Title
Structure Prediction and Virtual Screening: Application to G Protein-coupled Receptors

Author
Marko, Adam Christian

Publication Date
2009

Peer reviewed|Thesis/dissertation

Structure Prediction and Virtual Screening: Application to G Protein-Coupled Receptors

by

Adam Christian Marko

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Acknowledgements

Andrej Sali, for serving as my graduate advisor and steering me in the right direction.

Tom Ferrin, for useful input into this document, for support for two years, and for granting me some options to enable my continued growth as a bioinformatician.

Brian Shoichet, for useful input into this document, specifically docking procedures.

Peter Kolb, for detailed input into virtual screening, as well as additional docking advice, assistance with the ligand libraries, and molecular dynamics.

and

Julia Molla, for tying up the loose ends and dealing with all of the details.

Structure Modeling and Virtual Screening: Application to G Protein-coupled Receptors

Abstract

G Protein-coupled Receptors (GPCRs) make up the largest family of proteins in the human proteome. These receptors are the target of an estimated 40% of drugs, and the potential for additional therapeutics that target GPCRs is great. Here, we describe a procedure for modeling the Human Adenosine A1 Receptor, a protein for which no known crystal structure exists. This protein plays a role in many cellular processes, and may be involved in Creutzfeldt–Jakob Disease, a human prion disease. To this end, we also describe a virtual screening procedure used to find novel ligands of the Adenosine A1 Receptor. Finally, the resultant iterative process of docking and modification of the active site of the Adenosine A1 Receptor is described.

Table of Contents

Introduction

Protein Structure Modeling

The use of predicted protein structures is of great interest due to the comparatively small number of solved structures. There are approximately 55,000 solved structures in the PDB, compared to over 6 million protein sequences, (Eswar, Webb et al. 2007). Protein structure modeling remains an important field of research, and has far reaching implications for the pharmaceutical, agricultural, and other commercial industries, in addition to other areas of biological research, such as phylogenetic studies (Schwede, Sali et al. 2009).

While several methods of computational protein structure prediction exist, homology modeling remains the most reliable (Irwin and Shoichet 2005). Homology modeling is a process by which a protein sequence of unknown structure, called the target, is modeled into a three dimensional structure based on information from a sequence that has a known structure, called the template. Typically, a sequence identity between the target and template sequence of >50% results in a homology model that is accurate enough to use in high-resolution computational experiments, such as virtual screening (Sanchez, Pieper et al. 2000).

Homology modeling can be broken down into four main steps:.

1) Identification of homologous templates to the target sequence

   Existing databases of known structures are searched using any number of search algorithms, such as BLAST or PSI-BLAST. Typically, the sequence with the highest sequence identity to the target is chosen to serve as the template. In

many cases, there may be more than one template, and multiple templates can be used. The templates can be from different organisms other than that of the target (Chothia and Lesk 1986; Altschul et al 1997; Marti-Renom et al. 2000).

2) Generation of an alignment between the target and templates

Alignment software is used to create a sequence alignment between the target sequence and the template sequence. This alignment determines which residues are considered to be equivalent to guide the modeling software. There are numerous alignment algorithms and programs that can perform pairwise or multiple sequence alignments (Tramontano et al. 2001). In addition, due to ambiguities in alignments, some algorithms allow the generation of "suboptimal alignments", which explore probabilistically unfavorable alignment space (Marko, Stafford et al. 2007).

3) Using a homology modeling program to create the model

Modeling software generates a three dimensional structure based on the input alignment (Sali and Blundell 1993). Once the initial structure has been generated, additional steps can be taken to potentially refine the model by optimizing side-chains or performing molecular dynamics, for example.

4) Scoring the prediction

Though a model has been generated, it still remains a prediction. As a result, it is necessary to predict its errors. Multiple methods exist for scoring predicted protein structures, and they include statistics-based scoring, trained neural

networks, or scoring based on physical principles (Baker and Sali 2001; Eramian, Eswar et al. 2008).

Membrane Protein Structure Modeling

Membrane proteins remain a challenge to crystallize. In fact, only 1% of all structures in the PDB are membrane proteins. In light of the lack of structures, the ability to predict membrane protein structure will be of great importance for the foreseeable future (Caffrey 2003; Elofsson and von Heijne 2007).

Three dimensional structure predictions of membrane proteins were attempted before any high-resolution structure was solved. Early models of bacteriorhodopsin and G Protein-coupled Receptors were made using information from low-resolution experiments, such as electron microscopy. *Ab initio* structure prediction of membrane proteins remains difficult for a number of reasons, such as the inherent size of membrane proteins and correctly modeling the membrane part of the structure (Elofsson and con Heijne 2007).

Some methods for predicting membrane protein structure places primary importance on identifying the transmembrane helical segments. A typical transmembrane segment contains a stretch of predominantly hydrophopic residues. These residues must be long enough to span the lipid bilayer as an alpha helix. There are also recurring sequence motifs that can increase the accuracy of an alignment, and thereby the predicted structure (Elofsson and von Heijne 2007).

For a computational experiment such as virtual ligand screening, a reliable three-dimensional protein structure is required. At the current state of the art, comparative

modeling typically yields the most accurate model of a membrane protein. However, as a result of the relatively few number of solved membrane proteins, the field of membrane structure prediction is still in its early stages. Yet, if a template can be found that is greater than 30% sequence identity, the resultant models are comparable in accuracy to globular proteins of similar sequence identity (Forrest, Tang et al. 2006; Elofsson and von Heijne 2007).

G Protein-Coupled Receptors

G Protein-Coupled Receptors, or GPCRs, are membrane-bound receptors found only in eukaryotes. GPCRs play essential roles in the recognition and transmission of cellular signals. These receptors make up the largest family of proteins in humans, at approximately 800 sequences. There are five human GPCR families, including rhodopsin, secretin, adhesion, glutamate, and Frizzled/Taste2 (Fredriksson, Lagerstrom et al. 2003).

GPCRs remain an important focus area in structural biology because of their clinical relevance. They account for nearly 40% of the prescription pharmaceuticals on the market. Some notable examples of drugs that target GPCRs are Zyprexa, Clarinex, Zantac, and Zelnorm (Fillmore, 2004).

Challenges involved with modeling GPCRs

Given the importance of GPCRs in the human proteome alone, it is of great interest to the scientific community to be able to structurally characterize these proteins. Unfortunately, obtaining crystal structures of GPCRs is difficult, thus few GPCRs have

available structures (Table I). In light of the complications involving crystallization of these proteins, other structure-determination methods are desirable.

Table I. X-ray structures of GPCRs.

Protein	PDB Code	Resolution	Species
Rhodopsin	1GZM, 1HZX, 1JGJ, 1L9H, 1U19, 2I35, 2I37, 2J4Y, 2ZIY, 3C9L, 3DQB	2.2 - 4.2 Å	Bovine, squid
β2-Adrenergic Receptor	2R4R, 2R4S, 2RH1, 3D4S	2.4 - 3.4 Å	Human
β1-Adrenergic Receptor	2VT4	2.7 Å	Turkey
Adenosine A2A Receptor	3EML	2.6 Å	Human

Template-based protein structure prediction, also called homology modeling, is the most appropriate way to predict the structures of membrane proteins. Though recently membrane structures have been predicted by use of sparse restraint sets, these methods are new and have not been tested in a blind prediction sense. Furthermore, these methods are limited to membrane proteins of about 250 amino acid residues in length and are somewhat coarse-grained (Barth, Wallner et al. 2009).

To build a GPCR model of high accuracy, a homology model must be constructed. Comparative modeling is currently the only method that will allow a computational experiment such as virtual screening to be performed.

Virtual Screening

Homology models are useful in structure-based drug discovery, facilitating the investigation of ligand-protein interactions in an effort to find novel ligands and improve their potency. One technique, “virtual screening”, computationally tests large libraries of organic molecules for those that complement the structure of a protein binding site.

While this is useful when there are known crystal structures, homology models can accelerate the virtual screening process and can support decision making in the event that crystal structures do not exist. In addition, homology models can also provide valuable insight before experimental high-throughput screening begins. (Schwede, Sali et al. 2009; Michino, Abola et al. 2009; Fan et al. 2009).

A Viable Model for a Clinically Relevant GPCR

To generate a model that is suitable for virtual screening, the target protein must have at least one template that is of reasonable (>30%) sequence identity.

Before the first experimental structures of GPCRs were determined, models aided in the selection and infroduction of GPCR ligands to the clinic. Thus, it is of interest to be able to select a GPCR target for modeling that has no known crystal structure and plays a role in some human disease(Engel, Skoumbourdis et al. 2008). A case could be made for many GPCRs as targets for therapeutic development; such examples include the Histamine Receptor (inflammation), the Orexin Receptor (sleep regulation), and the Calcitonin Receptor (blood calcium level).

Studies have shown that the Adenosine A1 receptor (A1R) is implicated in the etiology of several diseases, including Creutzfeldt–Jakob disease (CJD) and Alzheimers disease (AD). As a result of this, the Adenosine A1 Receptor was chosen as a candidate for structure modeling and computational docking of ligands with the intent of discovering novel small molecules.

The Adenosine A1 Receptor

The Adenosine A1 receptor is a Rhodopsin-like GPCR that is important in a number of human cellular processes (Townsend-Nicholson, Baker et al. 1995). Adenosine, the endogenous ligand of the A1 receptor, is involved in regulating multiple metabolic processes. The Adenosine receptors mediate adenosine function, and induce the inhibition of adenylyl cyclase activity, which is the enzyme that synthesizes cAMP from ATP.

Stimulation of the A1 receptor activates phospholipase C and D, and several potassium and calcium channels. The A1 receptors are found in several tissues and are found in highest density in the central nervous system. The proteins are expressed in cerebral cortex, hippocampus, cerebellum, thalamus, and brainstem. In the brain, adenosine modulates neuronal activity by decreasing presynaptic release of various neurotransmitters.

The adenosine A1 receptor plays an inhibitory role in the glutamenergic system, and glutamate is a potential mediator of degeneration of prion diseases. Because of the A1 role in the glutamenergic system, it is a potential target for therapeutic development for prion diseases, and thereby was chosen as the GPCR to model (Rodriguez, Martin et al. 2006).

Creutzfeldt–Jakob Disease

Creutzfeldt-Jakob Disease (CJD) is a prion disease, which, like all prion diseases, is characterized by neuronal loss, spongiform change, and accumulation of prion protein (Prusiner 1994) .

CJD is a form of neurodegenerative encephalopathy. It is a transmissible human prion disease, affecting approximately 1 in 1,000,000 individuals. CJD is believed to occur by the conformational change of normal human prions (prion protein, or PRP). The normal prions are produced in the rough endoplasmic reticulum, where they then travel to the cellular membrane. There, they encounter rogue prions, which have already changed conformation, from a mostly helical structure to nearly 50% beta sheet. The rogue prions then form fibrils, and these fibrils accumulate in the nervous system (ninds.nih.gov/disorders/cjd/)..

CJD is always fatal, and its symptoms include dementia, memory loss, and myoclonus. The death of brain cells is caused by the buildup of protein aggregates. The aggregates are made up of PrP, and they cause round or oval vacuoles between 1 and 50 microns in diameter in brain tissue. CJD is transmissible, and the incubation time is unknown. It should be noted that some human prion diseases, such as Kuru, have a mean incubation time of 14 years but in some cases can take up to 40 years. In the late 1990's, several unrelated residents of the United States state of Kentucky were diagnosed with CJD, and it was later revealed that these individuals had all regularly consumed squirrel brains. Thus, PrP in the squirrel brains may have been transmitted to humans through consumption (Will, Ironside et al. 1996).

Due to the association between Adenosine A1 receptor, the glutamenergic system and prion diseases(Rodriguez, Martin et al. 2006), research was conducted into the levels of A1 receptor in PrP related diseases. These studies looked at CJD and AD in humans, and Bovine Spongiform Encephalopathy (Mad Cow Disease, or BSE) in mice .

The levels of the Adenosine A1 receptor in the frontal cortex of 12 patients with CJD and 6 age-matched controls were measured. In addition, levels of A1 in BSE-infected mice were studied at different post incubation times to monitor changes in A1R levels with disease progression. An increase of A1 levels of 190% was found in cerebral cortex in CJD and in the mouse BSE model at advanced stages of the disease [Fig.1]. Increased activity of the receptor was also observed when compared to the controls. There was *no change* in Adenosine A2 Receptor levels in CJD patients [Fig. 2]. Therefore, the A1R, as opposed to the closely related A2, may play a role in CJD progression. Furthermore, increased A1 levels were observed in patients with Alzheimer's disease. Given this information, the A1 receptor presents itself as a therapeutic target for prion diseases.

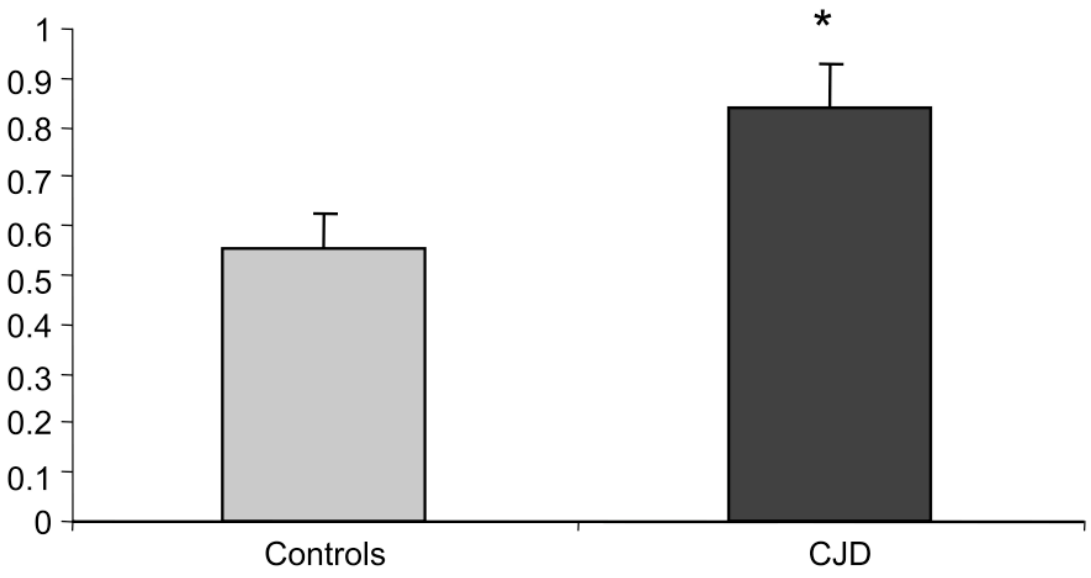

Figure 1. Adenosine A1 levels in human frontal cortex. Error bars are +/- standard deviation. Student t-test $p < 0.05$. (Rodriguez, Martin et al. 2006)

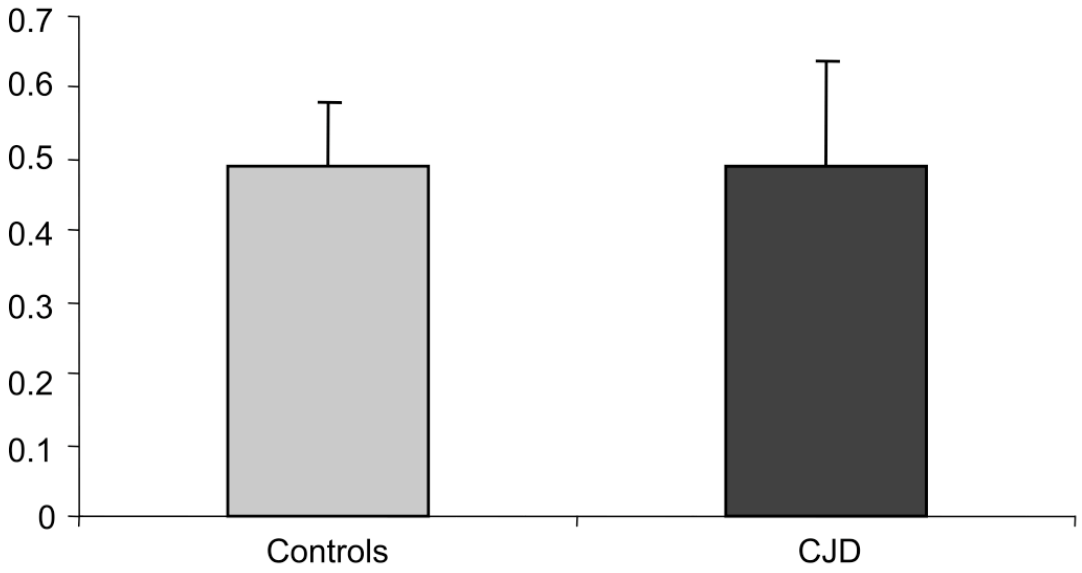

Figure 2. Adenosine A2 levels in human frontal cortex. No change is observed. (Rodriguez, Martin et al. 2006)

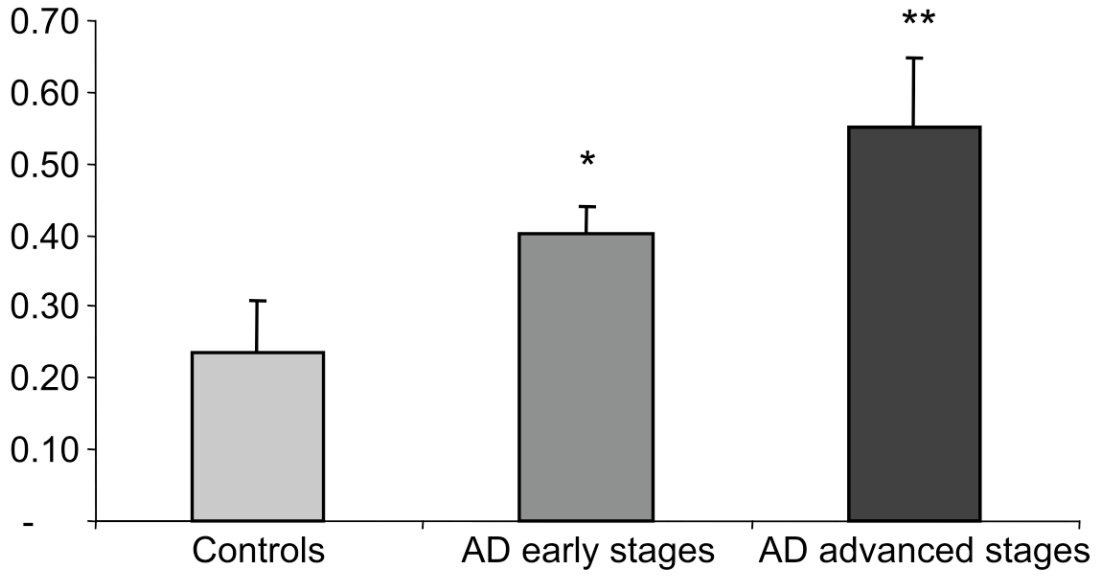

Figure 3. Adenosine A1 levels in human frontal cortex in Alzheimer's Disease patients. Error bars are +/- standard deviation,
Student t-test p<0.05 (*) p<0.01 (**)(Rodriguez, Martin et al. 2006)

A radioligand binding experimental assay is available for the Adenosine A1 Receptor, and it is also accessible for real time imaging through the use of a Positron Emission Tomography (PET) scan using selective radioactive A1R ligands (Rodriguez, Martin et al. 2006). The PET scan is a technique that produces a three-dimensional image of processes in the body. The system detects pairs of gamma rays emitted indirectly by a positron-emitting tracer, in this case an A1R ligand.

PET scanning could allow superior diagnoses of prion diseases in human patients, since routine laboratory findings are often not helpful in diagnosing CJD. There is no dysfunction of major organ systems besides the central nervous system, and cerebrospinal fluid (CSF) will not show an increase in immunoglobulins.

The Adenosine A1 Receptor as a case study

Modeling Criteria

The Adenosine A1 Receptor is an ideal target for comparative modeling and virtual screening. It meets the following criteria:

(1) Appropriate templates exist: Solved X-ray structures of two human and one non-human are available, including the human adenosine A2 receptor (3EML), the human B2 Adrenergic Receptor (2R4R), and the turkey B1 Adrenergic Receptor (2VT4). The sequence identity between the Human A1 and A2 is the highest of the solved GPCR structures, at approximately 50%. This is high enough sequence identity to achieve an accurate model suitable for virtual screening.

(2) Clinical Relevance: The A1 receptor is associated with the human diseases of Crutzfeldt Jakob Disease as well as Alzheimers disease. In addition, it is involved in a number of important physiological processes, such as sleep regulation, therefore it may be the target of additional drugs.

(3) Experimental assays: Radioligand binding can be used as an assay to test novel molecules. In addition, the A1 Receptor is accessible for real time studies through PET imaging.

The Adenosine A1 Receptor meets three important criteria for a meaningful comparative modeling and virtual screening study.

Structure Prediction of the Adenosine A1 Receptor

There are several different, though closely related, initial approaches to generating a comparative model of the A1R to be used for docking. All steps, however, require that at least one suitable template be found. Several potential templates exist [Table 2]. The 3 initial approaches are as follows:

(1) ModPipe, an automated protein structure prediction protocol

(2) Structure prediction using a multiple sequence alignment (multiple sequence alignment) of multiple templates as input to the program Modeller(Sali and Blundell 1993)

(3) Structure prediction using a pairwise sequence alignment as input to the program Modeller

Table II. Sequence Identities of available templates of Adenosine A1 Receptor target. Sequence identities recovered from NCBI BLAST.

GPCR/PDB code	**Adenosine A2 Receptor / 3EML**	**β1-Adrenergic Receptor / 2VT4**	**β2-Adrenergic Receptor / 2RH1**	**Squid Rhodopsin / 2ZIY**	**Bovine Rhodopsin / 1GZM**
A1R Sequence Identity	52%	34%	33%	22%	18%

The first method of modeling, ModPipe, uses an automated protocol to search a database of PDB sequences for any number of closely related templates. These templates are then aligned to the target sequence using the Salign module (Madhusudhan, Marti-Renom et al. 2006; Madhusudhan, Webb et al. 2009) of MODELLER. While ModPipe

can find multiple templates for the target sequence, it does not create multiple sequence alignments. That is to say, a different model is created for each target-template alignment. Therefore, the potential benefits of multiple templates are not realized (Rychlewski and Fischer 2005).

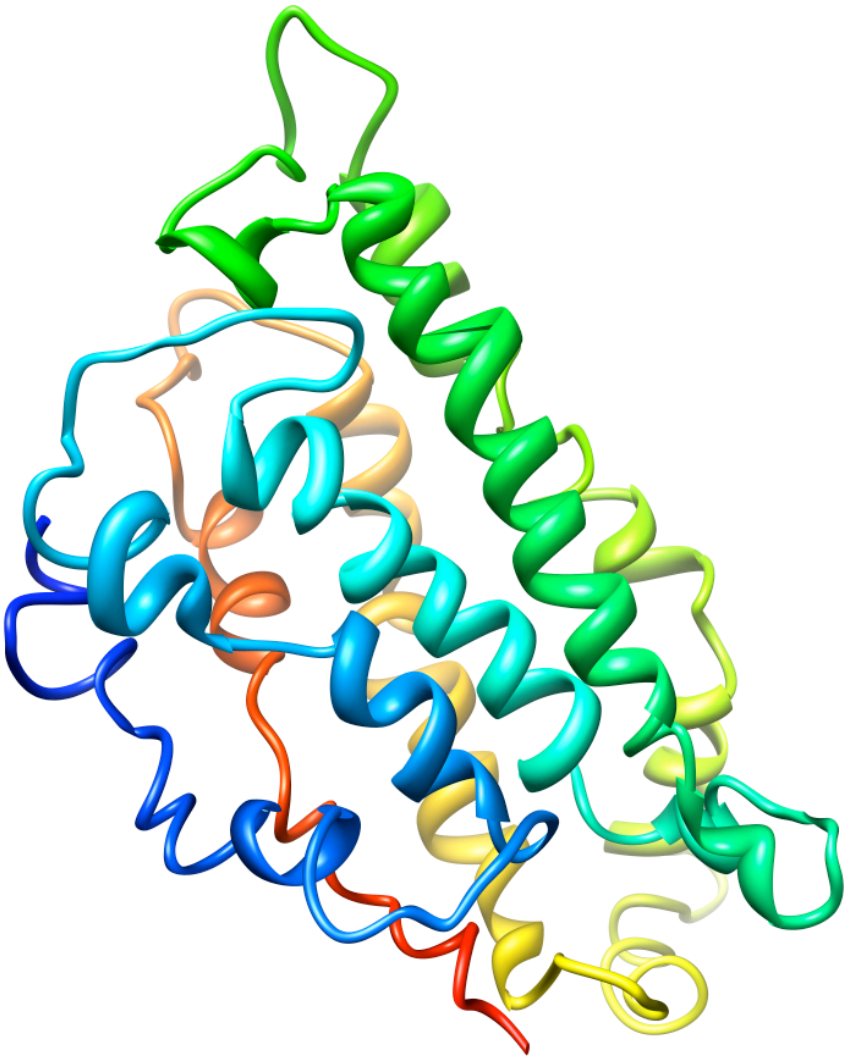

Figure 4. Adenosine A1 Receptor model resulting from ModPipe run. Note the misaligned helices. This model would not be suitable for docking. See ModPipe jobfile in Appendix

From the ModPipe results, the best scoring model, and also the model with the most alignment coverage, was the alignment between Adenosine A1 and the Human B2aR

structure (2RH1). The next most favorable model was constructed with bovine rhodopsin (1GZM) as the template. This is unsurprising, considering the small number of GPCR structures available in the PDB.

The resultant model clearly had at least one misaligned helix (Fig 4, See Appendix for alignment). This could be due to the fact that the template (2RH1) is too divergent from Adenosine A1. As a result of these misaligned helices, which affected the active site, this model was not considered to be appropriate for docking calculations.

Shortly after applying ModPipe, the Human Adenosine A2 (3EML) structure was released. Since this structure was of higher sequence identity than the B2aR template (52% vs 33%), it was reasonable to create a new alignment including the solved Adenosine A2 structure. A multiple sequence alignment (MULTIPLE SEQUENCE ALIGNMENT), containing the Adenosine A1 target sequence, the Adenosine A2 sequence (template), and the B2aR sequence (template), was created using MUSCLE (Edgar 2004). This alignment was then used as input into the program Modeller.

The resultant structure demonstrated improved helix packing and therefore the alignment was superior to the original, single template ModPipe Salign alignment [Fig. 5]. This alignment was selected to build the first model for the computational docking run. However, the initial model was not constructed with any of the ligands that were in the crystal structure of A2. Since the Adenosine A2 structure was solved with the antagonist: *4-{2-[(7-amino-2-furan-2-yl[1,2,4]triazo lo[1,5- a][1,3,5]triazin-5-yl)amino]ethyl}phenol* or ZMA, water molecules, and steric acid, these heteroatoms

were included in the updated model. This ideally would create a more accurate model, since these atoms will be taken into account when the prediction is generated.

Figure 5. Adenosine A1 Receptor model resulting from multiple sequence alignment. Note the marked improvement in helix alignment and packing. This model was used for the 1st set of docking calculations. See appendix for alignment

Due to inherent variability in the modeling routine, and ensemble of 100 models was constructed using the same alignment. Since more than one structure was generated, a method must be used to select the most native structure from the ensemble of models. While multiple methods for assessing protein structure exist, very few are effective at assessing membrane proteins.

One of the fastest and most accurate potentials for assessing protein structure is DOPE (Discrete Optimized Protein Energy), which is included as an assessment method as part of Modeller (Shen and Sali 2006). However, DOPE is trained to work on globular proteins, therefore its usefulness for assessing GPCRs, which are membrane proteins, is suspect. In light of this fact, an attempt was made to extract globular regions from the modeled GPCR structures that could then be scored with DOPE.

To obtain these globular regions, a 6 Å sphere of residues was selected around the C7, C11, and C12 atoms of the ZMA ligand. This selection was then written out into a PDB file. This resulted in 100 small PDB formatted files that approximated globular proteins. These approximations were scored with DOPE and DOPE_HR (DOPE high resolution) and the top 5 scoring models were visually inspected. The model that was chosen as the best among the top 5 also had the extracellular disulfides in a favorable conformation (Hanson and Stevens – complete the ref).

In addition to constructing models from the multiple sequence alignment, a pairwise alignment was created using only the target Adenosine A1 sequence with the Adenosine A2 structure as the template. The same protocol was used as for the multiple sequence alignment, where the same atoms around the ZMA ligand were selected, written out as PDB files, and then scored with DOPE.

When the DOPE and DOPE_HR scores of both the multiple sequence alignment and pairwise selections were compared, the scores were better in the models built with the multiple sequence alignment. As a result, the model from the multiple sequence alignment was chosen as the initial model for virtual screening.

When the models were built from the multiple sequence alignment, side chain conformational space was sampled through the generation of 100 models. Upon close visual inspection of the active site of the multiple sequence alignment model, residue ASN 254 appeared in a sterically unfavorable conformation. The amide group was bent back towards the main chain. This prompted additional optimization. ASN 254 was selected and optimized via a selection and residue modeling and optimization procedure that is part of Modeller. [See Modeller optimization input file in Appendix.]

This modified loop optimization resulted in a more favorable conformation, resulting in a model that was chosen for docking. Unfortunately, this model did not result in favorable docking hits. Known ligands did not bind to the model, which indicated that the active site was likely incorrect.

Since the multiple sequence alignment model was not successful for docking, the pairwise alignment model was selected as the model to base the virtual screening upon.

The initial pairwise model was generated including the heteroatoms of ZMA and stearic acid, both of which are visible in the 3EML PDB structure. Following the same procedure as the multiple sequence alignment model, 100 models were generated with Modeller and the globular selections were scored with DOPE. The top 5 scoring models were visually inspected and the one that appeared the most conformationally plausible was selected to perform the initial docking run. However, this model did not result in known ligands binding.

Given the poor results of docking, the model was further inspected. It was determined that several side-chain conformations were unfavorable with respect to the ZMA ligand.

As per the protocol with the multiple sequence alignment model, side-chains in the active site were selected for optimization. The side-chains were selected by taking a 6 A sphere around the C7, C11, and C12 of the ZMA molecule. This subset of residues were then optimized in Modeller. [See Modeller jobfile, Appendix, for specific parameters.]

After extensive optimization, the resultant model was visually inspected, and residues Leu250 and Asn254 did not appear to be in a normal conformation. Asn254 exhibited bending of the amide group towards the main chain. Leu250 also demonstrated a potentially unlikely conformation, though it was much less pronounced than that of Asn254. To rectify this, these residues were selected while all other residues were held fixed, in addition to the ZMA ligand being included and fixed. This selection was then further optimized. Afterwards, this model was used for the virtual screening run. Unfortunately, this model did not result in the binding of known ligands.

Virtual Screening of the Adenosine A1 Receptor

In this study, the docking calculations for each protein model (see Structure Prediction below) were done with DOCK 3.5.54 (Kuntz, Blaney et al. 1982). This program uses spheres to guide the placement of the ligand atoms in the binding site. During a docking calculation, the heavy atoms of the molecule being docked are matched with the spheres in the binding site. The binding affinity is then estimated by summing the electrostatic and van der Waals interaction energies--correcting for the desolvation penalty, which arises from the transfer of a ligand from water into the low-dielectric environment of the protein. For efficiency reasons, these energy terms are precalculated and stored on grids (Lorber and Shoichet 2005).

To evaluate the setup of the docking calculation, the 534 known adenosine A1 receptor ligands from the WOMBAT (Olah et al. 2005) database were docked and visually inspected for the correctness of their binding poses. In the quest for novel ligands of the adenosine A1 receptor, we docked the 2.7 M compounds of the lead-like subset of the ZINC 8 database (Irwin and Shoichet 2005). These molecules had been chosen to fulfill the following criteria: xLogP < 3.5, molecular weight < 350 g/mol and number of rotatable bonds ≤ 7. For every ligand, up to 1000 conformations had been precalculated with the program OMEGA(Open Eye Scientific Software 2008) and the partial charges for its atoms had been assigned with the program AMSOL(Hawkins et al. 2003).

To analyze and rationalize the binding modes of the known ligands, small fragments were docked with the program SEED (Majeux, Scarsi et al. 1999; Majeux, Scarsi et al. 2001). The fragments were chosen such that they would represent the smallest entities that would likely bind to the receptor. In the present study, we used benzene, adenine, [1,2,4]triazolo[1,5-a][1,3,5]triazin-7-amine and furane. The docking approach implemented in the program SEED determines optimal positions and orientations of small to medium-size molecular fragments in the binding site of a protein. Apolar fragments are docked into hydrophobic regions of the receptor, while polar fragments are positioned such that at least one intermolecular hydrogen bond is formed. Each fragment is placed at several thousand different positions with multiple orientations (for a total of in the order of 10^6 conformations), and the binding energy is estimated whenever severe clashes are not present (usually about 10^5 conformations). The binding energy is the sum of the van der Waals interaction and the electrostatic energy. The

latter consists of screened receptor-fragment interaction, as well as values of receptor and fragment desolvation (Scarsi,1997).

An Iterative Approach to Structure Modeling and Docking

Since all three previous methods to construct a viable model for docking failed, a new procedure was warranted. A discussion about active site modeling options yielded a "semi-manual" method for modeling. Known antagonists of the Adenosine A1 receptor would be manually placed into the active site of the model resulting from the pairwise alignment. Next, a CHARMM minimization would be performed on the ligand only, while the residues were held fixed. After the minimization, the residues of the active site would be optimized. This procedure would be repeated iteratively until there was little or no discernable change in the active site residue conformation and the position of the ligand.

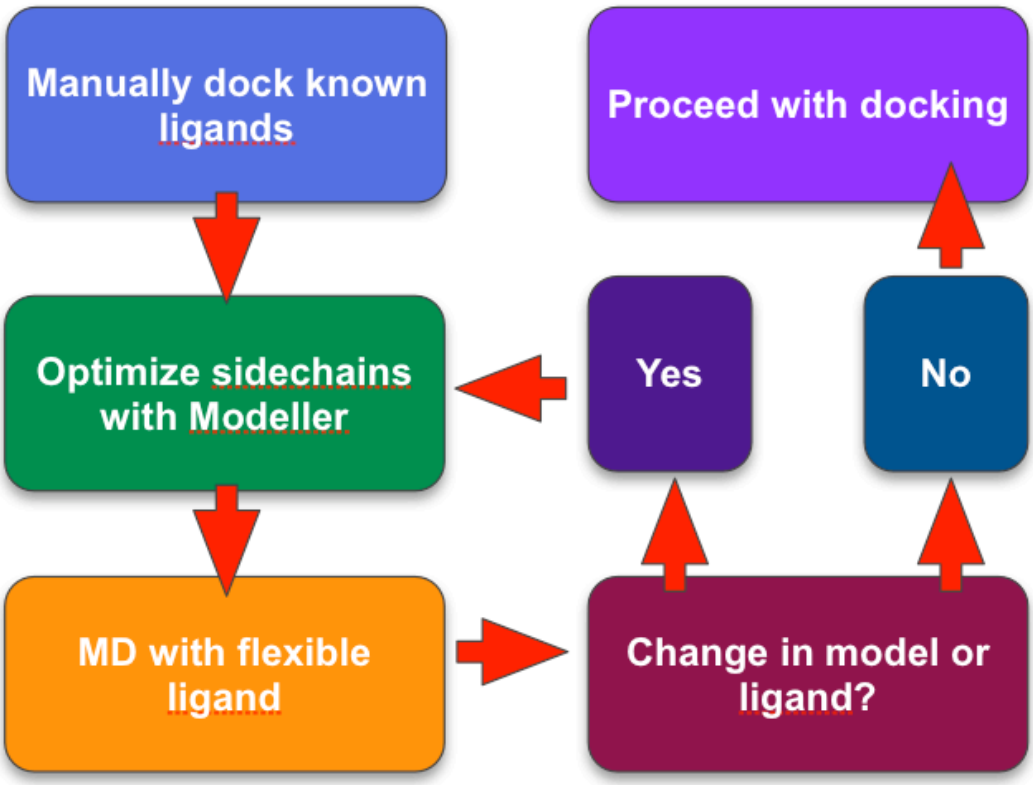

Figure 6. Flow chart describing iterative optimization/docking procedure.

Active Site Residues and Mutagenesis Study

The residues of the active site to be modeled with the iterative procedure were selected if they were within 6 A of the ligand heavy atoms. In addition, the selected residues were compared to a mutagenesis study of Adenosine A1 agonists and antagonists (Tables III and IV). Residues that caused major changes in binding affinity (up to 100 fold decrease) through alanine substitution were checked against the selection of residues within 6 A of the ligand. In all cases, the residues that caused the change in binding affinity (after alanine substitution) were included in the selection. Through comparison to this study, a greater degree of confidence was obtained in choosing the active site residues most likely to be important in binding.

Table III. A1R Agonist mutational analysis change in binding affinity. These residues were included in the selection of optimized residues for the docking model.

Residue/Mutation	Ligand	Effects
T/A 91	NECA, CADO, (R)-PIA,CPA	Decrease >100 fold
Q/A 92	NECA, CADO, (R)-PIA,CPA	Decrease >23 fold

Table IV. A1R Antagonist mutational analysis change in binding affinity.

Residue/Mutation	Ligand	Effects
V/A 87	N0840	Decrease 3 fold
L/A 88	N0840	Decrease >100 fold
T/A 91	N0840	Decrease >100 fold
Q/A 92	N0840	Decrease >100 fold

Once these residues were selected, the iterative process began. After the first iteration, it was noted that some of the helices surrounding the active site were "kinked" or broken. On closer inspection, it was revealed that this bending was due to the optimization of the prolines in the helices. When the proline residues were excluded from the optimization, the helices maintained their integrity.

After the broken helix problem was remedied, an additional problem became apparent. In some cases, the side-chains were causing a steric clash with the ligand. To quantify this, in the first iteration, there were 17 atom-atom contacts within 3 A distance between the ligand atoms and the protein side-chains. The number of clashes could potentially mean that the restraint term within Modeller that determines appropriate side-chain-ligand distances was too strict for our particular docking case. To test this, the soft sphere restraint within the jobfile, was changed from 1.0 (arbitrary units), to 5.0 units, in 1.0 unit increments. At each increment, the model was visually inspected and the

interface distances were calculated. The ideal soft sphere restraint unit was found to be 3.0. The default value of 1.0 was not strict enough, while the value of 5.0 caused unnatural side-chain conformations (side-chains faced into the backbone, for example). The optimizations, therefore, were run using the soft sphere restraint of 3.0, and resulted in only 2 atom-atom contacts within 3 angstrom distance of the molecule and ligand, versus 17 atom-atom contacts when using the default soft sphere restraint of 1.0.

Now that an appropriate protocol was established, the iterative procedure described above began on a known antagonists of the Adenosine A1 receptor: Zinc identifier: 13589664. This ligand is chemically dissimilar to ZMA, which is the ligand that is crystallized with the Adenosine A2 structure. A ligand that was different than ZMA was chosen given the repeated unsuccessful docking attempts using a model that was built with the ZMA ligand.

This model demonstrated continuous improvements in each step of the iteration. That is to say, the ligand did not move out of the active site during minimization and no side-chains adopted abnormal conformations. Upon the 3rd iteration of minimization-side chain optimization, the side-chains failed to change conformation appreciably, as was the case with the pose of the ligand. At this point, the modeling protocol was terminated and the model was deemed appropriate for a virtual screening run. Early docking results are promising for this model, since more of known ligands have bound favorably as compared to the original multiple sequence alignment and pairwise models. The computational docking runs are ongoing as of September 03 2009.

Discussion

Despite advances in globular protein structure prediction, the accurate prediction of GPCR structure and ligand interactions remains a challenge. The iterative process described here holds potential as a scaffold by which to base a standardized protocol on. This particular study required a considerable amount of human intervention and subjectivity, but this is not uncommon given the current state of the art. Ideally, the entire process would be automated, and not require a manual placement of known binding partners. However, modeling and virtual screening of GPCRs is likely to improve as more structures are solved.

The Adenosine A1 Receptor is important in a large number of human cellular processes, and is likely involved in Creutzfelt-Jakob Disease and Alzheimer's Disease. It is not unreasonable to suggest that the A1 Receptor could play a role in additional human diseases and disorders, so it is clearly an important target for therapeutic development. Since the computational docking part of this study is ongoing, the actual test of the effectiveness of this method will only come after the docking calculations are complete and the top scoring ligands are tested with an experimental assay.

Future Directions

Membrane protein structure prediction remains challenging, even in the case of homology modeling, since so few templates exist as compared to that of globular proteins. Minor successes have been demonstrated in the field of *ab initio* membrane protein structure prediction, but the field has not had any immediate breakthroughs. Modeling of membrane proteins and GPCRs will improve as structures are solved, as well as with incorporation of additional information, such as more detailed lipid modeling.

A note on alternative software

SCWRL is one of the most popular side-chain modeling software programs available. The most recent version of SCWRL (4.0) has a new rotamer library, and models side-chains quickly (Wang, Canutescu et al. 2008). However, during this study, several comparisons were performed and SCWRL appeared to model side-chains in a manner that resulted in similar conformations to those generated by Modeller, albeit faster. Modeller was chosen over SCWRL because it allowed finer control of individual optimization features, such as the soft sphere restraint. In addition, all of the homology modeling was done using Modeller, and the software is developed in our lab.

Bibliography

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Baker D., Sali A. (2001) "Protein structure prediction and structural genomics." Science.Oct 5;294(5540):93-6.

Barth, P., B. Wallner, et al. (2009). "Prediction of membrane protein structures with complex topologies using limited constraints." Proc Natl Acad Sci U S A **106**(5): 1409-14.

Caffrey, M. (2003). "Membrane protein crystallization." J Struct Biol **142**(1): 108-32.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-7.

Elofsson, A. and G. von Heijne (2007). "Membrane protein structure: prediction versus reality." Annu Rev Biochem **76**: 125-40.

Engel, S., A. P. Skoumbourdis, et al. (2008). "A virtual screen for diverse ligands: discovery of selective G protein-coupled receptor antagonists." J Am Chem Soc **130**(15): 5115-23.

Eramian, D., N. Eswar, et al. (2008). "How well can the accuracy of comparative protein structure models be predicted?" Protein Sci **17**(11): 1881-93.

Eswar, N., B. Webb, et al. (2007). "Comparative protein structure modeling using MODELLER." Curr Protoc Protein Sci **Chapter 2**: Unit 2 9.

Forrest, L. R., C. L. Tang, et al. (2006). "On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins." Biophys J **91**(2): 508-17.

Fredriksson, R., M. C. Lagerstrom, et al. (2003). "The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints." Mol Pharmacol **63**(6): 1256-72.

Irwin, J. J. and B. K. Shoichet (2005). "ZINC--a free database of commercially available compounds for virtual screening." J Chem Inf Model **45**(1): 177-82.

Kuntz, I. D., J. M. Blaney, et al. (1982). "A geometric approach to macromolecule-ligand interactions." J Mol Biol **161**(2): 269-88.

Lorber, D. M. and B. K. Shoichet (2005). "Hierarchical docking of databases of multiple ligand conformations." Curr Top Med Chem **5**(8): 739-49.

Madhusudhan, M. S., M. A. Marti-Renom, et al. (2006). "Variable gap penalty for protein sequence-structure alignment." Protein Eng Des Sel **19**(3): 129-33.

Madhusudhan, M. S., B. M. Webb, et al. (2009). "Alignment of multiple protein structures based on sequence and structure features." Protein Eng Des Sel **22**(9): 569-74.

Majeux, N., M. Scarsi, et al. (1999). "Exhaustive docking of molecular fragments with electrostatic solvation." Proteins **37**(1): 88-105.

Majeux, N., M. Scarsi, et al. (2001). "Efficient electrostatic solvation model for protein-fragment docking." Proteins **42**(2): 256-68.

Marko, A. C., K. Stafford, et al. (2007). "Stochastic pairwise alignments and scoring methods for comparative protein structure modeling." J Chem Inf Model **47**(3): 1263-70.

Michino, M., E. Abola, et al. (2009). "Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008." Nat Rev Drug Discov **8**(6): 455-63.

Prusiner, S. B. (1994). "Biology and genetics of prion diseases." Annu Rev Microbiol **48**: 655-86.

Rodriguez, A., M. Martin, et al. (2006). "Adenosine A1 receptor protein levels and activity is increased in the cerebral cortex in Creutzfeldt-Jakob disease and in bovine spongiform encephalopathy-infected bovine-PrP mice." J Neuropathol Exp Neurol **65**(10): 964-75.

Rychlewski, L. and D. Fischer (2005). "LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction." Protein Sci **14**(1): 240-5.

Sali, A. and T. L. Blundell (1993). "Comparative protein modelling by satisfaction of spatial restraints." J Mol Biol **234**(3): 779-815.

Sanchez, R., U. Pieper, et al. (2000). "Protein structure modeling for structural genomics." Nat Struct Biol **7 Suppl**: 986-90.

Schwede, T., A. Sali, et al. (2009). "Outcome of a workshop on applications of protein models in biomedical research." Structure **17**(2): 151-9.

Shen, M. Y. and A. Sali (2006). "Statistical potential for assessment and prediction of protein structures." Protein Sci **15**(11): 2507-24.

Tramontano, A., Leplae, R., Morea V. (2001) "Analysis and assessment of comparative modeling predictions in CASP4." Proteins: Suppl 5:22-38.

Townsend-Nicholson, A., E. Baker, et al. (1995). "Localization of the adenosine A1 receptor subtype gene (ADORA1) to chromosome 1q32.1." Genomics **26**(2): 423-5.

Wang, Q., A. A. Canutescu, et al. (2008). "SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling." Nat Protoc **3**(12): 1832-47.

Will, R. G., J. W. Ironside, et al. (1996). "A new variant of Creutzfeldt-Jakob disease in the UK." Lancet **347**(9006): 921-5.

Appendix

Modeller optimization jobfile. Note the soft sphere restraint set to 3.0, as opposed to default 1.0. In addition, the residues are individually selected and added to a larger selection, this is necessary since they are discontinuous.

```
from modeller import *
from modeller.automodel import *
env = environ()
env.io.hetatm=True
env.io.atom_files_directory = ['.','../atom_files']
env.libs.topology.read(file='$(LIB)/top_heav.lib')
env.libs.parameters.read(file='$(LIB)/par.lib')
# give more weight to soft sphere restraints
env.schedule_scale = physical.values(default=1.0,soft_sphere=3.0)
log.minimal()
class MyLoop(loopmodel):
        def select_loop_atoms(self):
                s=selection()
                s.add(self.residue_range('16','16'))
                s.add(self.residue_range('58','58'))
                s.add(self.residue_range('63','63'))
                s.add(self.residue_range('65','66'))
                s.add(self.residue_range('69','69'))
                s.add(self.residue_range('70','70'))
                s.add(self.residue_range('83','85'))
                s.add(self.residue_range('87','88'))
                s.add(self.residue_range('91','92'))
                s.add(self.residue_range('170','173'))
                s.add(self.residue_range('175','177'))
                s.add(self.residue_range('180','181'))
                s.add(self.residue_range('184','184'))
                s.add(self.residue_range('247','247'))
                s.add(self.residue_range('250','251'))
                s.add(self.residue_range('253','254'))
                s.add(self.residue_range('257','258'))
                s.add(self.residue_range('264','265'))
                s.add(self.residue_range('270','271'))
                s.add(self.residue_range('273','274'))
                s.add(self.residue_range('277','278'))
                s=s.by_residue()
                return s
m = MyLoop(env,
           inimodel='A1-13672416.pdb',
```

```
          sequence='A1-13672416-FULL-R1.pdb')
m.loop.starting_model = m.loop.ending_model = 1
# variable target function method with conjugate gradients
m.library_schedule = autosched.slow
m.max_var_iterations = 500
m.loop.md_level = refine.very_slow  # loop refinement method
m.repeat_optimization = 20
m.max_molpdf = 6000
m.make()
```

Modpipe input file, used to construct the 1st model

Multiple sequence alignment (pir format) used to construct 2nd docking model. Contains Adenosine A1 receptor sequence, and Human Adenosine A2 and B2 Adenosine Receptor as templates.

```
>P1;adenosineA1-ZMA
sequence:adenosineA1-ZMA:1:A:350:A::::
--MPPSISAFQAAYIGIEVLIALVSVPGNVLVIWAVKVNQALRDATFCFIVSLAVADVAVGALVIPLA---I
LINIGPQTYFHTCLMVACPVLILTQSSILALLAIAVDRYLRVKIPLRYKMVVTPRRAAVAIAGCWILSFVVG
LTPM-FGWNNLSAVERAWAANGSMGEPVIKCEFEKVISMEYMVYFNFFVWVLPPLLLMVLIYLEVFYLIRKQ
LNKKVSASSGDPQKYYGKELKIAKSLALILFLFALSWLPLHILNCITLFCPSC-HKPSILTYIAIFLTHGNS
AMNPIVYAFRIQKFRVTFLKIWNDHFRCQPAPPIDEDLPEERPDD......wwwwwwwwwwwwwwwwwwwwww
wwwwwwwwwwww*
>P1;3EML-ligands.pdb
structure:3EML-ligands.pdb:3:A:576:A::::
-IMGSSV------YITVELAIAVLAILGNVLVCWAVWLNSNLQNVTNYFVVSLAAADIAVGVLAIPFA---I
TISTGFCAACHGCLFIACFVLVLTQSSIFSLLAIAIDRYIAIRIPLRYNGLVTGTRAKGIIAICWVLSFAIG
LTPM-LGWNNCGQSQ-------GCGEGQVACLFEDVVPMNYMVYFNFFACVLVPLLLMLGVYLRIFLAARRQ
LRSTLQ-----------KEVHAAKSLAIIVGLFALCWLPLHIINCFTFFCPDCSHAPLWLMYLAIVLSHTNS
VVNPFIYAYRIREFRQTFRKIIRSHVLRQ----------------.....wwwwwwwwwwwwwwwwwwwwww
wwwwwwwwwwww*
>P1;2rh1A
structure:2rh1A:32:A:+288:A::::
WVVGMGI---------VMSLIVLAIVFGNVLVITAIAKFERLQTVTNYFITSLACADLVMGLAVVPFGAAHI
LMKMWTFGNFWCEFWTSIDVLCVT-ASIETLCVIAVDRYFAITSPFKYQSLLTKNKARVIILMVWIVSGLTS
FLPIQMHWYRATHQEAI-----NCYAEETCCDF---FTNQAYAIASSIVSFYVPLVIMVFVYSRVFQEAKRQ
LKFCL------------KEHKALKTLGIIMGTFTLCWLPFFIVNIVHVIQDNLIRKEVYI--LLNWIGYVNS
```

```
GFNPLIYC-RSPDFRIAFQELLC----------------------*
```

Pairwise sequence alignment (pir format) used to construct 3rd docking model. Contains Adenosine A1 receptor sequence, and Human Adenosine A2 as a template.

```
>P1;adenosineA1-ZMA-pw

sequence:adenosineA1-ZMA-pw:1:A:350:A::::

-MPPSISAFQAAYIGIEVLIALVSVPGNVLVIWAVKVNQALRDATFCFIVSLAVADVAVGALVIPLAILINI

GPQTYFHTCLMVACPVLILTQSSILALLAIAVDRYLRVKIPLRYKMVVTPRRAAVAIAGCWILSFVVGLTPM

FGWNNLSAVERAWAANGSMGEPVIKCEFEKVISMEYMVYFNFFVWVLPPLLLMVLIYLEVFYLIRKQLNKKV

SASSGDPQKYYGKELKIAKSLALILFLFALSWLPLHILNCITLFCPSC-HKPSILTYIAIFLTHGNSAMNPI

VYAFRIQKFRVTFLKIWNDHFRCQ......wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww*

>P1;3EML-ligands.pdb

structure:3EML-ligands.pdb:3:A:576:A::::

IMGSSV------YITVELAIAVLAILGNVLVCWAVWLNSNLQNVTNYFVVSLAAADIAVGVLAIPFAITIST

GFCAACHGCLFIACFVLVLTQSSIFSLLAIAIDRYIAIRIPLRYNGLVTGTRAKGIIAICWVLSFAIGLTPM

LGWNNCGQSQ-------GCGEGQVACLFEDVVPMNYMVYFNFFACVLVPLLLMLGVYLRIFLAARRQLRSTL

Q-----------KEVHAAKSLAIIVGLFALCWLPLHIINCFTFFCPDCSHAPLWLMYLAIVLSHTNSVVNPF

IYAYRIREFRQTFRKIIRSHVLRQ......wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww*
```

ModPipe input script (.conf file)

```
# - Base directory for creating temporary working directory

TMPDIR                          /scratch/amarko/human/a12348

# - ModPipe Repositories

DATDIR                          /netapp/home/amarko/GPCR/adenosinea1/data

# - MODELLER executable

MODELLER                     modCVS

# - Database tag; used to set names for profiles

NRDBTAG                      uniprot90

# - Non-redundant sequence database (should be in binary form)

NRSEQDB                      /netapp/database/uniprot/sequences/uniprot90.hdf5

# - Template sequence database

TEMPLATESEQDB                /netapp/home/amarko/seqDB/pdb_95.hdf5

# - Database of structure profiles

XPRF_LIST
        /netapp/home/amarko/GPCR/profiles/PSSM/pdb95_gpcrv2_prf.list

XPRF_PSSMDB
        /netapp/home/amarko/GPCR/profiles/PSSM/pdb95_gpcrv2_prf.pssm

# - PDB repository

PDB_REPOSITORY
"/netapp/home/amarko/GPCR/nochimera/NC:/netapp/home/eashwar/work/adam/pdb:/netapp/databas
e/pdb/remediated/uncompressed_files"

# - TAR (UNIX) executable (should be able to handle -z option)

TAREXE                       gtar

# - Profile update flag

PRFUPDATE                    OFF

# - Cutoff value for length of alignments

MINALNLEN                      30

# - Number of alignments per alignment

NUMMODELS                    1

# - Scheme to select the best model calculated for each alignment

SELECT_MODEL_BY                MOLPDF

RETURN_MODELS                  BEST

# -- Modes for profile calculation

PRF_BUILD_PROFILE            ON

PRF_PSI_BLAST                ON
```

```
# -- NCBI Blast database (Specify only the base filename without extension)
#    This will be used to calculate the PSI-Blast profile
NCBISEQDB                    /netapp/database/uniprot/sequences/uniprot
# -- Whether to include ligands/waters from the template in the modeling process
HETATOMS                     ON
WATERS                         OFF
# -- Parameters for clustering alignments
#    The condition is OVLP > CUT && PCOVLP > CUT && NONOVLP < CUT &&
#    PCNONOVLP < CUT && IDCOL > CUT && PCIDCOL > CUT.
#    -------++++++++++++++
#           ++++++++++++++---------
#    The '+' indicates the overlaping region & the '-' indicates
#    the non-overlapping region.
CLUSTERALI                     OFF
ALICLUST_OVLP                   0
ALICLUST_PCOVLP                60
ALICLUST_NONOVLP               20
ALICLUST_PCNONOVLP             20
ALICLUST_IDCOL                  0
ALICLUST_PCIDCOL               80
```

UCSF Library Release

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation,or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Adam C Marlow — September 11, 2009

Author Signature — Date