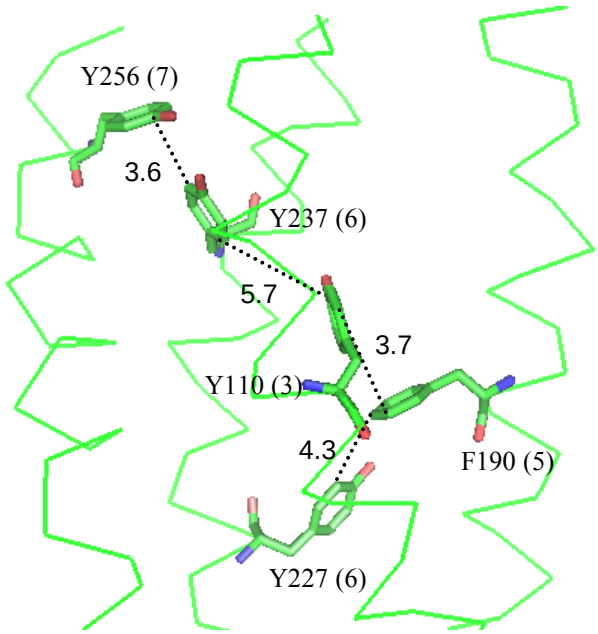


Figure 2.5 Aromatic interactions in TM regions of mMrgC11 receptor. Aromatic residues involved in the π-stacking through TM3, 5, 6 and 7 are shown with the closest C–C distance between two benzyl rings (Å).

in the mMrgC11 receptor. Thus activation in mMrgC11 might involve a different mechanism. As discussed in section 4.2 we find that the agonists to MrgC11 bind in the pocket located between TM3, 4, 5 and 6, which might affect the aromatic–aromatic interactions to help induce activation.

2.4.2 Description of the peptide binding sites

The predicted RFa binding site is located between TM3, TM4, TM5 and TM6 as shown in Figure 2.3. In contrast to 11-cis retinal in rhodopsin, we find that RFa orients vertically in the binding pocket. As seen in Figure 2.3, the aromatic rings stacked between TM3 and TM6 confines the ligand to the region between TM3, TM4, TM5 and TM6. A similar binding orientation has been suggested for the formylated peptide, *f*MLF[25], which binds parallel to the helix in the formyl peptide receptor (FPR). Since RFa is a small peptide ligand (like *f*MLF) it can be placed parallel in the pocket but for longer peptides, the additional amino acids might be kinked towards TM2 and TM7, having contact with these TMs mainly in the loop regions.

Predicted binding site of the dipeptides

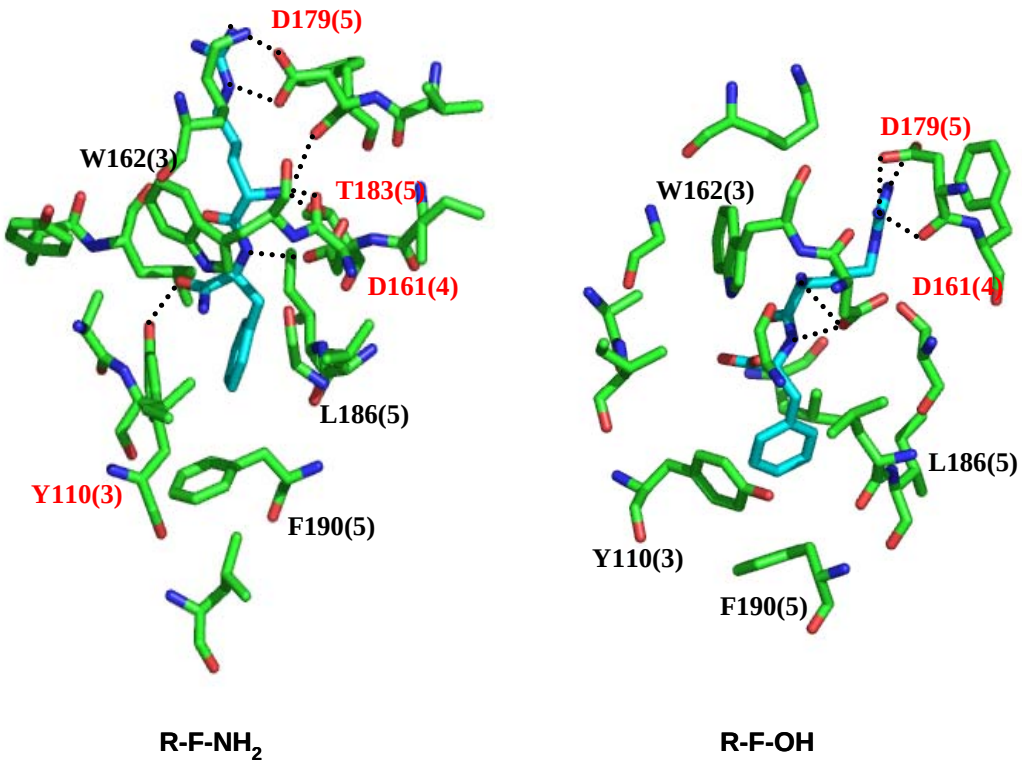


Figure 2.6 Predicted 5 Å binding pocket of the RFa and RF dipeptide agonists. The intermolecular hydrogen bonds calculated with explicit hydrogens using the same criteria as in Figure 2.4 are indicated by the dotted lines. A residue whose side chain participates in the hydrogen bond is specified in red, while one whose backbone is involved is in blue. The residues showing good hydrophobic interactions are specified in black. The top of the picture corresponds to the extracellular regions.

The detailed interactions of bound dipeptides with mMrgC11 receptors are described in Figure 2.6. The binding mode of R-F-OH (RF) is similar to R-F-NH_2 (RFa) although the side chain rotamers of certain residues are different. The common features are that the positively charged moieties are stabilized through the salt bridges and other hydrophilic interactions. Thus the Arg has a good electrostatic interaction with Asp179 (TM5) and the N-terminus has good electrostatic interaction with Asp161 (TM4). The N-terminus of RFa also forms a hydrogen bond with the hydroxyl group of Thr183 (TM5). In addition the C-terminus of RFa makes a hydrogen bond with the hydroxyl group of Tyr110 (TM3).

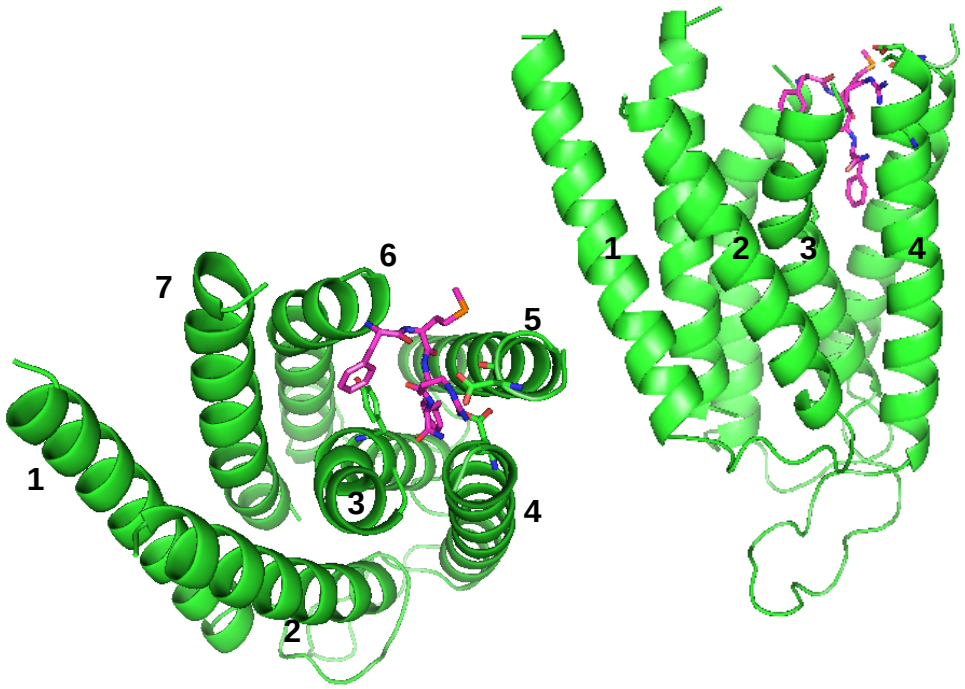


Figure 2.7 Predicted 3D structure for the FMRFa/mMrgC11 complex. The Cα atoms in the TM regions are traced in cartoon while the three key residues (Y110, D161, and D179) are shown in stick. The top view is from extracellular (EC) region and in the side view the EC region is at the top.

The phenyl group of the Phe is stabilized by several aromatic residues present in the binding pocket. Tyr110 interacts most closely with Phe of both dipeptides. For RFa the phenyl ring is in a sandwiched geometry with Tyr110 while for RF these two rings have the displaced T-shape. Phe190 (TM5) also has a good π-π interaction with Phe of the ligand, while Leu186 (TM5) also contributes a good hydrophobic environment for Phe.

Predicted binding sites of the tetrapeptide agonists, F-M-R-F-NH_2, (D)F-M-R-F-NH_2 and F-(D)M-R-F-NH_2

Three tetra-peptides known to be good agonists for mMrgC11[20] were docked into the binding region identified for RFa. The common C-terminal dipeptide part, which is parallel to the average helical axis with the C-terminus of the peptide toward the intracellular region, is bound similarly to RFa (or RF). The extra F-M peptide stretches out horizontally toward TM6 as shown

in Figure 2.7 for the F-(D)M-R-F-NH_2 (FdMRFa) case, where the chirality of Met is modified to be left-handed. In FdMRFa, the amide group of the C-terminus forms hydrogen bonds with the side chain of Asp161 (TM4) and the backbone carbonyl group of Gly158 (TM4). Phe at the C-terminus resides in good aromatic and hydrophobic environment formed by Tyr110 (TM3), Phe190 (TM5) and Leu186 (TM5). Arg is stabilized through the electrostatic interactions with Asp161 (TM4) and Asp179 (TM5). Thr183 (TM5) also interacts with the side chain of Arg. Asp161 (TM4) forms a hydrogen bond with a nitrogen atom of the backbone. Met located in the peripheral region between TM5 and TM6 is nearby such hydrophobic residues as Leu238 (TM6), Phe239 (TM6) and Ile187 (TM5), but has no specific interaction. The N-terminal Phe is sandwiched between Trp162 (TM4) and Tyr237 (TM6), leading to good aromatic interactions. The N-terminus is exposed to the extracellular region. Thus for longer peptide agonists the extra residues might be added starting from this N-terminal position. This might account for the binding of Met-Enk-RF-amide. This is all shown in Figure 2.8.

In F-M-R-F-NH_2 (FMRFa), the overall binding mode is similar to FdMRFa. Some differences are that Thr183 (TM5) no longer participates in the hydrogen bonding with the peptide and the side chain of the right-handed Met is closer to TM5 and interacts at the edge of aromatic ring of Phe180 (S–C distance = 4.0 Å). The preference of S atoms at the edge of aromatic ring has been observed in the study of the non-bond interaction involving sulfur atom of Met by analyzing the protein crystal structures[26].

(D)F-M-R-F-NH_2 (dFMRFa) shows similar interactions. Although the N-terminal Phe has a different chirality from the previous two ligands, it has a similar conformation of the side chain and fits in between Trp162 (TM4) and Tyr237 (TM6). In this case the Met leads to an intra-residue S…O interaction and an inter-residue interaction with Leu240, where the sulfur atom behaves as an electrophile [26].

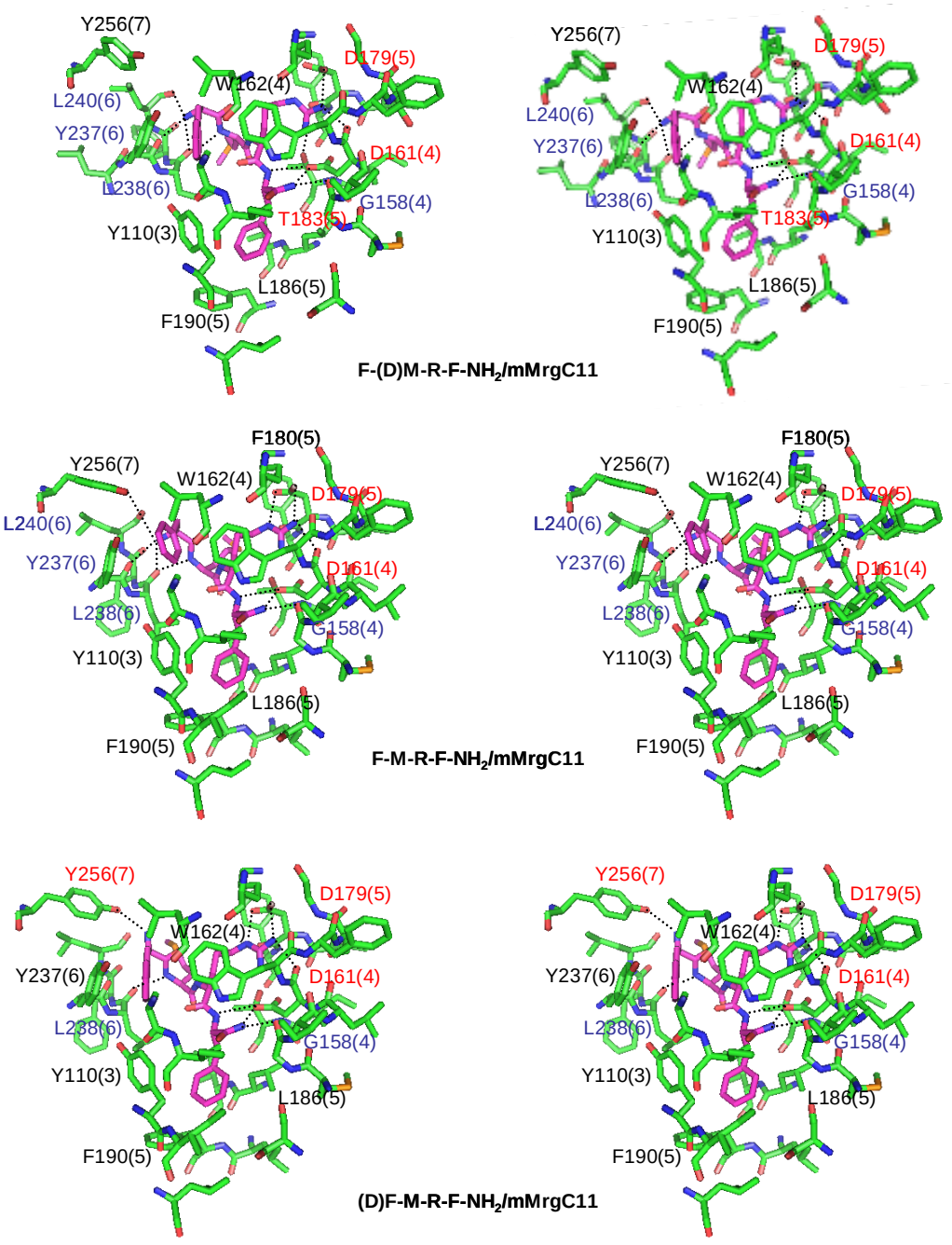


Figure 2.8 Predicted 5 Å binding site to mMrgC11 of the agonist tetra-peptides, F-(D)M-R-F-NH_2, F-M-R-F-NH_2 and (D)F-M-R-F-NH_2.

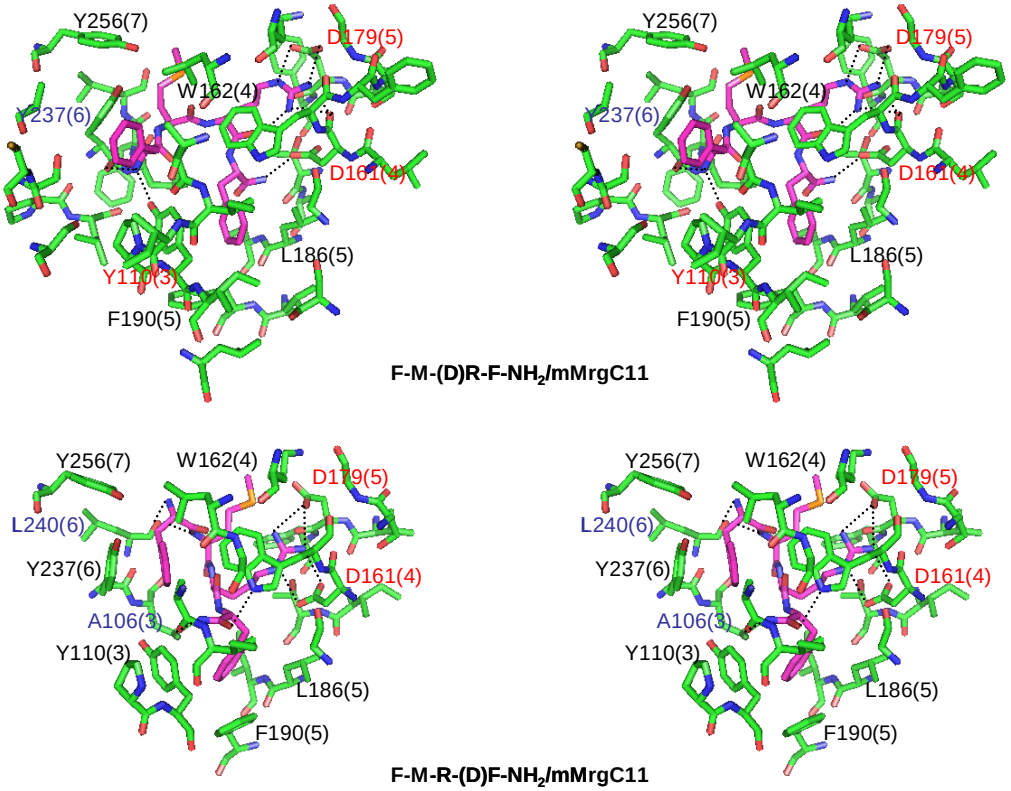


Figure 2.9 Predicted 5 Å binding pocket of the non-agonist tetra-peptides, F-M-(D)R-F-NH_2 and F-M-R-(D)F-NH_2 (neither case was observed experimental to bind even at 30 μM).

We calculate FdMRFa to bind strongest, with FMRFa and dFMRFa having binding energies just 7% and 11% weaker.

Predicted binding sites of the non-agonists, F-M-(D)R-F-NH_2 and F-M-R-(D)F-NH_2

The two other chirally modified FMRFa peptides, FMdRFa and FMRdFa, do not agonize mMrgC11. Our predicted 5 Å binding sites for them are shown in Figure 2.9. In both cases, the C-terminal Phe interacts with Tyr110 (TM3) and Phe190 (TM5) as seen for other agonists.

In F-M-(D)R-F-NH_2 (FMdRFa) the side chain of Arg is located near Asp161 (TM4) and Asp179 (TM5), with good electrostatic interactions. However the contact is less tight and the non-bond interaction energies with Asp161 and Asp179 decrease by 41% and 12% respectively,

compared with FdMRFa. We see the intra-residue S…O interaction for Met in this case. The N-terminal Phe loses π-π interaction with Trp162 (TM4).

In the other non-agonist, F-M-R-(D)F-NH_2 (FMRdFa), the side chain of Arg is between Asp161 (TM4) and Asp179 (TM5) and the interaction is weaker than in FdMRFa. The sulfur of Met shows the interaction with the backbone carbonyl group of Asp179. The N-terminal Phe is sandwiched with Trp162 (TM4) and Tyr237 (TM6). Overall these two non-agonist peptides show the similar binding characteristics to the agonist peptides, but the interaction energy is much weaker by 34% for FMdRFa and by 32% for FMRdFa, compared with FdMRFa.

Summary of binding sites

This study identified several residues critical for peptide binding in mMrgC11. The two aspartic acids, Asp161 (TM4) and Asp179 (TM5) contribute to good electrostatic interactions for the electropositive groups of the ligands; Arg for tetrapeptides and Arg and N-terminus for dipeptides. Several aromatic residues contribute to good π-π interactions. Tyr110 (TM3) and Phe190 (TM5) contact with the common C-terminal Phe of all five agonists. Tyr110 is highly conserved across MRG family of receptors. In the tetrapeptide agonists, the additional phenyl group interacts with Trp162 (TM4) and Tyr237 (TM6). As mentioned previously, these aromatic residues are well stacked in the receptor in the absence of a ligand and provide the interhelical interactions among TM3, TM5, TM6 and TM7. This coupling with two phenyl groups of the tetrapeptide ligand along with the strong electrostatic interaction of Arg with Asp161 (TM4) and Asp179 (TM5) is likely to induce the conformational change responsible for the activation.

2.4.3 Mutagenesis experimental results

Based on the predictions described above, we expect that Tyr110 (TM3) (highly conserved aromatic residue among Mrg family), Asp161 (TM4), and Asp179 (TM5) are all critical to binding. Thus we embarked on a series of mutation experiments to validate theses predictions.

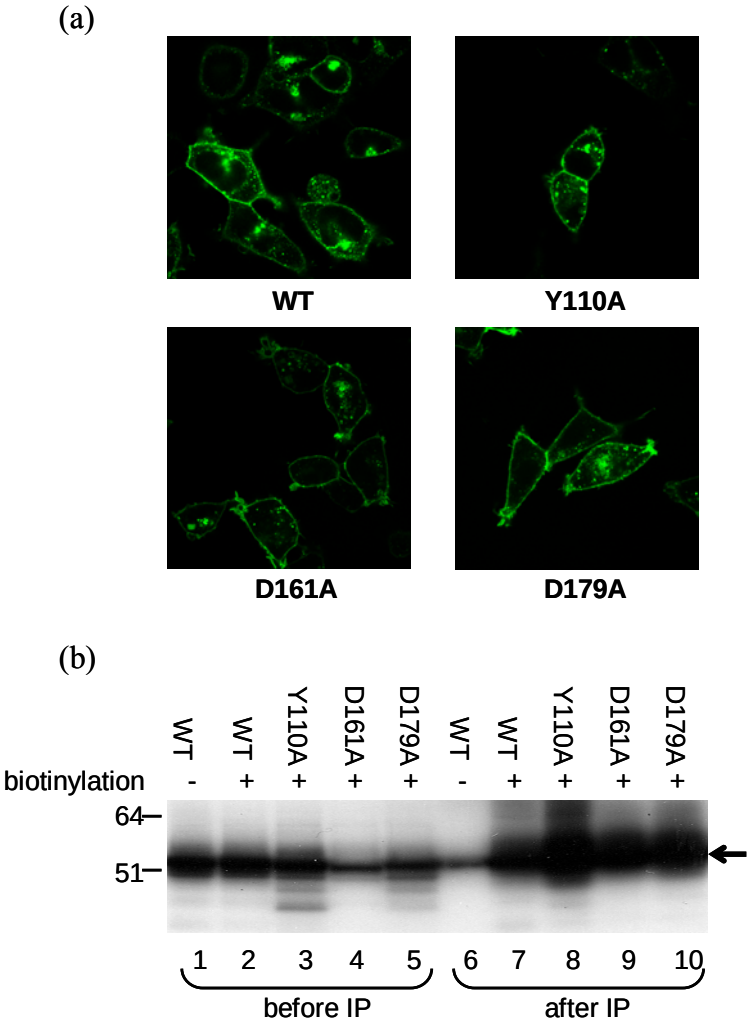


Figure 2.10 The expression of mMrgC11 wild type and mutant receptors in the Flp-In293 cells. (a) GFP images of wild type and mutant cells. The GFP was fused into the C-terminus of the receptor. (b) Biotinylation of the cell surface where receptors are localized and folded. The biotinylated cell extract is blotted with anti-GFP after immunoprecipitation (IP) with streptavidin. Lanes 1-5 are before IP and Lanes 6-10 are after IP. The molecular weight markers are shown on the left in kDa.

Table 2.1 The EC50 values of various peptide ligands determined by the intracellular calcium release assay with Flp-In293 cells expressing mMrgC11 receptor

Peptide	Sequence	EC50, nM	Han et. al[20]
RF	RF	1255 ± 239	632 ± 124
RFa	RF-NH_2	682 ± 186	460 ± 35
FMRF	FMRF	666 ± 228	544 ± 117
FMRFa	FMRF-NH_2	168 ± 26	114 ± 32
dFMRFa	(D)F-M-R-F-NH_2	276 ± 56	108 ± 1
FdMRFa	F-(D)M-R-F-NH_2	113 ± 18	11 ± 4
FMdRFa	F-M-(D)R-F-NH_2	inactive	inactive
FMRdFa	F-M-R-(D)F-NH_2	inactive	inactive
Bam15	VGRPEWWMDYQKRYG	292 ± 19	53 ± 2
γ1-MSH	YVMGHFRWDRF-NH_2	398 ± 189	17 ± 3
γ2-MSH	YVMGHFRWDRFG	340 ± 66	11 ± 5
NPFF	FLFQPQRF-NH_2	358 ± 25	54 ± 5

Inactive means that no activation was detected up to the highest concentration tested, 10μM.
Data represent the mean (±SEM) of four independent experiments.

Expression and localization of mMrgC11 wild type and mutant receptors

Based on the predictions, we carried out three sets of experiments in which key residues were mutated to alanine – Tyr110Ala, Asp161Ala and Asp179Ala. Figure 2.10(a) shows the GFP images for mMrgC11 wild type and for the three mutant receptors. All mutant cells show fluorescence signals as intense as the wild type and the cell boundaries are clearly identified. These images indicate that the mutant receptors are expressed at level similar to the wild type and are well localized at the cell membranes.

To determine whether the mutants properly fold across the cell membrane, we combined immunoprecipitation (IP) experiments with biotinylation. Lanes 1-5 in Figure 2.10(b) show total mMrgC11 receptor proteins including ones that are not biotinylated but present in cytosol and those that have not crossed properly through the membranes. These blots indicate again that all

three mutants are well expressed in the cells, although the expression levels of D161A and D179A mutants seem slightly lower. The results of blots after IP with streptavidin (lanes 6-10) show that the mutant receptors localized on the cell membranes take apical positions at similar amounts to the wild type. Since the band corresponding to the non-specific binding of streptavidin (lane 6) is much weaker, we conclude that the major portions of blots in lane 7-10 come from the biotin-specific binding. This suggests the mutant proteins folds properly on the membranes as well as the wild type protein.

Dose-dependent intracellular calcium release assay with stably expressed MrgC11 receptors

Table 2.1 shows the EC50 values of various peptide ligands determined by intracellular calcium assay experiment with Flp-In293 cells expressing the mMrgC11 receptor. The di- and tetra-peptides and some longer peptide agonists were selected from the ligands previously identified by Han *et al*.[20]. We obtained slightly higher EC50 values in our cellular system, compared to the previous measurements. This difference might result from a variety of sources such as different coupling efficiencies, different expression levels of receptor, and different cellular environment[27]. Nonetheless, the selectivity observed in this study is consistent with the previous results– for example; FMdRFa and FMRdFa still show no activity.

Out of the twelve ligands tested for the wild type receptor, we selected the six most potent ligands to measure the potencies for Y110A, D161A and D179A mutant receptors, as shown in Table 2.2.

We find that the Y110A mutant is not activated by any of the six tested ligands up to a concentration of 33 μM, indicating that Y110 is critical for binding and activation.

The D179A mutants show no potency for the three tetrapeptide ligands, while the other three are activated only under 10 times higher concentration of the ligand.

Table 2.2 Binding constants (EC50 values in nM) of mutant mMrgC11 receptors from intracellular calcium assays

	Binding[a]	Wild Type	Y110A[b]	D161A[b]	D179A[b]	Y110F[c]	Y110W[c]
FdMRFa	100%	113 ± 18	>33000	>33000	>33000	714 (6.3)	334 (3.0)
FMRFa	93%	168 ± 26	>33000	+ (18000)	>33000	1795 (10.7)	1531 (9.1)
dFMRFa	88%	276 ± 56	>33000	>33000	>33000	1500 (5.4)	1513 (5.5)
Bam15		292 ± 19	>33000	>33000	+ (3000)	749 (2.6)	1713 (5.9)
. γ1-MSH		398 ± 189	>33000	>33000	+ (3000)	331 (0.8)	302 (0.8)
. γ2-MSH		340 ± 66	>33000	+ (33000)	+ (2000)	340 (1.0)	349 (1.0)
FMdRFa	67%	>10000					
FMRdFa	68%	>10000					

[a] Calculated binding energy relative to FdMRFa (absolute value = 117kcal/mol). [b] + means that activation starts at a given concentration. [c] Numbers in parentheses are the ratio with respect to the EC50 values of WT.

For mutants D161A we find that 4 of the 6 ligands no longer activate while that other two only activate for 100 times the concentration.

These results that mutation of Tyr110, Asp161 and Asp179 very strongly reduce or eliminate the activity of mMrgC11 receptor validate the predictions that these residues are involved in the ligand binding.

For a positive control experiment, the mutant of Asp81 in TM2 to Ala was transiently expressed in HEK293 cells along with the Y110A, D161A, and D179A mutant receptors also transiently expressed under the same condition. Then the intracellular calcium assay experiment was carried out with 0.33 μM of FMRFa. Except for the D81A mutant, the other three showed no activity.

We investigated the implication of the hydroxyl group on the Tyr110 in ligand recognition by replacing this tyrosine with phenylalanine or tryptophan. The potencies of γ1-MSH and γ2-MSH ligands are not affected by the absence of the hydroxyl group, indicating that the hydroxyl group does not contribute to ligand activation for these ligands. For the three tetrapeptide agonists

the Y110F or Y110W mutations leads to a factor of 5 to 10 reduction in the potency. This is consistent with our predicted structure (Figures 2.7 and 2.8) which does not have the hydroxyl group of Tyr110 interacting with the ligand, but instead forms a hydrogen bond with the carbonyl group of the backbone. The missing hydroxyl group should results in a dangling hydrogen bond donor which might induce an overall conformational change in the binding pocket to explain the loss in activity. Since mutation of Tyr110 to Ala totally extinguishes the activity for all six ligands, we conclude that the aromatic ring must be significant for all six cases.

To investigate whether the two non-agonist tetrapeptides, FMdRFa and FMRdFa are antagonists or weak binders (or non-binders), we saturated the receptors either with FMdRFa or with FMRdFa in three concentrations, 3.3, 16 and 33 μM and then measured the EC50 value for FdMRFa. The intensity of calcium signal remained on the same level as in the absence of FMdRFa or FMRdFa and the EC50 values did not change much (within standard deviation). This result shows that FMdRFa and FMRdFa do not block the efficacy of FdMRFa and at best bind only weakly to the receptor.

Summarizing, the experimental results show that Tyr110 (TM3), Asp161 (TM4) and Asp179 (TM5) are possibly in the binding site in agreement with the predictions. These predicted mutations focused on the dipeptide binding region. Using the binding region for the tetrapeptide, we now suggest that mutations of Trp162 (TM4), Phe190 (TM5), and Tyr237 (TM6) to Ala would also dramatically decrease binding. Additional validations could be to mutate either the receptor or the peptide ligand and to carry out other cell assay experiments such as radiolabelled ligand binding assays. Such studies should further improve our understanding of the structure and ligand binding site.

2.4.4 Prediction of the structure of the mMrgA1 receptor and the binding site for ligands

The 3D structure of mMrgA1 was predicted using MembStruk procedure described in this chapter. The CRMSD of Cα atoms between mMrgC11 and mMrgA1 is 2.49Å in the TM

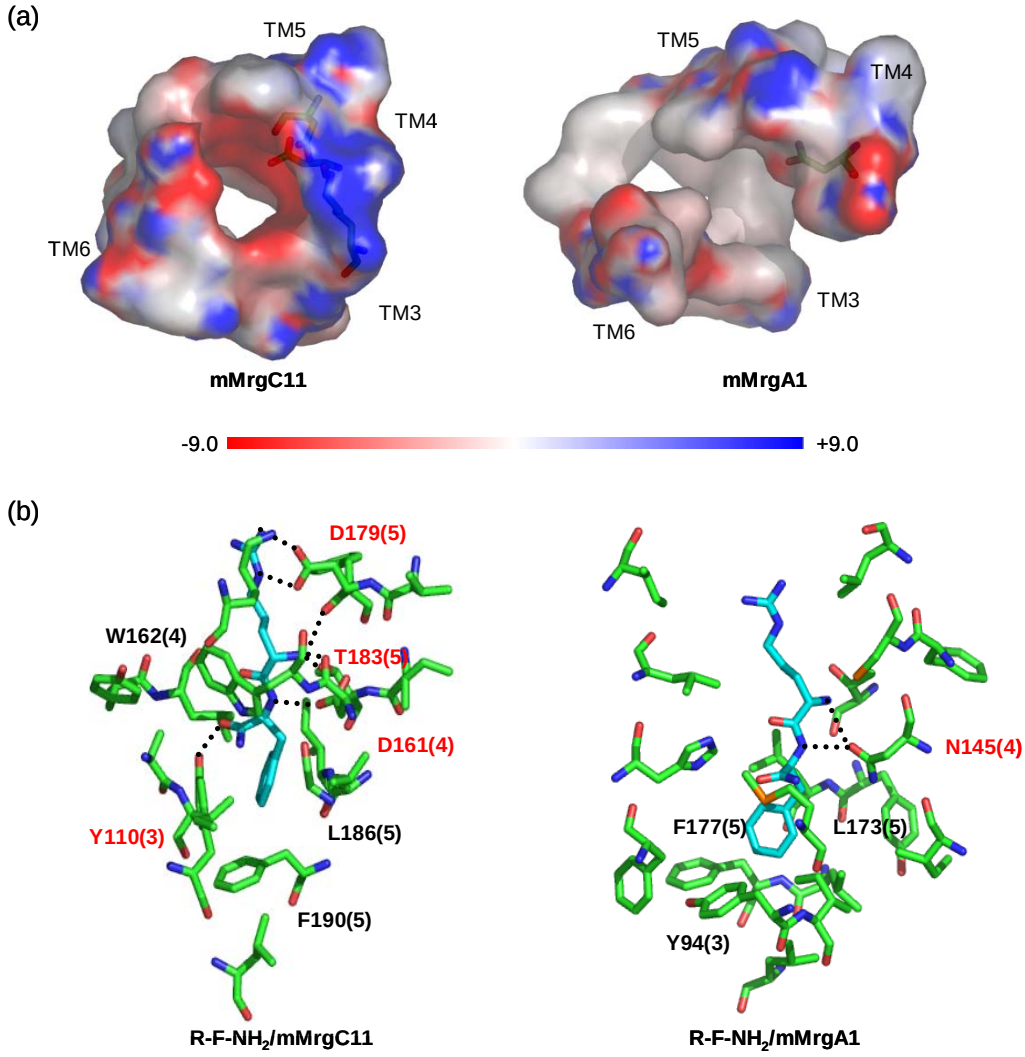


Figure 2.11 Comparison between mMrgC11 and mMrgA binding sites. (a) The electrostatic potential map of the binding pocket in mMrgC11 and mMrgA1. The residues within 5 Å from RFa ligand were selected for visualization. Asp161, Asp179 and Lys99 of mMrgC11 are specified in stick and Asn145 of mMrgA1 in stick. The electrostatic potential was computed using APBS and visualized on PyMOL. The van der Waals radii of DREIDING forcefield were used for APBS calculation. (b) The predicted 5 Å binding pocket of RFa in mMrgC11 and mMrgA1 receptor.

regions and 4.94 Å if the loops are included. The sequence identity between them is 53% for the TM regions and 46% for the entire sequence. It was observed experimentally that mMrgA1 receptor is activated much less potently by tetra-peptide ligands containing the RF-amide motif as compared to the mMrgC11 receptor, and that neither the amide nor acidic form of RF di-peptide activates mMrgA1[20].

We docked the RFa ligand into the mMrgA1 receptor by superimposing it with RFa-bound mMrgC11 receptor. The side chains of residues within 5 Å were reassigned using SCREAM and then the potential energy of the ligand-receptor complex structure was minimized. We found that Tyr94 (TM3), F177 (TM5) and L173 (TM5) (homologous residues of Tyr110, Phe190 and Leu186 in mMrgC11) form a hydrophobic pocket for Phe as in mMrgC11, but they are located slightly farther (the closet C–C distance between aromatic rings is 4 to 5 Å). Asn145 (TM4), the homologous residue of Asp161 in mMrgC11, is involved in the hydrogen bonding with the N-terminus. The Arg side chain of the peptide is surrounded with hydrophobic residues and does not have any favorable interaction with receptor. The calculated binding energy (positive value) predicts that it does not bind to mMrgA1 receptor.

Figure 2.11 shows the electrostatic potential maps of the binding pocket in mMrgC11 and of the corresponding region in mMrgA1. The electrostatic potential was calculated for the entire receptor using adaptive Poisson-Boltzmann solver (APBS)[28]. The binding pocket within 5 Å from RFa docked in mMrgC11 receptor is selectively presented here. We can see that the pocket of mMrgA1 is more hydrophobic than that of mMrgC11. In mMrgC11, two aspartic acids (Asp161 and Asp179) are located in the spot showing the fairly negative potential. We observed that in mMrgC11 the positively charged side chain of Arg and the N-terminus are favored in this region. In the absence of the ligand, Lys in TM3 (Lys99) compensates for this highly negative potential. For mMrgA1 these Asp residues are replaced by Asn. We expect that highly polar ligands such as a peptide containing an Arg residue might be unfavorable for the hydrophobic

character of the pocket in mMrgA1. This might explain why this ligand fails to bind strongly to the mMrgA1 receptor, explaining the low potency for RFa ligand to mMrgA1. This provides additional confirmation of our predicted binding site and protein structure for mMrgC11. We expect that the potency for these ligands to mMrgA1 might increase if these Asn residues are mutated to Asp, an experiment we intend to do soon.

2.4.5 Comparison of Mrg sequences

The 39 verified Mrg sequences were aligned using Clustal-W (v. 1.83) with the default parameters (protein gap open penalty = 10.0, protein gap extension penalty = 0.2, protein matrix = Gonnet) as shown in Figure S2.4. It includes 19 mouse, 13 rat, 1 monkey and 6 human Mrg receptors. The sequence identities range from 21% to 97 %. The mouse MrgF and rat MrgF have the highest sequence identity. The human MrgF also shows the relatively high sequence identity with rat and mouse orthologs (85% and 86% respectively). Across the 39 sequences we examined the sequence variations in the six key residues (Tyr110, Phe190, Asp161, Asp179, Trp162 and Tyr256) that we identified in this study. As mentioned before, Tyr110 is conserved throughout the Mrg sequences except for the 5 Mrgs that have the homologous Phe at the same position. Other five residues show various range of alteration;

- D161: 14 D or E, 7 N, 5 L, 2 A, 3 H, 2 P, 2 T, 1 Q, 1 K, 1 V and 1 Y.
- W162: 6 W, 14 G, 8 S, 4 N, 3 R, 2 A, 1 M and 1 E.
- D179: 14 D, 10 N, 5 I, 3 M, 2 A, 2 H, 2 W and 1 S.
- F190: 20 F, 8 C, 3 T, 2 M, 2 S, 1 A, 1 I, 1 L and 1 V.
- Y256: 20 Y or F, 3 C, 3 D, 3 Q, 2 H, 2 L, 2 N, 2 I, 1 S, and 1 T.

We observed that only rat MrgC has all six residues conserved and mouse MrgB1, mouse MrgB2 and rat MrgB2 have Tyr110, Asp161, Asp179, Phe190 and Y256 at their homologous positions. Trp162 are replaced with Gly for these MrgB receptors. It has been shown that γ2-MSH is the

most potent at activating rat MrgC and the active moiety recognized by rat MrgC is the C-terminal of γ2-MSH, F-R-W-D-R-F-G[29]. The rat MrgC also binds Met-Enkephalin RF-amide with F-M-R-F-NH_2 at the C-terminus. These experimental results suggest the similar characteristics in the binding site of rat MrgC receptor to that of mMrgC11, further supporting our predictions.

2.5 Summary and conclusions

We predicted the 3D structure of the mMrgC11 receptor and used it to predict the binding sites for a number of di- and tetra-peptide ligands. We find that in each case the peptide ligand binds in a pocket among TM3, 4, 5 and 6 oriented parallel to the helical axis. These predictions suggested that three residues (Tyr110 (TM3), Asp161 (TM4) and Asp179 (TM5) in the binding pocket) play a key role in the binding.

To test these predictions, we carried out several mutagenesis experiments. For 6 ligands exhibiting EC50 of 100 to 400 nM in wild type, we find that the EC50 for the Y110A, D161A and D179A mutant receptors are higher than 33 μM for 14 of 18 combinations and 50 to 100 times higher for the other 4 combinations. This validates the implication of these residues for the activation or binding of the ligand.

Since the peptide forms a zwitterion at pH 7 giving it relatively polar character and since the ligands that bind to MrgC11 contain an Arg whose side chain is positively charged at pH 7, it is plausible that the two aspartic acids in the binding pocket participate. On the other hand, mMrgA1 has increased hydrophobic character in the corresponding region (these Asp are replaced by Asn). This might be responsible for the low efficacy of the ligand.

Our predicted binding site also suggests additional mutation candidates to be tested, especially residues involving hydrophobic interaction such as Trp162, Leu186, Phe190, Tyr237, and Leu238.

This study indicates how collaboration between theory and experiment can provide insight into the structural characterization of these Mrg receptors to determine how they are related with function. This could lead to the design of small molecule antagonists to selectively inhibit these receptors as candidate drugs for treating pain. Such studies would be equally valuable for many other GPCR receptors, indicating that a systematic combination of computational tools along with biochemical experiments can provide an increased understanding membrane protein receptors and their activation.

References

1. 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.
2. Lembo, P.M.C., et al., *Proenkephalin A gene products activate a new family of sensory neuron-specific GPCRs.* Nat. Neurosci., 2002. **5**(3): p. 201-209.
3. Civelli, O., *GPCR deorphanizations: the novel, the known and the unexpected transmitters.* Trends Pharmacol. Sci., 2005. **26**(1): p. 15-19.
4. Vaidehi, N., et al., *Prediction of structure and function of G protein-coupled receptors.* Proc. Natl. Acad. Sci. U. S. A., 2002. **99**(20): p. 12622-12627.
5. Trabanino, R.J., et al., *First principles predictions of the structure and function of G-protein-coupled receptors: Validation for bovine rhodopsin.* Biophys. J., 2004. **86**(4): p. 1904-1921.
6. Mayo, S.L., B.D. Olafson, and W.A. Goddard, *Dreiding - a Generic Force-Field for Molecular Simulations.* J. Phys. Chem., 1990. **94**(26): p. 8897-8909.
7. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* J. Phys. Chem. B, 1998. **102**(18): p. 3586-3616.
8. Lim, K.T., et al., *Molecular dynamics for very large systems on massively parallel computers: The MPSim program.* J. Comput. Chem., 1997. **18**(4): p. 501-521.
9. Ding, H.Q., N. Karasawa, and W.A. Goddard, *Atomic Level Simulations on a Million Particles - the Cell Multipole Method for Coulomb and London Nonbond Interactions.* J. Chem. Phys., 1992. **97**(6): p. 4309-4315.
10. Hall, S.E., *Development of a structure prediction method for G-protein coupled receptors*, in *Division of Chemistry and Chemical Engineering*. 2005, California Institute of Technology: Pasadena.

11. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res., 1997. **25**(17): p. 3389-3402.

12. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice.* Nucleic Acids Res., 1994. **22**(22): p. 4673-4680.

13. Unger, V.M., et al., *Arrangement of rhodopsin transmembrane alpha-helices.* Nature, 1997. **389**(6647): p. 203-206.

14. Canutescu, A.A., A.A. Shelenkov, and R.L. Dunbrack, *A graph-theory algorithm for rapid protein side-chain prediction.* Protein Sci., 2003. **12**(9): p. 2001-2014.

15. *MODELLER6v2*, University of California San Francisco: San Francisco.

16. Cho, A.E., et al., *The MPSim-Dock hierarchical docking algorithm: Application to the eight trypsin inhibitor cocrystals.* J. Comput. Chem., 2005. **26**(1): p. 48-71.

17. Ewing, T.J.A. and I.D. Kuntz, *Critical evaluation of search algorithms for automated molecular docking and database screening.* J. Comput. Chem., 1997. **18**(9): p. 1175-1189.

18. *Cerius2 Modeling Environment, Release 4.0*, Accelrys Inc.: San Diego.

19. Zamanakos, G., *A fast and accurate analytical method for the computation of solvent effects in molecular simulations*, in *Division of Physics, Mathematics and Astronomy*. 2002, California Institute of Technology: Pasadena.

20. 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.* Proc. Natl. Acad. Sci. U. S. A., 2002. **99**(23): p. 14740-14745.

21. Ghosh, A., C.S. Rapp, and R.A. Friesner, *Generalized born model based on a surface integral formulation.* J. Phys. Chem. B, 1998. **102**(52): p. 10983-10990.

22. Okada, T., et al., *The retinal conformation and its environment in rhodopsin in light of a new 2.2 angstrom crystal structure.* J. Mol. Biol., 2004. **342**(2): p. 571-583.

23. McDonald, I.K. and J.M. Thornton, *Satisfying Hydrogen-Bonding Potential in Proteins.* J. Mol. Biol., 1994. **238**(5): p. 777-793.

24. Miura, S. and S.S. Karnik, *Constitutive activation of angiotensin II type 1 receptor alters the orientation of transmembrane helix-2.* J. Biol. Chem., 2002. **277**(27): p. 24299-24305.

25. Mills, J.S., et al., *Characterization of the binding site on the formyl peptide receptor using three receptor mutants and analogs of Met-Leu-Phe and Met-Met-Trp-Leu-Leu.* J. Biol. Chem., 2000. **275**(50): p. 39012-39017.

26. Pal, D. and P. Chakrabarti, *Non-hydrogen bond interactions involving the methionine sulfur atom.* J. Biomol. Struct. Dyn., 2001. **19**(1): p. 115-128.

27. Heringdorf, D.M.Z., et al., *Discrimination between plasma membrane and intracellular target sites of sphingosylphosphorylcholine.* Eur. J. Pharmacol., 1998. **354**(1): p. 113-122.

28. Baker, N.A., et al., *Electrostatics of nanosystems: Application to microtubules and the ribosome.* Proc. Natl. Acad. Sci. U. S. A., 2001. **98**(18): p. 10037-10041.

29. Grazzini, E., et al., *Sensory central neuron-specific receptor activation elicits and peripheral nociceptive effects in rats.* Proc. Natl. Acad. Sci. U. S. A., 2004. **101**(18): p. 7175-7180.

Supporting figures and tables

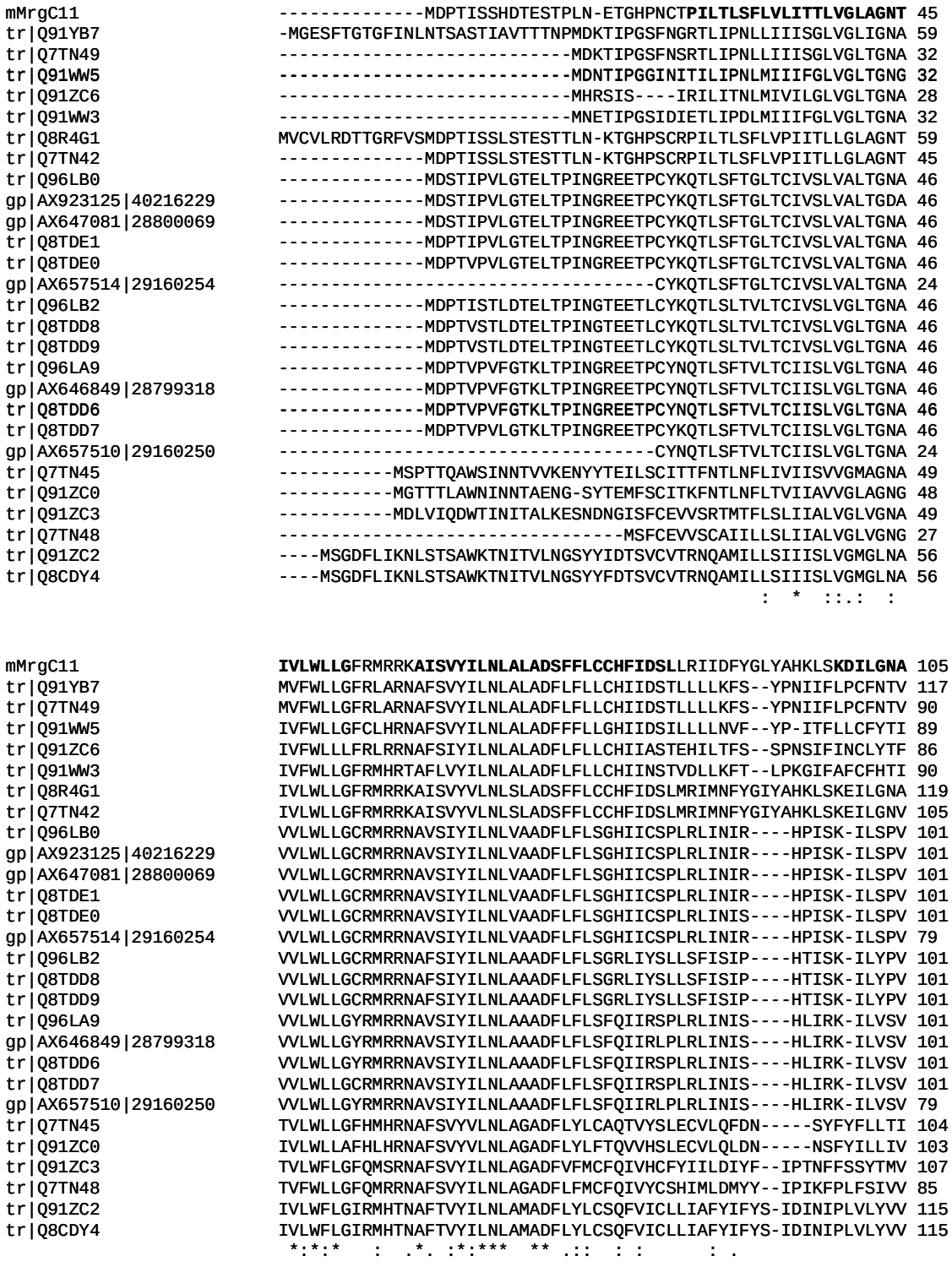


Figure S2.1 Multiple sequence alignment for mMrgC11 with 27 homologous sequences.

```
mMrgC11                 AIIPYISGLSILSAISTERCLCVLWPIWYHCHRPRNMSAIICALIWVLSFLMGILDWF-S 164
tr|Q91YB7               MMVPYIAGLSMLSAISTERCLSVVCPIWYRCRRPKHTSTVMCSAIWVLSLLICILNRYFC 177
tr|Q7TN49               MMVPYIAGLSMLSAISTERCLSVVCPIWYRCRRPKHTSTVMCSAIWVLSLLICILNRYFC 150
tr|Q91WW5               MMVLYIAGLSMLSAISTERCLSVLCPIWYHCHRPEHTSTVMCAVIWVLSLLICILNSYFC 149
tr|Q91ZC6               RVLLYIAGLSMLSAISIERCLSVMCPIWYRCHSPEHTSTVMCAMIWVLSLLLCILYRYFC 146
tr|Q91WW3               KRVLYITGLSMLSAISTERCLSVLCPIWYHCRRPEHTSTVMCAVIWVLSLLICILDGYFC 150
tr|Q8R4G1               AIIPYISGLSILSAISTERCLSVLWPIWYHCHRPRNMSAIICVLIWVLSFLMGILDWFFS 179
tr|Q7TN42               AFIPYISGLSILSAISTERCLSVLWPIWYHCHRPRNMSAIICVLIWVLSFLMGILDWFFS 165
tr|Q96LB0               MTFPYFIGLSMLSAISTERCLSILWPIWYHCRRPRYLSSVMCVLLWALSLLRSILEWMFC 161
gp|AX923125|40216229    MTFPYFIGLSMLSAISTERCLSILWPIWYHCRRPRYLSSVMCVLLWALSLLRSILEWMFC 161
gp|AX647081|28800069    MTFPYFIGLSMLSAISTERCLSILWPIWYHCRRPRYLSSVMCVLLWALSLLRSILEWMFC 161
tr|Q8TDE1               MTFPYFIGLSMLSAISTERCLSILWPIWYHCRRPRYLSSVMCVLLWALSLLRSILEWMFC 161
tr|Q8TDE0               MTFPYFIGLSMLNAISTERCLSILWPIWYHCRRPRYLSSVMCVLLWAPSLLRSILEWMFC 161
gp|AX657514|29160254    MTFPYFIGLSMLSAISTERCLSILWPIWYHCRRPRYLSSVMCVLLWALSLLRSILEWMFC 139
tr|Q96LB2               MMFSYFAGLSFLSAVSTERCLSVLWPIWYRCHRPTHLSAVVCVLLWALSLLRSILEWMLC 161
tr|Q8TDD8               MMFSYFAGLSFLSAVSTERCLSVLWPIWYRCHRPTHLSAVVCVLLWALSLLRSILEWMLC 161
tr|Q8TDD9               MMFSYFAGLSFLSAVSTERCLSVLWPIWYRCHRPTHLSAVVCVLLWALSLLRSILEWMLC 161
tr|Q96LA9               MTFPYFTGLSMLSAISTERCLSVLWPIWYRCRRPTHLSAVVCVLLWGLSLLFSMLEWRFC 161
gp|AX646849|28799318    MTFPYFTGLSMLSAISTERCLSVLWPIWYRCRRPTHLSAVVCVLLWGLSLLFSMLEWRFC 161
tr|Q8TDD6               MTFPYFTGLSMLSAISTERCLSVLWPIWYRCRRPTHLSAVVCVLLWGLSLLFSMLEWRFC 161
tr|Q8TDD7               MTFPYFTGLSMLSAISTERCLSVLWPIWYRCRRPTHLSAVVCVLLWGLSLLFSMLEWRFC 161
gp|AX657510|29160250    MTFPYFTGLSMLSAISTERCLSVLWPIWYRCRRPTHLSAVVCVLLWGLSLLFSMLEWRFC 139
tr|Q7TN45               LMFNYLAGFCMIAAISTERCLSVTWPIWYHCQRPRHTSATVCALFWAFSLLLSLLLGQGC 164
tr|Q91ZC0               TMFAYLAGLCMIAAISAERCLSVMWPIWYHCQRPRHTSAIMCALVWVSSLLLSLVVGLGC 163
tr|Q91ZC3               LNIAYLSGLSILTVISTERFLSVMWPIWYRCQRPRHTSAVICTVLWVLSLVLSLLEGKEC 167
tr|Q7TN48               LNIGYLCGMSILSAISIERCLSVMWPIWYRCQRPRHTSAVICTLLWVLALVWSLIEGKEC 145
tr|Q91ZC2               PIFAYLSGLSILSTISIERCLSVIWPIWYRCKRPRHTSAITCFVLWVMSLLLGLLEGKAC 175
tr|Q8CDY4               PIFAYLSGLSILSTISIERCLSVIWPIWYRCKRPRHTSAITCFVLWVMSLLLGLLEGKAC 175
                          . *: *:.:: .:* ** *.:  ****:*: *   *:  *  .*  :::  ::    .

mMrgC11                 GFLGETHHH-LWKN-VDFIITAFLIFLFMLLSGSSLALLLRILCGPRRKPLSRLYVTIAL 222
tr|Q91YB7               GFLDTKYEKDNRCLASNFFTAACLIFLFVVLCLSSLALLVRSFCGAGRMKLTRLYATIML 237
tr|Q7TN49               GFLDTKYEKDNRCLASNFFTAACLIFLFVVLCLSSLALLVRLFCGAGRMKLTRLYATIML 210
tr|Q91WW5               GFLNTQYKNENGCLALNFFTAAYLMFLFVVLCLSSLALVARLFCGTGQIKLTRLYVTIIL 209
tr|Q91ZC6               GFLDTKYEDDYGCLAMNFLTTAYLMFLFVVLCVSSLALLARLFCGAGRMKLTRLYVTITL 206
tr|Q91WW3               GYLDNHYFNYSVCQAWDIFIGAYLMFLFVVLCLSTLALLARLFCGARNMKFTRLFVTIML 210
tr|Q8R4G1               GFLGETHHH-LWKN-VDFIVTAFLIFLFMLLFGSSLALLVRILCGSRRKPLSRLYVTISL 237
tr|Q7TN42               GFLGETHHH-LWKN-VDFIVTAFLIFLFMLLFGSSLALLVRILCGSRRKPLSRLYVTISL 223
tr|Q96LB0               DFLFSGADS-VWCETSDFITIAWLVFLCVVLCGSSLVLLVRILCGSRKMPLTRLYVTILL 220
gp|AX923125|40216229    DFLFSGADS-VWCETSDFITIAWLVFLCVVLCGSSLVLLVRILCGSRKMPLTRLYVTILL 220
gp|AX647081|28800069    DFLFSGANS-VWCETSDFITIAWLVFLCVVLCGSSLVLLVRILCGSRKMPLTRLYVTILL 220
tr|Q8TDE1               DFLFSGANS-VWCETSDFITIAWLVFLCVVLCGSSLVLLVRILCGSRKMPLTRLYVTILL 220
tr|Q8TDE0               DFLFSGADS-VRCETSDFITIAWLVFLRVVLCGSSLVLLVRILCGSRKMPLTRLYVTILL 220
gp|AX657514|29160254    DFLFSGANS-VWCETSDFITIAWLVFLCVVLCGSSLVLLVRILCGSRKMPLTRLYVTILL 198
tr|Q96LB2               GFLFSGADS-AWCQTSDFITVAWLIFLCVVLCGSSLVLLIRILCGSRKIPLTRLYVTILL 220
tr|Q8TDD8               GFLFSGADS-AWCQTSDFITVAWLIFLCVVLCGSSLVLLIRILCGSRKIPLTRLYVTILL 220
tr|Q8TDD9               GFLFSGADS-AWCQTSDFITVAWLIFLCVVLCGSSLVLLIRILCGSRKIPLTRLYVTILL 220
tr|Q96LA9               DFLFSGADS-SWCETSDFIPVAWLIFLCVVLCVSSLVLLVRILCGSRKMPLTRLYVTILL 220
gp|AX646849|28799318    DFLFSGADS-SWCETSDFIPVAWLIFLCVVLCVSSLVLLVRILCGSRKMPLTRLYVTILL 220
tr|Q8TDD6               DFLFSGADS-SWCETSDFIPVVWLIFLCVVLCVSSLVLLVRILCGSRKMPLTRLYVTILL 220
tr|Q8TDD7               DFLFSGADS-SWCETSDFIPVAWLIFLCVVLCVSSLVLLVRILCGSRKMPLTRLYVTILL 220
gp|AX657510|29160250    DFLFSGADS-SWCETSDFIPVAWLIFLCVVLCVSSLVLLVRILCGSRKMPLTRLYVTILL 198
tr|Q7TN45               GFLFSKFDY-SFCRYCNFIATAFLIVIFMVLFVSSLALLAKIICGSHRIPVTRFYVTIAL 223
tr|Q91ZC0               GFLFSYYDY-YFCITLNFITAAFLIVLSVVLSVSSLALLVKIVWGSHRIPVTRFFVTIAL 222
tr|Q91ZC3               GFLYYTSGP-GLCKTFDLITTAWLIVLFVVLLGSSLALVLTIFCGLHKVPVTRLYVTIVF 226
tr|Q7TN48               GFLFDTNGP-GWCETFDLIATAWLIVLIVVLLGSSLALVINIFCGLYRIPVTRLYVTIVF 204
tr|Q91ZC2               GLLFNSFDS-YWCETFDVITNIWSVVFFGVLCGSSLTLLVRIFCGSQRIPMTRLYVTITL 234
tr|Q8CDY4               GLLFNSFDS-YWCETFDVITNIWSVVFFGVLCGSSLTLLVRIFCGSQRIPMTRLYVTITL 234
                        . *            :.:     :.:  :*  *:*.*:   . *  .  .:*::.** :
```

Figure S2.1 (*continued*)

```
mMrgC11                 TVMVYLICGLPLGLYLFLLYWFGVHLHYPFCHIYQVTAVLSCVNSSANPIIYFLVGSFRQ 282
tr|Q91YB7               TVLVFLLCGLPFGIHWFLLIWIKIDYGKFAYGLYLAALVLTAVNSCANPIIYFFVGSFRH 297
tr|Q7TN49               TVLVFLLCGLPFGIHWFLLIWIKIDYGKFAYGLYLAALVLTAVNSCANPIIYFFVGSFRH 270
tr|Q91WW5               SILVFLLCGLPFGIHWFLLFKIKDDFHVFDLGFYLASVVLTAINSCANPIIYFFVGSFRH 269
tr|Q91ZC6               TLLVFLLCGLPCGFYWFLLSKIKNVFTVFEFSLYLASVVLTAINSCANPIIYFFVGSFRH 266
tr|Q91WW3               TVLVFLLCGLPWGITWFLLFWIAPGVFVLDYS---PLLVLTAINSCANPIIYFFVGSFRQ 267
tr|Q8R4G1               TVMVYLICGLPLGLYLFLLYWFGIHLHYPFCHIYQVTVLLSCVNSSANPIIYFLVGSFRH 297
tr|Q7TN42               TVMVYLICGLPLGLYLFLLYWFGIHLHYPFCHIYQVTVLLSCVNSSANPIIYFLVGSFRH 283
tr|Q96LB0               TVLVFLLCGLPFGIQWALFSRIHLDWKVLFCHVHLVSIFLSALNSSANPIIYFFVGSFRQ 280
gp|AX923125|40216229    TVLVFLLCGLPFGIQWALFSRIHLDWKVLFCHVHLVSIFLSALNSSANPIIYFFVGSFRQ 280
gp|AX647081|28800069    TVLVFLLCGLPFGIQWALFSRIHLDWKVLFCHVHLVSIFLSALNSSANPIIYFFVGSFRQ 280
tr|Q8TDE1               TVLVFLLCGLPFGIQWALFSRIHLDWKVLFCHVHLVSIFLSALNSSANPIIYFFVGSFRQ 280
tr|Q8TDE0               TVLVFLLCGLPFGIQWALFSRIHLDWKVLFCHVHLVSIFLSALNSSANPIIYFFMGSFRQ 280
gp|AX657514|29160254    TVLVFLLCGLPFGIQWALFSRIHLDWKVLFCHVHLVSIFLSALNSSANPIIYFFVGSFRQ 258
tr|Q96LB2               TVLVFLLCGLPFGIQFFLFLWIHVDREVLFCHVHLVSIFLSALNSSANPIIYFFVGSFRQ 280
tr|Q8TDD8               TVLVFLLCGLPFGIQFFLFLWIHVDREVLFCHVHLVSIFLSALNSSANPIIYFFVGSFRQ 280
tr|Q8TDD9               TVLVFLLCGLPFGIQFFLFLWIHVDREVLFCHVHLVSIFLSALNSSANPIIYFFVGSLRQ 280
tr|Q96LA9               TVLVFLLCGLPFGILGALIYRMHLNLEVLYCHVYLVCMSLSSLNSSANPIIYFFVGSFRQ 280
gp|AX646849|28799318    TVLVFLLCGLPFGILGALIYRMHLNLEVLYCHVYLVCMSLSSLNSSANPIIYFFVGSFRQ 280
tr|Q8TDD6               TVLVFLLCGLPFGILGALIYRMHLNLEVLYCHVYLVCMSLSSLNSSANPIIYFFVGSFRQ 280
tr|Q8TDD7               TVLVFLLCGLPFGILGALIYRMHLNLEVLYCHVYLVCMSLSSLNSSANPIIYFFVGSFRQ 280
gp|AX657510|29160250    TVLVFLLCGLPFGILGALIYRMHLNLEVLYCHVYLVCMSLSSLNSSANPIIYFFVGSFRQ 258
tr|Q7TN45               TVLVFIFFGLPIGICVFLLPWIHMMLSSFF---YEMVTLLSCVNSCANPIIYFFVGSIRH 280
tr|Q91ZC0               TVVVFIYFGMPFGICWFLLSRIMEFDSIFFNNVYEIIEFLSCVNSCANPIIYFLVGSIRQ 282
tr|Q91ZC3               TVLVFLIFGLPYGIYWFLLEWIREFHDNKPCGFRNVTIFLSCINSCANPIIYFLVGSIRH 286
tr|Q7TN48               TVLVFLLCGLPYGIYWFLLEWTEKFNYNLPCGFHPVTVLLSCVNSCANPIIYFLVGSIRH 264
tr|Q91ZC2               TVLVFLIFGLPFGIYWILYQWISNFYYVEICNFYLEILFLSCVNSCMNPIIYFLVGSIRH 294
tr|Q8CDY4               TVLVFLIFGLPFGIYWILYQWISNFYYVEICNFYLEILFLSCVNSCMNPIIYFLVGSIRH 294
                        :::*::  *:* *:   *                   *:.:**. ******::**:*:


mMrgC11                 H-RKHRSLKR---VLKRALEDTPEEDEYTDSHL-HKTTEISESRY---------- 322
tr|Q91YB7               --QKHQTLKM---VLQRALQDTPETAEN--------TVEMSSSKVEP-------- 331
tr|Q7TN49               --QKHQTLKM---VLQRALQDTPETAEN--------TVEMSSSKVEP-------- 304
tr|Q91WW5               R-LKHQTLKM---VLQNALQDTPETAKI--------MVEMSRSKSEP-------- 304
tr|Q91ZC6               R-LKHQTLKM---VLQSALQDTPETPEN--------MVEMSRNKAEL-------- 301
tr|Q91WW3               R-LNKQTLKM---VLQKALQDTPETPEN--------MVEMSRNKAEP-------- 302
tr|Q8R4G1               R-KKHRSLKM---VLKRALEETPEEDEYTDSHV-QKPTEISERRC---------- 337
tr|Q7TN42               R-KKHRSLKM---VLKRALEETPEEDEYTDSHV-QKPTEISERRC---------- 323
tr|Q96LB0               R-QNRQNLKL---VLQRALQDTPEVDEGGGWLP-QETLELSGSRLEQ-------- 322
gp|AX923125|40216229    R-QNRQNLKL---VLQRALQDTPEVDEGGGWLP-QETLELSGSRLEQ-------- 322
gp|AX647081|28800069    R-QNRQNLKL---VLQRALQDTPEVDEGGGWLP-QETLELSGSRLEQ-------- 322
tr|Q8TDE1               R-QNRQNLKL---VLQRALQDTPEVDEGGGWLP-QETLELSGSKLEQ-------- 322
tr|Q8TDE0               L-QNRKTLKL---VLQRDLQDTPEVDEGGWWLP-QETLELSGSRLEQ-------- 322
gp|AX657514|29160254    R-QNRQNLKLDSMCRRTALYKTIRSRESYSLSREQQREDPTHDSILS-------- 304
tr|Q96LB2               R-QNRQNLKL---VLQRALQDASEVDEGGGQLP-EEILELSGSRLEQ-------- 322
tr|Q8TDD8               R-QNRQNLKL---VLQRALQDASEVDEGGGQLP-EEILELSGSRLEQ-------- 322
tr|Q8TDD9               R-QNRQNLKL---VLQRALQDTPEVDEGGGWLP-QETLELSGSRLEQ-------- 322
tr|Q96LA9               R-QNRQNLKL---VLQRALQDKPEVDKGEGQLP-EESLELSGSRLGP-------- 322
gp|AX646849|28799318    R-QNRQNLKL---VLQRALQDKPEVDKGEGQLP-EESLELSGSRLGP-------- 322
tr|Q8TDD6               R-QNRQNLKL---VLQRALQDKPEVDKGEGQLP-EESLELSGSKLGP-------- 322
tr|Q8TDD7               R-QNRQNLKL---VLQRALQDKPEVDKGEGQLP-EESLELSGRRLGP-------- 322
gp|AX657510|29160250    R-QNRQNLKL---VLQRALQDKPEVDKASATRS-RTRTTSTSSASTPPRPT---- 304
tr|Q7TN45               HRLQRQTLKL---LLQRAMQDTPEEE-GGERGPSQKSEDLEVVRCSS-------- 323
tr|Q91ZC0               HRLRWQSLKL---LLQRAMQDTPEEE-SGERGPSQRSGELETV------------ 321
tr|Q91ZC3               HRFQRKTLKL---LLQRAMQDSPEEEECGEMGSSRRPREIKTVWKGLRAALIRHK 338
tr|Q7TN48               HRFQRKTLKL---LLQKAMQDTPEEEECGEMGS---------------------- 294
tr|Q91ZC2               RRFRRKTLKL---LLQRAMQDTPEEEQSGNKSSSEHPEELETVQSCS-------- 338
tr|Q8CDY4               RRFRRKTLKL---LLQRAMQDTPEEEQSGNKSSSEHPEELETVQSCS-------- 338
                           . :.**      :  : .  .
```

Figure S2.1 (*continued*)

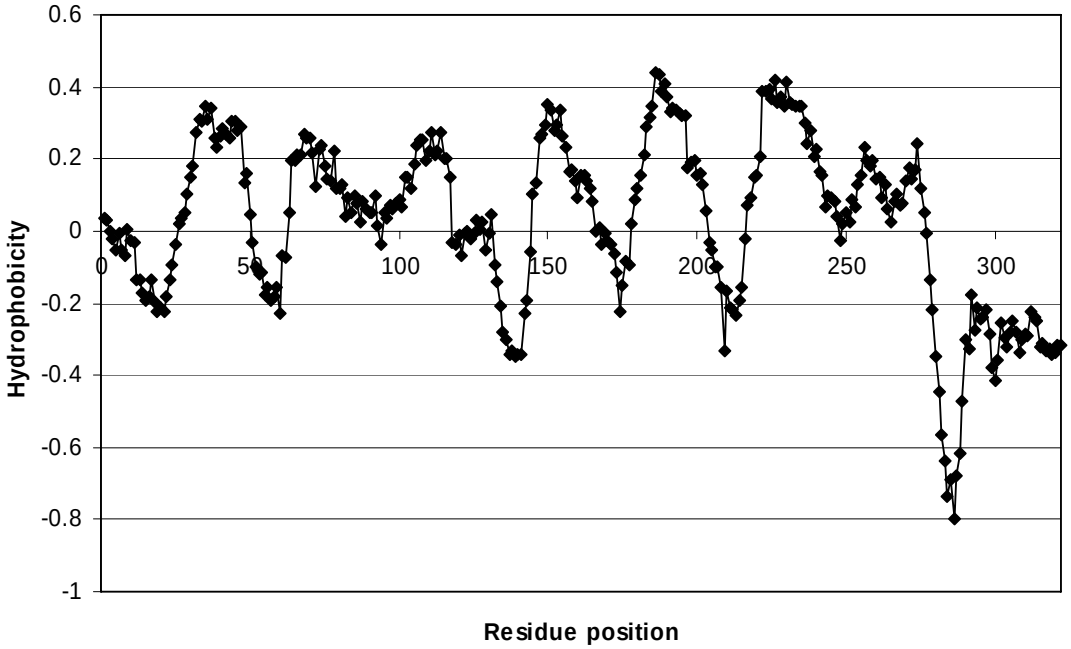


Figure S2.2 Hydrophobicity profile for mMrgC11 sequence set (window size = 12).

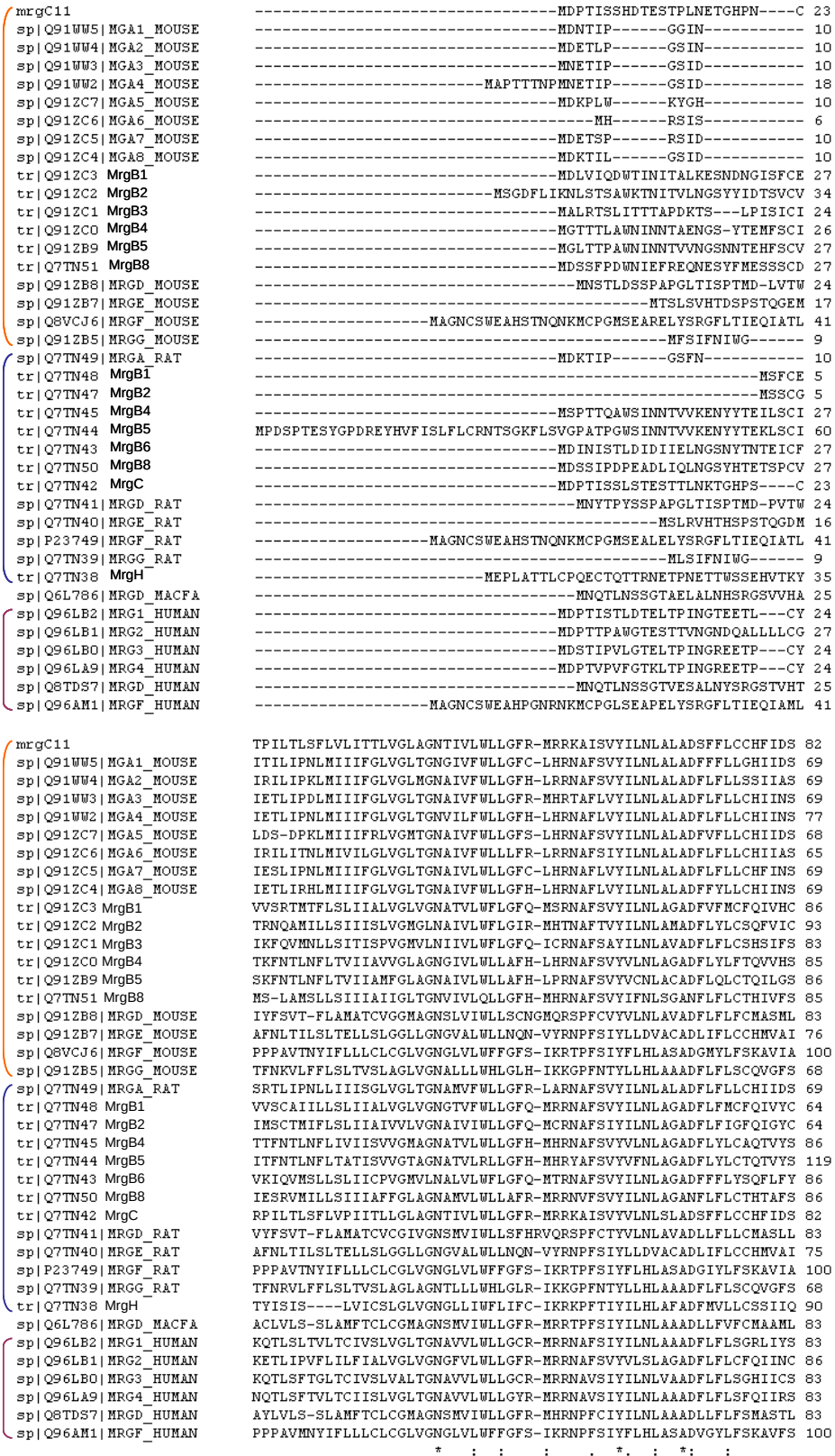


Figure S2.3 Multiple alignment of 39 verified Mrg sequences, including 19 mouse (orange), 13 rat (navy), 1 monkey and 6 human (violet) receptors. The positions of six key residues are specified in red boxes.

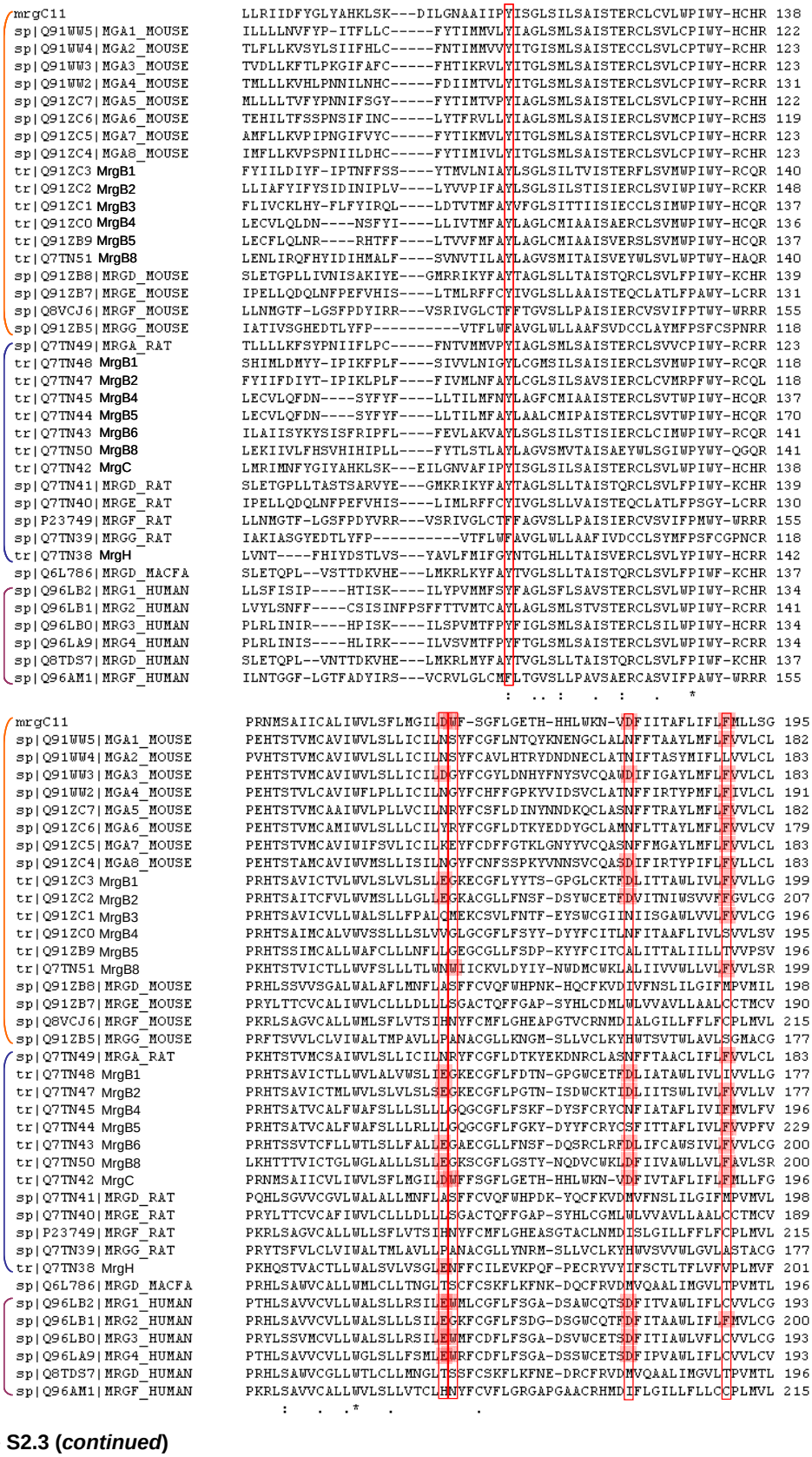


Figure S2.3 (*continued*)

Figure S2.3 (*continued*)

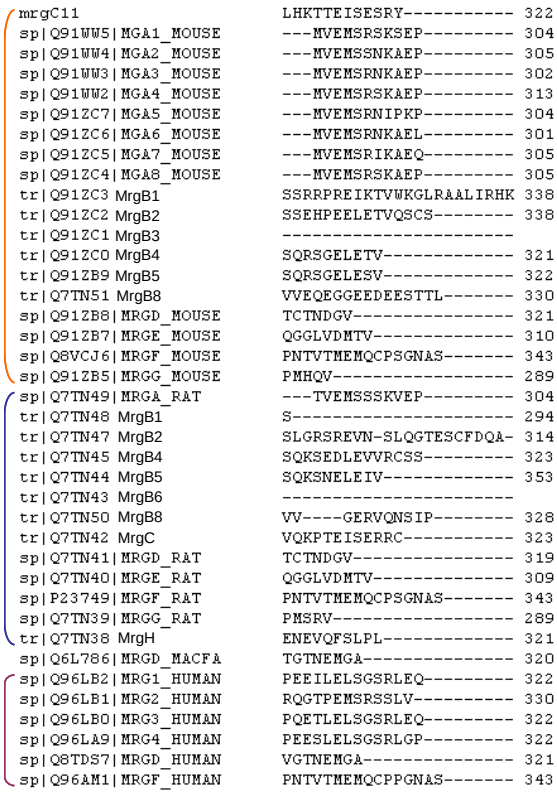


Figure S2.3 (*continued*)

Table S2.1 Hit sequences from independent BLAST search of each TM

Sequence	Identity (%)[a]	TM1	TM2	TM3	TM4	TM5	TM6	TM7
tr|Q91YB7	46	x	x	x	x		x	x
tr|Q7TN49	49	x	x	x	x		x	
tr|Q91WW5	46	x	x	x	x		x	x
tr|Q91ZC6	47	x	x		x	x	x	x
tr|Q91WW3	47	x	x	x			x	
tr|Q8R4G1	88	x	x	x	x	x	x	x
tr|Q7TN42	88	x	x	x	x	x	x	x
tr|Q96LB0	51	x	x	x			x	x
gp|AX923125|40216229	51			x	x		x	x
gp|AX647081|28800069	51	x	x	x	x		x	
tr|Q8TDE1	51	x	x	x	x		x	x
tr|Q8TDE0	49	x	x		x		x	x
gp|AX657514|29160254	46	x	x	x	x		x	x
tr|Q96LB2	53	x			x		x	x
tr|Q8TDD8	53	x			x		x	x
tr|Q8TDD9	53	x			x		x	x
tr|Q96LA9	51	x		x	x		x	x
gp|AX646849|28799318	50	x		x	x		x	x
tr|Q8TDD6	50	x		x	x		x	x
tr|Q8TDD7	50	x		x	x		x	x
gp|AX657510|29160250	50	x		x	x		x	x
tr|Q7TN45	43	x	x			x	x	x
tr|Q91ZC0	42	x						x
tr|Q91ZC3	44	x	x	x	x	x	x	x
tr|Q7TN48	48	x	x				x	x
tr|Q91ZC2	42		x		x		x	x
tr|Q8CDY4	42		x		x		x	x
tr|Q91ZB9	41	x						x
tr|Q96LB1	49	x	x		x	x	x	x

[a] w.r.t. the sequence of mMrgC11

Table S2.1 (*continued*)

Sequence	Identity (%)[a]	TM1	TM2	TM3	TM4	TM5	TM6	TM7
tr|Q91ZC5	45	x	x	x			x	x
tr|Q7TN39	23	x						
tr|Q91ZC4	42	x	x	x	x		x	x
tr|Q91WW4	42	x	x	x	x		x	
tr|AAH64040	40	x	x	x	x		x	
tr|Q8NGK7	37	x					x	
tr|Q8TDS7	37	x						
tr|Q91ZB8	35		x					
tr|Q91ZC7	45		x	x			x	
tr|Q91WW2	42	x	x	x			x	x
tr|Q7TN47	42	x					x	x
tr|Q7TN50	39	x	x					
tr|Q7TN41	34		x					
tr|Q91ZB5	23	x						
tr|Q91ZB7	32		x		x		x	
tr|Q7TN40	30		x					
tr|Q7TN43	44		x	x				
tr|Q91ZC1	36		x					
tr|Q7TN51	34		x		x			
sp|MRG_HUMAN	29		x					
tr|Q7TN44	40		x					
tr|Q8IXE2	31				x			x
tr|Q7TN46	42				x		x	x
tr|Q8N7J6	33				x			x
sp|MRGF_HUMAN	31				x			x
sp|MRGF_RAT	31							x
sp|MRGF_MOUSE	31							x

[a] w.r.t. the sequence of mMrgC11

Table S2.2 Calculated energies (in kcal/mol) of configurations generated in combinatorial rotations of TM3, 5 and 6; the rotational angle of TM3 was scanned for 360 degrees (in 30 degree increments)

TM5	TM6	-150	-120	-90	-60	-30	0	30	60	90	120	150	180
0	**0**	613.5	603.1	627.3	644.4	631.5	575.9	615.7	604.7	656.5	702.2	674.7	645.0
0	**-30**	632.9	585.6	626.5	633.4	652.2	579.3	593.9	593.6	635.3	706.1	683.0	593.0
0	**30**	610.7	597.1	633.8	618.1	642.8	589.1	608.6	593.7	634.9	679.9	670.2	621.6
0	**-60**	627.0	606.7	669.5	641.1	686.5	598.7	638.6	641.9	657.2	742.7	678.5	663.4
0	**60**	609.9	588.9	653.4	626.7	631.6	586.4	694.7	593.8	636.3	676.6	729.2	612.2
-30	**0**	514.4	566.8	599.9	568.0	595.5	555.5	594.9	571.3	570.8	623.7	624.1	616.5
-30	**-30**	552.1	569.9	580.8	557.2	612.1	566.0	567.9	569.3	571.6	604.1	606.5	561.1
-30	**30**	545.4	549.1	605.0	542.8	597.2	569.4	593.8	574.5	581.1	624.9	629.9	577.5
-30	**-60**	555.4	560.5	643.8	552.6	610.8	596.2	593.6	593.3	643.2	686.4	678.4	609.2
-30	**60**	534.2	558.9	610.2	553.3	600.0	547.8	599.5	569.3	560.0	627.8	631.9	580.9
30	**0**	593.3	554.0	616.0	581.1	571.7	551.1	585.6	594.1	607.7	645.8	649.7	598.3
30	**-30**	609.5	538.9	609.6	531.6	551.6	509.1	573.2	600.4	576.0	606.7	607.8	567.2
30	**30**	595.0	539.8	609.3	553.5	555.3	546.1	640.1	597.0	581.9	624.6	618.1	585.3
30	**-60**	841.2	569.8	646.4	596.6	583.8	566.5	618.5	604.0	594.8	650.5	745.2	661.6
30	**60**	624.0	575.4	627.3	555.8	572.1	571.7	583.4	658.1	555.4	635.2	667.7	607.3
-60	**0**	550.2	526.2	546.3	547.7	536.7	478.3	584.3	561.9	583.6	634.3	587.1	565.2
-60	**-30**	565.5	479.0	513.7	544.1	537.9	511.8	540.8	552.3	596.1	580.7	564.4	531.2
-60	**30**	558.0	551.4	532.3	519.8	555.2	488.3	560.7	555.2	557.7	612.9	569.3	543.0
-60	**-60**	557.4	524.0	562.2	553.9	587.8	574.4	615.6	558.8	606.9	609.3	573.2	601.3
-60	**60**	520.5	556.1	545.6	533.6	545.3	536.8	561.1	571.6	588.4	621.0	606.1	586.5
60	**0**	605.5	564.7	632.0	609.2	588.3	551.0	600.1	575.0	604.0	673.8	655.2	634.4
60	**-30**	611.1	542.6	601.8	549.0	569.0	563.0	1082.5	557.3	580.7	747.9	641.6	560.2
60	**30**	614.4	617.6	626.6	545.2	592.0	556.3	698.1	574.1	578.9	693.8	631.5	624.1
60	**-60**	613.3	588.0	664.1	554.4	619.8	587.1	616.1	583.0	591.5	694.0	658.7	634.5
60	**60**	616.3	638.9	681.2	569.1	599.2	607.3	606.9	591.2	590.9	704.2	658.1	623.5

Table S2.3 Calculated binding energy (in kcal/mol) and its component contribution for ligands in mMrgC11; the binding energy of RFa in mMrgA1 is also included on the last row for comparison

Ligand	B.E.	Coulomb	VDW	Hbonds	Desolvation	EC50, nM
FMRFa	-109	-83	-34	-81	90	168 ± 26
dFMRFa	-103	-75	-35	-79	86	276 ± 113
FdMRFa	-117	-90	-33	-82	88	113 ± 37
FMdRFa	-78	-46	-51	-56	75	inactive
FMRdFa	-80	-60	-43	-61	84	inactive
acetylated RFa	-97	-67	-19	-45	34	
acetylated dRFa	-82	-48	-25	-40	31	
acetylated RdFa	-75	-49	-27	-29	30	
RF	-71	-80	-19	-47	75	1255 ± 478
RFa	-74	-86	-19	-62	94	682 ± 371
RFa/mMrgA1	58	-0.03	-15	-20	93	inactive

Chapter 3

Molecular Dynamics Simulation of Mouse MrgC11 Receptor with Bound F-(D)M-R-F-NH_2 in Explicit Lipid/Water Environment

3.1 Introduction

GPCR belongs to one of the membrane protein families embedded into the lipid bilayers, while the intra- and extracellular regions are exposed to the aqueous media. The membrane environment influences the function of membrane proteins, through electrostatic and steric interaction as well as through the membrane's internal pressure. Therefore the proper environment should be taken into account in the molecular simulation. However the resulting calculation, incorporating proteins, lipid bilayers, water molecules and ions needs to handle with 50,000 atoms even for the small proteins and this large simulation size poses a major computational challenge. Thanks to advances in computing power and availability of an efficient parallel molecular dynamics (MD) code, computational biologists have succeeded in performing the required calculations. Recently an all-atom molecular dynamics simulation of a complete virus system composed of 1 million atoms was presented by the Schulten group in the University of Illinois at Urbana-Champaign, using a parallel molecular dynamics program NAMD[1].

In chapter 2, in order to reduce the computation cost, the minimally required molecular components were considered in predicting the protein structure and the binding site. However our predicted mMrgC11 receptor structure was sufficiently accurate to identify binding sites for selective ligands, i.e. chirally modified tetrapeptides of F-M-R-F-NH_2. Therefore our structure prediction and docking methods might be good enough to predict the interaction between ligand

and GPCR. Nevertheless, it is worth performing MD studies for the mMrgC11/ligand complex structure in more realistic environments. These could provide the validation for our predicted structure and also information about the dynamic behavior, which might lead to understanding the role of conformational change on receptor activation.

Here we have carried out the all-atom MD simulation for mMrgC11/F-(D)M-R-F-NH_2 complex structure in explicit lipid and water environments, using NAMD 2.5 program[2]. In the following sections the detailed simulation procedure and the structural characteristics observed in a 7ns simulation run are described, focusing on the behavior of the ligand in the binding pocket and the conformational change on the transmembrane (TM) domains.

3.2 Simulation procedure

3.2.1 Setup of lipid and water environment

A molecular graphics program, Visual Molecular Dynamics (VMD) was used for the simulation setup. The Biograf file of the final optimized mMrgC11/F-(D)M-R-F-NH_2 complex structure was split into separate ligand and the receptor files. The hydrogen atoms were removed and the structure files were converted into PDB format compatible in VMD. The hydrogen atoms were then re-assigned with the estimated coordinates based on entries of internal coordinates present in the CHARMM topology dictionary. The N-terminus was acetylated (residue name: ACP in the CHARMM topology dictionary) and the C-terminus was capped with the N-methylamide group (residue name: CT3). The PDB and PSF files for the receptor and the ligand were then combined, generating a single PDB and PSF file respectively.

The complex structure was replaced for the mid-plan perpendicular to the TM helical axis to be positioned at z = 0. The equation of the mid-plane (Ax+By+Cz+D = 0) was calculated for the receptor using MembComp program[3]. Briefly, the hydrophobic center which showed the maximum hydrophobicity on the hydrophobicity profile was previously determined for each TM

helix. The plane of intersection was aligned to these seven points utilizing a least square approach. The origin of the plane was the geometric center of the centers defined for each helix. With this equation, the coordinates of the complex structure were transformed. Here the plane was moved to z = 0 and the vector normal to the plane became the z-axis. Also the origin of the plane was set to the geometric center of TM α-helices.

Next, the complex structure was then superimposed on the 75 Å x 75 Å slab of a solvated palmitoyl-oleoyl-phosphatidylcholine (POPC) lipid bilayer patch and the lipids and water molecules overlapping with the protein were removed (POPC within 1 Å and waters within 5 Å of the protein). The system was fully solvated with water by adding a ~30 Å thick slab from an equilibrated water box. The VMD autoionize plugin was then used to randomly place the ions necessary to neutralize the system. The resulting system was composed of 47,651 atoms; 4,180 receptor atoms, 74 ligand atoms, 4,288 lipid atoms, 39,087 water atoms and 10 chlorine atoms.

3.2.2 Molecular dynamics simulations

All simulations were performed with the parallel molecular dynamics code NAMD 2.5[2] using the CHARMM22 force field[4, 5] for proteins, the CHARMM27 parameters for the lipids and the TIP3P water model[6]. The simulated system was kept at constant temperature of 310 K by using Langevin dynamics for all non-hydrogen atoms, with a Langevin damping coefficient of 1 ps^{-1}. A constant pressure of 1 atm was maintained by using the Langevin piston method with a period of 200 fs and decay timescale of 200 fs.

Simulation was carried out with an integration time step of 1 fs. The bonded interaction was computed every time step; short-range nonbonded interaction every two time steps; and long-range electrostatic interaction every four time steps. A cutoff of 12 Å was used for van der Waals and short-range electrostatic interactions and a switching function started at 10 Å for van der Waals interactions to ensure a smooth cutoff. The simulation was performed under periodic

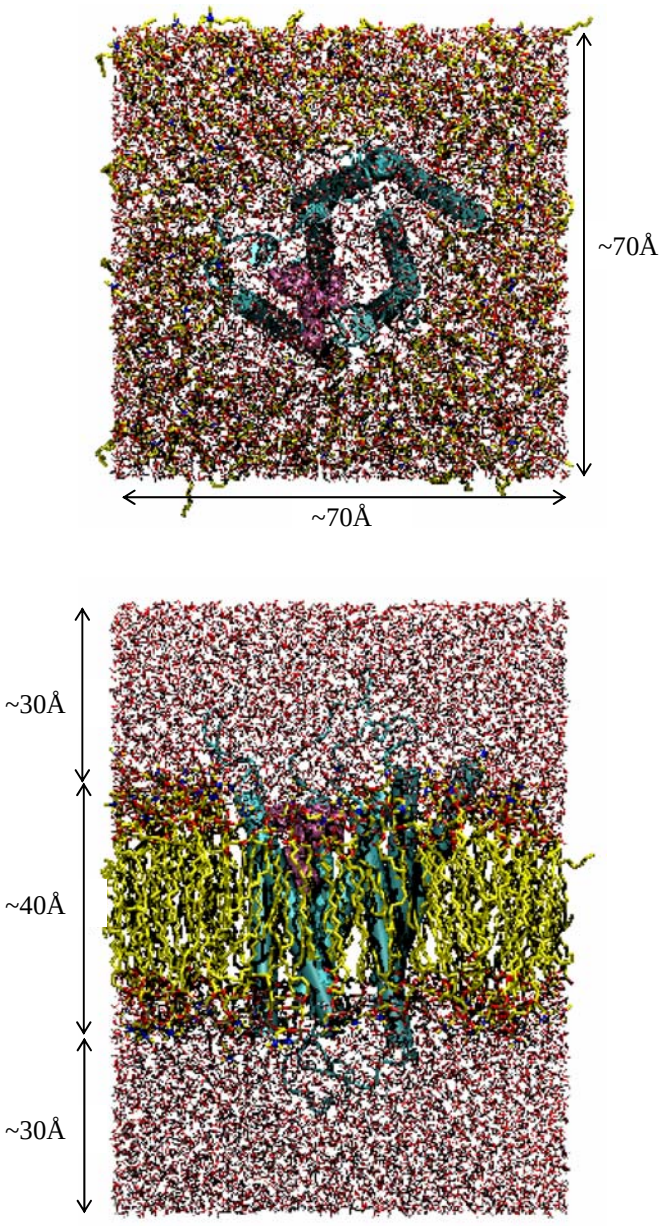


Figure 3.1 Fully solvated mMrgC11/F-(D)M-R-F-NH_2 complex in the membrane. It shows the final system built for NAMD run. The receptor (cyan) is shown in cartoon representation, the ligand (mauve) in VDW, lipids (yellow for carbon) in licorice and waters in line.

boundary condition with full electrostatics employed by using the Particle Mesh Ewald (PME) method.

Prior to full dynamics, the system was subjected to 5,000 steps of conjugate gradient energy minimization, followed by 100 ps of equilibration, while the coordinates of the receptor/ligand complex were fixed. In the equilibration, the system was gradually heated up from 0 K to 310 K by using Langevin dynamics with a damping coefficient of 5 ps^{-1} and the target temperature reached after ~7 ps. The system was again subjected to 5,000 steps of conjugate gradient energy minimization without any restraint. In energy minimization, the nonbonded interaction and electrostatic interaction were computed every time step. Lastly, the full dynamics simulation was carried out as described above.

The simulation was performed on a Linux-based cluster of Dual Intel Xenon 2.4 (or 3.06) GHz processors with 1 GB of memory per CPU. The first and second minimization took about 3 and 4 hours respectively with a single 3.06 GHz processor and the equilibration about 8 hours with 6 3.06 GHz processors. The 7 ns production run took about 17 days with 12 2.4 GHz processors.

3.3 Results and discussion

3.3.1 Comparison between initial and final structures

The final structure after a 7 ns equilibration was minimized with conjugate gradient for 5,000 steps. The minimized mMrgC11/F-(D)M-R-F-NH_2 complex structure was superimposed with the initial structure by aligning TM Cα atoms of the receptor. The RMSD for TM Cα atoms was 2.50 Å. As expected, the loop regions were floppier (RMSD = 7.00 Å) and the most dramatic change was the closure of the binding site by the extracellular loop 2 (EC2). The formation of the 'lid' in the binding site by the EC2 was observed in bovine rhodopsin structure where the disulfide bond formed between two Cys residues in EC1 (closer to TM3) and EC2 stabilizes the

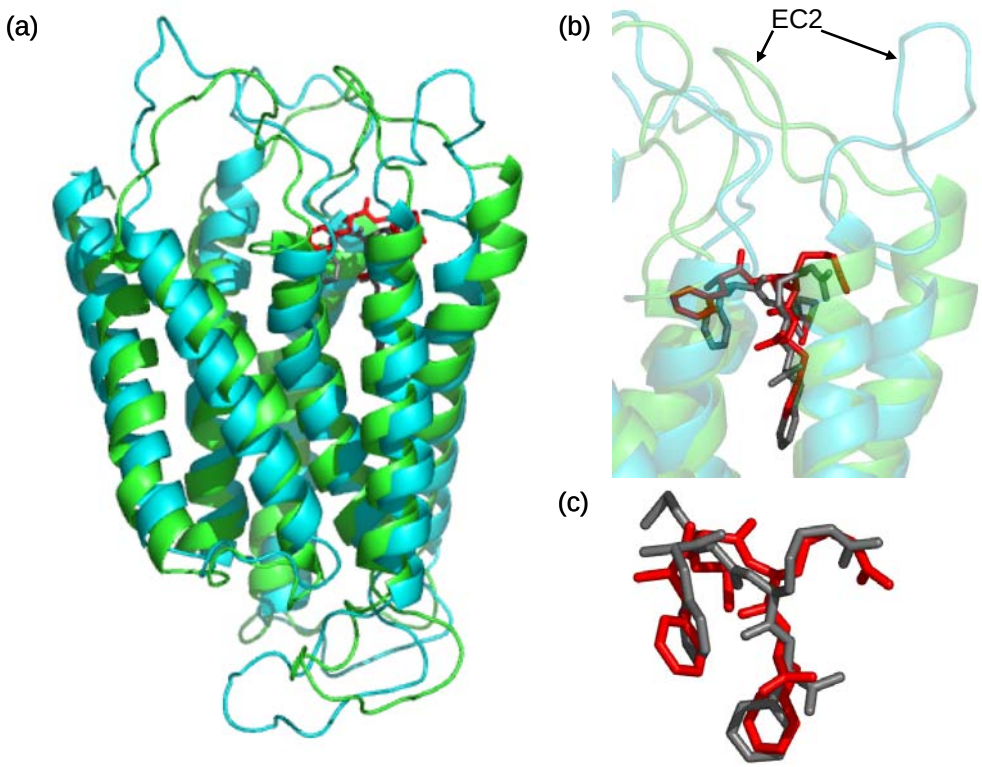


Figure 3.2 The mMrgC11/F-(D)M-R-F-NH_2 complex structure after 7 ns run. (a) Two complex structures at 0ns (cyan) and 7ns (green) are superimposed by aligning TM Cα atoms between them. The ligands are colored in black for 0ns and in red for 7 ns. The water, lipid and ion molecules are removed for clarity. (b) The ligands are in close-up after the residues in 5 Å binding pocket are aligned (RMSD for ligand = 2.48 Å). (c) Two ligands at 0 and 7 ns are aligned with heavy atoms (RMSD = 1.83 Å).

closed conformation of the EC2[7]. These Cys residues are conserved in several GPCRs including amine receptors, and the presence of a 'plug' in the binding crevice was also suggested in the β2 adrenergic receptor from the inaccessibility of quenchers to a fluorescent ligand[8]. The mMrgC11 receptor does not have the corresponding Cys residues and no disulfide bond is expected. However two oppositely charged residues, Lys96 (EC1) and Glu169 (EC2) are located at the similar sites and induce the closed conformation of EC2. If formation of a plug by EC2 is a general feature for all GPCRs, a key event in receptor activation would involve significant conformational change of EC2 allowing for rapid access of a ligand to the binding pocket.

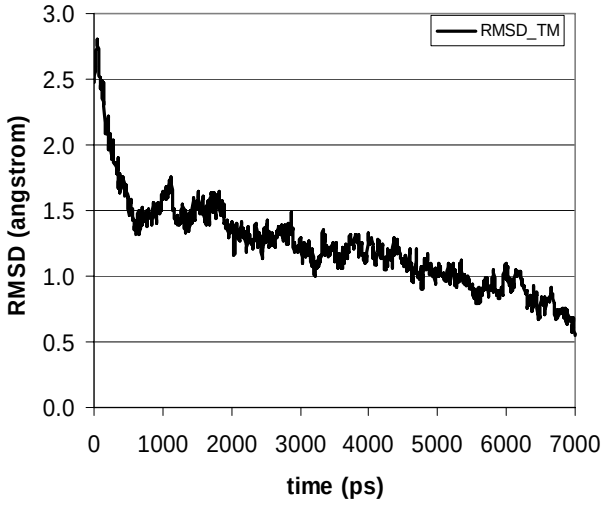
The tetrapeptide ligand moved by ~1.5 Å towards the exracellular regions after 7 ns equilibration, but not out of the binding pocket as shown in Figure 3.2(b). This kind of upward movement was also observed for the epinephrine agonist in the β2-adrenergic receptor after 4ns of MD simulation in the presence of full membrane and water[9]. This behavior was distinct from rigidity shown in an antagonist case.

The conformation of F-(D)M-R-F-NH_2 was examined by aligning the heavy atoms of the ligand (Fig. 3.2(c)). Two Phe and Arg were relatively rigid since interactions with four aromatic residues (Tyr110, Phe190, Trp162 and Tyr256) and two Asp residues (Asp161 and Asp179) restrained their movement as predicted in chapter 2. The Met was labile and its side chain underwent a large conformation change, leading to 1.83 Å of RMSD for the ligand.

3.3.2 Dynamic behavior in receptor conformation during MD simulation

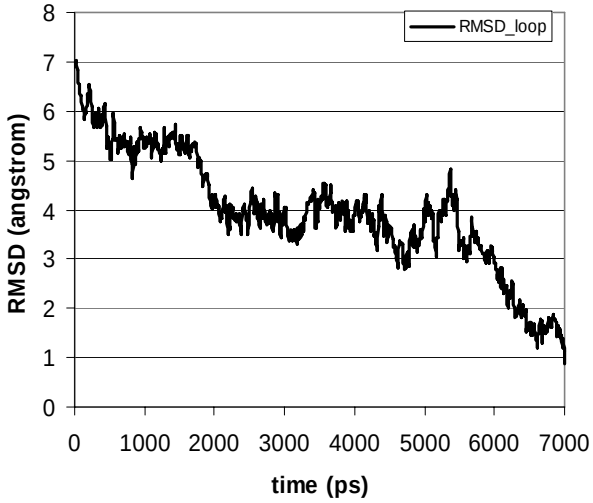
The RMSD of Cα atoms in the receptor was evaluated every 10 ps and plotted in Figure 3.3. Since the loop parts were much flexible, only the Cα atoms in TM regions were used in alignment. The RMSD plot indicates that the TM regions became well equilibrated after 7 ns. To explore conformational fluctuation for each TM, the corresponding Cα atoms were aligned respectively and then the RMSD of each TM was computed along the time. All TMs showed the similar plot (monotonous decrease) to Figure 3.3(a) of the whole TM regions. The large conformational change after 7 ns simulation was observed for TM6 and 7 (the RMSD values are 2.33 Å and 2.36 Å respectively). This conformational flexibility may be relevant to GPCR activation. The ligand binding is thought to trigger a cascade of structural changes in the receptor molecule that are capable of inducing activation of the associated G proteins. Here flexibility actually means low conformational barrier, leading to an ultimate structural change. The conformational change in TM6 of the rhodopsin or the β2 adrenergic receptor was supported by several structural and photophysical experiments[10, 11]. Also the EPR study in rhodopsin

(a)



TM	RMSD (Å)
1	1.43
2	1.79
3	1.58
4	1.25
5	1.88
6	2.33
7	2.36

(b)



Loop	RMSD (Å)
IC1	2.20
EC1	6.92
IC2	5.96
EC2	11.22
IC3	5.15
EC3	7.35

Figure 3.3 The RMSD fluctuation of Cα atoms with respect to the final 7 ns structure. The RMSD is calculated every 10ps by aligning the Cα atoms in TM regions. (a) The TM regions are selected. The RMSD values in table are calculated for the initial structure after aligning each TM respectively. (b) The whole loop part is selected in graph and each loop in table for RMSD calculation.

suggests that movement of the cytoplasmic end of TM7 relative to TM1 may occur in response of photoactivation[12].

The dramatic conformation change in loop regions was clearly demonstrated in the RMSD plot of Figure 3.3(b). Based on the RMSD value for each loop region (see table next to the RMSD plot), we can see that the extracellular loops underwent a larger conformational change. The most prominent change was for EC2 from an open to the closed conformation. The complete closure occurred after 6 ns and stayed until the end of simulation. Some conformational fluctuation was observed for EC1 and EC3 during the simulation. The dynamic behavior of the extracellular loop might be obvious since the ligand is bound in the upper half of TM regions from the extracellular region and directly perturbs the conformation of the residues close to the ligand.

Overall the significant change in the conformation of the receptor was seen in the mMrgC11/F-(D)M-R-F-NH_2 complex structure. Since F-(D)M-R-F-NH_2 is an agonist, the conformational change (from the inactive conformation to the active one) might be an apparent consequence.

3.3.3 Binding mode of F-(D)M-R-F-NH_2 after equilibration

The binding mode of the tetrapeptide F-(D)M-R-F-NH_2 after 7 ns equilibration is shown in Figure 3.4. The C-terminus amide group maintains the hydrogen bond with the side chain of Asp161 (TM4). The C-terminus F is still positioned in a stabilizing aromatic and hydrophobic environment formed by Tyr110 (TM3), Phe190 (TM5), Leu186 (TM5) and additionally Ile107 (TM3). The R is stabilized through the electrostatic interaction with Asp161 and an additional hydrogen bond with the side chain of Thr183 (TM5). However Asp179 in TM5 moved a little away from the R, but within the range where the electrostatic interaction was still effective (distance between NH1of R and OD2 of Asp179 = 6.41 Å). The water molecules actually intervened in interaction between Asp179 and the ligand, and mediated a hydrogen bond between the side chain of Asp179 and the backbone carbonyl group of the ligand (Fig. 3.5). The

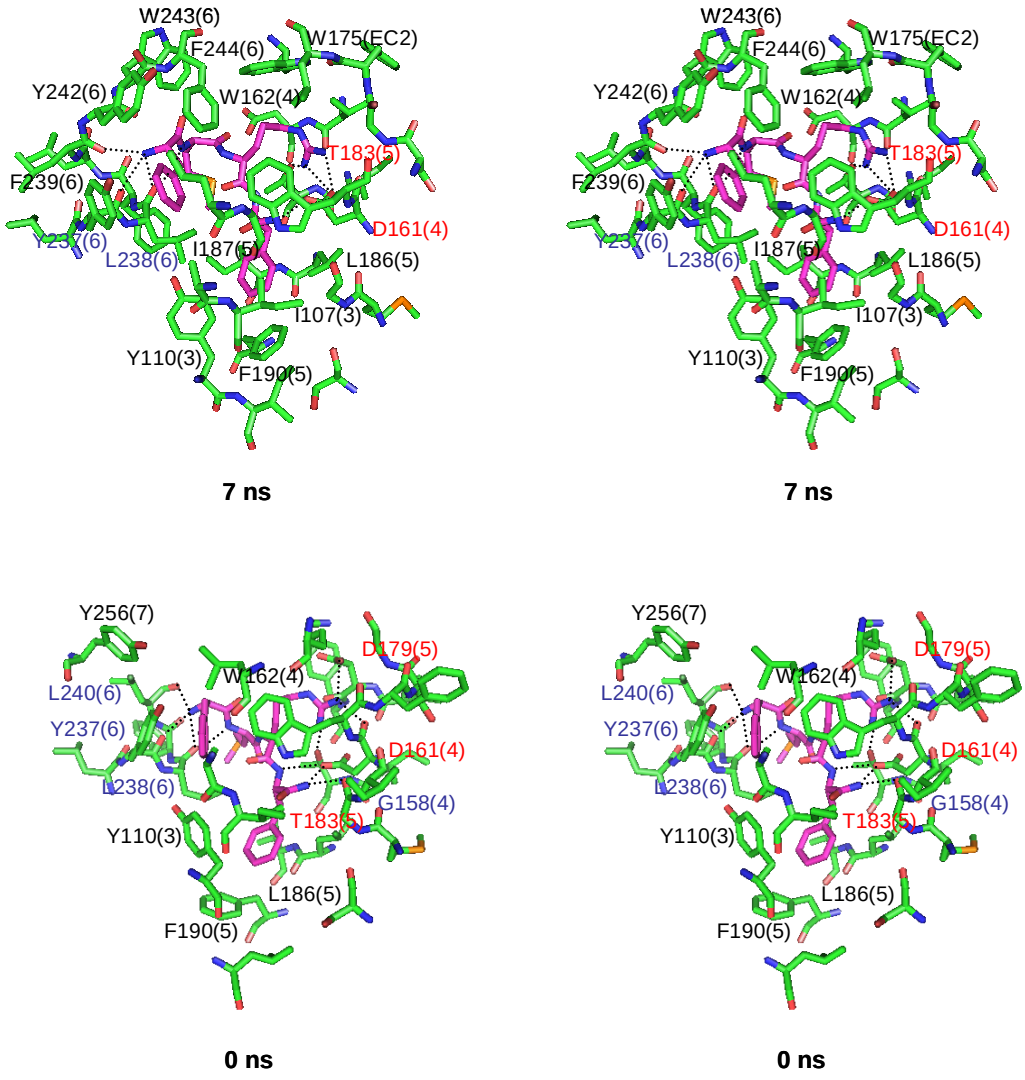


Figure 3.4 The 5 Å binding site of F-(D)M-R-F-NH_2 in mMrgC11 receptor. The intermolecular hydrogen bonds (calculated with explicit hydrogens using the same criteria as in Figure 2.4) are indicated by the dotted lines. A residue whose side chain participates in the hydrogen bond is specified in red, while one whose backbone is involved is in blue. The residues showing good hydrophobic interactions are specified in black. The top of each picture corresponds to the extracellular regions.

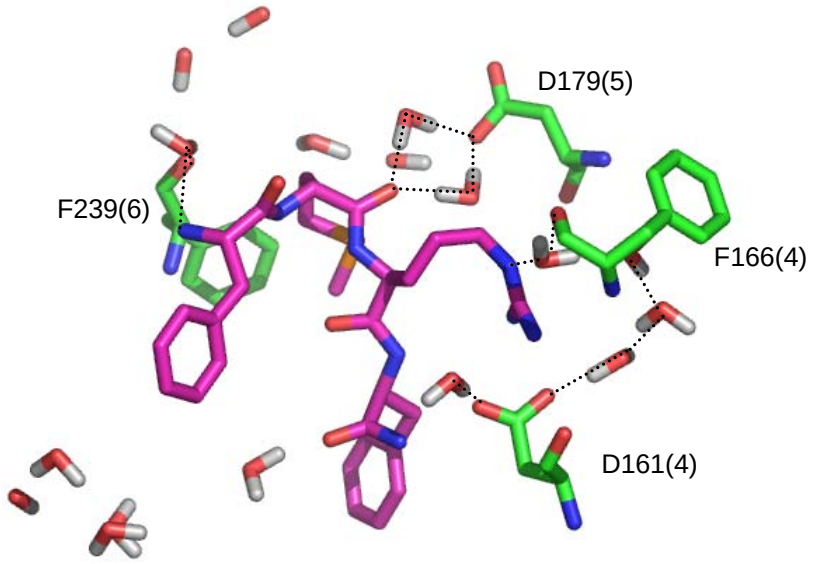


Figure 3.5 Water molecules in 5 Å binding pocket. The water-involved hydrogen bonds are indicated by the dotted line.

carboxylate group of Asp179 was solvated with more water molecules and also stabilized by the positively charged quaternary amine group of a lipid molecule (distance between N of quaternary amine and OD1 of Asp179 = 4.16 Å). This relatively weak interaction of Asp179 compared to Asp161 might be validated by our experimental observation in chapter 2. For the D161A mutant, four of the six agonists were rendered inactive, while the remaining two were only active at 100 times higher concentrations. Similarly, the D179A mutant showed no affinity for the three tetrapeptide agonists, while the other three were activated only at 10 times higher concentration of the ligand. This indicates that Asp161 should interact more effectively with the agonists than Asp179.

The N-terminal F remained sandwiched between Trp162 (TM4) and Tyr237 (TM6) (the closest C-C distances between two aromatic rings are 4.64 Å and 3.39 Å respectively). Two more aromatic residues in TM6, Tyr242 and Phe244 (located close to the extracellular loop) came into

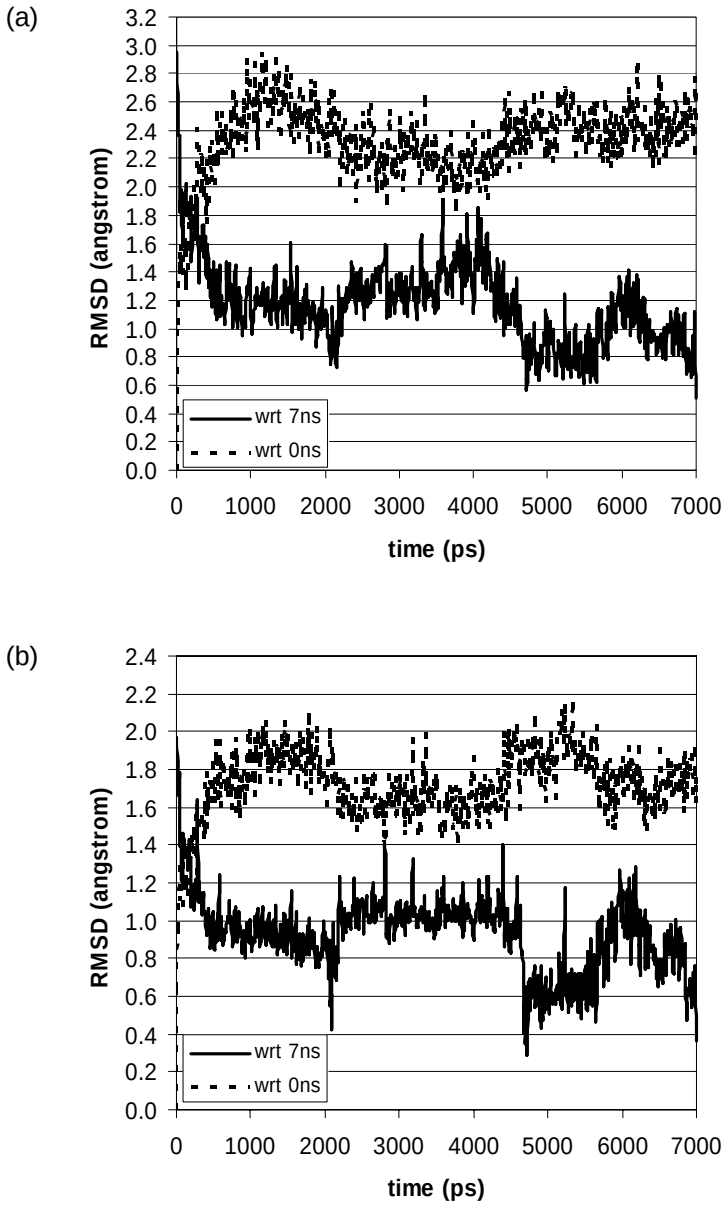


Figure 3.6 The RMSD fluctuation for ligand heavy atoms. (a) The residues in 5 Å binding pocket are aligned and then the RMSD for the ligand is computed. (b) The ligand parts are aligned each other. The real line is for the RMSD with respect to the final structure after a 7 ns equilibration and the dotted one for that with respect to the initial structure.

the binding pocket (the closest C-C distances between two aromatic rings are 4.22 Å and 3.40 Å respectively) and yielded the additional favorable aromatic interaction with the N-terminal F.

The water molecules filled the void in the binding pocket, forming hydrogen bonds with polar atoms, and some of them mediated an intermolecular hydrogen bond between the receptor and the ligand as observed for Asp179. The backbone atoms of Phe166 (TM4) and Phe239 (TM6) form the water-mediated hydrogen bonds with the side chain of R and the N-terminus of the tetra-peptide respectively.

3.3.4 Time profile of receptor-ligand interactions

Ligand conformation in the binding site

The RMSD of the ligand was evaluated every 10 ps in MD simulation. In Figure 3.6(a), the residues within the 5 Å binding pocket were used in alignment to give information about the ligand configurations in the binding site throughout time. The ligand was configurationally flexible and the RMSD of the final 7 ns minimized structure was 2.48 Å with the initial structure.

The ligand conformation itself fluctuated throughout the MD simulation and the RMSD values were ~1.5-2.0 Å with respect to the initial conformation. This indicates that the major contribution of configurational change shown previously is the conformational variation of the ligand itself. From the correlation between two RMSD plots (Fig. 3.6(a) and Fig. 3.6(b)) we can see that the ligand is confined within the binding pocket for 7ns, but exhibits conformational flexibility.

Intermolecular hydrogen bonds

The intermolecular hydrogen bonds between the receptor and the ligand were determined with the same criteria used in chapters 2 and 4 (see Fig. 4.3) for the initial and the final minimized structures. The distance between the donor and acceptor atoms was computed for every hydrogen bond pair and plotted along the time in Figure 3.7.

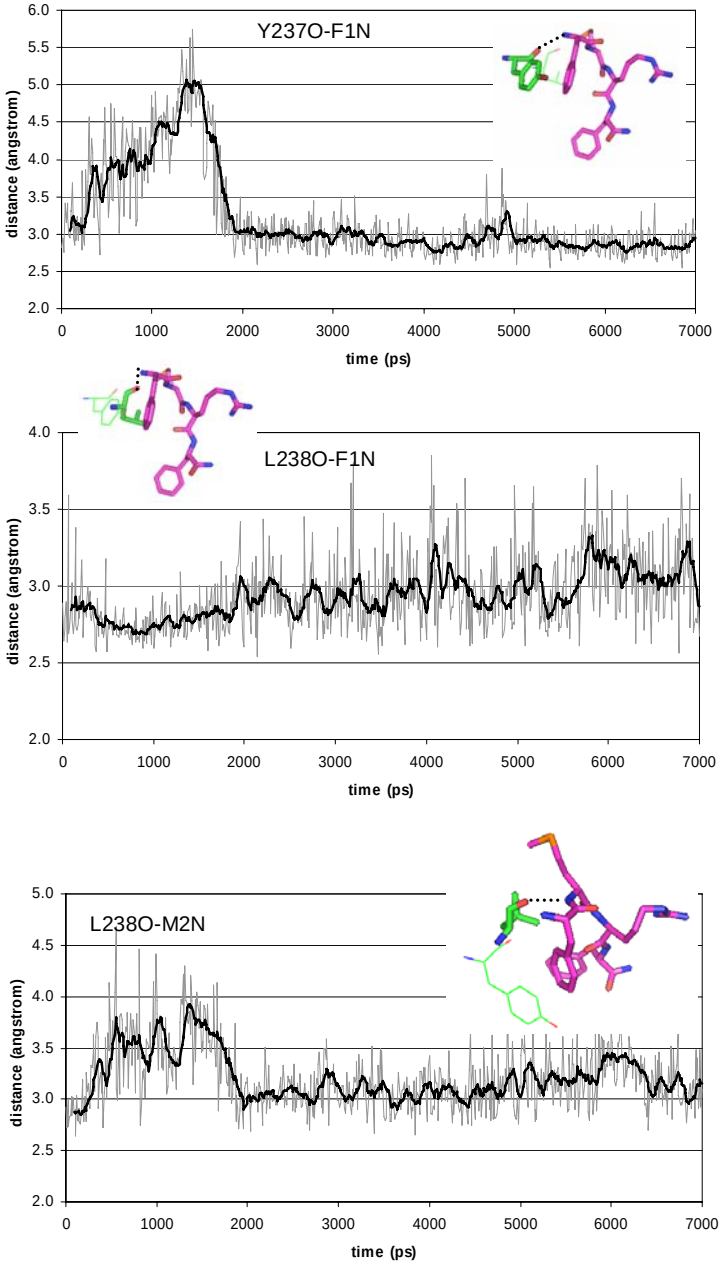


Figure 3.7 Time profile of intermolecular hydrogen bond distance. The distance is measured every 10 ps (grey). The moving average per 100 ps is in black. The hydrogen bond pair is indicated in this way: receptor part-ligand part with the index of [residue name][residue number][atom name participating in H-bond]. The ligand (carbon in purple) and the receptor residue (carbon in green) involved in the hydrogen bond is shown in stick at the picture. The corresponding hydrogen bond is indicated in the dotted line.

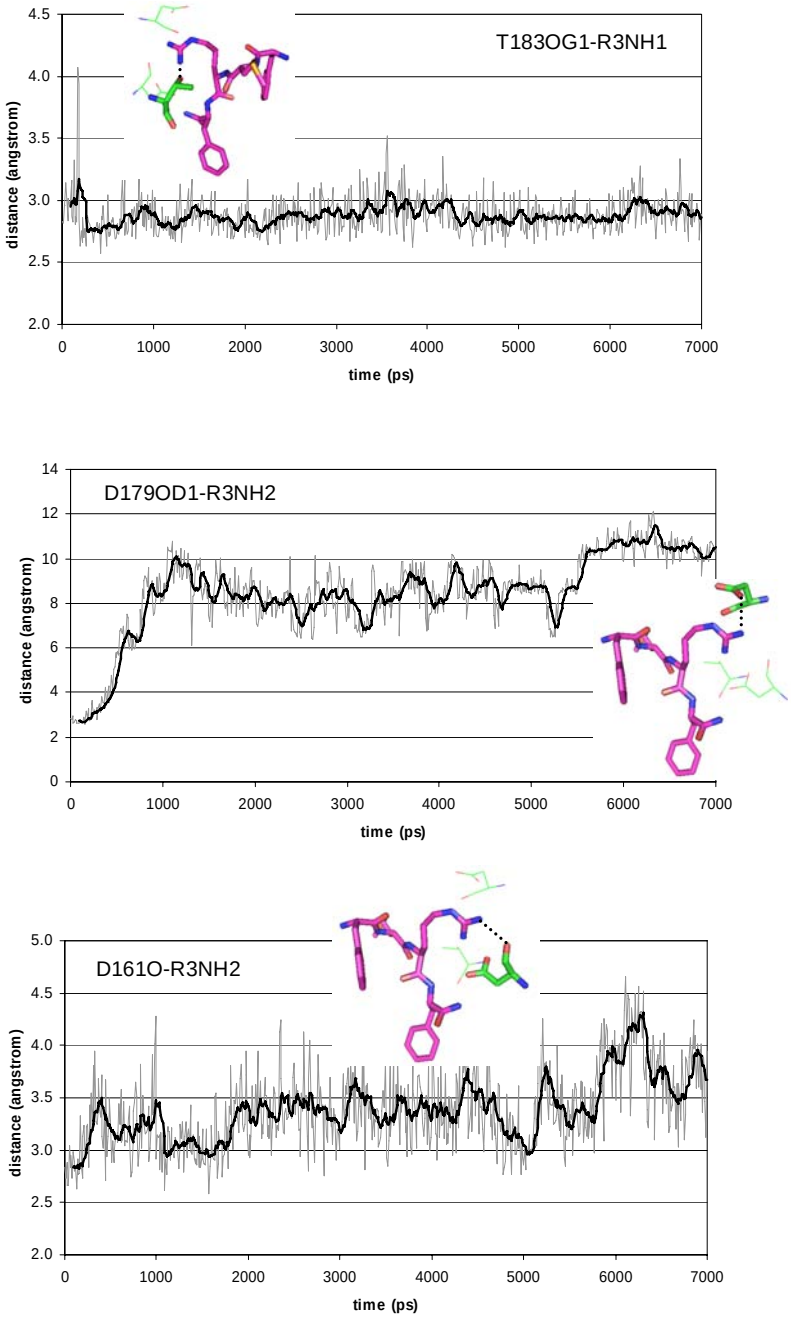


Figure 3.7 (*continued*)

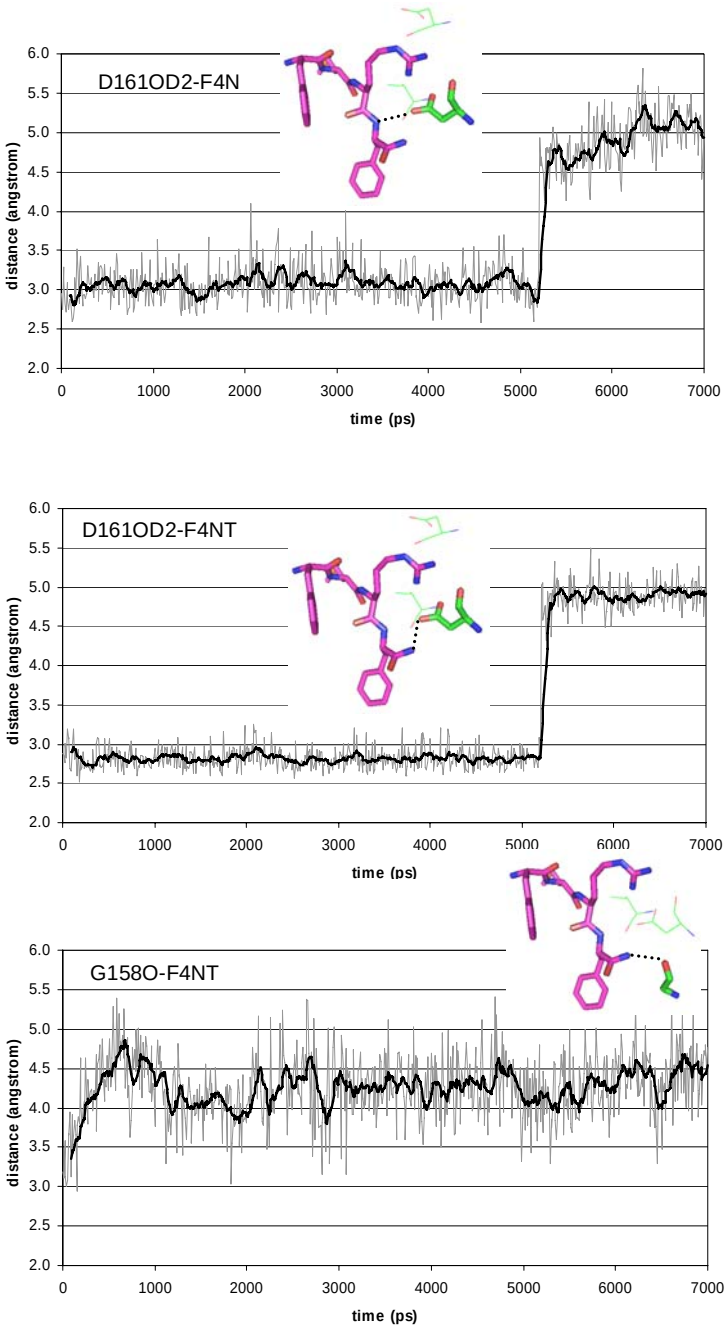


Figure 3.7 (*continued*)

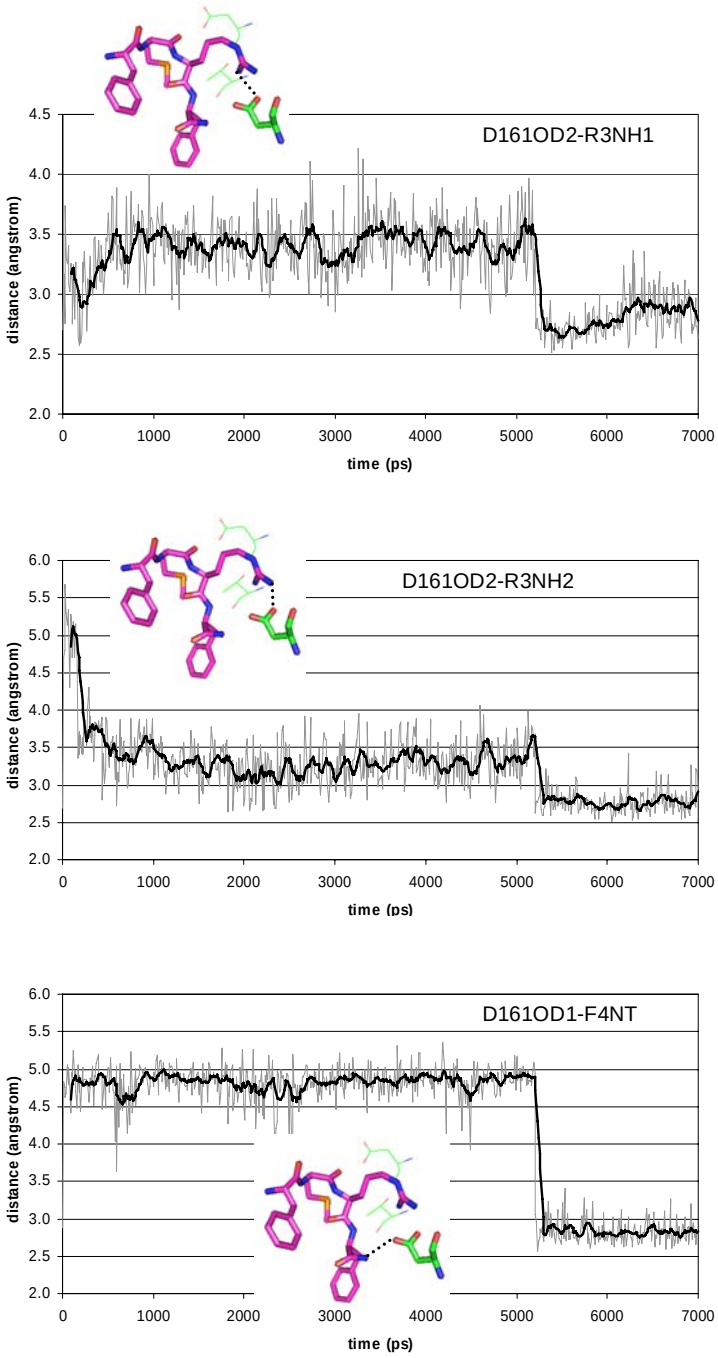


Figure 3.7 (*continued*)

The hydrogen bonds with the residues in TM6 (i.e. between the N-terminus and Tyr237 (Y237O-F1N) and between the backbone amide group and Leu238 (L238-M2N)) were interrupted for the first 2ns and then became stable. This observation indicates that some conformational rearrangement in (or near) TM6 occurs during these time frames. This fluctuation is well correlated with entry of Phe244 into the binding site as will be shown in Figure 3.8. The intrusion of Phe244 as well as Tyr242 perturbs the initial conformations, but the favorable interactions previously present are recovered after re-organization.

The direct hydrogen bond of the ligand with Asp179 became loose at the early stage due to a water molecule stepping in between them. However Asp179 still made contact with the ligand through the water-mediated hydrogen bond and electrostatic interaction. The side chain of R mostly interacted with the side chain of Asp161 in the ideal configuration after equilibration. The torsion $\chi 2$ of the side chain in Asp161 was shown to be in relatively low barrier and the C-terminus amide group of the ligand switched a hydrogen bond partner between two carboxylate oxygen atoms of Asp161 after ~5 ns.

The hydrogen bond between the side chain of Thr183 and the side chain of R remained stable throughout the simulation period.

Inter-aromatic interaction

The time evolution of the centroid-to-centroid distance between two interacting aromatic rings was explored. The interaction of Tyr110 (TM3) and Phe190 (TM5) with the C-terminus F kept steady during 7 ns simulation. As mentioned previously, two more aromatic residues in TM6, Tyr242 and Phe244 participated in interaction with the N-terminal F after equilibration. Tyr242 came close at the early time step, but it did not interfere in the overall conformation of the binding site. However intrusion of Phe244 actually affected the present binding mode (note that Tyr237 and the N-terminal F become apart and back together) and then after conformational rearrangement all favorable aromatic interactions were recaptured.

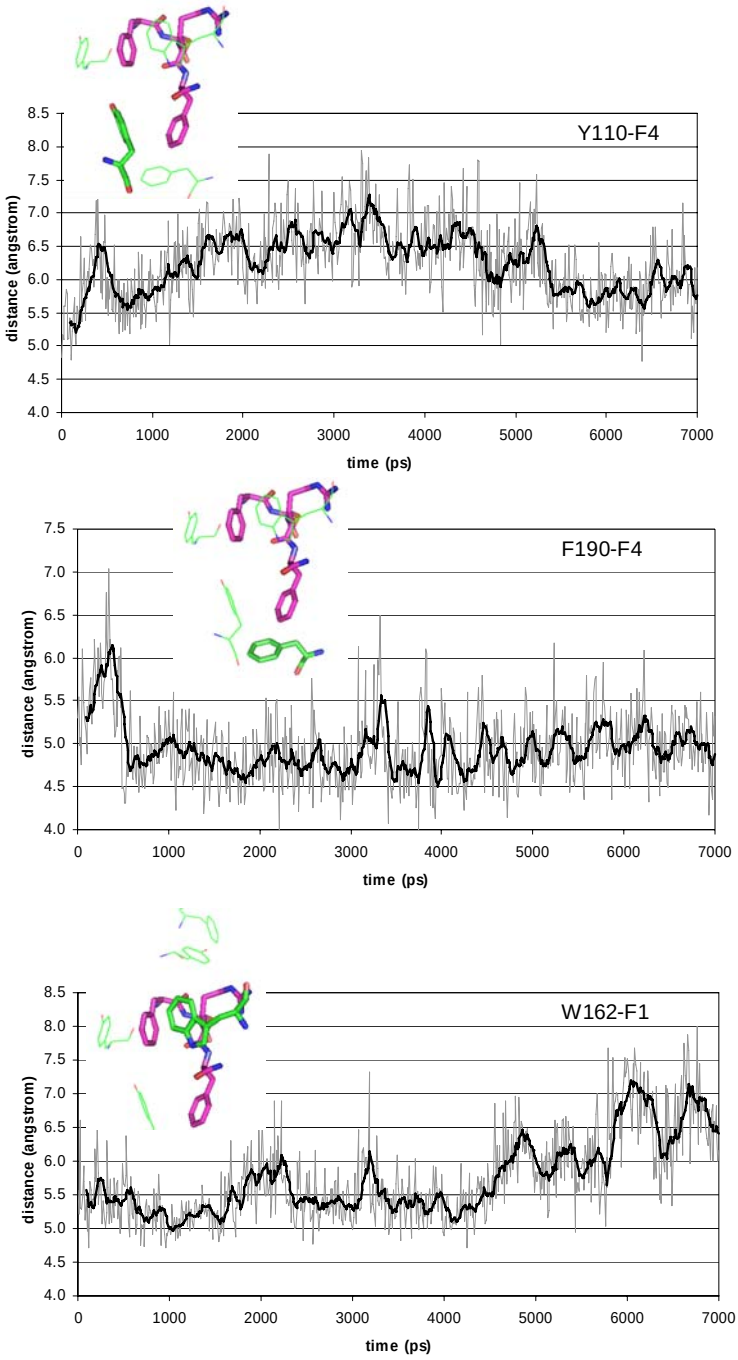
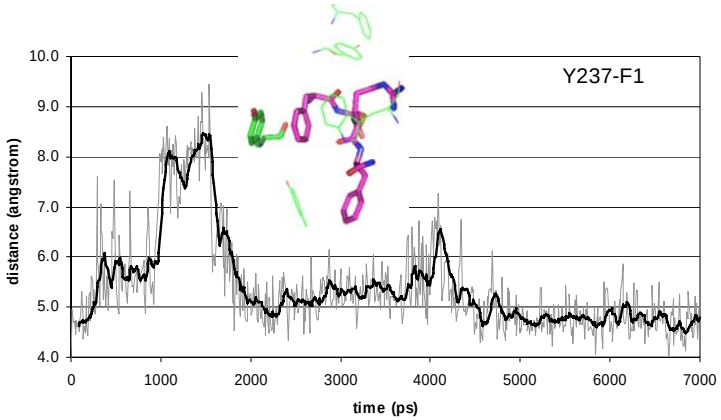
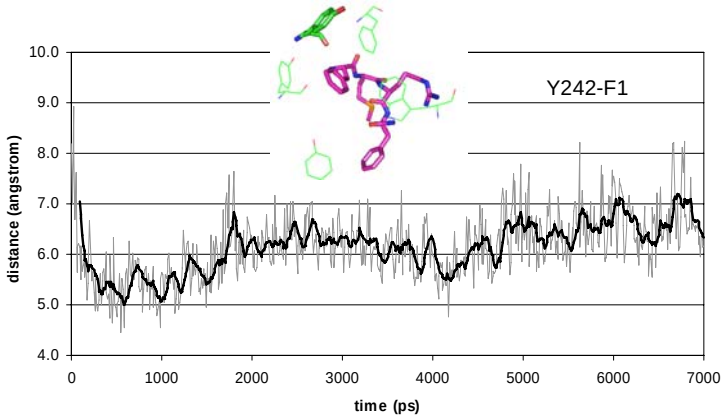


Figure 3.8 Time profile of centroid-to-centroid distance between two aromatic residues. F1 and F4 denote the N-terminal and C-terminal F of the tetrapeptide ligand respectively. The distance is measured every 10 ps (grey). The moving average per 100 ps is in black. The ligand (carbon in purple) and the receptor residue (carbon in green) are shown in stick at the picture. For tryptophan, the center of the six-membered ring is considered.





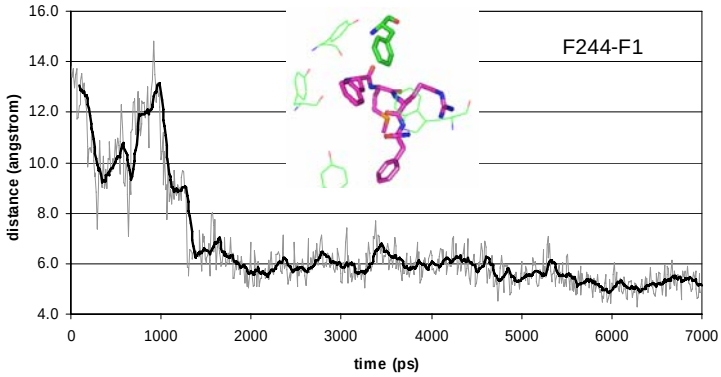


Figure 3.8 (*continued*)

The centroid distance at 7 ns showed that Trp162 (TM4) moved a little away from the ligand, and the closet C-C distance between aromatic rings of Trp162 and the N-terminal F was 4.64 Å. However Trp162 was not completely out of interaction with this F.

The stability of aromatic interactions identified in the previous prediction implies the accuracy of our predicted structure for the binding site. The new interactions with Tyr242 and Phe244 indicate that the explicit membrane and water simulation might be necessary to obtain correct conformations for residues on the boundary of the TM and the loop.

3.3.5 Time profile of inter-helical interactions

The nonbond distances between residues on different helices were analyzed to understand how the dynamics in the explicit lipid and water environment affect the inter-helical interactions. The inter-helical hydrogen bonds were identified for the initial and the final 7 ns minimized structure (Fig. 3.9) and the distances for these hydrogen bond pairs were measured throughout the MD simulation. The comparison of two hydrogen bond networks demonstrates some dynamic behavior on the inter-helical interactions. The initial hydrogen bond network was subjected to re-arrangement during the MD simulation. We can also see some hydrogen bond pairs preserved after 7ns; Tyr63 (TM2) (one of the residues conserved in the Mrg receptor family (with 39 sequences available on Swiss-Prot and TrEMBL))–Ser112 (TM3) and Arg215 (TM6)–Val277 (TM7). Moreover the hydrogen bond between Tyr63 and Ser112 remained stable throughout the MD run as shown in Figure 3.10 and it may play a role in maintaining helix packing.

The hydrogen bond between Asn66 (TM4) and Trp151 (TM4) (the highly conserved residues in the family A of GPCRs that form an interhelical hydrogen bond in rhodopsin) became loose, but not totally apart. Asn66 partly formed a hydrogen bond with Ser115 (TM3) during the MD run.

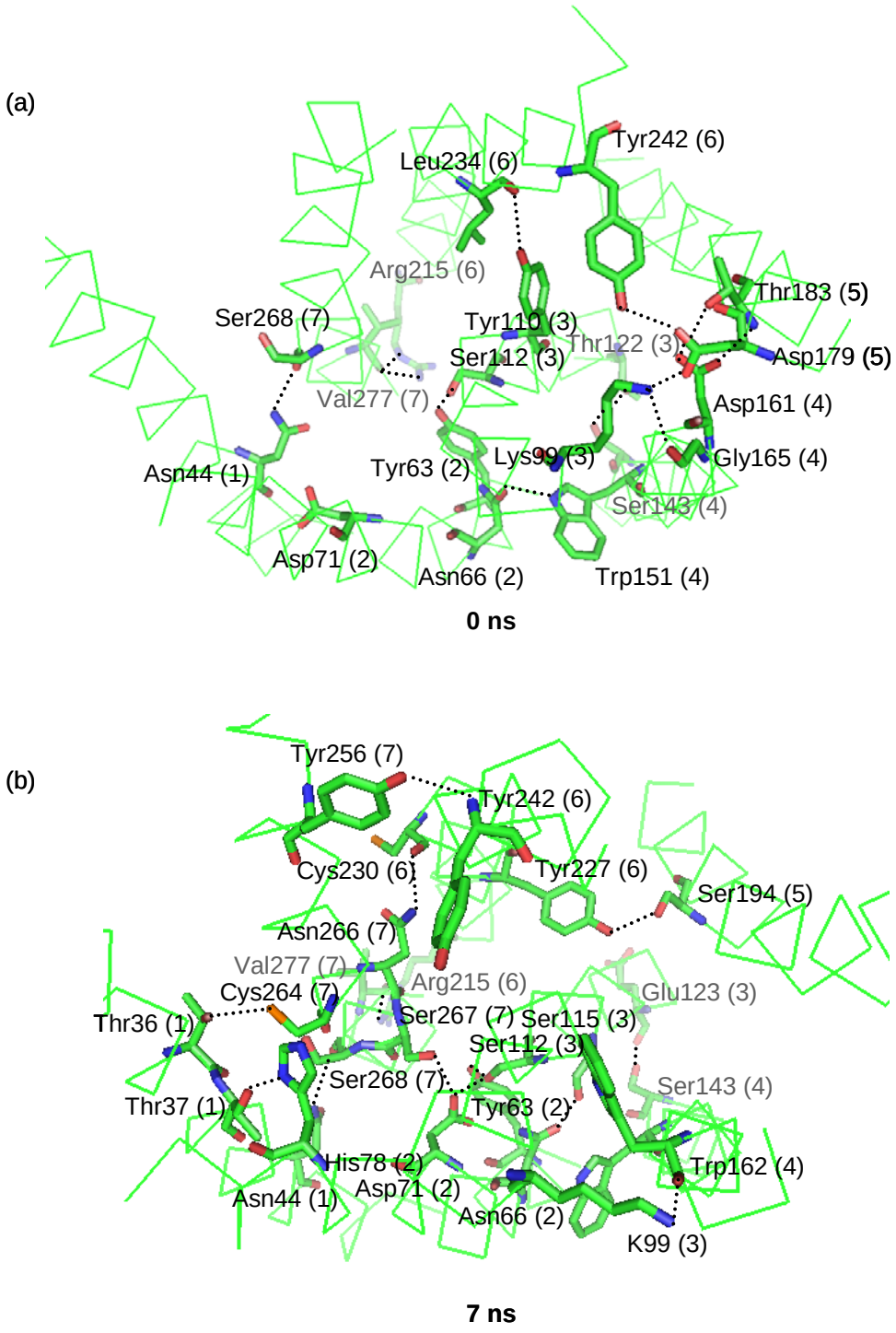


Figure 3.9 Interhelical hydrogen bond networks in the mMrgC11 receptor. It is viewed form the extracellular region. The interhelical hydrogen bonds (dashed lines) are specified with residues participating in hydrogen bond. (a) for the initial structure. (b) for the 7ns minimized structure. The HBPLUS program was used to calculate hydrogen bonds (maximum D-A distance = 3.9 Å, minimum D-H-A angle = 90.0°).

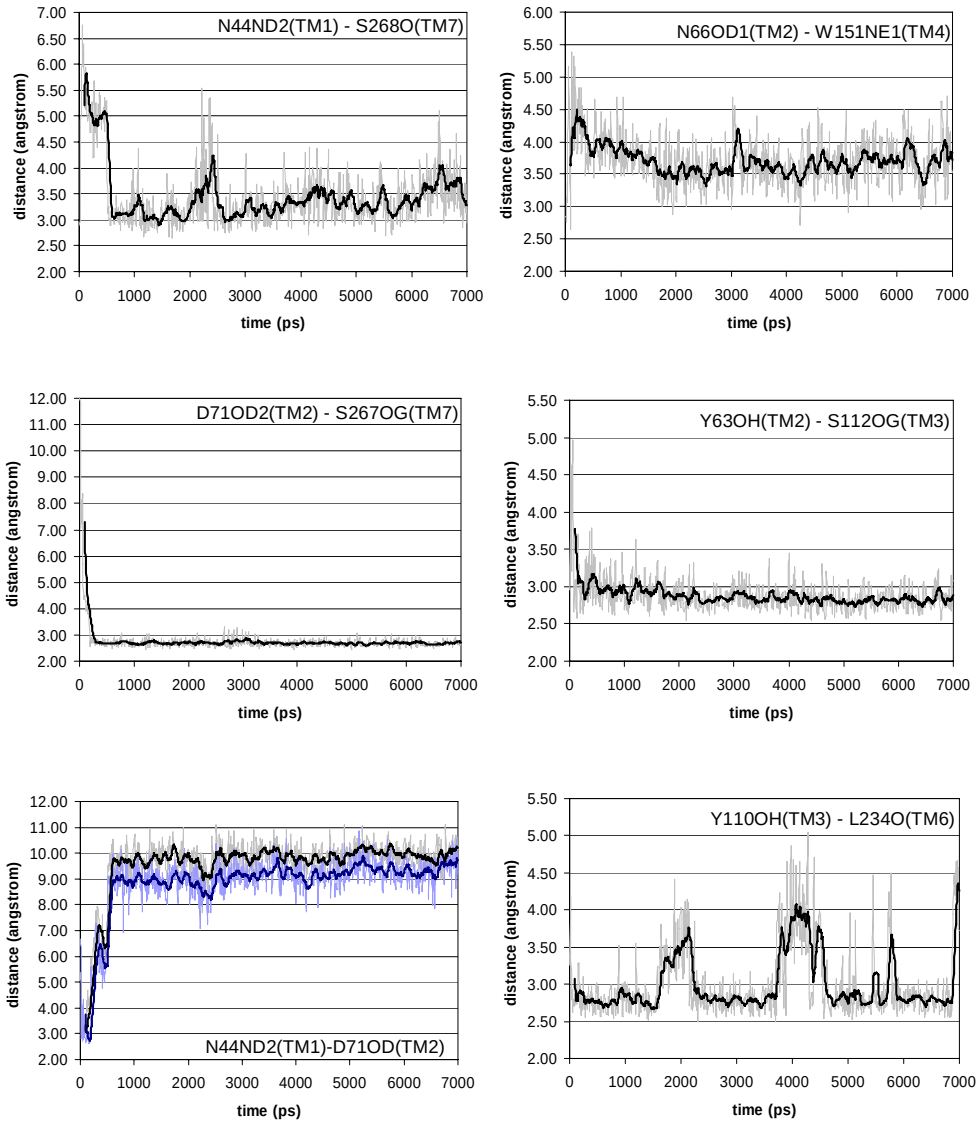


Figure 3.10 The time profile of the distance between residues residing in different helices. The same index rule ([residue name][residue number][atom name]) is used as in Figure 3.7.

Initially Asn44 (TM1) (highly conserved in the family A GPCRs) formed a hydrogen bond with the Ser268 carbonyl group of the backbone in TM7 as shown in Figure 3.9. This hydrogen bond was loosened for the early 500 ps (this may be an equilibration period needed for the protein structure to be adjusted from perturbation of the lipid and water molecules) and then the pair remained close enough for the hydrogen bond formation. At t = 0, Asp71 (TM2) that forms an inter-helical hydrogen bond with this Asn in rhodopsin was in the proximity, but was not in hydrogen bond contact in the mMrgC11 receptor. In the MD run the distance between Asp71 and Asn44 became larger, leading to ~9-10 Å. Instead Asp71 moved close to TM7 and formed a stable hydrogen bond with Ser267. The approaching of TM2 and TM7 in the activated state was suggested for the angiotensin receptor II type 1 from mutation-induced constitutive activation, and later the *in situ* measurement of TM2 movement in the angiotensin receptor was also reported[13, 14]. Based on our simulation, it might be proposed that TM2 first moves further from TM1 on activation and then towards TM7. During 7ns, the concerted formation of hydrogen bonds in TM1, TM2 and TM7 that exists in the inactivated rhodopsin structure[7, 15] was not observed in our predicted mMrgC11 receptor, suggesting that the receptor structure was in the activated conformation.

Lastly we examined the distance between Tyr110 (TM3) and Leu234 (TM6). Tyr110 was one of the residues interacting with the ligand and underwent conformational fluctuation. This kind of flexibility between TM3 and TM6 may help induce the receptor activation.

3.4 Summary and conclusions

We performed the all-atom MD simulation of mMrgC11/F-(D)M-R-F-NH_2 structure in the explicit lipid and water environment. The analysis of the 7 ns MD trajectory clearly demonstrated that our predicted structure of the mMrgC11 receptor and its binding site of F-(D)M-R-F-NH_2 was stable in the full membrane system. The conformational flexibility of the side chain and small structural change in TM regions were present, but no significant instability was detected.

Moreover the initial interactions of the ligand with the key residues (Asp161 and four aromatic residues, Tyr110, Phe190, Trp162 and Tyr237) were preserved throughout the entire MD run except for Asp179 in TM5. Nevertheless Asp179 interacted with the ligand through water-mediated hydrogen bond and electrostatic interaction. These findings validate our structure prediction method, indicating that the MembStruk predicted structures are fairly accurate.

In addition we observed some dynamic behavior in protein structure. In the TM regions, TM6 and TM7 showed relatively large conformational change and it suggested the possibility of their implication in receptor activation. The loops underwent large structural fluctuations, and the most dramatic change was seen in EC2. Interestingly the electrostatic interaction of two oppositely charged residues, Glu169 (EC2) and Lys96 (EC1) pulled them each other, resulting in the closed conformation of EC2 that is similarly shown in rhodopsin. Two more aromatic residues in TM6, Tyr242 and Phe244 were newly identified to contact the N-terminal F of the ligand after the equilibration, securing the ligand in the binding site. They could be additional mutation candidates to be tested for the further validation. These observations indicate that the explicit membrane and water simulation might be necessary to obtain correct conformations for the loops, including residues on the boundary of the TM and the loop.

An extended simulation along with incorporation of G protein into our receptor structure where the intracellular loops are now fully equilibrated could be explored to examine the reciprocal effect of the G protein and the mMrgC11 receptor on the conformational change in activation. It would definitely provide the better understanding on the GPCR activation process.

References

1. Freddolino, P.L., et al., *Molecular dynamics simulations of the complete satellite tobacco mosaic virus.* Structure, 2006. **14**(3): p. 437-449.

2. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of Computational Chemistry, 2005. **26**(16): p. 1781-1802.

3. Hall, S.E., *Development of a structure prediction method for G-protein coupled receptors*, in *Division of Chemistry and Chemical Engineering*. 2005, California Institute of Technology: Pasadena.

4. Mackerell, A.D., et al., *Self-Consistent Parameterization of Biomolecules for Molecular Modeling and Condensed Phase Simulations.* Faseb Journal, 1992. **6**(1): p. A143-A143.

5. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.

6. Jorgensen, W.L., et al., *Comparison of Simple Potential Functions for Simulating Liquid Water.* Journal of Chemical Physics, 1983. **79**(2): p. 926-935.

7. Okada, T., et al., *The retinal conformation and its environment in rhodopsin in light of a new 2.2 angstrom crystal structure.* Journal of Molecular Biology, 2004. **342**(2): p. 571-583.

8. Tota, M.R. and C.D. Strader, *Characterization of the Binding Domain of the Beta-Adrenergic-Receptor with the Fluorescent Antagonist Carazolol - Evidence for a Buried Ligand-Binding Site.* Journal of Biological Chemistry, 1990. **265**(28): p. 16891-16897.

9. Spijker, P., et al., *Dynamic behavior of fully solvated beta 2-adrenergic receptor, embedded in the membrane with bound agonist or antagonist.* Proceedings of the

National Academy of Sciences of the United States of America, 2006. **103**(13): p. 4882-4887.

10. Farrens, D.L., et al., *Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin.* Science, 1996. **274**(5288): p. 768-770.

11. Ghanouni, P., et al., *Agonist-induced conformational changes in the G-protein-coupling domain of the beta(2) adrenergic receptor.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(11): p. 5997-6002.

12. Altenbach, C., et al., *Structure and function in rhodopsin: Mapping light-dependent changes in distance between residue 65 in helix TM1 and residues in the sequence 306-319 at the cytoplasmic end of helix TM7 and in helix H8.* Biochemistry, 2001. **40**(51): p. 15483-15492.

13. Groblewski, T., et al., *Mutation of Asn(111) in the third transmembrane domain of the AT(1A) angiotensin II receptor induces its constitutive activation.* Journal of Biological Chemistry, 1997. **272**(3): p. 1822-1826.

14. Miura, S. and S.S. Karnik, *Constitutive activation of angiotensin II type 1 receptor alters the orientation of transmembrane helix-2.* Journal of Biological Chemistry, 2002. **277**(27): p. 24299-24305.

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

Chapter 4

Virtual Ligand Screening of Chemical Libraries for Mouse MrgC11 Receptor: Combination of QSPR and Docking Methods[1]

4.1 Introduction

High-throughput screening (HTS) of chemical libraries is the widely adopted method for finding novel lead compounds in drug discovery. It enables a large number of compounds to be screened using highly automated, robotic techniques. Although HTS makes it possible in principle to test all available compounds, it is not necessarily feasible for a number of practical reasons. One of reasons is the cost of such screenings: even though the robotics and miniaturization have significantly reduced the unit cost, the huge number of compounds now available from many companies means that the overall expense can be significant. Moreover, as the available databases get larger and larger, the hit rates in HTS dramatically decrease. A possibility to avoid these problems is not to screen the whole compound set in the library experimentally, but only a small subset, which is likely to bind to the target protein receptor. This pre-selection can be performed by virtual screening (VS), which uses computer-based methods to select most promising compounds from the ligand databases for experimental assays. Virtual screening can be carried out by searching databases for molecules fitting either a known pharmacophore (ligand-based) or a three-dimensional structure of macromolecular target (structure-based). In the case of GPCRs, the limited availability of the structural data has forced the computational design of ligands to heavily rely on ligand-based drug design techniques.

[1] This work was carried out in collaboration with the Tropsha group of the University of North Carolina.

Indeed, the natural ligands can provide a good starting point, leading to useful pharmacophore models that can be used for virtual screening to identify lead structures with novel scaffolds[1]. The application of this method has been successfully demonstrated in the discovery of subtype selective agonists to the somatostatin receptor[2] and non-peptide antagonists to the urotensin II receptor[3]. Structure-based screening should be potentially more powerful than the ligand-based method since by exploiting structural information taken directly from the active site, it is possible to discover ligands with both diverse chemotypes and binding modes. However, it still suffers from docking/scoring inaccuracy, and in addition it requires the knowledge of the 3D structure of the target protein. Therefore, it has mostly been applied to targets for which a high resolution X-ray crystal structure is known. However, along with the deciphering of human genome, computational chemists are facing an overwhelming number of potential targets for which very little experimental 3D information is available. Therefore it will be very important in the near future to be able to use not only X-ray or NMR structures, but also protein models for structure-based virtual screening of chemical libraries.

The structure-based virtual screening mainly relies on a fast and accurate docking/scoring function that can be used to identify the correct binding mode. Theoretically, the most accurate estimate of the binding affinity can be obtained using force-field based methods. Examples include free energy perturbation (FEP)[4] or linear interaction energy (LIE) approaches[5]. However, the computational cost of such methods is too high to afford calculation in a high-throughput fashion. Therefore the huge chemical libraries should be filtered through a rapid pre-screening tool to identify the most promising compounds prior to engaging more computationally intensive docking approaches. The ligand-based similarity searching technique could be used for this purpose. In this approach, the ligand structures are typically represented by multiple chemical descriptors and the statistical data modeling techniques are used to establish quantitative correlation between descriptors and target properties of interest, such as binding constants or

specific biological activities[6]. Recently the Tropsha group in the University of North Carolina had developed a novel structure-based chemoinformatics approach to search for complimentary ligands based on receptor information (CoLiBRI)[7]. CoLiBRI is based on a representation to characterize both receptor active sites and their corresponding ligands in the same universal, multidimensional, chemical descriptor space. Mapping of both binding pockets and corresponding ligands onto the same multidimensional chemistry space would preserve the complementarity relationships between the binding sites and their respective ligands.

In this study, we carried out virtual screening for the mouse MrgC11 receptor, one of orphan GPCR receptors as an effort to identify small molecule ligands that behave as selective agonists or antagonists. Despite of the success of orphan GPCR-natural ligand pairing through reverse pharmacology many scientists focused on discovering new drugs appear to be bypassing the conventional deorphanizing step due to the difficulty in developing peptide libraries to look for the ligand. They perform initial high-throughput assays to find synthetic small-molecule agonists, which then can be used to explore the physiological aspects of the receptor. Here we first pre-screened compounds in the chemical database using the CoLiBRI and the resulting candidates were subsequently docked using the MSCDock method. The ‘hit’ compounds from docking were experimentally tested with the intracellular calcium release assay. In the following sections, we describe the computational methods in details and discuss the screening results.

4.2 Materials and methods

4.2.1 Pre-screening of compounds in chemical libraries

Pre-screening the compounds of the chemical libraries was carried out using the CoLiBRI program. CoLiBRI is based on the quantitative structure-property relationship (QSPR) method. It generates the molecular descriptors that capture key properties of the molecules, using the transferable atom equivalent (TAE)/RECON method. The TAE/RECON method that was developed by Breneman and co-workers[8] rapidly generates molecular electron density

Table 4.1 Electron-density-derived TAE descriptors; $\rho(r)$ represents the electron density distribution[9]

Integral Electronic Properties		
Energy		
Electronic population		
Volume		
Surface area		
Surface electronic properties (extrema, surface integral averages and histogram bins are available for each property)		
SIEP	Surface integral of electrostatic potential	
EP	Electrostatic potential	$EP(r) = \sum_{\alpha} \frac{Z_{\alpha}}{\lvert r - R_{\alpha} \rvert} - \int \frac{\rho(r')dr'}{\lvert r - r' \rvert}$
DRN	Electron density gradient normal to 0.002 e/au^3 electron-density isosurface	$\nabla\rho \cdot n$
G	Electronic kinetic energy density	$G(r) = -(1/2)(\nabla\psi^* \cdot \nabla\psi)$
K	Electronic kinetic energy density	$K = -(1/2)(\psi^*\nabla^2\psi + \psi\nabla^2\psi^*)$
DKN	Gradient of the *K* electronic kinetic energy density normal to surface	$\nabla K \cdot \boldsymbol{n}$
DGN	Gradient of the *G* electronic kinetic energy density normal to surface	$\nabla G \cdot \boldsymbol{n}$
F	Fukui F^+ function scalar value	$F^+(r) = \rho_{HOMO}(r)$
L	Laplacian of the electron density	$L(r) = -\nabla^2\rho(r) = K(r) - G(r)$
BNP	Bare nuclear potential	$BNP(r) = \sum_{\alpha} \frac{Z_{\alpha}}{\lvert r - R_{\alpha} \rvert}$
PIP	Local average ionization potential	$PIP(r) = \sum_{i} \frac{\rho_i(r)\lvert\varepsilon_i\rvert}{\rho(r)}$

distributions and evaluates the electronic surface properties, which are used for generating descriptors. It contains a library of the atomic types in a form which can transfer electron density properties. The RECON program reconstructs the electronic density properties of a molecule by assigning the closest match from a library of atom types for each atom in the molecule. The additivity principle is applied to calculate molecular descriptors by summing up the individual descriptor type values for all atoms in the molecule, using the RECON method. Therefore it is possible to derive pseudo-molecular descriptors for any group of atoms, e.g., active site fragment, making the TAE descriptors well suited for our approach. Table 4.1 shows a complete list of TAE descriptors. The local average ionization potential (PIP) of the molecule, one example of the electronic surface properties is shown onto its 0.002 e/au^3 (electrons per cubic Bohr) electron-

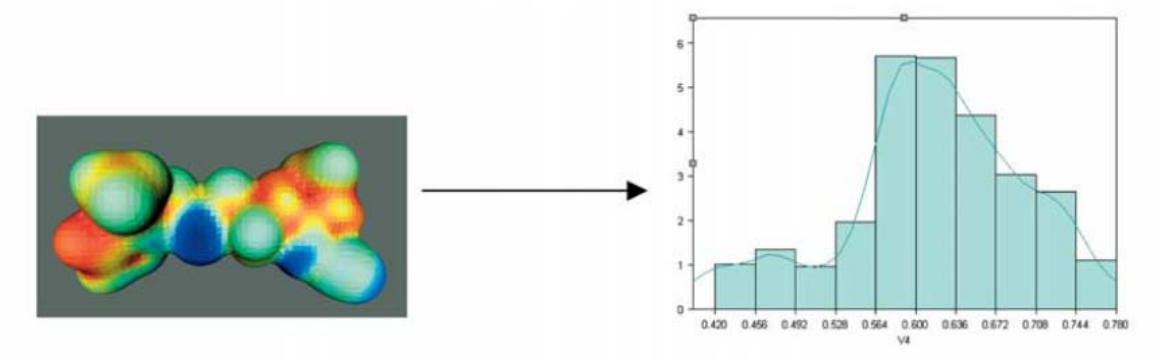


Figure 4.1 TAE local average ionization potential (PIP) surface property and its histogram distribution[9].

density surface in Figure 4.1. The distribution of this property is then presented as a histogram such as that shown on the right side of the figure. Each bin of the histogram is used as a descriptor, as well as statistical information such as maximum, minimum, and average of each surface property.

A computational geometry technique known as Delaunay tessellation is utilized to isolate receptor atoms that make contacts with bound ligands. Let us consider a collection of randomly distributed points in 2D (Fig. 4.2). By analogy, the red and blue dots represent the ligand atoms and the receptor atoms in the binding site, respectively. Delaunay tessellation partitions the space occupied by these points into a set of space filling, irregular triangles (tetrahedrons in 3D) with the original points as vertices. Therefore this method identifies all nearest neighbor triplets of vertices, including two types of interfacial triplets as shown in Figure 4.2: one ligand atom point and two receptor atom points; two ligand atom points and one receptor atom point. Applied to the 3D receptor-ligand complex case, it will generate three types of interfacial quadruplets: one ligand atom and three receptor atoms; two ligand atoms and two receptor atoms; three ligand atoms and one receptor atom. Therefore it provides a way of detecting all receptor atoms that are nearest neighbors of ligand atom. The TAE descriptors are then generated for a pseudo-molecule composed of these receptor atoms.

Using the TAE/RECON method, multiple descriptors as listed on Table 4.1 are generated for the receptor binding sites and their corresponding ligands so that each chemical entity is

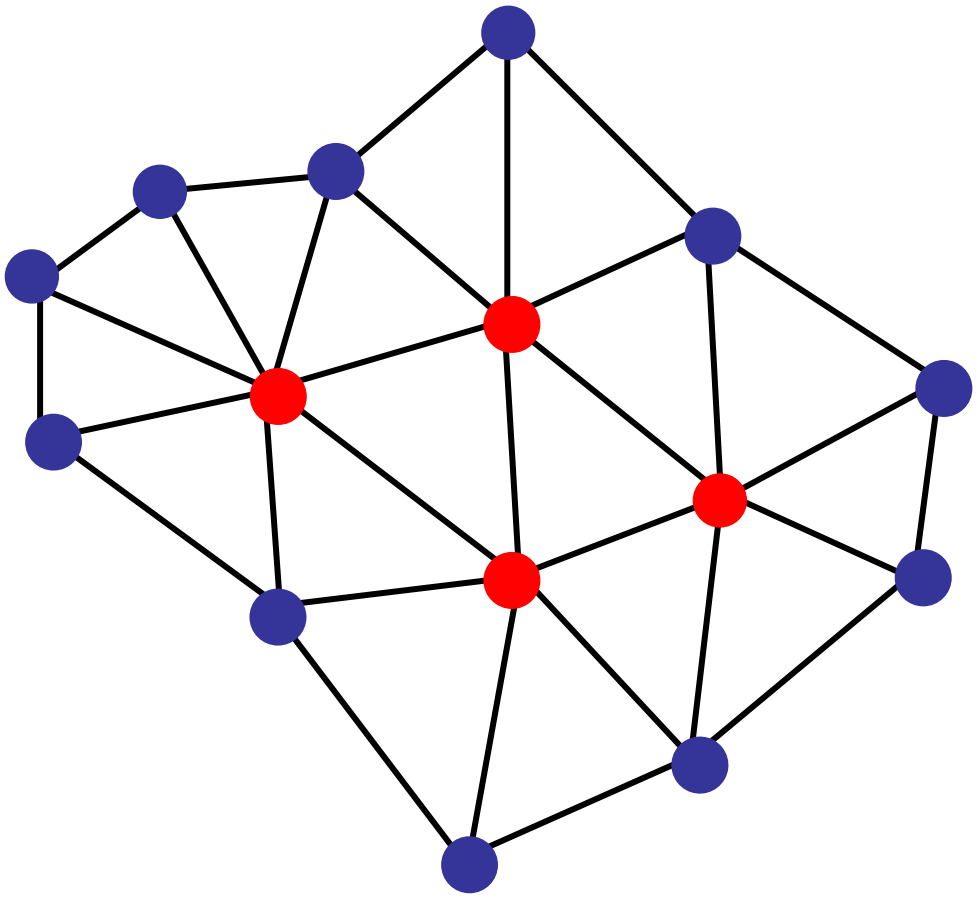


Figure 4.2 Delaunay tessllation of a collection of random points in 2D (modified from reference7)

represented as a vector in a multidimensional TAE/RECON chemical space. Since every descriptor may not be important for determining receptor-ligand complementarity, the subset of descriptors that best reflect this complementarity is determined, using a leave-one-out (LOO) cross-validation approach, in which each data value is left out in turn and a model derived using the remainder of the data. The overall procedure for selecting an optimal subset is as follows:

(1) A subset of n_{Var} descriptors (n_{Var} is a predefined number between 1 and the total number of available descriptor) is randomly selected.

(2) One of the receptors is chosen in the training set and the k nearest neighboring (kNN) receptors are selected in the n_{Var}-dimensional descriptor

space of the binding site. The coordinates of the chosen receptor's virtual ligand in the ligand space are predicted based on the relative orientation of ligands known to bind with the kNN receptors. This step is repeated until every receptor in the training set is eliminated once and all the receptor's virtual ligands are predicted. This resulting set of virtual ligands is called a CoLiBRI model.

(3) The predictive mean rank (PMR) for the model is calculated. It is related to the chemical similarity of the virtual ligands to the known ligands. The similarities are evaluated as Euclidean distances in the n_{Var}-dimensional descriptor space:

$$Dist_{i,j} = \sqrt{\sum_{d=1}^{n_{Var}} (X_{id} - X_{jd})^2}\ , \qquad \text{(Eq. 4.1)}$$

where X_{id} and X_{jd} are the dth selected descriptor for ligand i and j. The higher rank means the larger deviation of the model.

(4) Step 2 and 3 are repeated for all possible k values ($2 \le k \le$ total number of ligand - receptor pairs). The k values that leads to the lowest PMR value is chosen as optimal.

(5) The selection of n_{Var} descriptors is optimized based on simulated annealing. For a model built using randomly-sampled n_{Var} descriptors, the value of the fitness function, the inverse of its PMR value is calculated. By changing a fraction of the currently used descriptors to other randomly selected of n_{Var} descriptors, a new CoLiBRI model is generated for the new trial set (repeat steps 1 to 4) and the new corresponding fitness function is calculated. The new trial set is accepted or rejected based on the Metropolis criterion. This

Monte Carlo approach is continued as the temperature is lowered until the termination condition is satisfied.

At the end, both an optimum *k* value and an optimal subset of n_{Var} descriptors are determined and produce a model with the best predictive ability. More detailed mathematical expression is described in reference 7.

Now the CoLiBRI model is ready to be used for the ligand screening. First the target receptor is positioned in the selected descriptor subspace and its *k* nearest neighboring receptors from the training set are found. The known ligands of these *k* nearest neighboring receptors are then used to estimate the location of the target receptor's virtual ligand in the descriptor space in the same way as step 2 above. All ligands in the chemical library are ranked based on their distance to this predicted virtual ligand point (using Eq. 4.1), and the ligands with the smallest distance are considered as the most probable hit.

In our study the CoLiBRI models were generated for the dipeptide binding site using the same training set (670 complex structures from PDBbind[10]) used in reference 7 plus the predicted mMrgC11/R-F-OH complex structure.

4.2.2 Chemical libraries

Three sets of chemical libraries were screened in this study;

(1) The first set: An older version of the database from ChemDiv with 451,345 compounds was pre-screened using the CoLiBRI method. The multiple CoLiBRI models that predict complementarity were generated, varying the n_{Var} value, a number of selected descriptors used in generating a CoLiBRI model as described in section 2.1. The compound within the top 1,000 by at least one model was selected and total 3,900 compounds were collected for the next docking step.

(2) The second set: It was taken from a newer version (fall, 2004) of the database from ChemDiv with 513,000 compounds. We selected compounds that were consistently predicted to be within the top 1,000 by all models. This resulted in 442 hits.

(3) The third set: The 23 drug compounds known for producing pain relief were docked without any pre-screening. It includes some opiates (e.g. Demerol), local anesthetics (e.g. Lidocaine) and capsaicin (an agonist of vanilloid receptors in dorsal root ganglion (DRG)). All possible protonation states were considered, leading to a total of 43 ligand structures for docking.

For the pre-screened compounds from the first and second set, hydrogen atoms were added and Gasteiger charges[11] were assigned using Concord program. No further optimization was carried out before docking. For the third ligand set, Gasteiger charges were assigned and the structures were optimized in gas phase using conjugate gradient minimization using the DREIDING force field (FF)[12] on Cerius2[13].

The pre-screening of ChemDiv database for the di-peptide binding site was performed in collaboration with the Tropsha group of the University of North Carolina.

4.2.3 Molecular docking

MSC-Dock program was used for docking the pre-screened ligands. We used the Dock-Diversity Completeness protocol (DDCP). As described in chapter 2, DDCP attempts to generate a complete set of ligand configuration families with a fixed coordinate diversity. In this study the diversity was set to 0.6 Å. The rejection ratio (defined as the fraction of new configuration that belongs to previously generated families to the fraction that leads to a new family) was set to 2.2. The 50 families were selected with the best energies (by DOCK4.0 energy score) in the first phase and an average of six members in each family was generated in the second enrichment

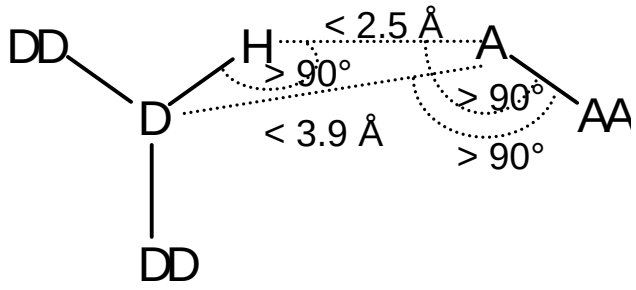


Figure 4.3 Geometric criteria for the hydrogen bonds. D is the donor heavy atom, H the hydrogen, A the acceptor, DD donor antecedent (i.e. an atom two covalent bonds away from the hydrogen) and AA acceptor antecedent.

phase. The final ~300 configurations were ordered by DOCK4.0 energy score and re-clustered with 0.6 Å of diversity to generate a new set of families. The top 5 family heads (a member with the best energy in each family) were conjugate gradient minimized (100 steps or 0.1 kcal/mol/Å of RMS force) with the ligand atoms movable and the receptor atoms fixed. Then the binding energies were then calculated for these 5 optimized ligand-receptor complex configurations. The calculated binding energy (BE) is defined by

BE = E (ligand in fixed protein) – E (ligand in water),

where the E (ligand in fixed protein) is the potential energy of the ligand calculated in the ligand-receptor complex with the coordinates of the receptor fixed. This potential energy includes the internal energy of the ligand and the interaction energy of the ligand with the receptor. E (ligand in water) is the potential energy of the free ligand in its docked conformation (snap bind energy) and its solvation energy calculated using the analytical volume generalized born (AVGB) continuum solvation method (cavity_params_1.3)[14]. The final best ligand-receptor structure was selected as the one with the most negative binding energy.

4.2.4 Selection of final hits

The ligands in the final docked conformation were sorted by three criteria; the binding energy, the van der Waals interaction energy and the energy of hydrogen bond between the

receptor and the ligand. The intermolecular hydrogen bond was determined by the geometric criteria shown in Figure 4.3[15] and its energy was evaluated using the DREIDING FF. For the first ligand set, the top 100 ligand compounds were chosen by each sorting criterion. Then we selected the compounds that were consistently within the top 100 by at least two criteria. This led to total 52 compounds. These selected compound structures were further optimized in the protein-ligand complex. The side chain conformation of receptor residues within 5 Å from the ligand was optimized using the SCREAM program and then the entire receptor-ligand complex structure was conjugate gradient minimized with 0.1 kcal/mol/Å of RMS force. This receptor-ligand complex was further refined using one cycle of annealing MD heating from 50 K to 600 K and cooling down back to 50 K in 50 K steps, with 1 ps of equilibration between temperature jumps. Here only the ligand and the receptor side chains within 5 Å of the binding pocket were allowed to move during the annealing cycle. At the end of the annealing cycle, the system was minimized to an RMS force of 0.3 (kcal/mol)/Å. The binding energy was then re-calculated for the final complex structure in the same way as described above.

The compounds in the second set were also sorted by three same criteria and the common compounds within the top 40 were selected. The 40th best binding energy is the halfway between the highest one and zero. This resulted in 21 compounds, which were optimized further as in the first set.

The pain-related compounds in the third set were sorted by their binding energy and the top 10 compounds were chosen, then the same post-optimization was carried out.

Both the protein and the ligand were described using the DREIDING FF and the protein charges were from CHARMM22[16]. All calculations used the MPSIM program[17], with nonbond interactions evaluated using the cell multipole method[18]. All simulations were performed in gas phase with the dielectric constant of 2.5.

After post-optimization, the residues of the receptor having either an intermolecular hydrogen bond or good van der Waals contact with the ligand were identified. By putting the priority on compounds having good contacts with the key residues—Tyr110 (TM3), Asp161 (TM4) and Asp179 (TM5)—26 compounds were finally chosen for experimental test. They included four outliers in the docking step to expand diversity and two pain-related compounds, capsaicin and ibuprofen.

4.2.5 Intracellular calcium release assay

The intracellular calcium release assay experiment was carried out to test activity for 26 compounds (the details are described in chapter 2). One of the known peptide agonists, F-M-R-F-NH_2 (EC50 = 168nM) was used as a control compound. To test agonistic activity, cells expressing stably mMrgC11 receptor proteins were treated with compounds in two different concentrations, 100 μM and 10 μM. To check antagonistic activity, cell sample was pre-incubated for >5min with a compound in 100 μM and 10 μM concentration and then were treated with 1 μM of F-M-R-F-NH_2. The inhibitory constant 50% (IC50), the concentration reducing the activity of 400 nM F-M-R-F-NH_2 by half was measured for the compounds showing the antagonistic effect in two ways. First, cells were pre-incubated with a compound in various concentrations and F-M-R-F-NH_2 was added later. Secondly, the compound was added to the cell sample together with F-M-R-F-NH_2 at the same time and the intracellular calcium release was measured.

4.2.6 Virtual screening of tetra-peptide binding site

The virtual screening for the tetra-peptide binding site was independently carried out in a similar way. Since the loops were in the ensemble of conformations as shown in chapter 3, the extracellular loops in the mMrgC11 receptor were not included in screening. The dataset of 800 ligand-receptor complexes from the PDBbind Database (PDB entry codes are listed in the

supporting information of reference 7) was divided into the training (used for model building; 525 structures) and the test (used for model validation; 275 structures) sets using the sphere exclusion method[19]. In building CoLiBRI models, six predicted Mrg complex structures were included in the training set; mMrgC11/(D)F-M-R-F-NH_2, mMrgC11/F-M-R-F-NH_2, mMrgC11/F-(D)M-R-F-NH_2, mMrgC11/R-F-NH_2, mMrgC11/R-F-OH and rat MrgA/adenine complex. The CoLiBRI models differ depending on the number of descriptors (4 to 40) and the content of a given number (10 content variations). Among these 370 models the top 100 models were chosen based on the PMR values for the test set of 275 receptors.

The first set of chemical library used in the previous dipeptide case was screened for the mMrgC11 receptor optimized with the bound F-(D)M-R-F-NH_2, which is the best known tetra-peptide agonist. Five F-M-R-F-NH_2 peptides (three agonists and two non-agonists), R-F-NH_2 and R-F-OH were included into the ChemDiv database, leading to total 451,352 compounds. The top 1,000 compounds were selected for each model. The models having (D)F-M-R-F-NH_2, F-M-R-F-NH_2 and F-(D)M-R-F-NH_2 as a hit after screening were identified, resulting in 92 out of 100 models. The 4,735 compound hits from the ChemDiv database were predicted by at least one of 92 models and the 16 compound hits were consistently predicted by all 92 models. However F-M-(D)R-F-NH_2 was also consistently recognized as a hit for all 92 models (false positive), indicating that the CoLiBRI model is not sensitive enough to completely distinguish between the chirally modified tetrapeptide agonists and non-agonists. Nevertheless identification of three agonists as hits provides some validation of the CoLiBRI models used in this study.

The 774 compound hits which were consistently predicted by at least 50 models were chosen for the next docking step. We also used MSC-Dock with the same parameters except for the diversity of 1.0 Å since the size (number of atoms) of hit compounds in the tetra-peptide binding site is larger than those in the di-peptide binding site.

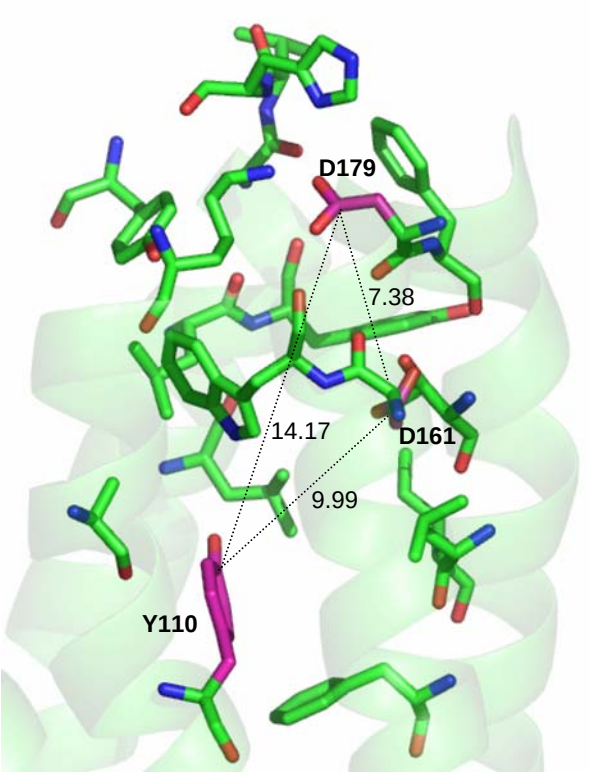


Figure 4.4 5 Å binding pocket of mMrgC11 receptor optimized with the di-peptide agonist, R-F-OH. Three key residues (Y110, D161 and D179) are identified and inter-residue distances are specified in Å for those residues.

Following the same scoring method and selection criteria (the top100 were selected for each criterion – binding energy, van der Waals interaction and hydrogen bond energy (the calculated binding energy = -41.77 to 431.47 kcal/mol; the 100th is approximately halfway between -41.77 to 0)), final 55 compounds were identified out of 774. Then these 55 complex structures were optimized in the same way as described section 2.4.

We docked F-(D)M-R-F-NH_2 with the same docking parameters and scoring method. The RMSD of the best configuration was 0.29 Å with respect to the previously predicted "true" bound configuration, validating our docking procedure.

4.3 Results and discussion

4.3.1 Hit compounds from virtual screening

Figure 4.4 shows the 5 Å binding site of the mMrgC11 receptor complexed with one of dipeptide agonists, R-F-OH. This dipeptide optimized structure was used for both pre-screening and docking. Three key residues were previously identified in the R-F dipeptide binding. Tyr110 had a good π-π interaction with F of the dipeptide, and two Asp residues, Asp161 and Asp179 interacted favorably with the sidechain of R and the N-terminus. The final hit compounds after virtual screening were listed in Figure S4.1 for the first ligand set and in Figure S4.2 for the second one. The ligand atoms forming hydrogen bonds with the receptor were specified. The contribution of each receptor residue to the van der Waals interaction was evaluated and the residues for which the absolute value of the interaction energy was larger than 3 kcal/mol were identified. Most of ligands had at least one aromatic ring, which replaced the phenyl ring of R-F dipeptide and interacted with nonpolar residues present inside the pocket such as Tyr110, Phe190 and Leu186. Some of ligands formed a hydrogen bond with Asp161 or/and Asp179, but none of the hydrogen bond partners were similar to the arginine sidechain.

By comparing the hit compounds from the first set with those from the second set, we could see that selection of the compounds consistently predicted by all CoLiBRI models provided a ligand with the higher binding energy showing better chemical contacts (i.e. contacts with all key residues) although the hit compounds showed less diversity. MOL282 (the ligand with the best binding energy in the second set) showed better binding by 6 kcal/mol than Mol2190 (the best one in the first set) and made contacts with Tyr110, Asp161 and Asp179.

Among the pain-related compounds, capsaicin and ibuprofen showed the best binding energy in docking. The binding energies were -45.11 and -43.14 kcal/mol respectively. The van der Waals interaction mainly contributed to the binding energy. Capsaicin formed a single hydrogen bond with Asp161 and ibuprofen does not have any contact with three key residues.

4.3.2 Experimental activity test

Table 4.2 Inhibitory constant 50% (IC50) of hit compounds (unit: μM)

	A	B
MOL282	46.5 ± 2.2[a]	74.6 ± 0.1[b]
capsaicin	26.0 ± 2.7[a]	N.A.
capsazepine	19.2 ± 5.9[b]	N.A.
dihydrocapsaicin	46.6	N.A.
N-vanillylnonamide	69.7 ± 17.7[b]	N.A.

A – pre-incubate a compound and then add 400 nM of F-M-R-F-NH_2, B – add a compound and 400 nM of F-M-R-F-NH_2 at the same time.
[a] mean ± SEM from triplicate independent measurements, [b] duplicate measurements
N.A.: no significant decrease in activity of F-M-R-F-NH_2 agonist is observed in >200 μM concentration.

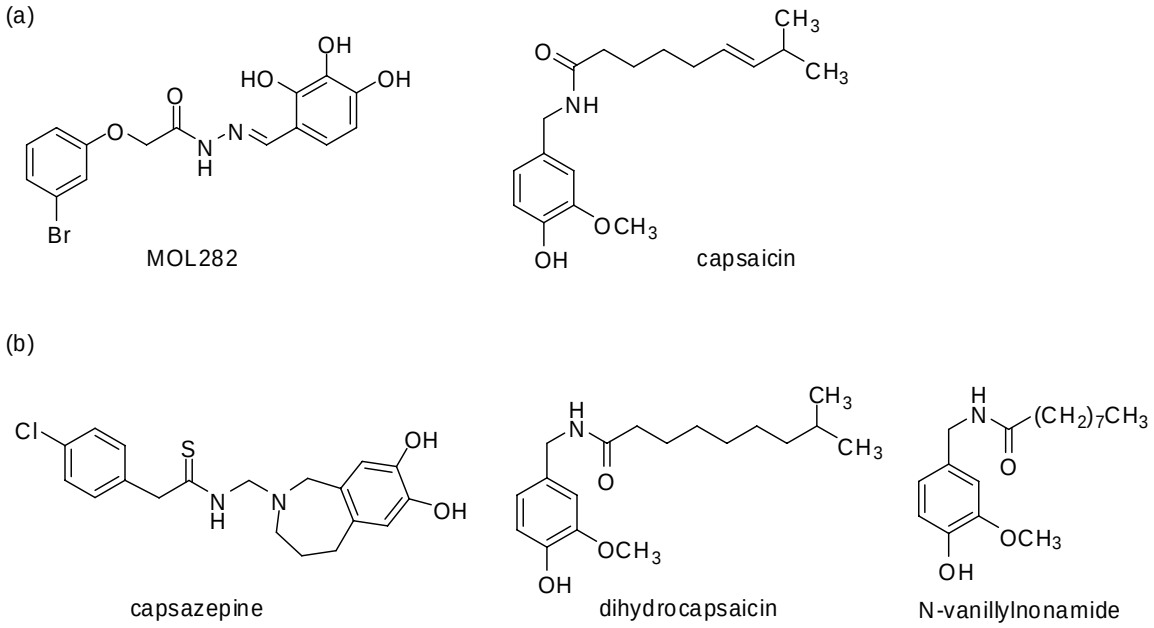


Figure 4.5 Compounds showing the inhibitory effect (a) from the hit compound set of VLS and (b) among the tested capsaicin analogs.

The agonistic activity for total 26 compounds (24 from the virtual screening plus capsaicin and ibuprofen) were tested using the intracellular calcium assay. The mMrgC11 receptor was activated by none of them up to 100 μM concentration. However some of them showed the inhibitory effect – blocking the activity of the known agonist, F-M-R-F-NH_2. Two compounds, MOL282 and capsaicin shown in Figure 4.5(a) blocked the activity of F-M-R-F-NH_2. The measured IC50s of MOL282 are 47 μM for pre-incubation case and 75 μM for simultaneous addition (Table 4.2). It means that MOL282 binds to the mMrgC11 receptor kinetically at the rate

comparable to F-M-R-F-NH_2. However capsaicin could not block the activity of the agonist when it was added together with the agonist at the same time, indicating that it is a slow binder than F-M-R-F-NH_2.

MOL282 was predicted to have the best binding energy from our virtual screening, and this experimental result provides the strong evidence that our predicted mMrgC11 structure is accurate enough to screen chemical libraries for potential ligands. Capsaicin is a well-known agonist of vanilloid receptor type 1 (VR1), which functions as a molecular integrator of painful chemical and physical stimuli[20]. Although Dong *et al.* claimed that mMrgAs and mMrgD were expressed in the $VR1^-$ sensory neurons[21], we could observe that capsaicin was able to inhibit the activity of a known agonist in the mMrgC11 receptor. Next we extended the experiment to capsaicin analogs, and five commercially available analog compounds were tested (capsazepine, dihydrocapsaicin, olvanil, N-vanillylnonamide and eugenol). Among five, three compounds showed antagonistic effect at the tens micromolar concentration. Their chemical structures are shown in Figure 4.5(b).

4.3.3 Refined docking of MOL282 and design of its derivatives

We docked the lead compound, MOL282 again into the mMrgC11 receptor in a more refined docking scheme. The conformations of MOL282 were extensively explored using the grid sampling method. Five torsion degrees of freedom were sampled by 60° steps from the initial optimized structure, leading to total 7,776 conformations. These conformations were ranked by the force field energy in gas phase and clustered with 1.0 Å of diversity. This resulted in the set of final 87 conformations. Each conformation was docked independently into the same binding region without further optimization.

The MSC-Dock with DDCP was used for docking as described in section 2.3. Here the top 25 families (instead of 5) were chosen and optimized with the receptor coordinates fixed. They were ranked by binding energy and then the top 10 configurations were determined. These 10

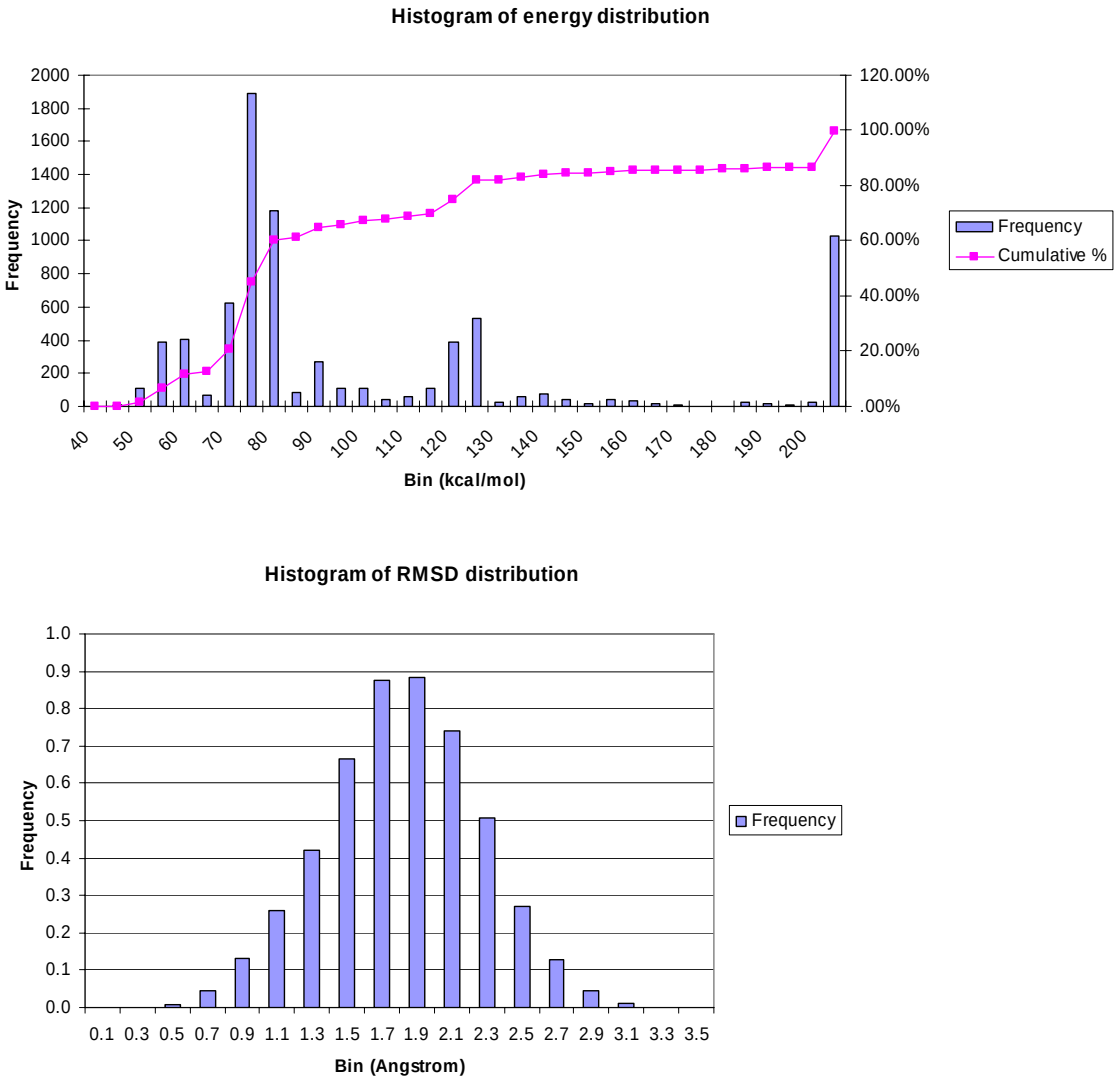


Figure 4.6 Histograms of energy and RMSD distribution for 7,776 conformations of MOL282 in grid search. The pair-wise RMSD is calculated with heavy atoms only.

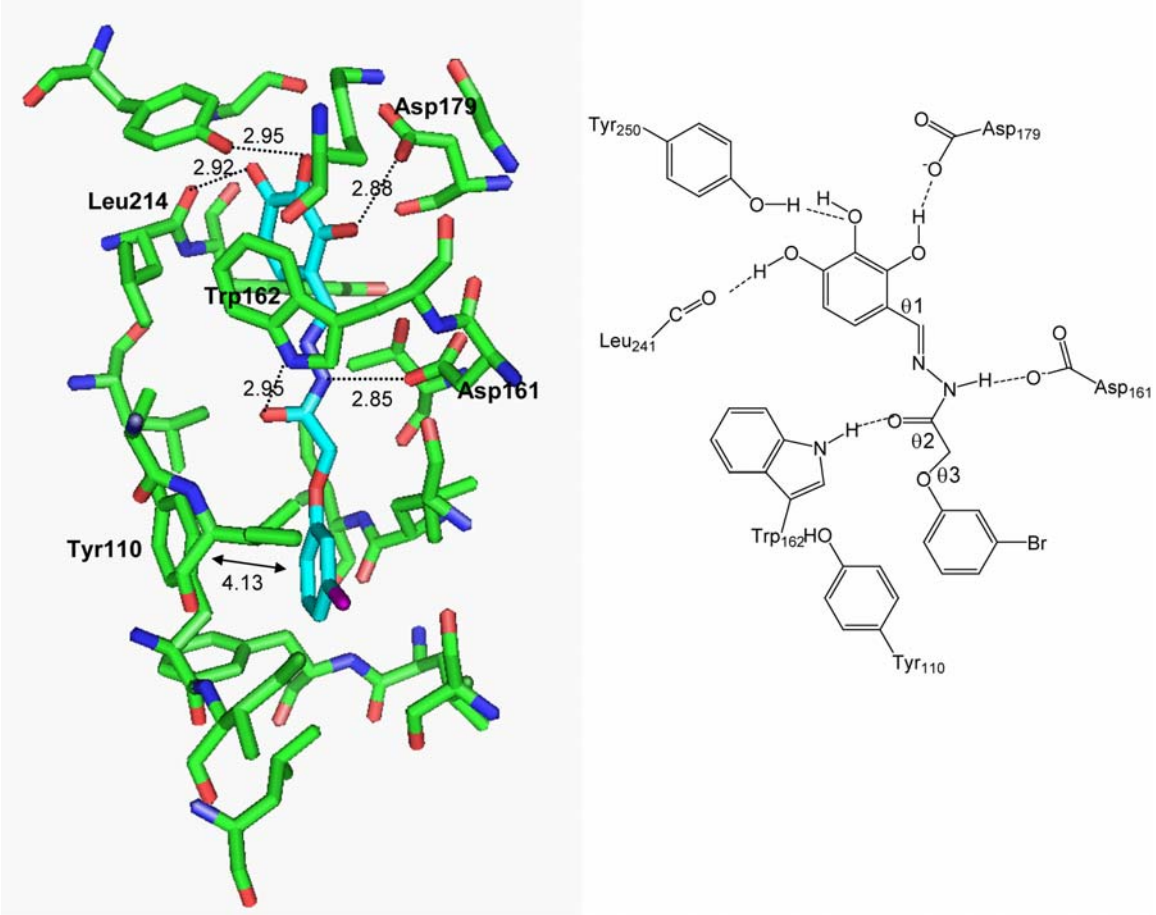
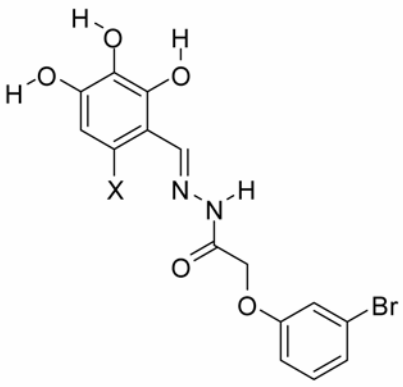


Figure 4.7 The 5 Å binding pocket of MOL282 in mMrgC11 receptor. The hydrogen bond and inter-aromatic ring distance are specified in Å.

receptor/ligand complex structures were further optimized with the conjugate gradient minimization while all atoms were movable. The final structure was then chosen with the best binding energy. Therefore we ended up with 87 optimized complex structures. No further optimization such as the sidechain replacement and annealing MD was carried out.

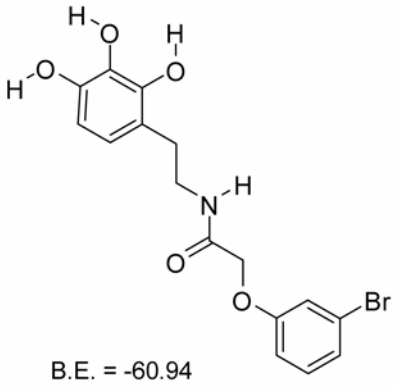
The best binding configuration across the 87 optimized structures is shown in Fig. 4.7. All three key residues interact with the ligand; Asp161 and Asp179 form hydrogen bonds with the ligand and Tyr110 participates in the π-π interaction with one of aromatic rings. Trp162, Leu241 and Tyr250 form the hydrogen bonds with the carbonyl group and the other hydroxyl groups of the ligand. However the ligand had the strain energy of ~15 kcal/mol (energy in gas phase with

(a)



X	B.E.
H	-58.81
CF3	-62.22
CH3	-63.32
ethyl	-63.93
isopropyl1	-63.33
isopropyl2	-61.70
t-butyl	-59.12

(b)



B.E. = -60.94

Figure 4.8 Suggested better binders derived from MOL282. The binding energy is in kcal/mol.

the dielectric constant of 2.5) in the docked conformation. Most strain resulted from the twist in θ1 torsion of Figure 4.7 (θ1=180° in the global minimum). To stabilize this twisted configuration, the substitution of a bulky group for the ortho hydrogen was suggested as shown in Figure 4.8(a). This bulky group also enhanced the van der Waals interaction with the receptor, leading to the increase of the binding energy. However as it became too bulky to occupy the void space in the binding pocket, it interfered binding (see the table in Fig. 4.8(a)).

Since nitrogen in C=N bond of MOL282 does not play a role in binding, C=N double bond was replaced by C-C single bond to reduce the strain seen in the docked configuration of MOL282 (Fig. 4.8(b)). This derivative of MOL282 binds to the mMrgC11 receptor similarly to MOL282, except that one of hydrogen bond partners was switched from Tyr250 to Lys99. The

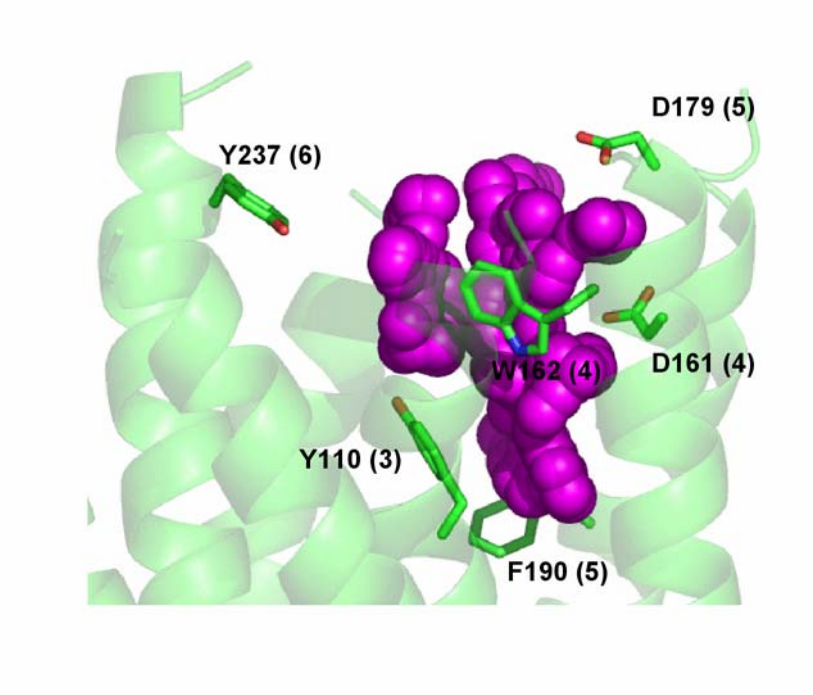


Figure 4.9 The 5 Å binding pocket of mMrgC11 receptor optimized with the tetra-peptide agonist, F-(D)M-R-F-NH_2. Six key residues identified in the previous prediction are shown in stick. The spheres representing the binding site of F-(D)M-R-F-NH_2 are colored by magenta.

strain energy of the ligand in the docked configuration decreased by ~7 kcal/mol and the snap binding energy slightly increased by ~5 kcal/mol, leading to the similar relaxed binding energy where the strain penalty was taken into account.

4.3.4 Virtual screening for F-(D)M-R-F-NH_2 bound site

The 5 Å binding site of F-(D)M-R-F-NH_2 is shown in Figure 4.9. Compared with the di-peptide binding site in Figure 4.4, the site is obviously wider. The buried surface was calculated using the Connolly MS program from Quantum Chemistry Program Exchange (QCPE) with a probe radius of 1.4 Å and a surface density of 5 dots/$Å^2$. The area for the buried part of F-(D)M-R-F-HN_2 was 466 $Å^2$, which was larger than 263 $Å^2$ for R-F-OH. The N-terminal F-(D)M part was extended towards TM6 and TM7, covering the additional TM regions. Tyr237 (TM6) is one of the key residues newly identified in the tetra-peptide binding site.

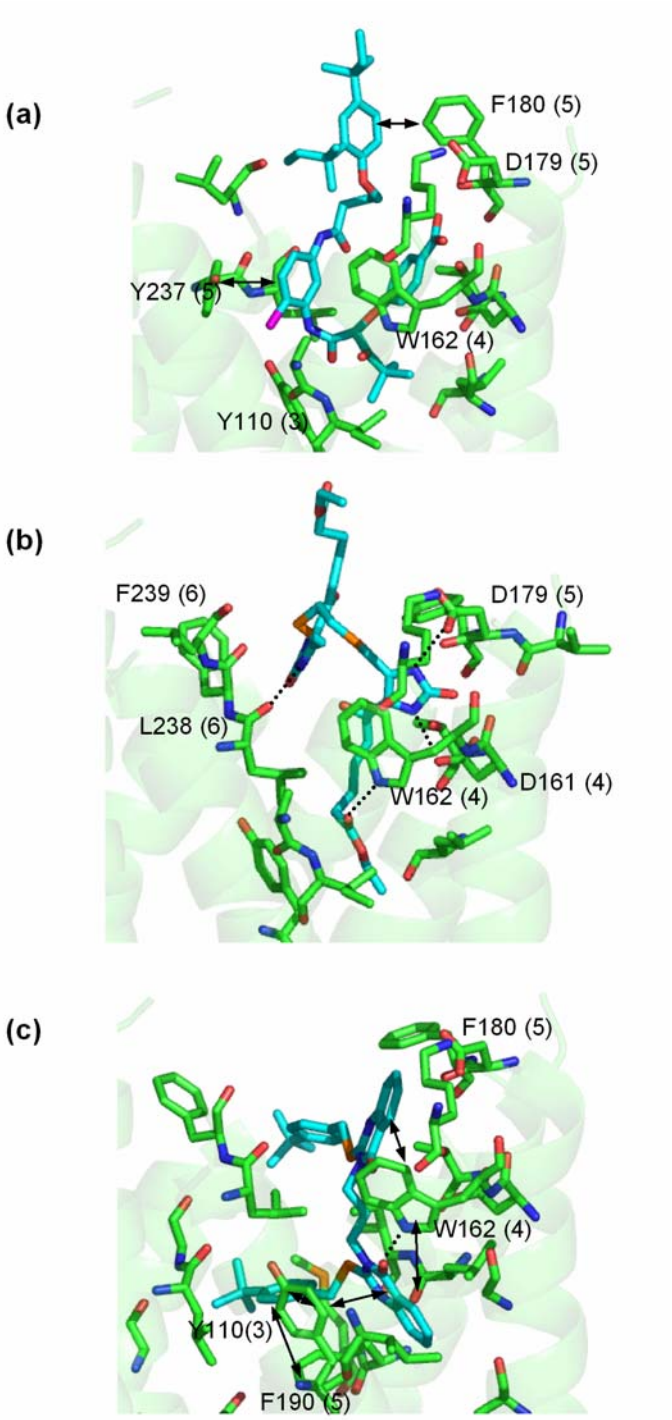


Figure 4.10 The 5 Å binding site of the best three hit compounds (a) comp242755 (b) comp241282 (c) comp391008. The intermolecular hydrogen bond is indicated by the dotted line and the aromatic interaction by the two-sided arrow.

The chemical structures of the final 55 hit compounds are shown in Figure S4.3, where the residues making a hydrogen bond or having a good van der Waals interaction (interaction energy with a ligand is greater than 3 kcal/mol) are identified together. Most are bulky since the surface area is considered as one of the descriptors, and relatively nonpolar compounds. They belong to the different class of compounds compared with those screened previously in the di-peptide case.

The detailed binding modes of the compounds with the best (comp242755), the second best (comp241282) and the third best (comp391008) binding energy are described in Figure 4.10. In comp242755, three aromatic rings interact with Trp162 (TM3), Phe180 (TM5) and Tyr237 (TM6). The *t*-butyl group has a favorable hydrophobic interaction with Tyr110 (TM3). The side chains of Trp162 (TM4) and Asp179 (TM5) are involved in the formation of hydrogen bond. However the hydrogen bond with Asp179 is unlikely if the carboxylate group in the benzoic acid part of comp242755 is deprotonated (pKa of benzoic acid = 4.20 for water at 25 °C). Since the buried receptor site might provide the different dielectric medium, the neutral form of comp242755 could be taken into account.

In comp24282, two key residues, Asp161 and Asp179 form hydrogen bonds with the ligand. Only two residues are shown to have good van der Waals interaction, but the ligand form two more hydrogen bonds with Trp162 (TM4) and Leu238 (TM6).

The comp391008 interacts with the receptor mainly through the hydrophobic interactions. The aromatic groups are well stacked with Phe190 (TM5), Tyr110 (TM3), Trp162 (TM4) and Phe180 (TM5). Asp161 and Asp179 do not interact with the ligand and are stabilized through the hydrogen bond or electrostatic interaction with Thr183 and Lys99 respectively as shown in the apo protein.

Although the hit compounds do not form as many hydrogen bonds as F-(D)M-R-F-NH_2, the nonpolar character would relieve the desolvation penalty in aqueous solution to help binding to the buried pocket of the receptor.

4.4 Summary and conclusions

The virtual screening with the combination of QSPR and docking method was carried out for the predicted mMrgC11 receptor. The antagonist ligand, MOL282 (IC50 = 46.5 μM) that had the best calculated binding energy was identified by mining ChemDiv database for the di-peptide binding site. The interactions with Asp161, Asp179 and Tyr110 shown in the agonist binding were also observed in MOL282. The novel ligands were derived from MOL282 in getting rid of the strain energy in its docked conformation. The identification of MOL282 as a hit provides the strong validation of our predicted binding site and low trial and error in the experiment (only 24 compounds were tested) demonstrates efficiency of our virtual screening method.

The different class of compounds was identified in virtual screening for the tetra-peptide binding site, having a large contribution of van der Waals interaction to the binding affinity. The experimental test of some of the top compounds would be needed to provide further validation.

The hit compounds identified in this study are certainly good staring points in designing new agonists or antagonists for the mMrgC11 receptor, and variation on the functional group in the series of ligands could be used to characterize the binding pocket. Moreover chemical characteristics of the hit compounds could provide some clues in deorphanizing Mrg receptors.

References

1. Klabunde, T. and G. Hessler, *Drug design strategies for targeting G-protein-coupled receptors.* Chembiochem, 2002. **3**(10): p. 929-944.
2. Rohrer, S.P., et al., *Rapid identification of subtype-selective agonists of the somatostatin receptor through combinatorial chemistry.* Science, 1998. **282**(5389): p. 737-740.
3. Flohr, S., et al., *Identification of nonpeptidic urotensin II receptor antagonists by virtual screening based on a pharmacophore model derived from structure-activity relationships and nuclear magnetic resonance studies on urotensin II.* Journal of Medicinal Chemistry, 2002. **45**(9): p. 1799-1805.
4. Kollman, P., *Free-Energy Calculations - Applications to Chemical and Biochemical Phenomena.* Chemical Reviews, 1993. **93**(7): p. 2395-2417.
5. Aqvist, J., V.B. Luzhkov, and B.O. Brandsdal, *Ligand binding affinities from MD simulations.* Accounts of Chemical Research, 2002. **35**(6): p. 358-365.
6. Tropsha, A., *Recent Trends in Quantitative Structure-Activity Relationships*, in *Burger's medicinal chemistry and drug discovery*, D.J. Abraham, Editor. 2003, John Wiley & Sons: New York. p. 49-77.
7. Oloff, S., et al., *Chemometric analysis of ligand receptor complementarity: Identifying complementary ligands based on receptor information (CoLiBRI).* Journal of Chemical Information and Modeling, 2006. **46**(2): p. 844-851.
8. Breneman, C.M., et al., *Electron-Density Modeling of Large Systems Using the Transferable Atom Equivalent Method.* Computers & Chemistry, 1995. **19**(3): p. 161-&.
9. Breneman, C.M., et al., *New developments in PEST shape/property hybrid descriptors.* Journal of Computer-Aided Molecular Design, 2003. **17**(2): p. 231-240.

10. Wang, R.X., et al., *The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures.* Journal of Medicinal Chemistry, 2004. **47**(12): p. 2977-2980.

11. Gasteiger, J. and M. Marsili, *Iterative Partial Equalization of Orbital Electronegativity - a Rapid Access to Atomic Charges.* Tetrahedron, 1980. **36**(22): p. 3219-3228.

12. 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.

13. *Cerius2 Modeling Environment, Release 4.0*, Accelrys Inc.: San Diego.

14. Zamanakos, G., *A fast and accurate analytical method for the computation of solvent effects in molecular simulations*, in *Division of Physics, Mathematics and Astronomy*. 2002, California Institute of Technology: Pasadena.

15. McDonald, I.K. and J.M. Thornton, *Satisfying Hydrogen-Bonding Potential in Proteins.* Journal of Molecular Biology, 1994. **238**(5): p. 777-793.

16. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.

17. Lim, K.T., et al., *Molecular dynamics for very large systems on massively parallel computers: The MPSim program.* Journal of Computational Chemistry, 1997. **18**(4): p. 501-521.

18. Ding, H.Q., N. Karasawa, and W.A. Goddard, *Atomic Level Simulations on a Million Particles - the Cell Multipole Method for Coulomb and London Nonbond Interactions.* Journal of Chemical Physics, 1992. **97**(6): p. 4309-4315.

19. Golbraikh, A., et al., *Rational selection of training and test sets for the development of validated QSAR models.* Journal of Computer-Aided Molecular Design, 2003. **17**(2): p. 241-253.

20. Szallasi, A. and P.M. Blumberg, *Vanilloid (capsaicin) receptors and mechanisms.* Pharmacological Reviews, 1999. **51**(2): p. 159-211.

21. 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.

Supporting Figures

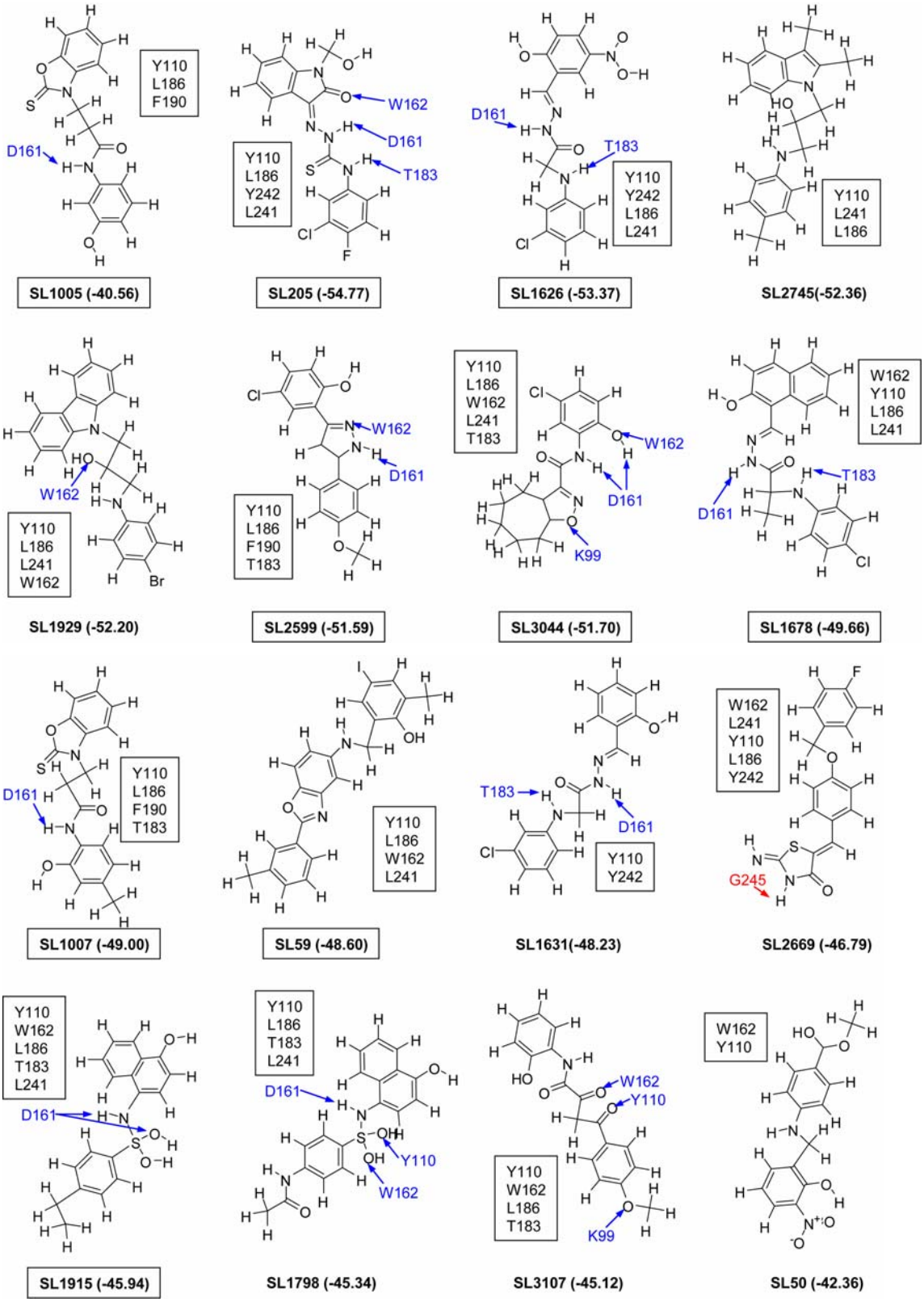


Figure S4.1 Hit compounds from the first ligand set after docking.

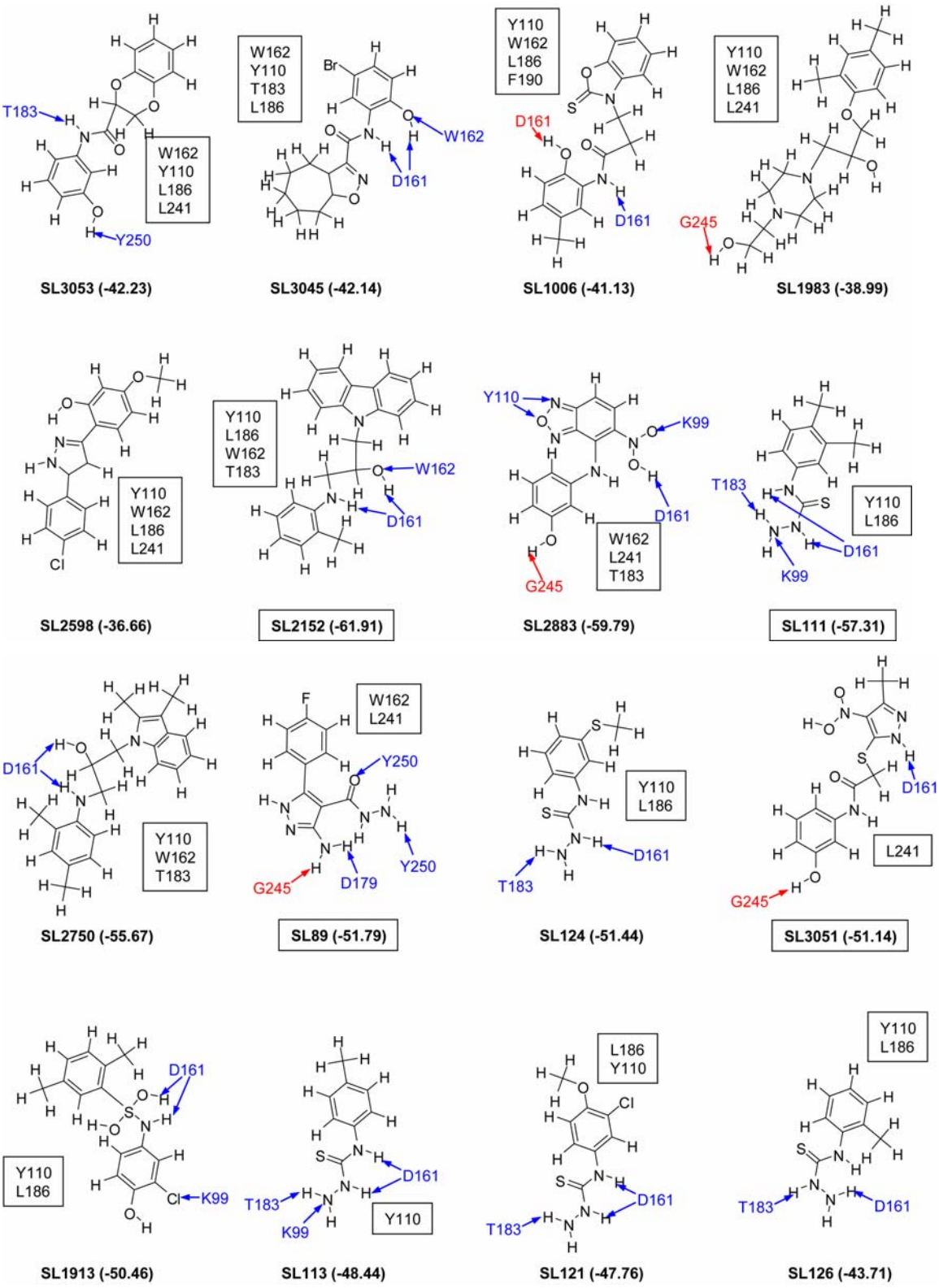


Figure S4.1 (*continued*) Hit compounds from the first ligand set after docking.

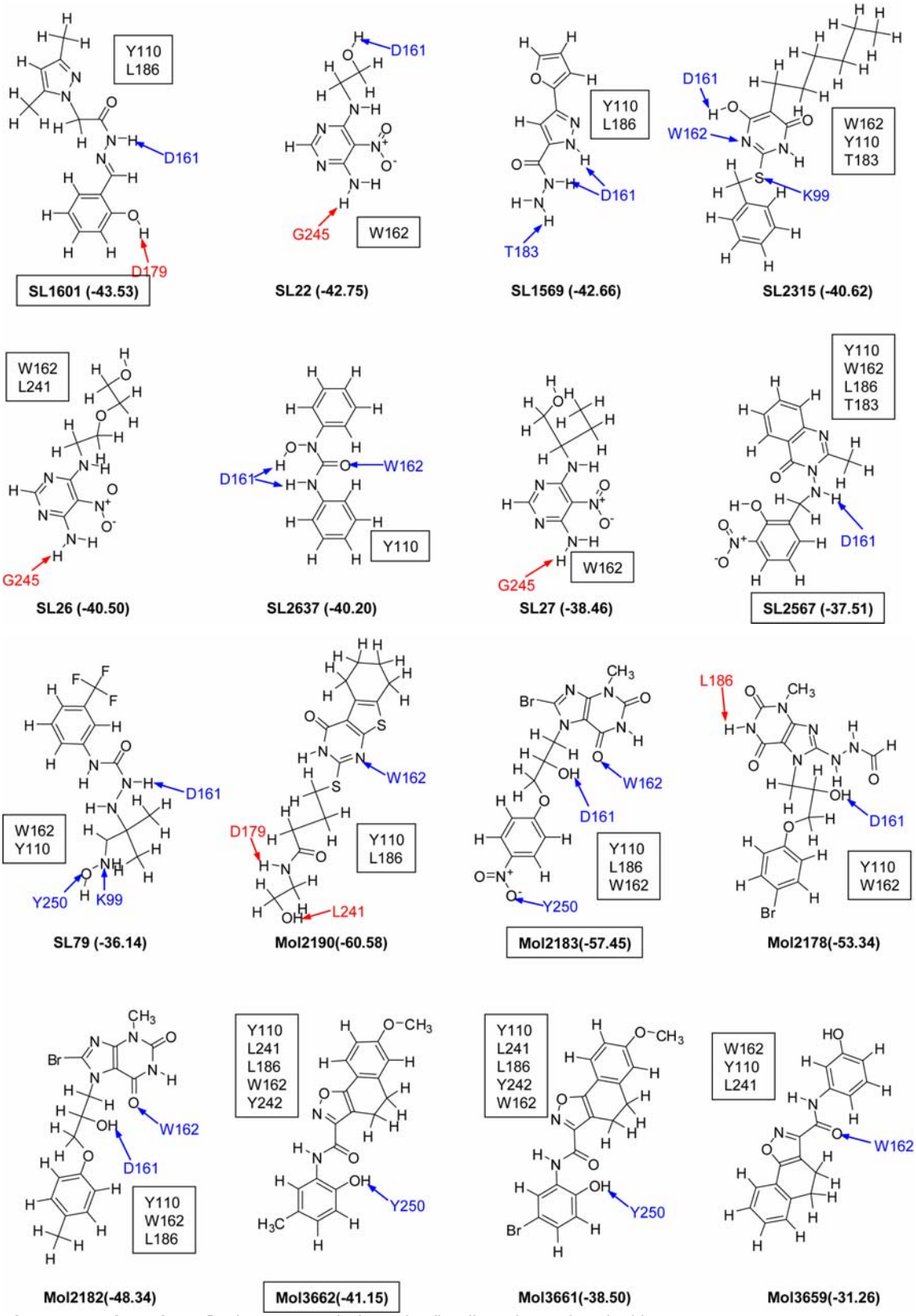


Figure S4.1 (*continued*) Hit compounds from the first ligand set after docking.

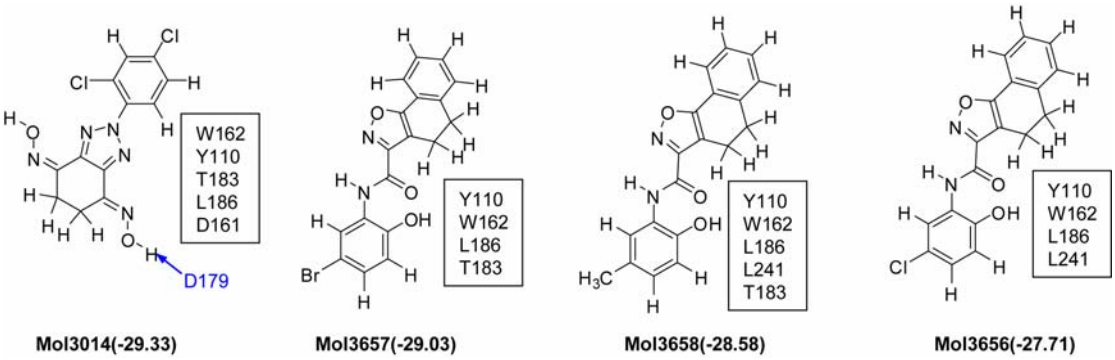


Figure S4.1 (*continued*) Hit compounds from the first ligand set after docking. The ligands whose names are enclosed by rectangular box were tested in experiment. The number in parenthesis corresponds to the calculated binding energy in kcal/mol. Residue in blue makes a hydrogen bond through its side chain with the atom indicted by the blue arrow. The residue in red has backbone atoms involved in the hydrogen bond. The residues in box have good van der Waals interactions with a ligand (E > 3 kcal/mol).

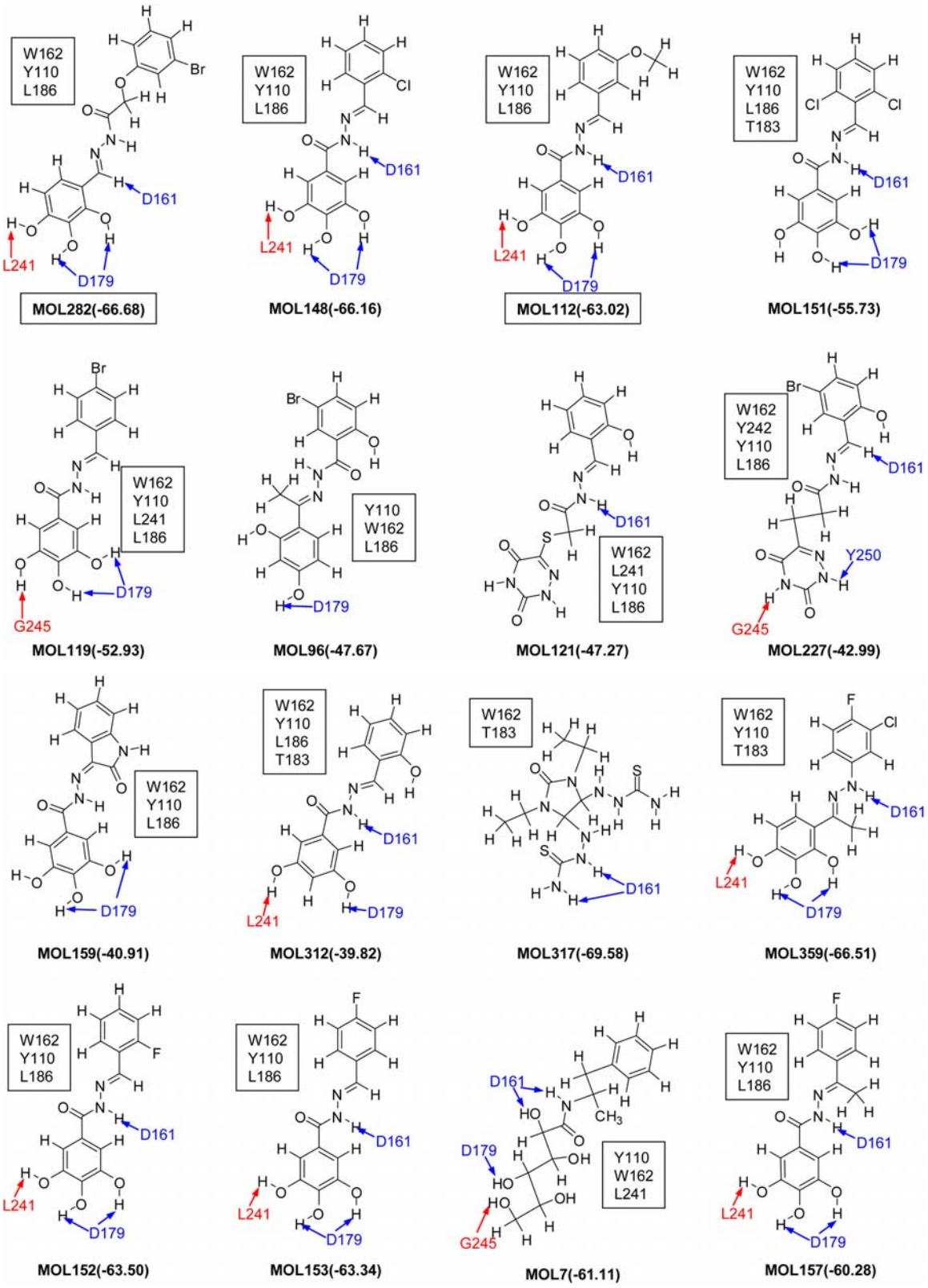


Figure S4.2 Hit compounds from the second set after docking.

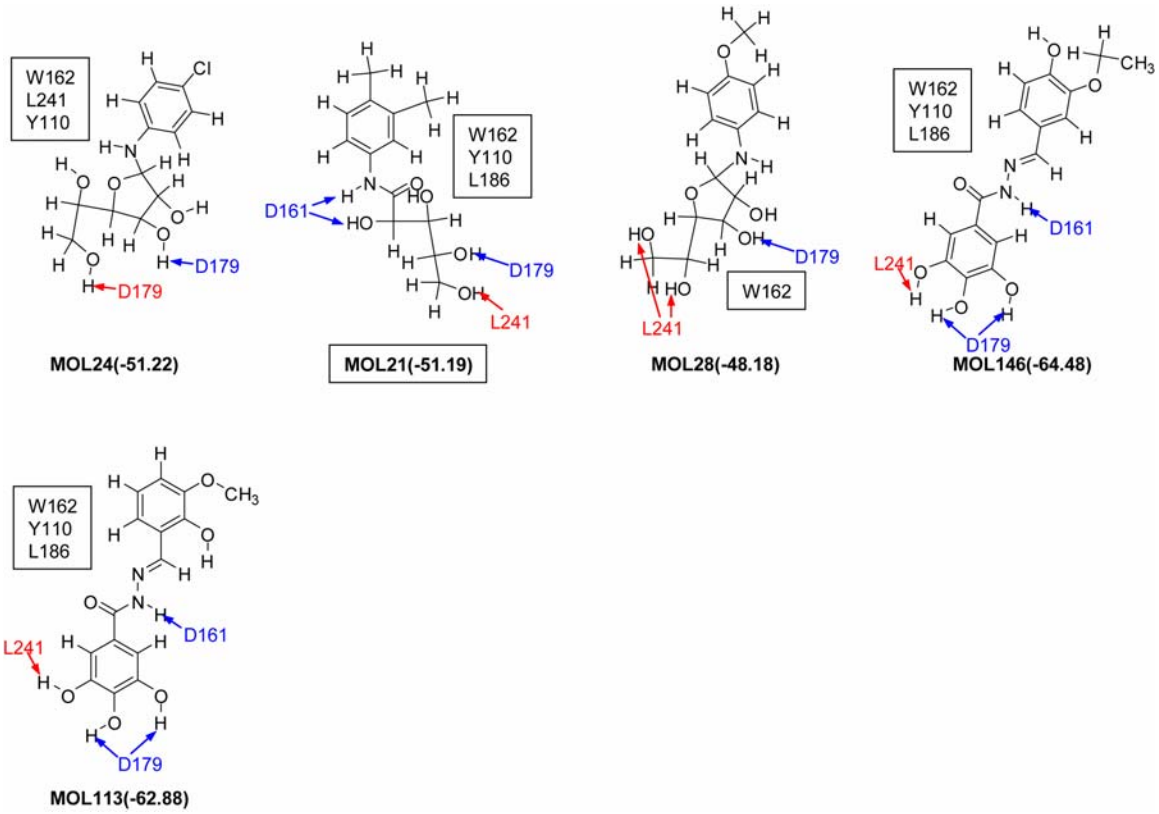


Figure S4.2 (*continued*) Hit compounds from the second set after docking.

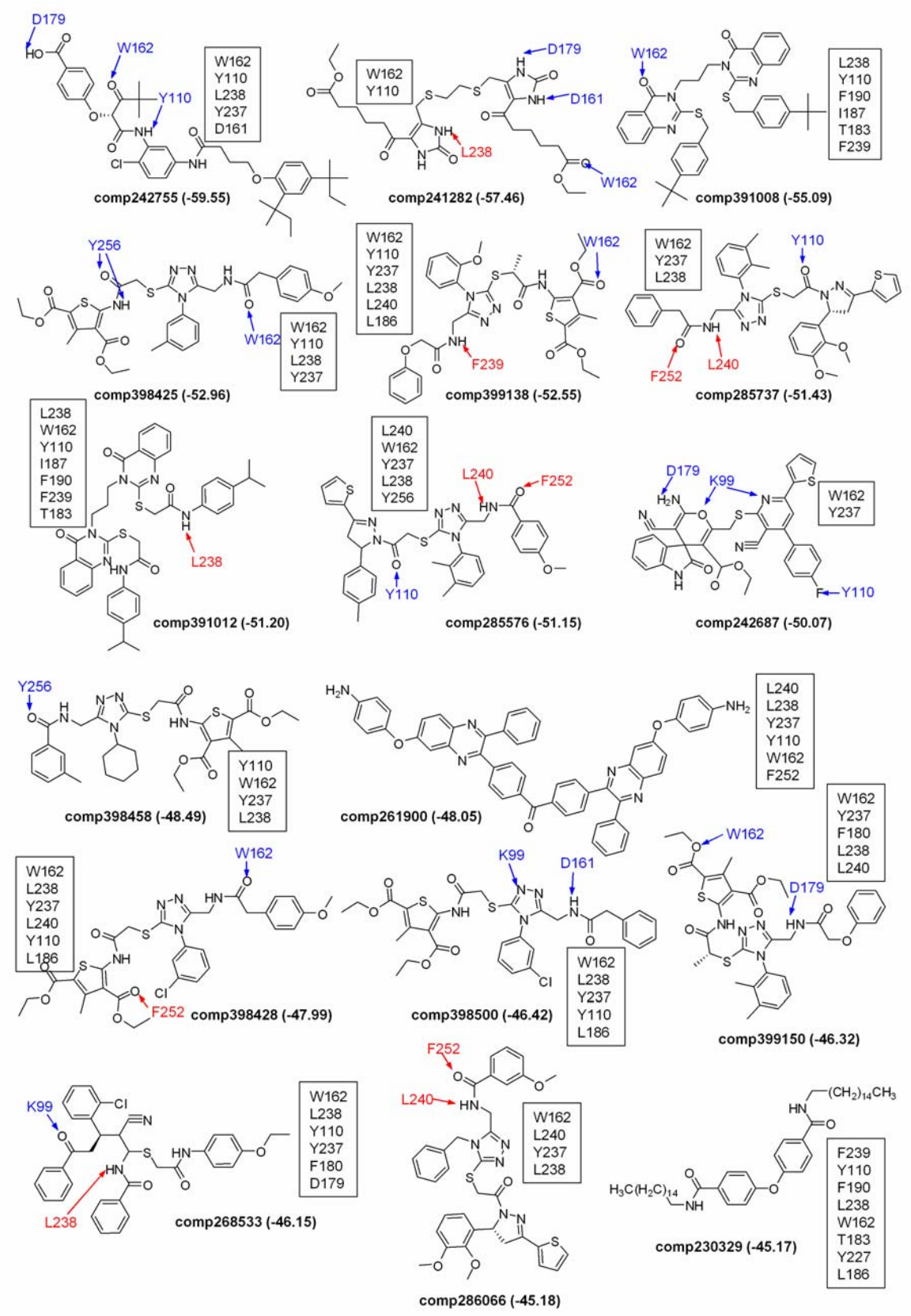


Figure S4.3 Hit compounds after virtual screening for the tetra-peptide binding site.

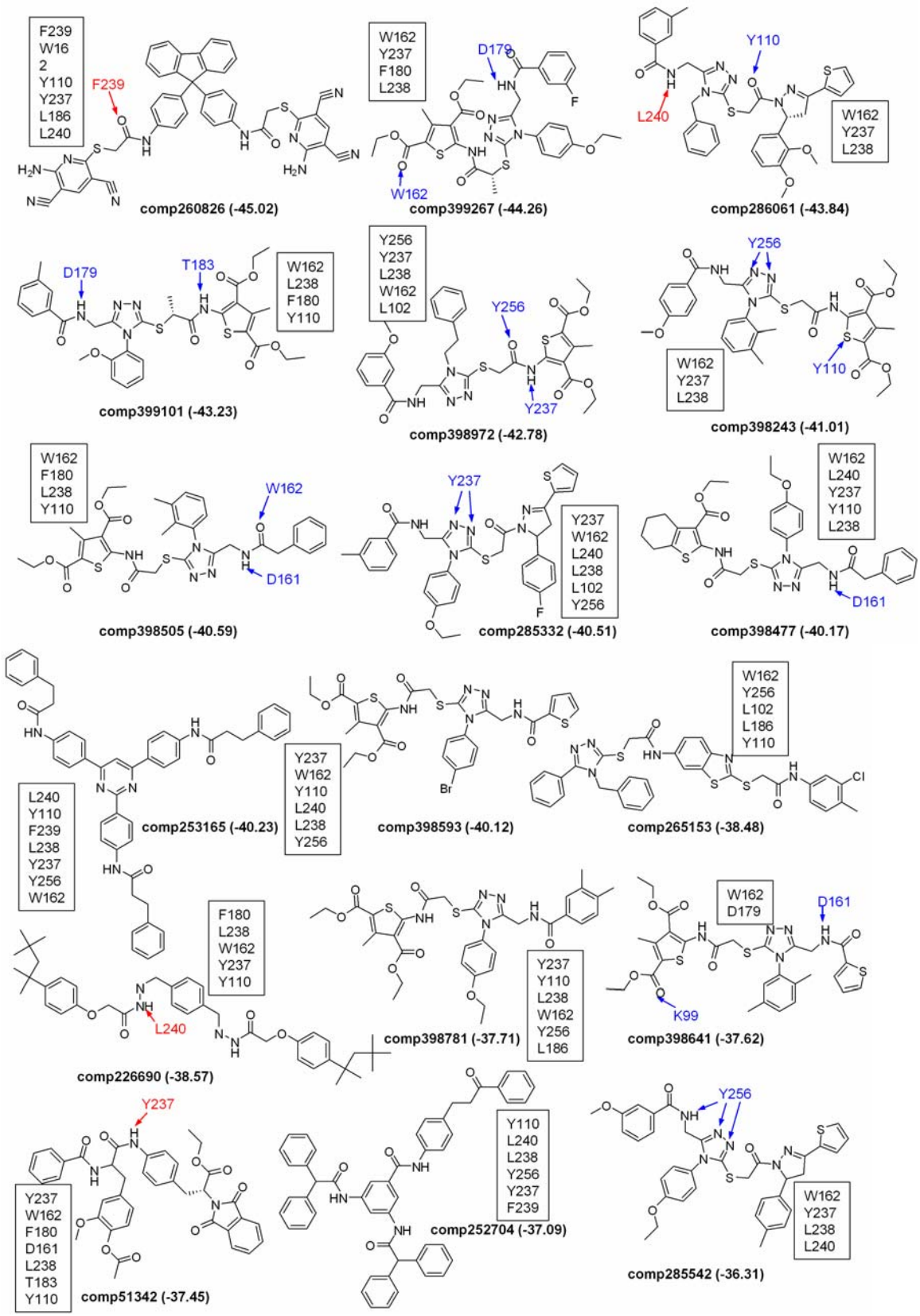


Figure S4.3 (*continued*) Hit compounds after virtual screening for the tetra-peptide binding site.

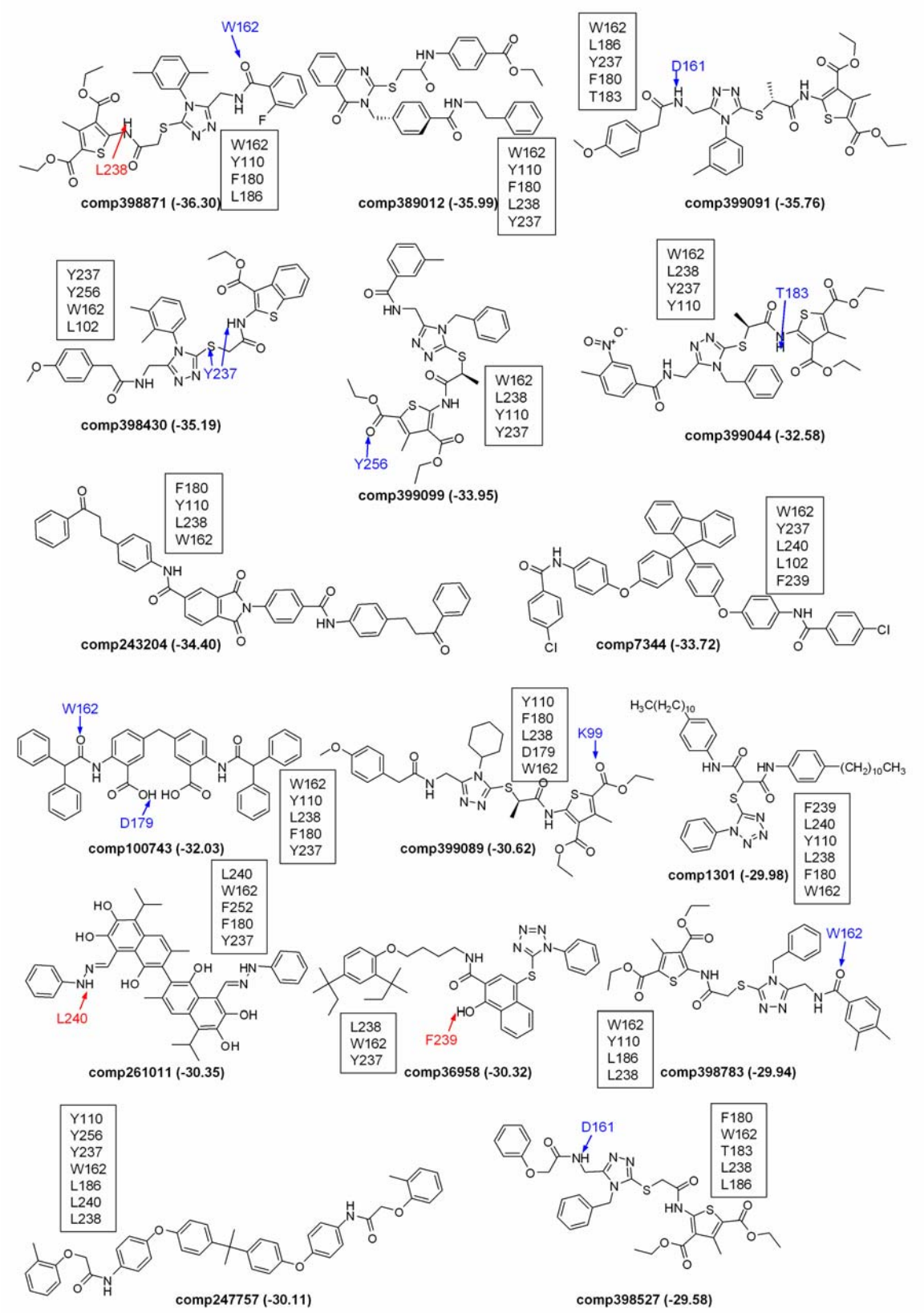


Figure S4.3 (*continued*) Hit compounds after virtual screening for the tetra-peptide binding site.

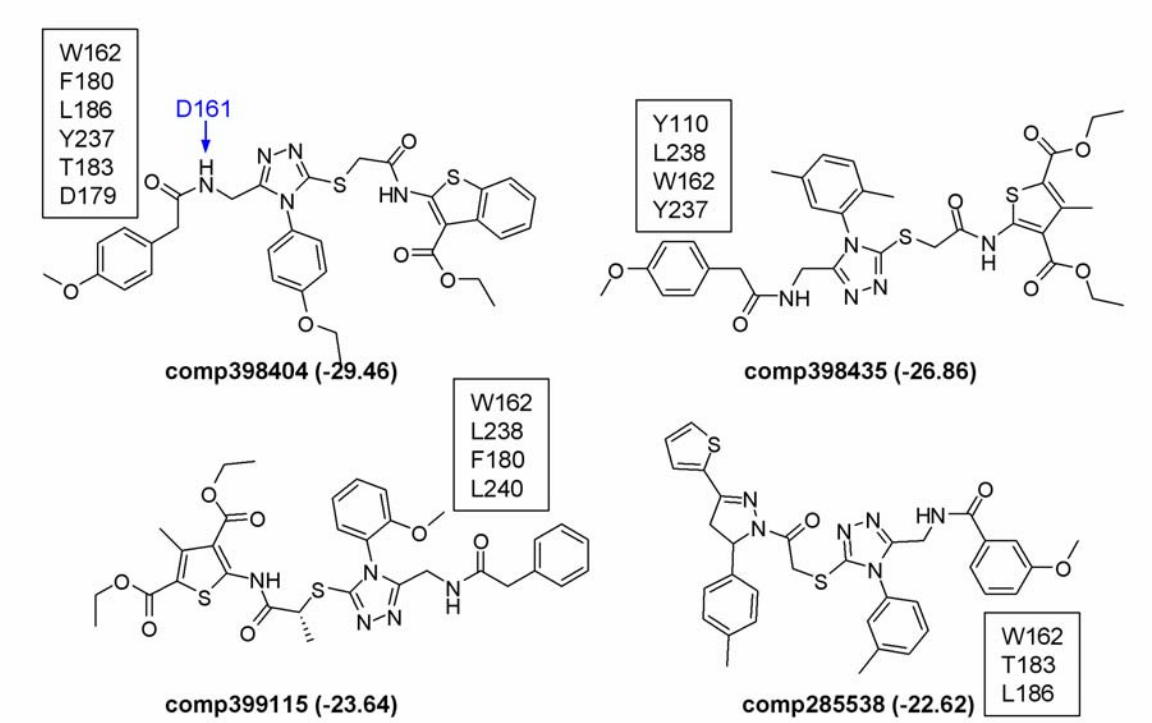


Figure S4.3 (*continued*) Hit compounds after virtual screening for the tetra-peptide binding site. The residues involved in the intermolecular hydrogen bonds or the van der Waals interactions are indicated in the same way as Figure S4.1.

Chapter 5

Prediction of the 3D Structure of Rat MrgA G Protein-Coupled Receptor and Identification of its Binding Site[1]

5.1 Introduction

Rat MrgA is one of a few Mrg receptors for which the small molecular (non-peptide) agonists have been identified. It has been shown to be activated by adenine (and not guanine). Indeed adenine activates rMrgA with a K_i value of 18 nM, potentially identifying it as the endogenous ligand[1]. In this chapter we predict the 3D structure of the rMrgA receptor, and we report the ligand binding site for adenine and related ligands. This work builds upon our previous studies in which we first predicted the 3D structures of mouse MrgC11 (mMrgC11) and MrgA1 (mMrgA1) receptors using the MembStruk computational method[2, 3]. These structures were validated by predicting the binding sites and energies for several tetrapeptides, identifying key residues, and then experimentally confirming the expected changes in binding resulting from mutations of these residues, as described in chapter 2.

For this study on rMrgA, we use these validated mMrgC11 and mMrgA1 structures as templates to predict through homology modeling the 3D structure of rMrgA receptor (it is 49 % and 77 % sequence identical to the mMrgC11 and mMrgA1 sequences). Then we used this structure of rMrgA in conjunction with the HierDock computational procedure to predict the binding site of all nine ligands to the rMrgA receptor for which experimental data are available.

[1] Portions of this chapter have been submitted from the *Journal of Computational Chemistry* for publication.

```
                TM1                                 TM2
rMrgA   RTLIPNLLIIISGLVGLTGNAMVFWLLGFRLARNAFSVYILNLALADFLFLLCHIIDSTL 60
mMrgA1  TILIPNLMIIIFGLVGLTGNGIVFWLLGFCLHRNAFSVYILNLALADFFFLLGHIIDSIL 60
mMrgC11 PILTLSFLVLITTLVGLAGNTIVLWLLGFRMRRKAISVYILNLALADSFFLCCHFIDSLL 60
          * .::::*   ****:** :*:***** : *:*:*********** :**  *:*** *
                              TM3                                 
rMrgA   LLLKF--SYPNIIFLPCFNTVMMVPYIAGLSMLSAISTERCLSVVCPIWYRCRRPKHTST 118
mMrgA1  LLLNV--FYP-ITFLLCFYTIMMVLYIAGLSMLSAISTERCLSVLCPIWYHCHRPEHTST 117
mMrgC11 RIIDFYGLYAHKLSKDILGNAAIIPYISGLSILSAISTERCLCVLWPIWYHCHRPRNMSA 120
         ::..   *.       : .  :: **:***:**********.*: ****:*:**.: *:
               TM4                 TM5                 
rMrgA   VMCSAIWVLSLLICILNRYFCGFLDTKYEKDNRCLASNFFTAACLIFLFVVLCLSSLALL 178
mMrgA1  VMCAVIWVLSLLICILNSYFCGFLNTQYKNENGCLALNFFTAAYLMFLFVVLCLSSLALV 177
mMrgC11 IICALIWVLSFLMGILDWF-SGFLGETHH--HLWKNVDFIITAFLIFLFMLLSGSSLALL 177
        ::*: *****:*: **: : .***.  :.  :     :*: :* *:***::*. *****:
                              TM6                                 
rMrgA   VRLFCGAGRMKLTRLYATIMLTVLVFLLCGLPFGIHWFLLIWIKIDYGKFAYGLYLAALV 238
mMrgA1  ARLFCGTGQIKLTRLYVTIILSILVFLLCGLPFGIHWFLLFKIKDDFHVFDLGFYLASVV 237
mMrgC11 LRILCGPRRKPLSRLYVTIALTVMVYLICGLPLGLYLFLLYWFGVHLHYPFCHIYQVTAV 237
         *::**. :  *:***.** *:::*:*:****:*:: ***  :  .       :* .: *
          TM7
rMrgA   LTAVNSCANPIIYFFVG 255
mMrgA1  LTAINSCANPIIYFFVG 254
mMrgC11 LSCVNSSANPIIYFLVG 254
        *:.:**.*******:**
```

Figure 5.1 Sequence alignment provided as an input for the homology modeling of rMrgA. The N-terminus (11 residues) and C-terminus (38 residues) were omitted because for such class A (rhodopsin-like) GPCRs especially for small ligands, they generally do not play a role in the binding of the ligand[4].

We also compare the putative binding site of rMrgA receptor with those of other known adenine-related GPCRs like adenosine receptors or purinergic receptors.

5.2 Materials and methods

5.2.1 Molecular modeling of receptor structure

We used MODELLER6v2[5] to build a homology model for the 3D structure of rMrgA receptor using the 3D structures for mMrgC11 and mMrgA1 as templates. The sequences of rMrgA receptor (TrEMBL accession number: Q7TN49) was aligned with mMrgC11 (TrEMBL accession number: Q8CIP3) and mMrgA1 (TrEMBL accession number: Q91WW5) using Clustal-W (version 1.82)[6] as shown in Figure 5.1. The sequence identity of rMrgA with mMrgC11 is 49%, while that for mMrgA1 is 77%, for the entire sequences. The TM regions have

44% to 76% identity (totaling 56%) between rMrgA and mMrgC11 and 77% to 88% identity between rMrgA and mMrgA1 (totaling 83%).

After predicting the overall 3D structure of rMrgA, the side chain conformations were re-assigned using the SCWRL3.0 side chain replacement program (~1.4 Å diversity)[7] and hydrogen atoms were added using the POLYGRAF software. The all-atom structure was optimized with the conjugate gradient minimization technique to an RMS in force of 0.5 kcal/mol/Å. Subsequently this minimized receptor structure was used as the starting point for gas phase NVT molecular dynamics (MD) simulations (using an internal dielectric constant of 2.5) at 300 K for 10 ps to account for changes in the backbone conformation. The conformation with the lowest total energy in the trajectory was selected and minimized to an RMS force of 0.5 (kcal/mol)/Å with conjugate gradients. All simulations used the DREIDING force field (FF)[8] with charges from CHARMM22[9] in the MPSim code[10]. The cell multipole method[11] was used for calculation of non bond interaction.

5.2.2 QM calculation of ligand tautomers

We docked to rMrgA the 9 molecules shown in Figure 5.2 (including adenosine phosphates), for all of which there are measured binding constants. The structures for these molecules were constructed using the Cerius2 build module[12]. The ligand conformations were minimized using conjugate gradients with the DREIDING FF and GASTEIGER charges[13]. For ligands with a significant number of torsions, such as 6-benzylaminopurine (6BAP), adenosine and adenosine phosphates, the X-ray crystal structures were obtained from the cambridge structural database and used as the starting conformation for docking without further optimization.

For 1-methyladenine (1MA) and 6BAP, several tautomeric forms are possible in addition to the direct substitution at N1 or N6 of adenine. For these systems we built all such tautomeric forms (see Figure 5.2) and calculated their relative stabilities using quantum mechanics (QM) (Jaguar v5.5 software[14]) to determine the dominant tautomeric form. The geometries were first

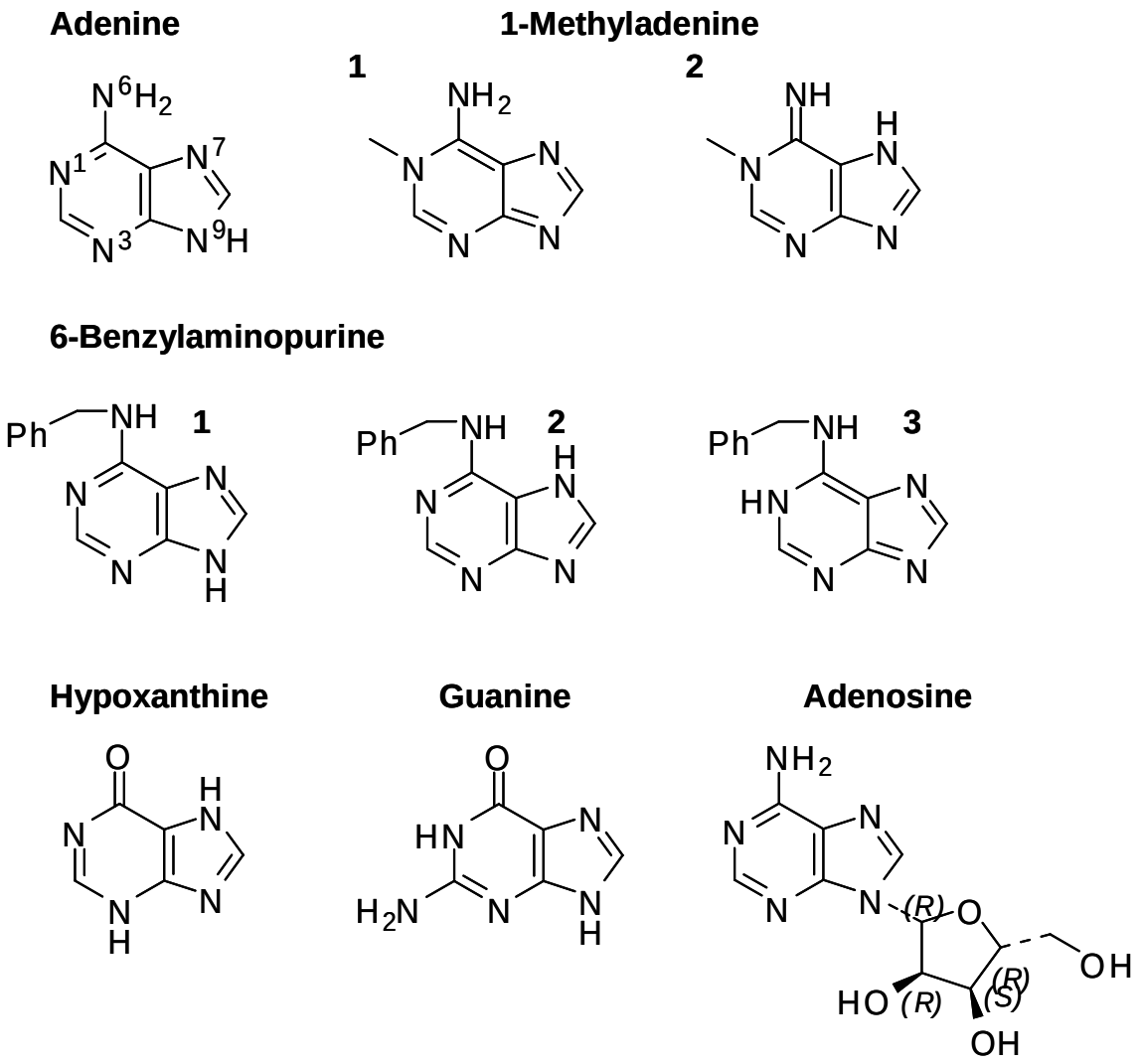


Figure 5.2 Ligand compounds used in docking studies for the rMrgA receptor. They are placed in order of experimental binding affinity from top-left to bottom-right. No binding was detected experimentally for the ligands of the third row. For 1-methyladenine (1MA) and 6-benzylaminopurine (6BAP), the most stable tautomeric forms are shown together.

optimized in the gas phase using the B3LYP flavor of Density Functional Theory with the 6-31G** basis set. The vibrational frequencies for thermodynamic quantities were calculated at the same level. The calculated frequencies were scaled by the factor 0.9614 appropriate for B3LYP/6-31G*. All thermodynamic quantities were computed at 298.15 K, based on standard ideal-gas statistical mechanics and the rigid-rotor harmonic oscillator approximations. We calculated the solvation energy in water using the Jaguar Poisson-Boltzmann methodology with

standard parameters (dielectric constant ε_{H2O} = 80.37, solvent probe radius R_{H2O} = 1.40 Å, and Dreiding van der Waals radii of atoms) for the final optimized QM structure. These results are in Table S5.1 of the supplementary information.

5.2.3 Prediction of the adenine binding site

Scanning the receptor to determine the putative binding region

To select the putative binding region, we used adenine (the best binder) to scan the entire receptor structure of rMrgA. To do this we first calculated the molecular surface using autoMS utility in DOCK4.0[15] with the default values for surface density (3.0 dots/Å^2) and probe radius (1.4 Å). Then we used SPHGEN in DOCK4.0 to generate spheres from each surface point to fill up the void space in the receptor. The receptor was partitioned into 41 cubic boxes each with sides of 10 Å such that all void spheres were included. The spheres inside each box were taken as an input for DOCK4.0 to define the docking region. The scoring energy grids of the protein were calculated using GRID in DOCK4.0, with a grid spacing of 0.3 Å and a nonbond cutoff distance of 10 Å. For each of the 41 regions, we performed rigid docking with the anchor search option in DOCK4.0. For each region, we sampled orientations until 100 passed the bump test and then we selected the ten top scoring orientations. For each of these 10 from each of the 41 boxes, we used MPSim to minimize the ligand conformation with the receptor coordinates fixed to obtain the final energy scores. Here we used the Dreiding FF. After scoring with MPSim, we calculated the percentage of buried surface for each of these 410 orientations using the Connolly MS program from Quantum Chemistry Program Exchange (QCPE). Of these, 103 had over 90 % of buried surface. From these we selected the best orientation for each box. Out of the 41 boxes, this led to seven possible binding regions with good energy and >90% buried surface. We then clustered the spheres near these seven regions, to obtain the two distinct putative binding sites shown in Figure 5.3.

Docking adenine and guanine into the predicted putative binding sites

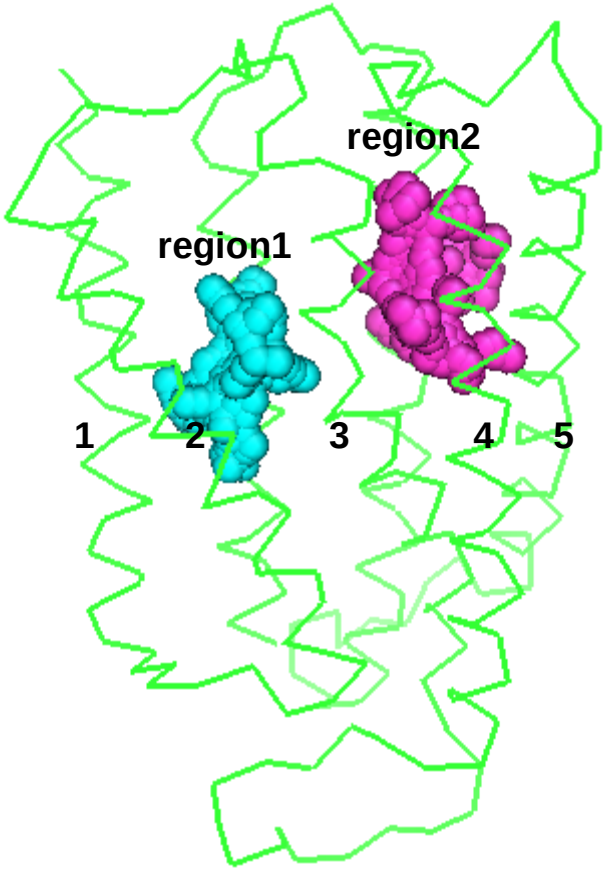


Figure 5.3 Putative binding sites predicted from the HierDock scanning procedure. Region 2 is in the TM3456 region that we find to bind adenine-like agonists. Region 1 is in the TM1237 region (it does not play a role in binding agonists, but might for antagonists).

The HierDock protocol was used to predict the binding site and energy of adenine to both binding regions. In the study on rMrgA we also used the modified HierDock protocol (MSC-Dock) described in chapter 2. Here we used a rejection ratio of 2.2 to define completeness (leading to 2,453 families that past the bump tests). We then enriched the top 75 families until there was an average of six members in each family (passing the bump tests). Then we scored these using MPSim (Dreiding FF) and selected the 30 best scoring family heads. These were minimized (conjugate gradients) using MPSim (50 steps or 0.1 kcal/mol/Å) with ligand movable and the receptor atoms fixed. Then the 5 best scoring ligands (total energy) were selected and the

side chain conformations of the residues of the receptor within 5 Å of the bound ligand were reassigned using the SCREAM side chain replacement program (This uses a side chain rotamer library of 1,478 rotamers with 1.0 Å resolution, with all atom DREIDING energy function to evaluate the energy for the ligand-receptor complex). The binding energies were then calculated for these 5 optimized ligand-receptor complex structures as the difference between the energy of the ligand in the fixed receptor and the energy of the ligand in solution. The energy of the free ligand was calculated for the docked conformation and its solvation energy was calculated using analytical volume generalized Born (AVGB) continuum solvation method[16]. The dielectric constants for the continuum solvation method were set to 78.2 for the external region and to 1.3 for the internal region.

Guanine shows no binding in the experiments (worse than ~100 μM). We docked it to the two putative binding regions determined from scanning the receptor (shown in Fig. 5.3).

5.2.4 Refinement of the binding mode of adenine

To account for changes in the backbone structure of the receptor due to ligand binding, we started with the docked structure and carried out annealing MD simulations allowing the ligand and residues within 10 Å in the binding pocket to move (with other residues fixed). The procedure was to heat the system from 50 K to 600 K and then to cool it back down to 50 K in steps of 50 K. The system was equilibrated for 1ps between changes in temperature. At the end of the annealing cycle, the system was minimized to an RMS force of 0.3 (kcal/mol)/Å and the side chains of the residues within 5 Å from the ligand was reassigned again with SCREAM.

5.2.5 Docking of other adenine derivatives

After optimizing the structure for adenine in the receptor, we re-clustered the spheres to define the binding site. Spheres within 1.0 Å from any atom in the docked adenine were selected out of the entire spheres generated for the final receptor structure that was previously optimized

with adenine. We then used the HierDock procedure described above to dock the adenine derivatives.

5.3 Results and discussion

5.3.1 Characteristics of receptor structure

The sequence identity of rMrgA receptor with bovine rhodopsin is ~18 % for TM regions (the averaged value obtained with the independent alignments for each TM). The RMSD of the coordinates of the Cα atoms between these two receptors is 3.72 Å in TM regions[17].

The RMSD of rMrgA with mMrgA1 (83 % sequence identity for TM regions) is 0.41 Å in the TM regions and the RMSD with mMrgC11 (56 % sequence identity for TM regions) is 2.59 Å in the TM regions. The predicted 3-D structure of rMrgA is shown in Figure 5.4(b) where it is superimposed with the predicted structures of mMrgA1 and mMrgC11.

Figure 5.5 shows the interhelical hydrogen bond network in TM regions formed in the rMrgA receptor;

The Asn31 (TM1) makes hydrogen bonds with the side chain of Asp58 (TM2) and the backbone carbonyl of Cys256 (TM7) at the same time and contributes to the interhelical stability among TM1, TM2 and TM7. This Asp-Asn pair is highly conserved across the family A of GPCRs, corresponding to Asp83 and Asn55 in bovine rhodopsin. There is a similar pattern in rhodopsin structure[18] where a carbonyl group of A299 in the backbone of TM7 is as the common hydrogen bond acceptor for Asn55.

The Tyr95 (TM3) is conserved throughout the Mrg receptor family (although 5 of 36 have a Phe conservative replacement at this position). Here the hydroxyl group of Tyr forms an interhelical hydrogen bond with a backbone carbonyl group of C218 in TM6.

The highly conserved Asn53 (TM2) and Trp136 (TM4) form a hydrogen bond as observed in rhodopsin (Asn78 and Trp161).

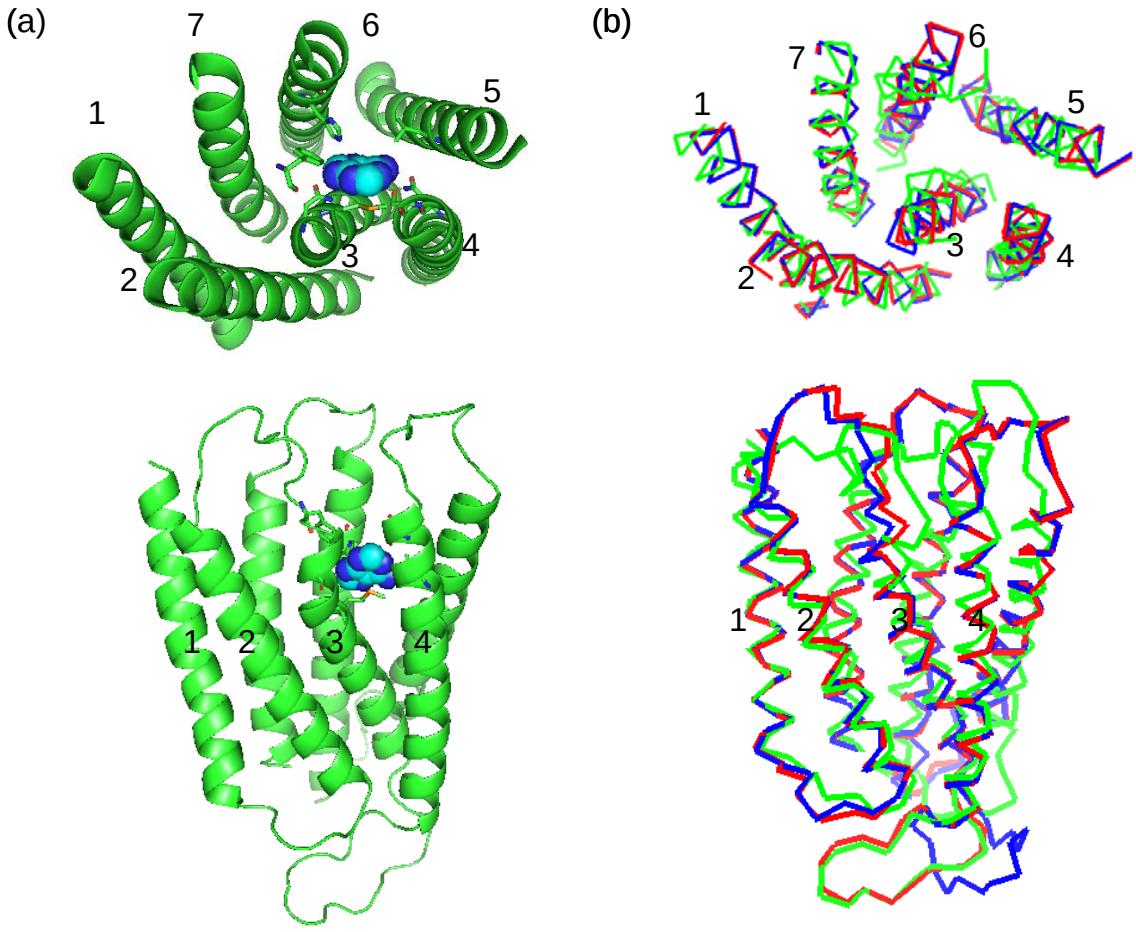


Figure 5.4 Predicted 3D structure of rMrgA receptor.

(a) Adenine (in spheres) is docked in rMrgA receptor. The residues within 5 Å of adenine are shown as sticks. (b) The rMrgA receptor (red) is overlapped with mMrgA1 (blue) and mMrgC11 (green). The top part shows the view from the extracellular side, while the bottom part shows the side view (with the extracellular part on top).

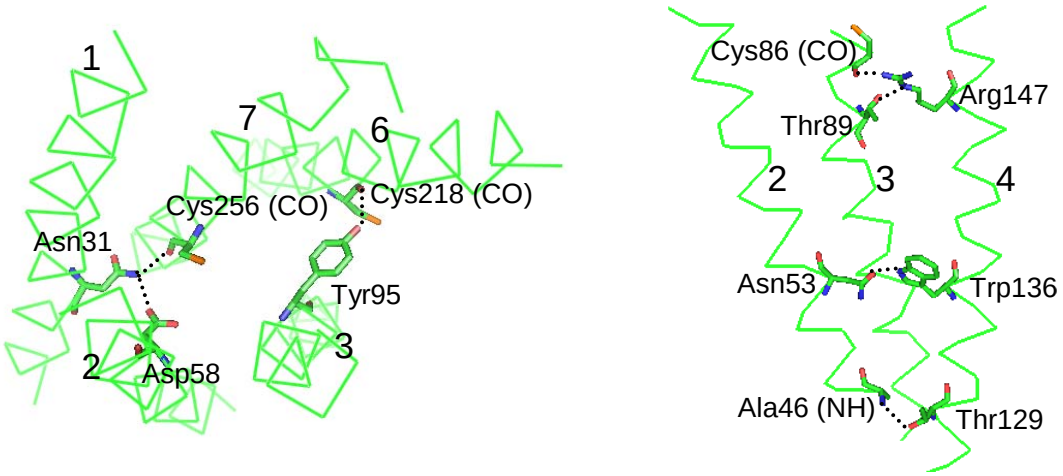


Figure 5.5 Interhelical hydrogen bonds (dashed lines) in rMrgA receptor, as identified using HBPLUS[19] (maximum D-A distance = 3.9 Å, minimum D-H-A angle = 90.0°).

One more hydrogen bond pair exists between Ala46 (TM2) and Thr129 (TM4) near the intracellular region.

In addition, the positively charged residue Arg147 (TM4) is oriented slightly towards the lipids and might contact with the negatively charged head group of the lipid molecule. We find that it forms the hydrogen bonds with Cys86 and Thr89 in TM3 that are one helical turn apart.

The highly conserved proline residues in TM6 and TM7 across the family A of GPCRs correspond to Pro221 (TM6) and Pro258 (TM7) in rMrgA receptor. They lead to bends of 15° and 18° in the α-helix structure.

The Pro94 (TM3) in rMrgA receptor corresponds to the double Gly in the middle of rhodopin. In both cases this leads to bending (19° for rMrgA and 13° for rhodopsin), making the overall backbone conformation of TM3 in these two receptors similar.

A major difference between rMrgA and most other family A GPCRs is that there is no Cys in the extracellular loop (EC) 2 or at the top of TM3. In rhodopsin and other amine receptors there are highly conserved cysteine residues in TM3 and in the EC2 that form a disulfide linkage

that constrains the structure of EC2. Thus for rMrgA receptor we find that EC2 has an open random coil conformation. (In rhodopsin this loop has a closed beta sheet structure).

5.3.2 QM results of ligand tautomers

The QM results of the free energies for the different tautomeric forms of 1MA and 6BAP are shown in Table S5.1. We find that in solution the free energy of 1MA1 is 1.87 kcal/mol lower. The relative abundance with respect to the tautomer with the lowest free energy was calculated from the free energy using the equation;

$$\frac{[tautomer]}{[tautomer]_{lowest}} = \exp\left(-\Delta G_{sol} \diagup RT\right),$$

where R is the gas constant (1.986 cal/mol·K) and T is the temperature (298.15 K). Thus we predict that the relative abundance of 1MA2 is only ~4 % of 1MA1. (In contrast 1MA1 is less stable than 1MA2 by 3.5 kcal/mol in the gas phase.)

There are three tautomers for 6BAP, but 6BAP1 is the most stable both in gas phase and in aqueous solution. Here the others forms have negligible abundance.

These calculations suggest that the majority species for 1MA or 6BAP have direct substitutions at the N1 or N6 of adenine. Therefore these forms were chosen for the docking studies.

5.3.3 Binding modes of adenine and other ligands

Location of the binding site

MSC-Dock predicts the adenine binding site lie between TM3, TM4, TM5 and TM6 as shown in Figure 5.4. This TM3-4-5-6 pocket (corresponding to region 2 in Fig. 5.3) is predicted to provide the binding site for the agonists to a number of other GPCRs (including dopamine, adrenergic, histamine). In addition the adenine is in a region similar to the β-ionone ring of 11-cis retinal in bovine rhodopsin (but the adenine leans more towards TM4 instead of TM6).

The scanning step also found a second binding site, denoted as region 1 in Figure 5.3. This other site is located in the interhelical hydrogen bond network between TM1, TM2 and TM7. In this site both adenine and guanine make a hydrogen bond with the highly conserved Asp58 in TM2, but the binding pocket is mostly hydrophobic except for this Asp residue. We found that the calculated binding energy of adenine in region 1 is only 66 % of that in region 2. The binding energy of guanine in region 1 was 73 % of that for adenine in region 2. Thus we conclude that this site is not the site for agonist binding (it could play a role for antagonists).

As discussed in section 3.1, Asp58 (TM2) plays a key role in stabilizing the TM1, 2, 7 triad, and it may be the site at which Na^+ binds for the allosteric regulation observed in human adenosine A1 receptor and α_{2A} adrenergic receptor[20, 21], making it unlikely to serve as the agonist binding site.

Based on these results we ruled out region 1 as a possible binding site.

Predicted Binding site of Adenine

Adenine is reported as the potential endogenous ligand for rMrgA receptor by Bender *et al.*[1]. The binding mode is detailed in Figure 5.6(a). The most critical residues for binding are Asn88 TM3 and Asn146 TM4. They each form bidentate hydrogen bonds with adenine, locking it tightly inside the pocket. The hydrogen bond partners of Asn146 are the same nitrogen atoms of adenine that participate in the DNA base pair. In addition Phe83 in TM3 and His225 in TM6 have good π stacking interactions with the purine ring. These features characterizing adenine binding site agree well with the empirical observations by Nobeli et al. to explain the molecular discrimination of adenine and guanine ligand moiety in complexes with proteins[22]. They observed that the protein aromatic residues stabilize an environment in which the ligand would have π stacking interaction with the side chain of these residues and that His is much more favorable for adenine. They found that amino acids with side chains like Asn that can form

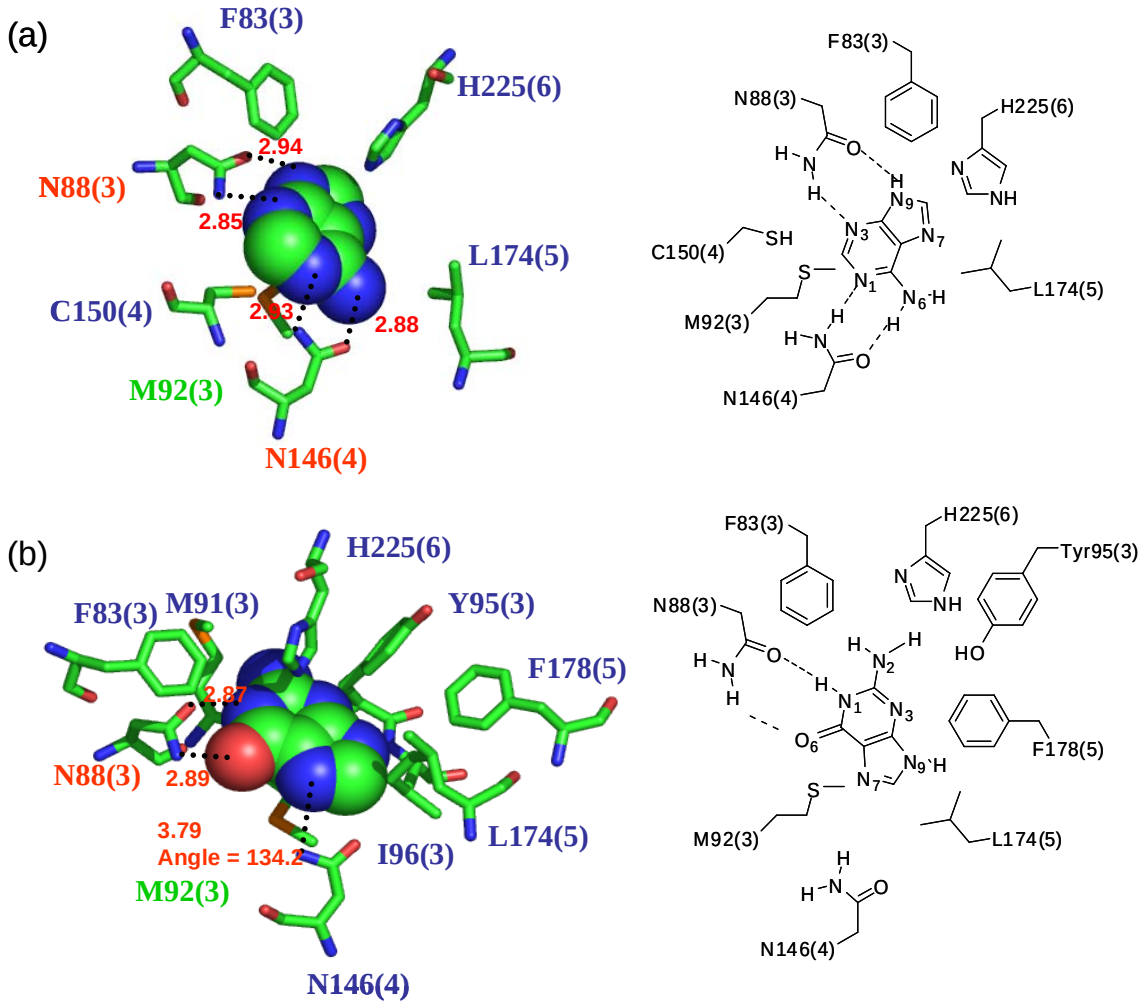


Figure 5.6 Predicted 5 Å binding pockets of adenine (top) and guanine (bottom) in the rMrgA receptor. The residue labels are colored according to the binding energy contributions from non bond interaction with the ligand:

red: greater than 10 kcal/mol contribution (best),

green: between 10 and 4 kcal/mol,

blue: worse than 4 kcal/mol (worst).

The hydrogen bonds are indicated by dotted lines with the distance between the donor and acceptor atoms. The number in parenthesis indicates the TM containing the residue.

simultaneously a donor hydrogen bonds and an acceptor hydrogen bond are favored for binding adenine.

The residues within the binding pocket in Figure 5.6 are grouped by color according to the intermolecular interaction energy with the ligand (red is strongest, blue is weakest). Here the intermolecular interaction energy includes Coulomb, van der Waals, and hydrogen bond terms. The most important are Asn88 and Asn146, which comes from strong hydrogen bond interactions. Met92 has moderate van der Waals interaction with adenine.

Predicted binding site of guanine

Changing the docked adenine structure to guanine, we find that the hydrogen bond donor and acceptor in the side chain of Asn146 does not match with the counterparts in guanine, resulting in a dramatic decrease in the predicted binding affinity (by 16 %) for guanine in this configuration. However N2 of guanine forms a new weak hydrogen bond with sulfur of Cys150.

Independently docking guanine, leads to a structure in which the guanine has the different orientation shown in Figure 5.6(b). Here its hydrogen bond interactions with Asn146 are not optimal. The carbonyl group of the Asn146 side chain loses a hydrogen bond partner and the Asn146 amine group does not make a good hydrogen bond. However the guanine retains similar interaction with the other residues.

Thus the predicted structure of rMrgA, explains the dramatic difference in bonding between adenine and guanine. Adenine can bind to both Asn in the active site leading to good hydrogen bonds for N1, N3, N6, and N9. In contrast guanine in the same configuration could make only half of these. As a result guanine binds in an alternate site where the sidechain of Asn88 form hydrogen bonds with the N1 and O6 atoms of guanine and Asn146 form a weak hydrogen bond with N7, but with binding that is 78 % weaker than for adenine. However if Tyr95 that is found nearby N2 and N3 of guanine is mutated to Gln, formation of two more

Table 5.1 Decomposition of total intermolecular interaction (kcal/mol) between ligand and rMrgA receptor, calculated for the residues within 5 Å of the ligand; the numbers in parentheses are the values relative to adenine

Ligand	Coulomb	VDW	Hbonds	TOTAL
Adenine	-2.37 (100)	-11.63 (100)	-28.17 (100)	-42.17 (100)
1MA	-1.20 (50)	-16.16 (138)	-23.87 (84)	-41.23 (97)
6BAP	-0.35 (14)	-29.90 (257)	-12.58 (44)	-42.82 (101)
HPX	-3.06 (129)	-12.64 (108)	-13.95 (49)	-29.65 (70)
Guanine	-3.27 (137)	-14.26 (122)	-16.59 (58)	-34.12 (80)
Adenosine	-1.23 (51)	-22.84 (196)	-12.95 (45)	-37.02 (87)

hydrogen bonds would be expected and might enhance the binding affinity in spite of the loss in van der Waals interactions. Indeed the predicted binding energy of guanine in the Tyr95Gln mutant is comparable to that of adenine in the wild type (99.9 % of adenine binding).

The total intermolecular interaction energy and its each component in the 5 Å binding pocket are tabulated in Table 5.1.

Predicted binding site of medium binders

For 1MA (Ki=4.4 μM) we also calculated two binding modes, one by perturbing adenine to 1MA, the other with independent docking. The binding modes of 1MA are described in Figure 5.7. The perturbed structure built by direct substitution at N1 in the docked adenine leads to a big clash between the bulky methyl group and Asn146. The independently docked 1MA is locked between Asn88 and Asn146 through hydrogen bonds with these two residues. However, this leads to slightly weakened bonding with Asn146 due to the loss of one of hydrogen bonds. This leads to a predicted binding affinity 83% of that to adenine. The methyl substituent of 1MA resides in the good hydrophobic environment.

For 6BAP, another mild binder (Ki = 58 μM), we find a docking orientation similar to that of 1MA. Here the large benzyl substituent has a close contact with Tyr95 with good π stacking interactions making the van der Waals term the dominant non bond interaction. 6BAP also forms

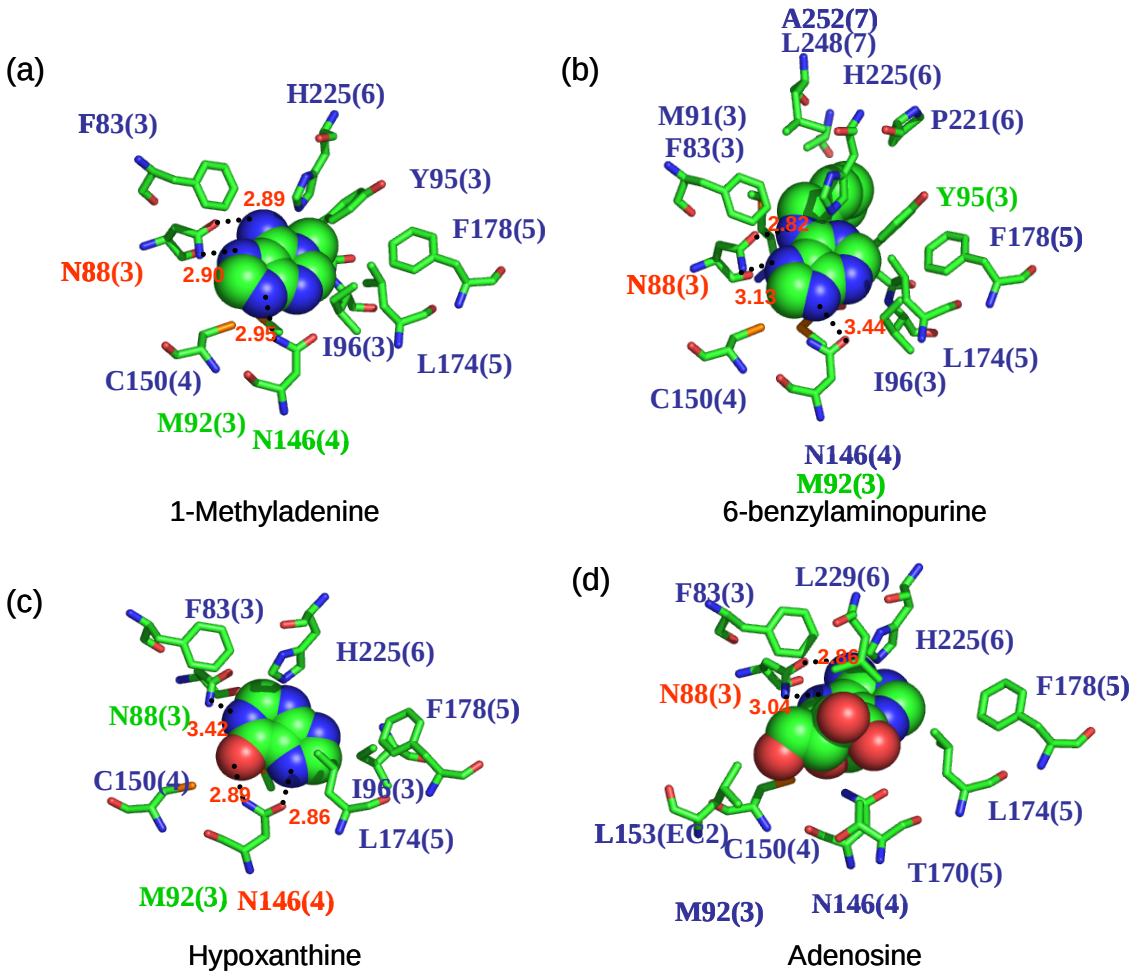


Figure 5.7 The 5 Å binding pockets for various ligands in the rMrgA receptor. The same color scheme is used as for Figure 5.6. (a) 1-Methyladenine, (b) 6-Benzylaminopurine, (c) Hypoxanthine, (d) Adenosine.

hydrogen bonds with Asn88 and Asn146, but the interaction with Asn146 is weaker than for adenine or 1MA. The loss of this interaction is partly compensated by the increased van der Waals interactions as shown in Table 5.1. The result is a binding affinity of 92 % of that of adenine.

Predicted binding site of poor binders

Hypoxanthine, one of the bad binders, makes nice contacts with Asn146 but has weak interactions with Asn88. Its hydrogen bond energy is comparable to 6BAP in Table 5.1, but the

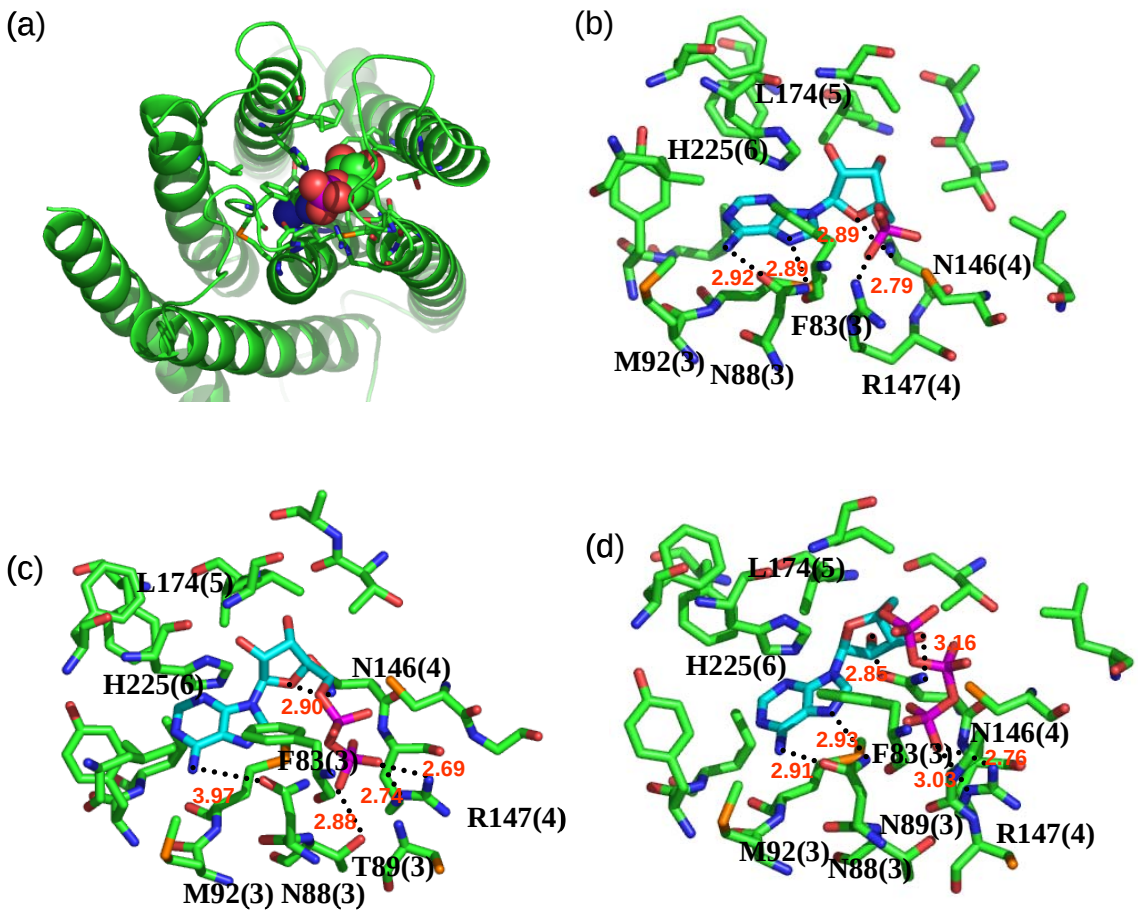


Figure 5.8 The 5 Å binding pockets of adenosine phosphates in the rMrgA receptor. (a) AMP, (b) AMP, (c) ADP, (d) ATP.

van der Waals interaction energy is insufficient to overcome the decreased hydrogen bond energy. The result is a binding affinity of 71 % of that of adenine.

For adenosine, we find that only Asn88 makes good hydrogen bond contacts with the ligand, with no other residues having good specific interactions. The result is a binding affinity of 71 % of that of adenine.

Predicted binding site of adenosine phosphates

Adenosine mono- and tri-phosphates (AMP and ADP) are observed to have binding constants to rMrgA in the range of 20-60 μM concentration. Our predicted structure is in Figure

5.8. We find that the adenine moiety forms good two hydrogen bonds with Asn88, but they have different glycosyl torsion angles. In both cases the sugar ring has a contact with Asn146. We find that the phosphate group points toward extracellular region and is stabilized by Arg147 in TM4 (on the boundary between the inside-bundle region and the membrane). This is only the positively charged residue located on the upper half of TM regions (excluding a Lys233 at the end of TM6). This further validates our prediction of binding site.

For neutral ligands such as adenine, the side chain of Arg147 leans more toward the membrane regions which might allow it to contact the head group of lipid as seen in the apo protein in Figure 5.5. However when the phosphate comes into the binding pocket, the Arg147 would move toward the pocket.

For adenosine diphosphate (ADP), the sugar ring interacts with Asn146 in the similar way to AMP but the adenine base does not interact strongly with Asn88 (see Fig. 5.8(c)). The phosphate group shows strong interaction with Arg147 and Thr89.

Comparison of calculated binding energy to ln K_i

The predicted binding energies for the various ligands are compared in Figure 5.9 with the experimental competition binding constant (inhibition constant) reported by Bender *et al.*[1] Of the nine compounds whose binding constants have been measured, we examined only the six neutral ligand with the fewest torsional degrees of freedom for docking (since the adenosine phosphates are highly negative-charged, the entropic effect in binding is no longer negligible and the uncertainty in calculated solvation energy increases). Figure 5.9 shows the good correlation between our calculated binding energy and the experimental inhibition constant, $ln\ K_i$. The calculating binding energy is for the minimized structure at 0 K, which ignores entropic effects. Except for adenosine all ligands are rigid with similar shapes so that the entropic contributions should be similar. This good correlation strongly validates our predicted structures and binding configurations.

Ligand	K_i, nM[a]	ln K_i	B.E.	B.E.[pert]
Adenine	18	2.89	100	
1-Methyladenine	4391	8.39	83	65
6-Benzylaminopurine	58328	10.97	92	74
Guanine	n.d.	-	78	83
Hypoxanthine	n.d.	-	71	52
Adenosine	n.d.	-	71	60

[a][1]
n.d.: not detectable up to the maximum concentration tried (~100 μM)
B.E.: relative binding energy (%) w.r.t adenine (52.02 kcal/mol)
B.E.[pert]: after being perturbed from docked adenine and optimized

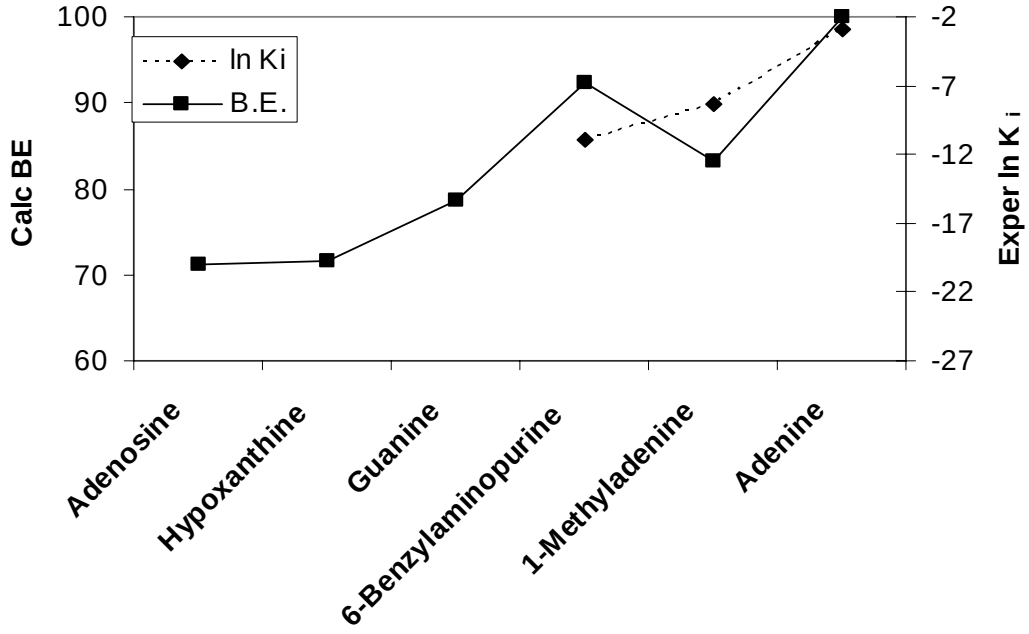


Figure 5.9 Comparison of calculated binding energies (left legend) with the experimental inhibition constants (right legend) for rMrgA ligands as described in the method section, the calculated energies are for the minimized structure (0K) without entropic contributions.

Table 5.2 Computational alanine-scanning results (SCAM) for adenine/rMrgA (energies in kcal/mol)[a]

	I.E.(WT)	ΔI.E.(Ala)	
Asn88	-17.787	17.161	41**%**
Asn146	-12.819	12.275	29**%**
Met92	-4.741	2.844	7%
Phe83	-1.545	1.433	3**%**
His225	-1.505	1.349	3**%**
Leu174	-0.665	0.432	1%
Tyr95	-0.422	0.344	0.8**%**
Ile96	-0.450	0.333	0.8%
Phe178	-0.298	0.265	0.6**%**
Cys150	-0.494	0.207	0.5%
Thr170	-0.273	0.168	0.4%
Met91	-0.321	0.161	0.4%
Arg147	-0.360	0.091	0.2%
Pro85	-0.393	0.086	0.2%
Leu177	-0.102	0.066	0.2%

[a] The intermolecular interaction energy (IE) for the wild type (WT, no mutation) is shown for all residues within 5 Å of the ligand. After mutating the residue to Ala and minimizing, we recalculated the IE of the ligand to this Ala, IE(Ala). The percentage change in binding of the mutant relative to the calculated total binding of WT is shown in the last column. These results show that the Ala mutations track well the calculated ligand-residue IE and confirm the important role of Asn88 (3), Asn146 (TM4), Met92 (TM3), Phe83 (TM3), and His225 (TM6) to the binding of adenine.

Effect of computational alanine-scanning mutations (SCAM) in the binding pocket

For the best binder, adenine, we carried out alanine scanning to assess the importance of various residues to binding. The residues within 5 Å of the ligand were each independently mutated to Ala and the energy for the ligand-protein complex was reoptimized (conjugate gradient minimization). Prior to the minimization we used SCREAM to reselect the side chain conformations of the other residues within 5 Å of the ligand. The results are summarized in Table 5.2.

As expected, the Asn88Ala and Asn146Ala mutations significantly reduce the binding affinity due to the loss of the hydrogen bonds. Mutation of either Phe83 or His225 abolishes the favorable van der Waals contacts.

The close correspondence between the contributions predicted for the wild type and the change in bonding calculated with the mutation to Ala, indicates that good estimates can be made without optimization of the coordinates.

5.3.4 Comparison of the adenine binding site in rMrgA to the nucleotide binding sites in adenosine receptors and purinergic receptors

We can compare the binding site of adenine to rat MrgA with the binding site of human A_1 and A_{2A} to adenosine (hA_1A and $hA_{2A}A$) receptors and human $P2Y_1$ to purinergic ($hP2Y_1$) receptor. These receptors all bind adenosine or ATP, with the adenine moiety in common, and all have been studied both experimentally and with modeling. The sequences of the adenosine receptors and the purinergic receptor were aligned separately with that of rMrgA receptor. The whole sequences were aligned first with Clustal-W while ensuring that specific highly conserved residues are matched to each other in the alignment: Asn at position 20 in TM1, Asp at position 13 in TM2, Arg in DRY sequence of TM3, Trp at position 12 in TM4, Pro at position 19 in TM6, Pro in NPXXY of TM7 (the number is counted from the starting residue of each TM in Figure. 5.10). Using the TM prediction of rMrgA receptor, the sequences for each TM were aligned independently. The averaged sequence identity of rMrgA receptor is ~22 % for hA_1A receptor and ~20 % for $hA_{2A}A$ receptor (considering only TM regions). For $hP2Y_1$ receptor, the TM sequence identity to rMrgA is ~24 %. The resulting TM sequence alignment is shown in Figure 5.10 where the key residues in adenosine receptors and $P2Y_1$ receptor identified from the binding or functional assay experiments are bolded and underlined[23, 24].

Recall that for rMrgA the adenine binding site mostly contacts with Asn88 (TM3), Asn146 (TM4) and Leu174 (TM5), with His225 (TM6) interacting closely with adenine.

TM1	MRGA_RAT	RTLIPNLLIIISGLVGLTGNAMVFWLLG	28
	AA1R_HUMAN	FQAAYIGI**E**VLIALVSVPGNVLVIWAVK	28
	AA2A_HUMAN	GSSVYITV**E**LAIAVLAILGNVLVCWAVW	28
	P2YR_HUMAN	QFYYLPAVYILVFIIGFLGNSVAIWMFV	28
TM2	MRGA_RAT	AFSVYILNLALADFLFLLCHIIDST	25
	AA1R_HUMAN	ATFCFIVSLAVADVAVGALVIPLAI	25
	AA2A_HUMAN	VTNYFVVSLAAADIAVGVLAIPFAI	25
	P2YR_HUMAN	GISVYMFNLALADFLYVLTLPALIF	25
TM3	MRGA_RAT	**F**LPCF**N**TV**MM**VP**YI**AGLSMLSAISTERC	28
	AA1R_HUMAN	TCLMVAC**P**VLI**LTQ**SSILALLAIAVDRY	28
	AA2A_HUMAN	GCLFIACF**V**LVL**T**QSSIFSLLAIAIDRY	28
	P2YR_HUMAN	MCKLQ**R**FI**FH**VNL**Y**GSILFLTCISAHRY	28
TM4	MRGA_RAT	KHTSTVMCSAIWVLSLLICIL**NR**YF**C**GF	28
	AA1R_HUMAN	PRRAAVAIAGCWILSFVVGLTPMFGWNN	28
	AA2A_HUMAN	GTRAKGIIAICWVLSFAIGLTPMLGWNN	28
	P2YR_HUMAN	KKNAICISVLVWLIVVVAISPILFYSGT	28
TM5	MRGA_RAT	LASNFFTAAC**L**IFL**F**VVLCLSSLALLVR	28
	AA1R_HUMAN	EFEKVISMEYMVYFNFFVWVLPPLLLMV	28
	AA2A_HUMAN	LFEDVVPMNYMVYFN**F**FACVLVPLLLML	28
	P2YR_HUMAN	FIYSMC**TT**VAM**F**CVPLVLILGCYGLIVR	28
TM6	MRGA_RAT	RLYATIMLTVLVFLLCGLPFGI**H**WFLLIWIK	31
	AA1R_HUMAN	KIAKSLALILFLFALS**W**LPL**H**IL**N**CITLFCP	31
	AA2A_HUMAN	HAAKSLAIIVGLFALCWLPL**H**II**N**CFT**F**FCP	31
	P2YR_HUMAN	KSIYLVIIVLTVFAVSYIPF**H**VM**K**TMNLRAR	31
TM7	MRGA_RAT	AYGLYLAA**L**VLT**A**VNSCANPIIYFFVG	27
	AA1R_HUMAN	PSILTYIAIFL**TH**GNSAMNPIVYAFRI	27
	AA2A_HUMAN	PLWLMYLA**I**VL**SH**TN**S**VVNPFIYAVRI	27
	P2YR_HUMAN	VYATYQVT**R**GLA**S**LNSCVDPILYFLAG	27

Figure 5.10 Sequence alignment of rat MrgA receptor with other receptors known to bind adenine components of ligands: human A_1 and A_{2A} adenosine receptors and human $P2Y_1$ purinergic receptor. The residues predicted to play an important role in ligand binding are in boldface and underlined.

In the putative A_{2A} binding site, the adenine moiety is recognized by TM3, TM5 and TM6[23]. The binding regions in TM3 overlap significantly throughout four receptors but we could not find any residue from adenosine or purinergic receptor that directly matches with Asn88 in rMrgA receptor. However, Gln92 in TM3 of hA1AR has the same functional group as Asn (shorter by one methylene) which was found to interact with the adenosine adenine moiety[25]. Asn146 in TM4 is a key residue in the adenine binding in rMrgA, but no similar residue is identified as a key residue in TM4 of adenosine or purinergic receptor. Arg157 in TM4 interacts with phosphate group of adenosine phosphates in rMrgA while Lys (TM6) and Arg (TM7) are involved in $P2Y_1$ receptor.

In conclusion, although similar residues recognize adenine, there is very little similarity in the location of the binding site of adenine in rMrgA receptor compared to adenosine and purinergic receptors. This suggests that rMrgA belongs to non-adenosine or non-purinergic receptor families even though adenine binds well and activates the receptor.

5.3.5 Comparison to other MrgA orthologs

We examined the sequences of the 8 mouse orthologs of rMrgA receptor to determine whether some might be good candidates for possible adenine binding receptors. These are collected together and compared to rMrgA in Figure S5.1. Among the eight mouse MrgA (mMrgA) receptors, we find that the mMrgA2 receptor has Asn residues at the same two positions in TM3 and TM4 as in rMrgA receptor. However, Bender *et al*. tested activation of the mMrgA2 receptor with adenine and found no activation[1]. Perhaps this is because mMrgA2 receptor does not have a proline in the middle of TM3 analogous to the Pro94 of for rMrgA receptor that we found to induce the bend in TM3. The change in the conformation of TM3 might put the Asn in TM3 of mMrgA2 receptor in the wrong orientation to bind sufficiently tightly with adenine to cause activation, explaining the lack of binding or activation by adenine mMrgA2 even though it has the same pair of Asn as rMrgA, This could be tested by mutating the Pro94 of

rMrgA to Val as in mMrgA2 to see if this causes a loss in activity or by mutating the Val94 of mMrgA2 to Pro to see if this leads to activity for adenine.

On the other hand, mMrgA5 receptor contains Pro in TM3 at the same position as in rMrgA and the Asn146 of rMrgA is also conserved. However, the Asn88 in TM3 of rMrgA is replaced with Tyr in mMrgA5 receptor. Here we suggest that mutation of Tyr87 to Asn in mMrgA5 might lead to adenine binding.

5.4 Summary and conclusions

We predicted the 3D structure of rMrgA receptor using homology to our MembStruk predicted mMrgA1 and MrgC11 structures and we predicted the binding sites for adenine and its derivatives using HierDock. The putative binding site is within TM3, 4, 5 and 6 with Asn88 in TM3 and Asn146 in TM4 serving as key residues in binding adenine. This Asn146 is homologous to Asp161 in mMrgC11 receptor that we previously identified as a key residue which was then validated experimentally. The side chain of Asn146 plays the role of the thymine in the same way as in the Watson-Crick hydrogen bond geometry of the A-T DNA base pair. It forms a bidentate hydrogen bond with both the N1 and N6 atom of adenine. The availability of the hydrogen bonds with these two Asn residues correlates with the binding affinity of the ligand.

These studies of the rMrgA receptor provide targets for mutagenesis experiments to further identify or validate important features in the binding site. This predicted binding site could be used to identify other small molecule ligands. Experimental tests of such ligands might help identify the endogenous ligand.

References

1. 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.
2. 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.
3. 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.
4. Bockaert, J. and J.P. Pin, *Molecular tinkering of G protein-coupled receptors: an evolutionary success.* Embo Journal, 1999. **18**(7): p. 1723-1729.
5. *MODELLER6v2*, University of California San Francisco: San Francisco.
6. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice.* Nucleic Acids Research, 1994. **22**(22): p. 4673-4680.
7. Canutescu, A.A., A.A. Shelenkov, and R.L. Dunbrack, *A graph-theory algorithm for rapid protein side-chain prediction.* Protein Science, 2003. **12**(9): p. 2001-2014.
8. 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.
9. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.

10. Lim, K.T., et al., *Molecular dynamics for very large systems on massively parallel computers: The MPSim program.* Journal of Computational Chemistry, 1997. **18**(4): p. 501-521.

11. Ding, H.Q., N. Karasawa, and W.A. Goddard, *Atomic Level Simulations on a Million Particles - the Cell Multipole Method for Coulomb and London Nonbond Interactions.* Journal of Chemical Physics, 1992. **97**(6): p. 4309-4315.

12. *Cerius2 Modeling Environment, Release 4.0*, Accelrys Inc.: San Diego.

13. Gasteiger, J. and M. Marsili, *Iterative Partial Equalization of Orbital Electronegativity - a Rapid Access to Atomic Charges.* Tetrahedron, 1980. **36**(22): p. 3219-3228.

14. *Jaguar 5.5*. 2003, Schrodinger: Portland, Oregon.

15. Ewing, T.J.A. and I.D. Kuntz, *Critical evaluation of search algorithms for automated molecular docking and database screening.* Journal of Computational Chemistry, 1997. **18**(9): p. 1175-1189.

16. Zamanakos, G., *A fast and accurate analytical method for the computation of solvent effects in molecular simulations*, in *Division of Physics, Mathematics and Astronomy*. 2002, California Institute of Technology: Pasadena.

17. Okada, T., et al., *The retinal conformation and its environment in rhodopsin in light of a new 2.2 angstrom crystal structure.* Journal of Molecular Biology, 2004. **342**(2): p. 571-583.

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

19. McDonald, I.K. and J.M. Thornton, *Satisfying Hydrogen-Bonding Potential in Proteins.* Journal of Molecular Biology, 1994. **238**(5): p. 777-793.

20. Barbhaiya, H., et al., *Site-directed mutagenesis of the human A(1) adenosine receptor: Influences of acidic and hydroxy residues in the first four transmembrane domains on ligand binding.* Molecular Pharmacology, 1996. **50**(6): p. 1635-1642.

21. Ceresa, B.P. and L.E. Limbird, *Mutation of an Aspartate Residue Highly Conserved among G-Protein-Coupled Receptors Results in Nonreciprocal Disruption of Alpha(2)-Adrenergic Receptor G-Protein Interactions - a Negative Charge at Amino-Acid Residue-79 Forecasts Alpha(2a)-Adrenergic Receptor Sensitivity to Allosteric Modulation by Monovalent Cations and Fully Effective Receptor G-Protein Coupling.* Journal of Biological Chemistry, 1994. **269**(47): p. 29557-29564.

22. Nobeli, I., et al., *On the molecular discrimination between adenine and guanine by proteins.* Nucleic Acids Research, 2001. **29**(21): p. 4294-4309.

23. Kim, S.K., et al., *Modeling the adenosine receptors: Comparison of the binding domains of A(2A) agonists and antagonists.* Journal of Medicinal Chemistry, 2003. **46**(23): p. 4847-4859.

24. Jiang, Q.L., et al., *A mutational analysis of residues essential for ligand recognition at the human P2Y(1) receptor.* Molecular Pharmacology, 1997. **52**(3): p. 499-507.

25. Rivkees, S.A., H. Barbhaiya, and A.P. Ijzerman, *Identification of the adenine binding site of the human A(1) adenosine receptor.* Journal of Biological Chemistry, 1999. **274**(6): p. 3617-3621.

Supporting figures and tables

Figure S5.1 Multiple sequence alignment of rat MrgA with mouse MrgAs using Clustal-W

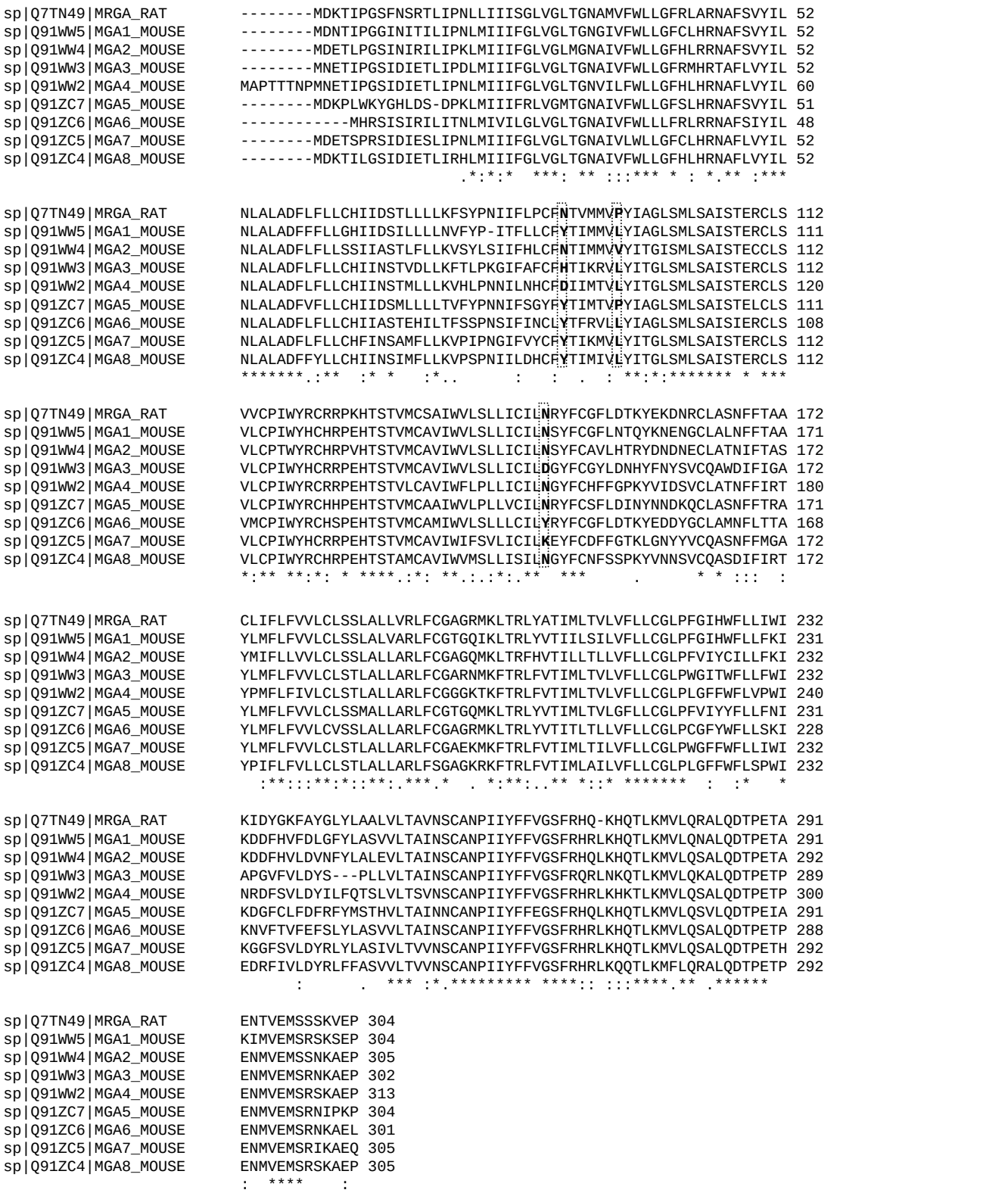


Table S5.1 The Gibbs free energies (kcal/mol) calculated from QM for various tautomeric forms of 1MA and 6BAP (numbered as shown in Figure 5.2)

Ligand	Ggas [a]	Gsol [b]	ΔGsol [c]	Relative abundance[d]
1MA1	-317838.40	-317864.54	0.00	1
1MA2	-317841.98	-317862.67	1.87	0.043
6BAP1	-462806.01	-462822.66	0.00	1
6BAP2	-462797.65	-462818.90	3.76	0.0017
6BAP3	-462788.10	-462812.30	10.36	2.5E-08

[a] Calculated using QM energy and vibrational frequencies for gas phase

[b] Calculated using Poisson-Boltzmann solvation in water

[c] relative to the most stable state

[d] abundance at 300K relative to the most stable

Appendix A

Stability of Oxidized Base and its Mispair in DNA: Quantum Mechanics Calculation and Molecular Dynamics Simulation

Abstract

5-formyluracil (FoU) is a potentially mutagenic lesion of thymine (T) produced in DNA by ionizing radiation and various chemical oxidants. The quantum mechanics (QM) calculation to compute pairing energies of FoU with a purine base was performed at the B3LYP/6-31G**//B3LYP/6-31G**++ level, considering various possible tautomeric, rotameric and ionized form of FoU. The pairing energies of FoU in keto form with either adenine (A) or guanine (G) are comparable to those of T. Although the tautomerism to enol provides triple hydrogen bonds with G, the energy penalty is not fully compensated by the extra hydrogen bond energy. These QM results lead to the conclusion that the ionization at N3 position of FoU would mainly account for the increased mispairing rate of FoU since the deprotonated FoU preferentially form H-bonds with G rather than A and therefore FoU has one more extra possibility of base pairing. The following molecular dynamics (MD) simulations for DNA dodecamers with normal A:T base pair, A:FoU base pair and G:FoU base mispair showed that hydrogen bonds in FoU paired with adenine remained stable in the duplex during the whole simulation, while G:FoU dodecamer showed slightly larger structural fluctuation since it contains non Watson-Crick pairs in the middle. The formyl group of FoU in the anti conformation affects the hydration pattern around the DNA structure. A water molecule that makes a bridge of H-bond between O7 of FoU and

O2P of phosphate seems to be responsible for the well-ordered solvent structure. The interesting result is that, even though the formyl group is located on the major groove side, its presence actually results in severe narrowing of minor grooves. No significant change in helical and backbone parameters is shown for A:FoU and G:FoU dodecamer except for the large shear in G-FoU pairs, which is obvious in Wobble-type geometry.

A.1 Introduction

The modification on DNA bases induces the formation of base mispairing during replication, which is fatal in keeping the genetic integrity in living organism. 5-formyluracil (FoU), one of well-known DNA base lesions is the oxidation product of thymine (T) by ionizing γ-radiation, Fenton-type reactions, and quinone-mediated UV-A photosensitization[1-3]. Privat and Sowers proposed that the electron-withdrawing formyl group increases the stability of the deprotonated form of FoU, which could exist in a non-negligible amount since FoU has a lower pK_a value close to physiological pH than T[4]. The deprotonated form of FoU would be mispaired with guanine (G) in a canonical Watson-Crick geometry. In the following replication process, G might form a correct pair with C and this leads to miscoding (i.e. starting with T, it ends up with C). Masaoka and co-workers also observed that the miscorporation ratio with deoxyguanosine monophosphate (dGMP) increased when FoU on the DNA template was substituted for T and this ratio also increased with increasing pH[5]. It supports the idea proposed by Privat and Sowers that the deprotonated form of FoU plays a key role in mispair mechanism.

The general repair steps carried out by DNA repair enzymes are detection, recognition and removal of mutagenic lesions from DNA. The pathway most commonly employed to remove incorrect bases (like uracil) or damaged bases (like 3-methyladenine) is called base excision repair (BER)[6]. Initially individual DNA glycosylases are targeted to distinct base lesions, which are flipped and cleaved out by the enzymes (damage-specific step) and then a damage-general step restores correct DNA base sequences. Several DNA glycosylases responsible for repair of FoU have been suggested, but the repair mechanism on the molecular level is not well understood yet. In *Escherichia coli*, FoU is reported to be removed from a DNA by the AlkA enzyme with efficiency comparable to that of 7-methylguanine, a good substrate for AlkA[7]. It was proposed that the electron deficient bases flip out from the DNA duplex to form strong π-donor/acceptor interaction with electron-rich aromatic amino acids present in the active site of AlkA. The MutS

proteins involved in methyl-directed mismatch repair also recognize FoU paired with G, but do not recognize it with A. The MutS complex with FoU:G inhibited the activity of AlkA to FoU and thus two independent repair pathways might exist[8]. Zhang *et al.* performed the trapping assay with $NaBH_4$ that is the clue for formation of Schiff base intermediate with NH_2 group in the enzyme at the abasic site[9]. They observed the trapped complex for the Nth, Nei and MutM protein in *E. coli* and also the cleaved bases for these enzymes. With the AlkA protein, no trapping was observed, but the repair mechanism by it cannot be excluded since other pathways without forming the Schiff base intermediate are plausible.

Even though the repair processes help maintain the genetic integrity, the cells are always vulnerable to having base lesions and the following mispairing. The presence of a non-natural base such as FoU would cause the structural changes in DNA double helix. When neutral FoU is mispaired with G and forms the non-canonical hydrogen bond; i.e. in this case, the Wobble type, this local change in H-bonding geometry can cause the overall changes in DNA double helix structure. One of the well-known examples showing the sequence-dependent conformational characteristic is the narrowing of minor grooves in the middle AT-tracts of DNA double helix. It is suggested that the N3 of A and O2 of T in AT base stacks form the electronegative pocket and then the counter cations, e.g., Na^+, are bound to that site and pull two bases closer together. Several studies have been carried out to show the correlation between the width of minor grooves and the location of counter ions in the simulation[10, 11]. One of the reasons why the width of the minor groove matters is that some drugs actually bind to the minor groove. For example, the antitumor antibiotic *netropsin* binds to the B-DNA double helix, especially at the AT base pair regions, without intercalating[12]. The hydration pattern is also critical for the stability of DNA structure and this pattern strongly depends on the sequence of DNA. The hydration spine in the AT-tracts is a good example[13]. Sometimes the hydration pattern also plays a crucial role in protein-DNA interaction. The similar hydration patterns of the protein-DNA interface in the trp

repressor-DNA complex and the naked DNA target were seen and it is proposed that both protein and DNA specially recognize each other's hydration pattern[14].

In this report, we examine the stability of FoU in free base pair system and when incorporated in DNA double helix. We compute pairing energies of various free DNA base pairs with the density functional theory, focusing mainly on mispairing of FoU with G. We also consider pairings of deprotonated form and enol tautomer of FoU with G in all possible hydrogen-bonding geometries. In order to see how stable the oxidized base and the following mispairs are in DNA double helix and how they affect the overall DNA conformation, the molecular dynamics (MD) simulations for DNA dodecamers with normal A:T base pair, A:FoU base pair and G:FoU base mispair are then carried out.

A.2 Computational Methods

A.2.1 Quantum Mechanics (QM) calculation of pairing energies in free DNA base systems

All QM calculations were performed using the Jaguar v4.1 quantum chemistry software[15]. The geometries for 1-methyl pyrimidine, 9-methyl purine bases and all pyrimidine-purine base pairs were first optimized in the gas phase at the B3LYP/6-31G** level. The vibration frequencies for thermodynamic quantities were also calculated at the same level. Then the 6-31G**++ basis set was used for the final geometry optimization starting from the 6-31G**-optimized geometry. Since the calculation of vibration frequencies is a quite time consuming, the diffuse function was not included in the first step. To validate the exclusion of the diffuse function in the calculation of vibration frequencies, we have considered the following combinations of basis sets and compared the calculated enthalpies of base pairing with experimental ones:

(1) 6-31G**/6-31G** (No diffuse function is included in both steps.)

(2) 6-31G**/6-31G**++ (The preliminary geometry optimization and the calculation of frequencies are done with 6-31G** basis set and then the geometry is re-optimized with 6-31G**++ basis set.)

(3) 6-31G**++/6-31G**++ (The diffuse function is included in both steps.)

No scaling factor was adopted in the frequency calculation. All thermodynamic quantities were computed at 300 K, based on standard ideal-gas statistical mechanics and the rigid-rotor harmonic oscillator approximation. The enthalpy (or free energy) for each species is defined as:

$$H_{300K}\,(or\ G_{300K}) = E_{0K} + ZPE + \Delta H_{0\to 300K}\,(or\ \Delta G_{0\to 300K}),$$

where the E_{0K} is the total energy of the molecules at 0 K calculated from QM, ZPE is the zero-point energy and, $\Delta H_{0\to 300K}$ (or $\Delta G_{0\to 300K}$) is the change of enthalpy (or free energy) from 0 K to 300 K.

The single point energy calculation was carried out for the free energy of solvation in water, G_{solv}, for the final optimized structure at the B3LYP/6-31G**++ level. The solvation free energies are computed with a self-consistent reaction field method by solving the Poisson-Boltzmann equation. For the dielectric constant of water, we used $\varepsilon_{H2O} = 80.37$ which is at 20 °C[16]. The probe radius was set to 1.40 Å. We used the default values for the van der Waals radii of atoms[17]. The free energy of the system in aqueous solution is given by

$$G_{aq} = G_{300K} + G_{solv}\,.$$

The calculations of pairing free energies were performed for various DNA base pairs, focusing on mispairing of FoU with G. Pairing of deprotonated form or enol tautomer of FoU with G was also considered in all possible hydrogen-bonding geometries. In this calculation, the basis set superposition error (BSSE), which is the artificial lowering in the complex energy relative to that of the separated monomers since the complex basis set is larger than that of each

monomer, should be taken into account. Since the free bases undergo the conformational change upon pairing, their relaxation energy terms were also incorporated into the estimation of the BSSE correction[18]. Therefore, the following BSSE-correction energy, E_{BSSE}, should be added to the "raw" pairing energy, ΔE_{0K}:

$$E_{BSSE} = \left[E_{AB}^{a}(A) - E_{AB}^{a\cup b}(A)\right] + \left[E_{AB}^{b}(B) - E_{AB}^{a\cup b}(B)\right]$$

$$\Delta E_{0K} = E_{AB}^{a\cup b}(AB) - \left[E_{A}^{a}(A) + E_{B}^{b}(B)\right]$$

where *a* and *b* are the basis sets for corresponding bases, *A* and *B*, and the *A*, *B* and *AB* on the subscript represent the geometries where the energies for the species inside the parenthesis were computed. The final equation form for BSSE-corrected free energies in gas and in aqueous solution is followed as:

$$\Delta G_{300K}(g) = \left\{G_{300K}(AB) - \left[G_{300K}(A) + G_{300K}(B)\right]\right\} + E_{BSSE}$$

$$\Delta G_{300K}(aq) = \left\{G_{aq}(AB) - \left[G_{aq}(A) + G_{aq}(B)\right]\right\} + E_{BSSE}.$$

A.2.2 Molecular dynamics (MD) simulation of DNA dodecamer system containing the FoU

The DNA dodecamer containing FoU has been crystallized recently[19]. It was a Dickerson-type dodecamer with the sequence d(CGCGAAT(FoU)CGCG) where one of the middle thymines was replaced with 5-formyluracil. The starting structure for our MD simulation was taken from one of these crystal structures (PDB ID: 1G8V). Three sets of simulations were carried out; one with normal Dickerson sequence, another with the FoU crystal structure and the other where A paired with FoU in the crystal structure was replaced by G. The formyl group in FoU could have a syn or an anti conformation to C4 atom. In the crystallographic study, the formyl group of one FoU adopts a syn conformation, but the other is distorted between the syn and anti conformation with almost equal occupancies. For our dodecamers, the formyl group of the FoU at each strand was assigned to be in the different conformation. The AMBER6 program

package was used for the simulations[20]. However, the FoU is a non-natural DNA base and the AMBER6 does not provide the charges and force-field (FF) parameters for it. Therefore we generated the charges and FF parameters with the consistent way used in the development of the PARM94 in AMBER.

Determination of charges and FF parameters for FoU

We took the thymine nucleoside structure (DTN) in AMBER6 as the initial structure and then changed the methyl group at the C5 position to formyl group. For the enol tautomeric form, the carbonyl group at C4 was converted to the hydroxyl group. We built the syn and the anti conformation of the formyl group separately. With those structures, the geometry optimization was performed at the HF/6-31G* level using Jaguar v4.1. The electrostatic potential (ESP) was calculated for the final geometry and was used as an input for the RESP module in AMBER6 to obtain the charges[21]. In AMBER6, sugar atoms have intermolecularly equivalent charges with the exception of C1' and H1' atoms. Those atoms were constrained to have the same charges given in AMBER6 during charge fitting. The sum of charges for hydrogen and oxygen in the hydroxyl group at the 3' and 5' terminal of nucleoside was constrained. The force-field atom types of modified part were assigned using the Antechamber module in AMBER7[22]. The consistent atoms with thymine kept the same force-field atomic types as in thymine. The common force-field parameters with thymine was taken from the Cornell *et al.* force field[23] given in AMBER6 and non-available parameters there were from the "general amber force field" (gaff.dat). Table A.3 summarizes the FF atomic types and the charges used in this simulation.

MD simulation procedures

The DNA dodecamer was embedded in a rectangular box of TIP3P water molecules extended by 10 Å in each direction of a DNA solute where there were approximately 4000 water molecules. The sodium cations were added to neutralize the system at the electronegative points determined by the electrostatic potential that was calculated at the crude grid points. Some water

molecules clashed with cations were replaced with those ions. Most cations were located near the negatively charged phosphate groups. First, the minimization was performed with the DNA under the harmonic constraint (500 kcal/molÅ^2) while only waters and sodium ions movable to relieve the bad contact between DNA solute and waters or cations. In the next, the constant pressure MD was carried out with isotropic position scaling during 25 ps while the system was gradually heated from 0 K to 277 K under 1 bar with the DNA still constrained (500 kcal/ molÅ^2). Near the end of the simulation the density of the system reached ~1 g/cc. One more 25 ps constant pressure MD was done at constant temperature of 277 K. While releasing the constraint of the solute, the whole system was minimized. Then without any constraint, the whole system was gradually heated up from 0 K to 277 K under the constant pressure of 1 bar. After the system was fully equilibrated in this way, the long-term constant volume MD simulation was performed. All the MD simulations were done with 2 fs integration step. The particle mesh ewald (PME) method was used for the long-range electrostatic interaction. The cutoff distance of 9 Å was used for van der Waals interaction of Lennard–Jones type.

The helical parameter analysis was done using the Curves 5.2 program[24]. The O7 of formyluracil was removed during analysis because the presence of O7 alters the definition of base axis system on the pyrimidine and affects the helical parameters especially related to bases.

A.3 Results and discussion

A.3.1 QM calculations of base pairing energies

Figure A.1 shows the hydrogen bonding patterns of FoU and deprotonated FoU with A or G. They are final QM-optimized structures obtained by the method described in the previous section. While the keto tautomer of FoU forms the hydrogen bonds of Watson-Crick type with A, the enol tautomer forms the Wobble type. The enol tautomer can form the Watson-Crick type of

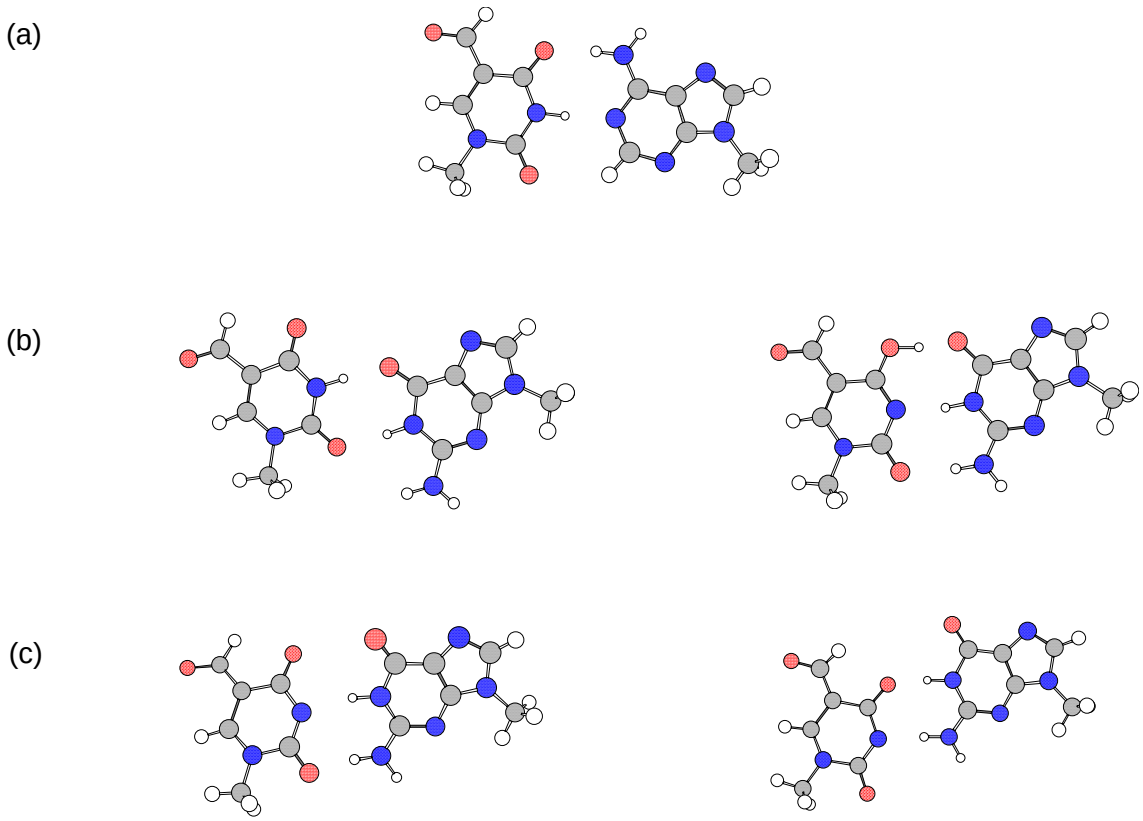


Figure A.1 Hydrogen bonding patterns of FoU with purine bases; QM-optimized structures, (a) A-FoU(keto) Watson-Crick, (b) G-FoU(keto) Wobble (*left*) and G-FoU(enol) Watson-Crick (*right*), (c) G-FoU(deprotonated) Watson-Crick (*left*) and Wobble (*right*).

mispairing with G through three hydrogen bonds. For the deprotonated FoU, both types of hydrogen bonding are possible with G.

The enthalpies of base pairing in gas phase for the canonical AT and GC pairs are shown in Table A.1. It can be seen that the exclusion of diffuse function on the frequency calculation does not make any difference on calculation of enthalpies. In all three cases, the calculated values agree fairly well with experimental ones, even though the slight improvement on the pairing enthalpy of GC is shown when the 6-31G**++ basis set is used for the final geometry optimization. In the case of the anion species like the deprotonated FoU, the diffuse function

Table A.1 Base pairing enthalpies in gas phase at 300 K for GC and AT calculated using B3LYP DFT method with three combinations of basis set[§]

	6-31G**/ 6-31G**	6-31G**/ 6-31G**++	6-31G**++/ 6-31G**++	Exptl[a]
AT (Watson-Crick)	-10.9	-10.5	-10.6	
AT (Hoogsteen)	-11.5	-11.1	-11.1	-13.0
GC (Watson-Crick)	-24.4	-23.3	-23.4	-21.0

[a] Reference [25], [§] unit: kcal/mol

should be included and the 6-31G**/6-31G**++ combination has been chosen for all other calculations in this study.

When the FoU is paired with A, the pairing free energies slightly increase in solution phase as well as in gas phase, compared with those of A-T Watson-Crick pair. Since the formyl group is an electron-withdrawing group, the inductive effect makes the charges on the pyrimidine ring deficient, and the hydrogen bond would become stronger if the FoU plays a role as a hydrogen donor. However, the FoU forms two hydrogen bonds with A both as a donor and as an acceptor. Therefore such an enhancement would be nullified and the pairing energy of A-FoU would become similar to that of AT. From the fact of the slight stabilization in A-FoU, it can be said that the hydrogen bond between H3 in FoU and N1 in A plays a more important role in the A-FoU pairing as previously shown by Kawahara et al.[26].

The FoU and G can form two hydrogen bonds in the Wobble geometry and their pairing free energy is comparable to that of FoU-A pair. The T-G pair also has the similar strength of hydrogen bond to T-A pair. It shows that the FoU and T can be paired with G as frequently as with A when they exist as the free bases.

Tautomerism of FoU

Table A.2 Pairing free energies in gas phase and in aqueous solution calculated using B3LYP DFT method with 6-31G**//6-31G**++ basis sets[a]

	$\Delta E(BSSE)$	ΔH_{300K}	$\Delta G_{300K}(g)$	G_{solv}	$\Delta G_{300K}(aq)$	$\Delta\Delta G_{300K}(aq)^{keto\rightarrow enol}$
TA [WC][e]	-11.9	-10.5	1.0	10.0	11.0	
TA [H][e]	-12.5	-11.1	0.5	10.3	10.9	
CG	-25.1	-23.3	-9.1	19.8	10.7	
FoUA	-12.4	-11.1	0.34	10.0	10.3	
FoUG	-12.9	-11.5	-0.05	10.3	10.2	
FoU´G[b]	-25.9 (-14.8)[c]	-25.2 (-14.4)[c]	-12.4 (-1.4)[c]	18.9	6.5 (15.9)[c]	9.4
TG	-12.8	-11.3	0.5	10.6	11.1	
T´G[b]	-27.1 (-15.3)[c]	-26.1 (-14.7)[c]	-13.6 (-1.5)[c]	19.5	5.9 (15.4)[c]	9.5
FoU⁻G [WC][e]	-23.4	-21.9	-10.1	19.2	9.1 (10.4)[c]	1.3*[d]
FoU⁻G [W][e]	-29.0	-27.5	-16.2	25.3	9.1 (10.4)[c]	1.3*[d]

[a] Unit: kcal/mol, [b] FoU´ and T´: enol tautomers of FoU and T, [c] ( ): considering energy penalty relative to the keto form of neutral FoU, [d] *: from *J. Phys. Chem. A* **105** 274 (5-formyluracil, T = 298 K, pH = 7.00), [e] WC: Watson-Crick; H: Hoogsteen; W: Wobble

The FoU and T can have enol tautomeric forms. In both FoU and T cases, the keto form is energetically more favorable than the enol form as shown in Table A.2 and the calculated equilibrium constants of tautomerism, which are defined as the concentration ratio of enol form to keto form, are 1.2 x 10^{-7} and 1.3 x 10^{-7} at 300 K in aqueous solution for FoU and T, respectively. However, one of enol tautomers could form three hydrogen bonds with G as shown in Figure A.1 and the barrier of tautomerism would be compensated by one extra hydrogen bonding. Actually the calculation results show that the pairing of enol form with G in gas phase is slightly favorable even after considering energy penalties (11.1 kcal/mol in E_{0K} for FoU and 11.8 kcal/mol in E_{0K} for T with respect to the keto form). On the other hand, it becomes unfavorable in aqueous solution because of large cost of solvation energy on pairing. If we assume that the DNA bases would be in the lower dielectric environment in oligonucleotides than in water, the pairing of enol form with G would be energetically plausible in the biological system.

Deprotonated form of FoU

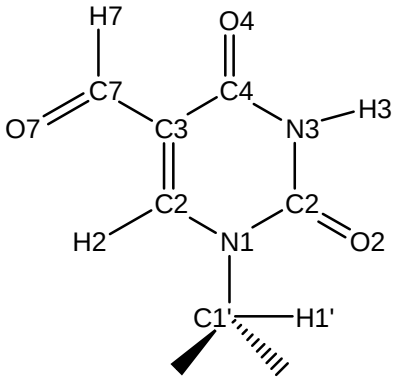
The smaller pKa of 5-formyl deoxyuridine predicts the existence of N3-deprotonated, negative species on a larger amount at the physiological pH[4]. At pH = 7 and T = 300 K, the 7.6% of 5-formyl deoxyuridine would dissociate into negative deoxyuridinium ion and proton while only 0.2% dissociates for deoxythymidine. The deprotonated FoU plays a role only as a hydrogen acceptor and therefore the pairing with A is expected to be extremely weak. Two possible geometries of pairing between FoU^- and G are shown in Figure A.1. The first one (geometry I) corresponds to Watson-Crick geometry, which could be optimal since it does not distort the overall backbone geometry in the normal DNA. The interesting thing is that both geometries have a big stabilization on pairing in gas phase and their pairing free energies are comparable to that of the triple hydrogen-bonding pair such as GC. In the geometry I the repulsion between two electronegative oxygens destabilizes the hydrogen bonding, and actually the purine and pyrimidine rings are no longer co-planar in this structure. The extra stability in the pair of deprotonated FoU with G could come from the ion and ion-induced dipole interaction since the permanent dipole for the isolated guanine does not point toward the negatively charged FoU. In aqueous solution, the solvation energy for the isolated FoU^- is quite huge and the final free energy in solution becomes comparable to those of neutral G-FoU Wobble pair and A-FoU Watson-Crick pair. Considering that the base pair is not fully exposed to water in the oligonucletide, these results support the mechanism that the ionization could allow formation of mispair with G during DNA replication and it would induce the transition mutation at the oxidized T site.

A.3.2 MD simulations of dodecamers

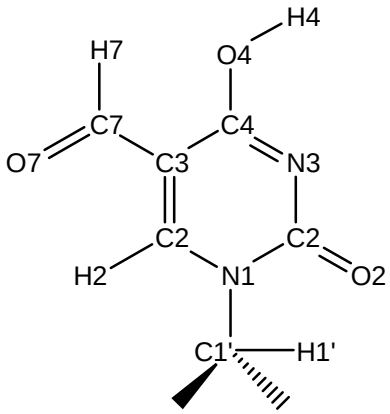
In the present MD simulation, we consider the most dominant keto form of FoU. The deprotonated FoU that might play a role in mispairing during the DNA replication step would turn into the thermodynamically most stable keto species.

AMBER force field parameters of FoU

Table A.3 The AMBER type force field parameters of 5-formyluracil – keto form (*top*) and enol form (*bottom*)



atom label	*FF atomic type*	*charges*	
		(anti)	*(syn)*
C1'	CT	0.166	0.142
H1'	H2	0.137	0.155
N1	N*	-0.010	-0.001
C6	CM	-0.196	-0.280
H6	H4	0.298	0.286
C5	CM	-0.055	-0.041
C7	C	0.396	0.396
O7	O	-0.518	-0.455
H7	HA	0.067	0.001
C4	C	0.412	0.522
O4	O	-0.527	-0.512
N3	NA	-0.295	-0.393
H3	H	0.305	0.319
C2	C	0.496	0.555
O2	O	-0.554	-0.573



atom label	*FF atomic type*	*charge*	
		(anti)	*(syn)*
C1'	CT	0.205	0.182
H1'	H2	0.104	0.121
N1	N*	-0.110	-0.083
C6	CM	-0.072	-0.178
H6	H4	0.243	0.237
C5	CM	-0.106	-0.094
C7	C	0.343	0.348
O7	O	-0.491	-0.457
H7	HA	0.074	0.038
C4	CA	0.620	0.715
O4	OH	-0.614	-0.587
H4	HO	0.469	0.454
N3	NC	-0.738	-0.775
C2	C	0.787	0.800
O2	O	-0.592	-0.600

Table A.4 The base pairing energies and the distances between H-bond donors and acceptors for the base pairs involved in the dodecamers of this work (FUA : anti conformer, FUS : syn conformer)

(Unit: kcal/mol)

	AMBER FF[a]	QM[b]
A:T	- 15.1	-11.9
G:C	-29.0	-25.1
A:FUA	-14.3	-12.4
A:FUS	-14.3	-12.5
G:FUA	-16.3	-12.9
G:FUS	-16.4	-12.7

(Unit: Å)

	AMBER FF[a]	QM[b]	X-ray[c]
A:T Watson-Crick			
N1-N3	2.85	2.91	2.82
N6-O4	2.79	3.01	2.95
G:C Watson-Crick			
N1-N3	2.85	2.98	2.95
N2-O2	2.74	2.94	2.86
O6-N4	2.78	2.83	2.91
A:FUA Watson-Crick			
N1-N3	2.86	2.89	
N6-O4	2.80	2.90	
A:FUS Watson-Crick			
N1-N3	2.86	2.88	
N6-O4	2.81	3.08	
G:FUA Wobble			
N1-O2	2.75	2.95	
O6-N3	2.80	2.77	
G:FUS Wobble			
N1-O2	2.74	2.83	
O6-N3	2.80	2.98	

[a] dielectric constant = 1; scaling of 1-4 vdW interaction = 0.5; scaling of 1-4 electrostatic interaction = 0.83; for deoxynucleosides [b] gas phase calculation; BSSE corrected; for 1-methylpyrimidines and 9-methylpurines [c] From experimental X-ray crystallographic data [26].

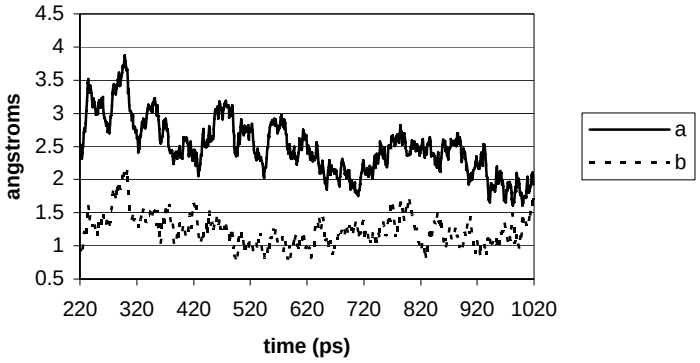


Figure A.2 Fluctuation in the root mean square deviation of coordinates (CRMSD) of DNA dodecamer containing FoUs (1G8V.pdb) during 1 ns simulation after equilibration; a : CRMSD with respect to the minimized DNA structure, b : CRMSD with respect to the mean structure over 220–1020 ps.

The force field (FF) atomic types and charges for FoU developed for this study are tabulated in Table A.3. To validate these parameters the base paring energies and geometries obtained with AMBER FF are compared with those from QM calculations as shown in Table A.4. The overall pairing energies are a little overestimated even in the cases of the canonical AT and GC pair, although this might result from the extra non-bond interaction between sugar rings in nucleosides. However, the extent of the overestimation for the pairs with formyluracil is comparable to the GC and AT cases. The hydrogen bond distances agree well with each other and the differences are within 0.3 Å. To check the stability of the DNA conformation during the MD simulation with newly implemented charges and FF parameters for FoU, the time evolution of the root mean square deviation of coordinates (CRMSD) was calculated for the dodecamer X-ray crystallographic structure containing FoU (1G8V) in Figure A.2. The CRMSD value with respect to the mean structure is 1.24 ± 0.24 Å and it is comparable to the one calculated for the DNA system with normal base sequence.

Hydrogen bond distance

The stability of DNA structure is directly related to the hydrogen bonds (H-bonds) between two base pairs. The bond distances between H-bond donor and acceptor atoms were measured

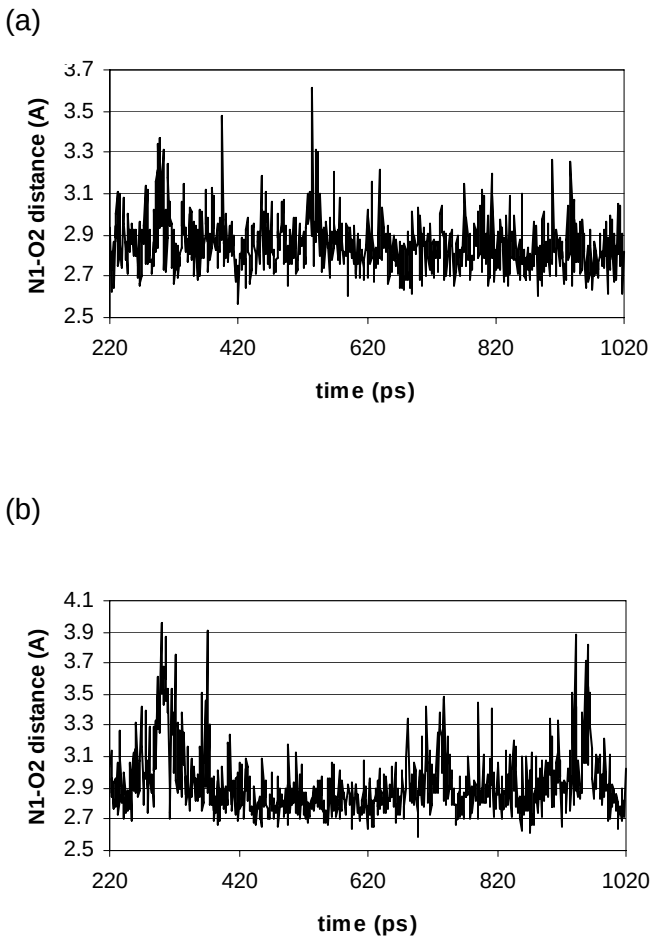


Figure A.3 The time profile of H-bond distance between N1 from G and O2 from FoU; (a) for the syn conformer of FoU at the 5th position and (b) for the anti conformer of FoU at the 8th position.

every 1 ps after 220 ps during the production period. For the DNA with the normal Dickerson sequence, the distances are within 3.1 Å in most times except for the bases at the terminal. The base pairs at the 5' terminal started unraveling around 420 ps and formed the H-bonds back in 50 ps later. The H-bonds at the 3' terminal broke around 620 ps and stayed unraveled until the end of the simulation. The floppiness of the bases at the terminal is usual since they have only the one-side stacking interaction. In the case of the dodecamer with A:FoU pairs, a similar phenomena were observed. The FoU in the middle of the dodecamer does not cause any instability in H-bonds and the DNA kept the stable conformation during the simulation. However, when the FoU

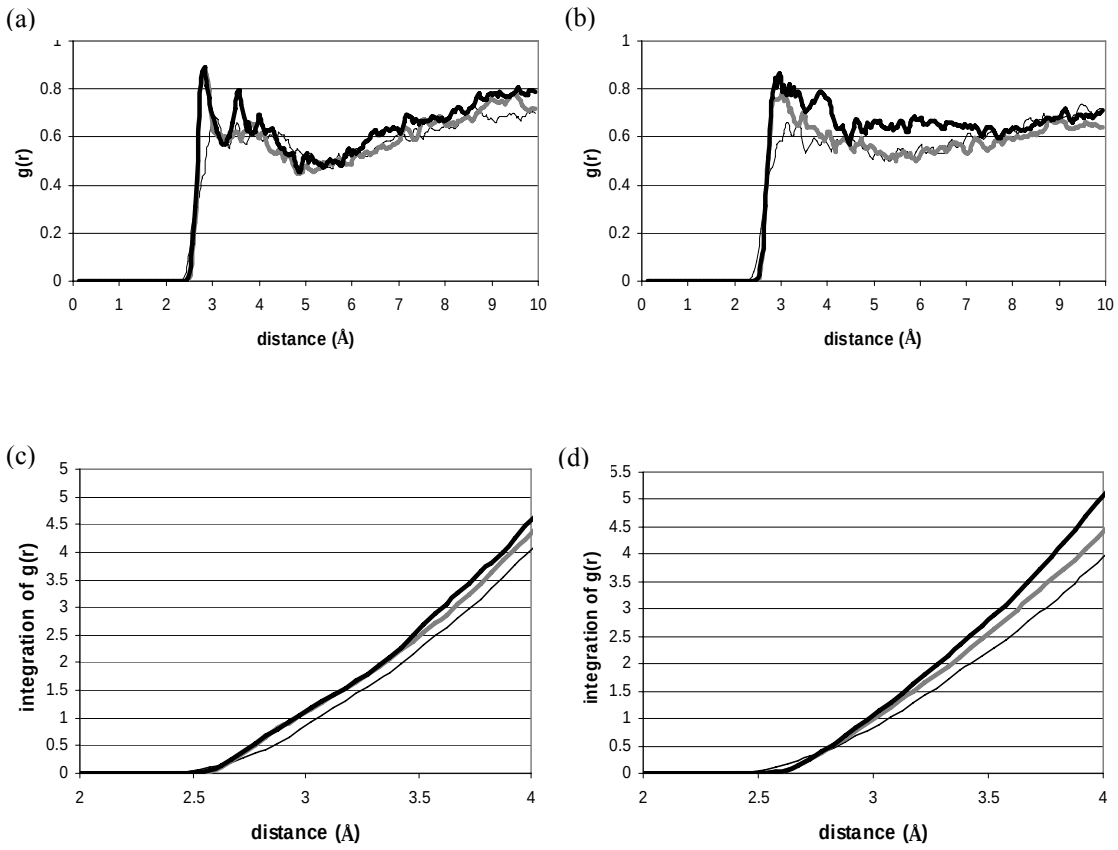


Figure A.4 [(a), (b)] Normalized radial distribution functions g(r) of water-oxygen [(c), (d)] and the number of waters in the solvation shell obtained by integrating g(r). The thick black line, the thick gray line and the thin black line show g(r) of the target atoms, O7 of FoU in the G:FoU case, O7 of FoU in the A:FoU case and H7 of T in the A:T case respectively. (a) and (c) : the formyl group of FoU is anti at the 8th base pair position; (b) and (d) : the formyl group is syn at the 5th base pair position. g(r) was normalized by the water of 1 g/cm^3. $n = 4\pi\rho\int g(r)r^2 dr$, where ρ is 0.033 molecules/Å^3 for water of 1 g/cm^3.

is paired with G, the large fluctuation in H-bonds for the G:FoU pair was observed, especially in the case of FoU with the formyl group in the anti conformation. For the syn conformer, the H-bonds were pretty steady.

DNA hydration

When the methyl group in thymine is substituted with the formyl group, this extra oxygen (O7) can play a role as a hydrogen bond acceptor. Figure A.4 shows the radial distribution function, g(r) of oxygens in water solvent.

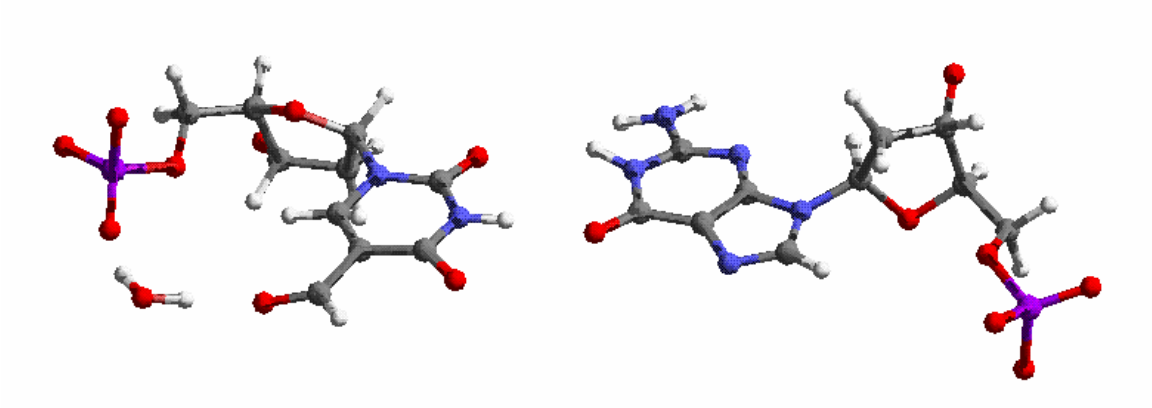


Figure A.5 The snapshot of guanine and formyl uracil with the formyl group in anti conformation (at 301 ps). The water makes a hydrogen bond bridge between O7 of FoU and O2P of the phosphate group. The O-O distance between water and O7 is 2.7 Å and the other O-O distance between water and O2P is 2.5 Å.

The first sharp peak that is not prominent in the thymine case is observed for the FoU case, especially when the formyl group is in the anti conformation. It shows that the waters around the oxygen of formyl group in anti are well ordered. This is because the water can make a bridge between O7 of FoU and O2P or O5' from the backbone when the formyl group is anti as shown in Figure A.5. In the syn conformation, the O7 is located away from these oxygen atoms and the O4 of FoU is too close to O7 atom for a water to make H-bond bridge. When g(r) is integrated over the first coordination shell ($r \sim 3.3$ Å) for FoU at the 8th position of G:FoU and A:FoU case, the number of waters is approximately 1.8 for both cases. We can clearly see that the more water molecules are around the FoU than the thymine.

Groove widths

The widths of the major and minor grooves for Dickerson crystal structure (PDB ID: 1BNA) and FoU crystal structure (PDB ID: 1G8V) were calculated using Curves program. They have the same crystal symmetry. If we ignore that the different experimental condition where the crystals were grown might affect the conformational differences, Figure A.6 shows definitely sequence-dependence of the groove widths. Although the overall shapes in the graphs are similar

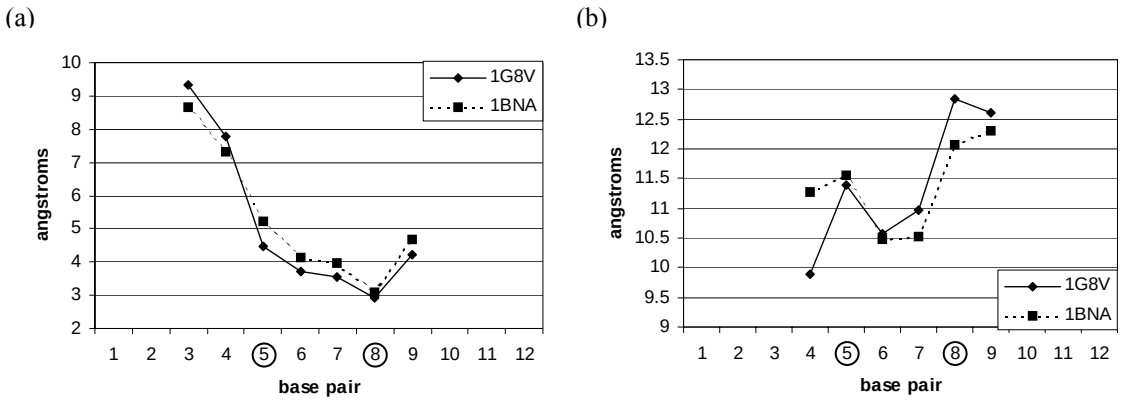


Figure A.6 The widths of minor (a) and major (b) grooves. The solid line is for the crystal structure with 5-formyluracil (1G8V) and the dotted line is for Dickerson crystal structure (1BNA). The positions where the sequence differences are shown are circled.

to each other, the widths of minor grooves for the FoU case become slightly narrower and the major grooves get wider.

The groove widths averaged over the MD simulation from 220 ps to 1020 ps are shown in Figure A.7. Since they are isolated DNA molecules immersed into the water box and less constrained, the absolute values of widths are larger than in the crystal structure. The dodecamer with the normal Dickerson sequence has an asymmetric distribution in minor groove throughout the sequence even though the sequence itself is symmetric. However, the dodecamer with FoU:A has a symmetric pattern, and the different conformation of formyl group does not seem to affect the width of minor groove. The replacement of T with FoU results in the significant decrease in the width of the minor groove. In the FoU:G case, the minor groove becomes narrower around the 5th base pair position where the syn FoU is located and on the other hand the minor groove becomes wider at the 8th base pair position where the anti FoU is located. The major grooves become narrower for the A:FoU DNA and this change is more prominent on the side of anti FoU. There is a huge increase in the width of the major groove near the 5th base pair for the G:FoU DNA.

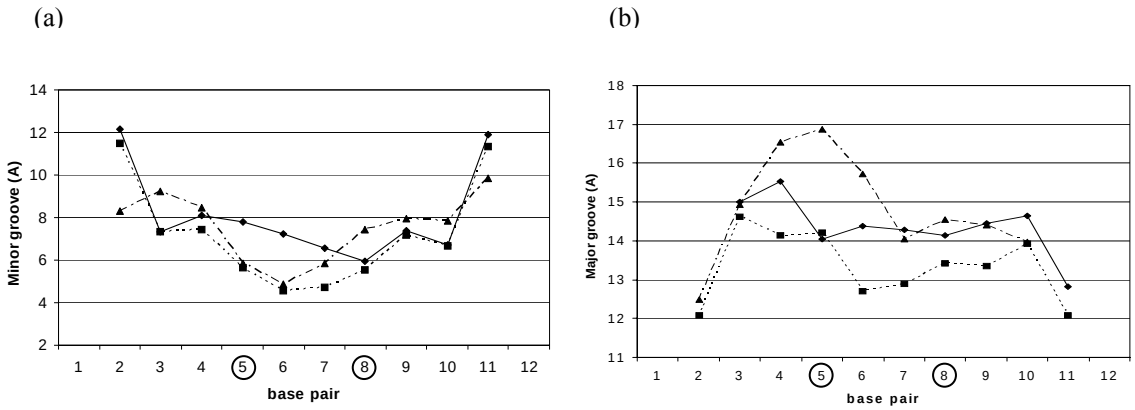


Figure A.7 The minor (a) and major (b) groove widths averaged over the MD simulation from 220 ps and 1020 ps. The diamond, square and triangle symbol are for the normal, A:FoU and G:FoU DNA dodecamer, respectively. The syn FoU is at the 5th position and the anti FoU is at the 8th position as indicated by circles.

Helical parameters

The global base-base parameters are analyzed and shown in Figure A.8. These parameters could be the indication of the stability in hydrogen bonding. The G:FoU DNA dodecamer has the large shearing at the 5th and 8th position where G:FoU mispairs are located. The G and FoU have the Wobble geometry and they are sliding each other compared to the pyrimidine-purine pair in Watson-Crick geometry. Therefore the large values in shear parameter reflect the Wobble geometry of G:FoU pair. Since the measurement is done in the 5'→3' direction, they have the opposite sign even though the sequence is symmetric. The conformation of formyl group does not make any difference in shearing.

Large buckling is detected at the 4th and 9th base pair in the A:FoU DNA dodecamer, compared with the case of normal Dickerson sequence. These pairs that flank the central AATT sequence have respectively positive and negative buckles that bend the center of these base pair away from the central tetramer. This may contribute to severe narrowing of the minor grooves in the A:FoU dodecamer. The similar huge positive buckling is shown only at the 4th base pair

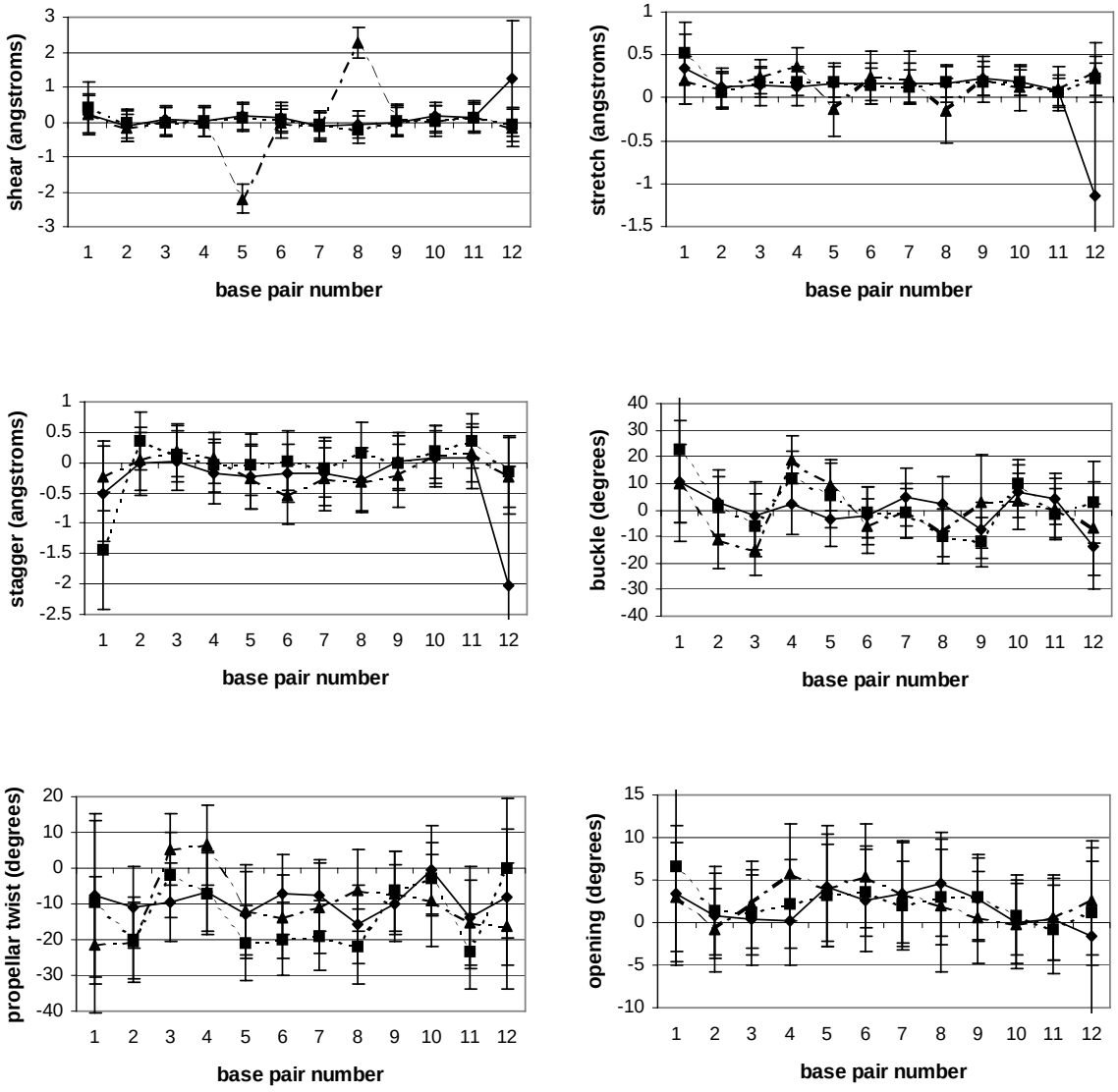


Figure A.8 The global base-base parameters. They are averaged values over the MD simulation from 220 ps and 1020 ps. The diamond, square and triangle symbol are for the normal, A:FoU and G:FoU DNA dodecamer, respectively.

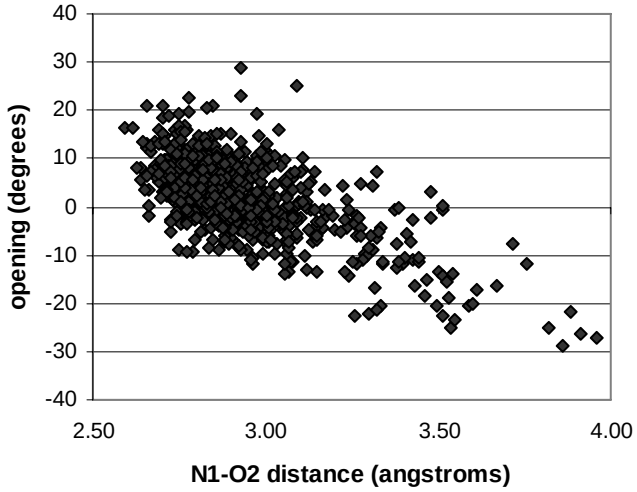


Figure A.9 The plot of N1-O2 distance versus the opening at the 8th G:FoU pair. The snapshot was taken every 1 ps during 220-1020 ps of MD simulation.

position in the G:FoU mispair case and it is reflected as the asymmetric narrowing of minor groove in the central part of the G:FoU dodecamer.

The large positive opening in base pair means opening in the major groove and thus narrowing of the minor groove. The deviation from the normal Dickerson sequence case that might explain the significant narrowing in the dodecamers with FoU present is not small, and in all three systems the width of minor groove shows the negative correlation with opening. Figure A.9 shows the correlation between the H-bond distance of N1-O2 at the 8th base pair in G:FoU dodecamer and the opening. In most times, the N1-O2 distance is near 2.9 Å and the opening fluctuates by 10º around zero. When the N1 and O2 get apart, the opening becomes more negative. This loose H-bond at the 8th base pair contributes to the larger width of minor groove than one at the 5th base pair as shown in Figure A.7.

Backbone parameters

The torsion angles for a polydeoxyribonucleotide chain and the pseudorotation phase angle P of a sugar ring are calculated for three dodecamer systems. The distinct sequence-dependent

aspects are not seen here. The preferred sugar puckering modes are O4'-endo, C1'-exo and C2'-endo that correspond to the "south" conformations. The structures keep B-type DNA conformations over the MD simulation. The phase angle P and the δ torsion (C5'-C4'-C3'-O3') at the 7th base residue and the 16th base residue show a little correlation with the opening of G and FoU base at the 8th base pairs. The 7th and 16th bases are right ahead of this G:FoU pair in the 5'→3' direction. The correlation coefficients for the P and δ torsion of the 7th base are 0.36 and 0.36, respectively. For the 16th base, they are 0.32 and 0.34 for the P and δ torsion.

A.4 Summary and Conclusion

We calculated pairing free energies of various free DNA base pairs at the B3LYP/6-31G**//B3LYP/6-31G**++ level, focusing on mispairing of 5-formyluracil which is an oxidative form of thymine. The free energy of keto FoU with G is comparable to that with A in both gas phase and solution phase while the pairing of enol FoU with G in solution phase is most unfavorable due to large cost of solvation free energy on pairing in addition to the barrier on tautomerism. The N3-deprotonated FoU forms strong hydrogen bonding with G in gas phase, which is energetically comparable to the triple hydrogen bonding of GC pair. The calculation in aqueous phase shows that mispairings of both neutral keto and deprotonated FoU with G are as probable as normal base pairings.

Considering that the neutral FoU could be mispaired as frequently as T with G from the aspect of energetics, we conclude that the ionization at N3 position of FoU would mainly account for the increased misparing rate of FoU since the deprotonated FoU preferentially form H-bonds with G rather than A and therefore FoU has one more extra possibility of base pairing.

The 1 ns MD simulations were then carried out for three DNA dodecamer-explicit water systems; one with normal Dickerson-type sequence, another with two thymines replaced by formyluracil in the Dickerson sequence and the other where these formyluracils are paired with

guanines. Even though the formyl group is on the side of the major groove, its presence actually leads to the severe narrowing of the minor grooves. In the case of G:FoU dodecamer, the same kind of narrowing was shown especially around the 5th base pair where the syn FoU makes a pair with G. The slightly wider minor groove at the 8th base pair position where the formyl group of FoU has anti conformation, is correlated with loosening of H-bonds between G:FoU.

The formyl group of FoU in the anti conformation affects the hydration pattern around the DNA structure. A water molecule makes a bridge of H-bond between O7 of FoU and O2P of phosphate and it provides the well-ordered water structure. No significant change in backbone parameters is shown for A:FoU and G:FoU dodecamer.

Overall the incorporation of FoU paired with A does not cause the significant structural change in DNA double helix except for the narrowing of the minor groove. On the other hand, the G:FoU dodecamer shows relatively larger fluctuation since it contains non Watson-Crick pairs. It might be worth studying how this kind of conformational distortion affects the interaction with a DNA-binding protein, for example like the DNA repair enzyme. More detailed molecular level description also should be investigated to explain the effect of the formyl group on the width of minor grooves.

References

1. Ames, B.N., M.K. Shigenaga, and T.M. Hagen, *Oxidants, Antioxidants, and the Degenerative Diseases of Aging.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(17): p. 7915-7922.
2. Kasai, H., et al., *5-Formyldeoxyuridine - a New Type of DNA Damage Induced by Ionizing-Radiation and Its Mutagenicity to Salmonella Strain Ta102.* Mutation Research, 1990. **243**(4): p. 249-253.
3. Bjelland, S., et al., *Cellular effects of 5-formyluracil in DNA.* Mutation Research-DNA Repair, 2001. **486**(2): p. 147-154.
4. Privat, E.J. and L.C. Sowers, *A proposed mechanism for the mutagenicity of 5-formyluracil.* Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis, 1996. **354**(2): p. 151-156.
5. Masaoka, A., et al., *Oxidation of thymine to 5-formyluracil in DNA promotes misincorporation of dGMP and subsequent elongation of a mismatched primer terminus by DNA polymerase.* Journal of Biological Chemistry, 2001. **276**(19): p. 16501-16510.
6. Mol, C.D., et al., *DNA repair mechanisms for the recognition and removal of damaged DNA bases.* Annual Review of Biophysics and Biomolecular Structure, 1999. **28**: p. 101-128.
7. Masaoka, A., et al., *Enzymatic repair of 5-formyluracil I. Excision of 5- formyluracil site-specifically incorporated into oligonucleotide substrates by AlkA protein (Escherichia coli 3- methyladenine DNA glycosylase II).* Journal of Biological Chemistry, 1999. **274**(35): p. 25136-25143.
8. Terato, H., et al., *Enzymatic repair of 5-formyluracil II. Mismatch formation between 5-formyluracil and guanine during DNA replication and its recognition by two proteins*

involved in base excision repair (AlkA) and mismatch repair (MutS). Journal of Biological Chemistry, 1999. **274**(35): p. 25144-25150.

9. Zhang, Q.M., et al., *Identification of repair enzymes for 5-formyluracil in DNA - Nth, Nei, and MutM proteins of Escherichia coli.* Journal of Biological Chemistry, 2000. **275**(45): p. 35471-35477.
10. Hamelberg, D., et al., *Flexible structure of DNA: Ion dependence of minor-groove structure and dynamics.* Journal of the American Chemical Society, 2000. **122**(43): p. 10513-10520.
11. Young, M.A., B. Jayaram, and D.L. Beveridge, *Intrusion of counterions into the spine of hydration in the minor groove of B-DNA: Fractional occupancy of electronegative pockets.* Journal of the American Chemical Society, 1997. **119**(1): p. 59-69.
12. Kopka, M.L., et al., in *Structure and motion : membranes,nucleic acids & proteins*. 1985, Adenine Press. p. 461-483.
13. Shui, X.Q., et al., *The B-DNA dodecamer at high resolution reveals a spine of water on sodium.* Biochemistry, 1998. **37**(23): p. 8341-8355.
14. Schwabe, J.W.R., *The role of water in protein DNA interactions.* Current Opinion in Structural Biology, 1997. **7**(1): p. 126-134.
15. Jaguar. 2000, Schrodinger Inc.: Portland.
16. *CRC Handbook of Chemistry and Physics*. 60th ed, ed. R.C. Weast. 1979, Boca Raton, FL: CRC Press.
17. Tannor, D.J., et al., *Accurate First Principles Calculation of Molecular Charge-Distributions and Solvation Energies from Ab-Initio Quantum- Mechanics and Continuum Dielectric Theory.* Journal of the American Chemical Society, 1994. **116**(26): p. 11875-11882.

18. Xantheas, S.S., *On the importance of the fragment relaxation energy terms in the estimation of the basis set superposition error correction to the intermolecular interaction energy.* Journal of Chemical Physics, 1996. **104**(21): p. 8821-8824.

19. Tsunoda, M., et al., *Crystallization and preliminary X-ray analysis of a DNA dodecamer containing 2 '-deoxy-5-formyluridine; what is the role of magnesium cation in crystallization of Dickerson-type DNA dodecamers?* Acta Crystallographica Section D-Biological Crystallography, 2001. **57**: p. 345-348.

20. Case, D.A., et al., *AMBER6*. 1999, University of California, San Francisco.

21. Cieplak, P., et al., *Application of the Multimolecule and Multiconformational Resp Methodology to Biopolymers - Charge Derivation for DNA, Rna, and Proteins.* Journal of Computational Chemistry, 1995. **16**(11): p. 1357-1377.

22. Case, D.A., et al., *AMBER7*. 2002, University of California, San Francisco.

23. Cornell, W.D., et al., *A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules.* Journal of the American Chemical Society, 1995. **117**(19): p. 5179-5197.

24. Lavery, R. and H. Sklenar, *Curves 5.2*. 1997.

25. Yanson, I.K., A.B. Teplitsky, and L.F. Sukhodub, *Experimental Studies of Molecular-Interactions between Nitrogen Bases of Nucleic-Acids.* Biopolymers, 1979. **18**(5): p. 1149-1170.

26. Kawahara, S., et al., *Ab initio and density functional studies of substituent effects of an A-U base pair on the stability of hydrogen bonding.* Journal of Physical Chemistry A, 1999. **103**(42): p. 8516-8523.