COMPUTATIONAL PREDICTION OF PROTEIN SMALL MOLECULE INTERFACES USING ROSETTA

By

Kristian Wallace Kaufmann

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

August, 2011

Nashville, Tennessee

Approved:

Dr. Jens Meiler

Dr. Randy D. Blakely

Dr. Michael P. Stone

Dr. Brian O. Bachmann

To my parents and siblings for having encouraged me through this process

and to Thuy for her patience

ACKNOWLEDGEMENTS

This work would not have been possible without the financial support the Vanderbilt Department of Chemistry, Molecular Biophysics Training Program at Vanderbilt university, or the National Institute of Drug Abuse Ruth L. Kirschstein Individual Prodoctoral fellowship.

I am especially grateful to my dissertation advisor Dr. Jens Meiler without whom I could not have done this work. Many of members of the Meiler lab have assisted me during my time most especially, Drs. Eric Dawson and Ralf Mueller. I am also indebted to the RosettaCommons and the many scientists involved in the development of Rosetta. Without the contributions of these scientists my work would not have been possible. I would also like to thank Dr. Randy Blakely, Dr. Julie R. Field, and Dr. L. Keith Henry for their direction help on my work modeling the serotonin transporter. Finally I would also like to thank Drs. Michael Stone and Brian Bachmann of serving on my committee and providing guidance on my research.

I would also like to thank my family for supporting me through the years. A doctorate is a difficult and taxing endeavor without the support of both family and friends this work would not have been accomplished.

STATEMENT OF DISSERTATION

The work undertaken during a doctorate is a collaborative effort as is evidenced by the many authors on the papers from which this dissertation is based. The writing and work contained in the following is my own with the following exceptions. Gordon Lemmon wrote the protein and ligand docking literature overview in Chapter 1. Kristin Glab and Ralf Mueller obtained the statistics used to develop the small molecule rotamer library in Chapter 2. Nicole Shen and Jens Meiler with my help wrote the first draft of Chapter 3. Even though I did not perform the above work I was heavily involvement in its development. The rest of the writing and work is my own.

Table of Contents

LIST OF FIGURES

Chapter Figure Page

LIST OF TABLES

Chapter Table Page

CHAPTER I

INTRODUCTION TO THE ROSETTA PROTEIN MODELING SUITE[1]

ROSETTA is a unified software package for protein structure prediction and functional design. It has been applied to predict protein structures with and without the aid of sparse experimental data, perform protein-protein and protein-small molecule docking, design novel proteins, and redesign existing proteins for altered function. ROSETTA allows for rapid tests of hypotheses in biomedical research which would be impossible or exorbitantly expensive to perform via traditional experimental methods. Thereby ROSETTA methods gain increasing importance in the interpretation of biological findings e.g. in genome projects and in the engineering of therapeutics, probe molecules, and model systems in biomedical research.

ROSETTA like all structure prediction algorithms must perform two tasks. First, ROSETTA must explore or sample the relevant conformational space and in the case of design sequence space. Second, ROSETTA must accurately rank or evaluate the energy of the resulting structural models. For this purpose, ROSETTA implements (mostly) knowledge guided Metropolis Monte Carlo sampling approaches coupled with (mostly) knowledge based energy functions. Knowledge based energy functions assume that most molecular properties can be derived from available information, in this case the Protein Data Bank (PDB) (Bernstein, Koetzle et al. ; Berman, Battistuz et al.).

[1] Published as "Practically Useful: what the ROSETTA protein modeling suite can do for you." Kaufmann KW, Lemmon GH, De Luca SL, Sheehan JH, Meiler J *Biochemistry* **2010** 49 2987-2998. Reprinted with permission of publisher

ROSETTA Conformational Sampling Strategies

The majority of conformational sampling protocols in ROSETTA use the Metropolis Carlo algorithm to guide sampling. Gradient based minimization is often employed for last step refinement of initial models. Since each ROSETTA protocol allows degrees of freedom specific for the task, ROSETTA can perform a diverse set of protein modeling tasks (Wang, Bradley et al. 2007).

Sampling backbone degrees of freedom

ROSETTA separates large backbone conformational sampling from local backbone refinement. Large backbone conformational changes are modeled by exchanging the backbone conformations of 9 or 3 amino acid peptide fragments. Peptide conformations are collected from the PDB for homologous stretches of sequence (Simons, Kooperberg et al. 1997) which capture the structural bias for the local sequence (Bystroff, Simons et al. 1996). For local refinement of protein models ROSETTA utilizes Metropolis Monte Carlo sampling of phi, psi angles calculated not to disturb the global fold of the protein. Rohl (Rohl, Strauss et al. 2004) provides a review of the fragment selection and backbone refinement algorithms in ROSETTA.

Sampling side chain degrees of freedom

Systematic sampling of sidechain degrees of freedom of even short peptides quickly becomes intractable (Levinthal 1968). ROSETTA drastically reduces the number of conformations sampled through use of discrete conformations of side chains observed in the PDB (Dunbrack and Karplus 1993; Kuhlman and Baker 2000). These ”rotamers” capture allowed combinations between side chain torsion angles as well as the backbone phi, psi angles and thereby reduce the conformational space (Dunbrack and Karplus 1993). A Metropolis Monte Carlo

simulated annealing run is used to search for the combination of rotamers occupying the global minimum in the energy function (Kuhlman and Baker 2000; Leaver-Fay, Kuhlman et al.). This general approach is extended to protein design by replacing a rotamer of amino acid A with a rotamer of amino acid B in the Monte Carlo step.

ROSETTA Energy Function

Simulations with ROSETTA can be classified based on whether amino acid side chains are represented by super atoms or centroids in the low-resolution mode or at atomic detail in the high-resolution mode. Both modes come with tailored energy functions that have been reviewed previously by Rohl (Rohl, Strauss et al. 2004).

ROSETTA knowledge based centroid energy function

The low-resolution energy function treats the side chains as centroids (Simons, Kooperberg et al. 1997; Simons, Ruczinski et al. 1999). The energy function models solvation, electrostatics, hydrogen bonding between beta strands, and steric clashes. Solvation effects are modeled as the probability of seeing a particular amino acid with a given number of alpha carbons within an amino acid dependent cutoff distance. Electrostatic interactions are modeled as the probability of observing a given distance between centroids of amino acids. Hydrogen bonding between beta strands is evaluated based on the relative geometric arrangement of strand fragments. Backbone atom and side chain centroid overlap is penalized and thus provides the repulsive component of a van der Waals force. A radius of gyration term is used to model the effect of van der Waals attraction. All probability profiles have been derived using Bayesian statistics on crystal structures from the PDB. The low resolution of this centroid-based energy function smoothes the energy landscape at the expense of its accuracy. The smoother energy landscape allows structures

which are close to the true global minima to maintain a low energy even with structural defects that a full atom energy function would stiffly penalize.

ROSETTA knowledge based all-atom energy function

The all-atom high-resolution energy function used by ROSETTA was originally developed for protein design (Kuhlman and Baker 2000; Kuhlman, Dantas et al. 2003). It combines the 6-12 Lennard Jones potential for van der Waals forces, a solvation approximation (Lazaridis and Karplus 1999), an orientation dependent hydrogen bonding potential (Kortemme, Morozov et al. 2003), a knowledge based electrostatics term, and a knowledge based conformation dependent amino acid internal free energy term (Dunbrack and Karplus 1993). An important consideration when constructing this potential was that all energy terms are pairwise decomposable. The pairwise decomposition of each of the terms limits the total number of energy contributions to ½ N (N-1) when N is the number of atoms within the system. This limitation allows pre-computation and storage of many of these energy contributions in the computer memory which is necessary for rapid execution of the Metropolis Monte Carlo sampling strategies employed by ROSETTA during protein design and atomic-detail protein structure prediction.

Protein Structure Prediction

The central tenet of structural biology is that structure determines function. Thus, the structure of a protein is critical for a full understanding of its function. Experimental structure determination techniques such as X-ray crystallography, nuclear magnetic resonance, electron paramagnetic resonance, and electron microscopy require significant investments of effort to produce structures of a molecule. Conversely, the advent of the genomic revolution created an explosion in the number of known sequences for biopolymers. For example, from October 2008 to March 2009 approximately

eight million (!) new, non-redundant sequences were added to the BLAST database. ROSETTA remedies the shortfall in structural information by predicting high probability structures for a given amino acid sequence.

De Novo folding simulation

The "protein folding problem" given an amino acid sequence predict the tertiary structure it folds into is considered the greatest challenge in computational structural biology. The ROSETTA *de novo* structure prediction algorithm has been reviewed and described in detail elsewhere (Simons, Kooperberg et al. 1997; Simons, Ruczinski et al. 1999; Rohl, Strauss et al. 2004; Bradley, Misura et al. 2005). Briefly, ROSETTA begins with an extended peptide chain. Insertion of backbone fragments rapidly "folds" the protein using the low resolution energy function and sampling approaches (Figure 1). ROSETTA attempts approximately 30,000 nine residue fragment insertions followed by a another 10,000 three residue fragment insertions to generate a protein model (Rohl, Strauss et al. 2004). Usually 20,000-50,000 models are folded for each individual protein (Bradley, Misura et al. 2005). The resulting models can either undergo atomic-detail refinement or if computational expense is an issue, clustering based on C_α-RMSD (Bonneau, Strauss et al.; Das, Qian et al. 2007) can reduce the number of models before performing refinement. The clustering parameters are chosen by the user to generate clusters that maintain the same overall fold (i.e. C_α-RMSD < 5) while maximizing coverage of the structure space sampled. The lowest energy models and the structures at the center of the clusters enter atomic-detail refinement (read below). In 2009, Das et al. implemented an addition to the existing *de novo* protein folding protocol that allowed for accurate prediction of homomeric proteins (Das, André et al. 2009). They combined elements of ROSETTA *de novo* structure prediction (Raman, Vernon et al. 2009) with protein-protein docking (Mandell and Kortemme 2009) to develop ROSETTA Fold-and-Dock. Fold-and-Dock alternates between cycles of symmetric fragment insertion as in ROSETTA *de novo* prediction, and

rigid body docking between the two partially assembled models. Following complex assembly, the

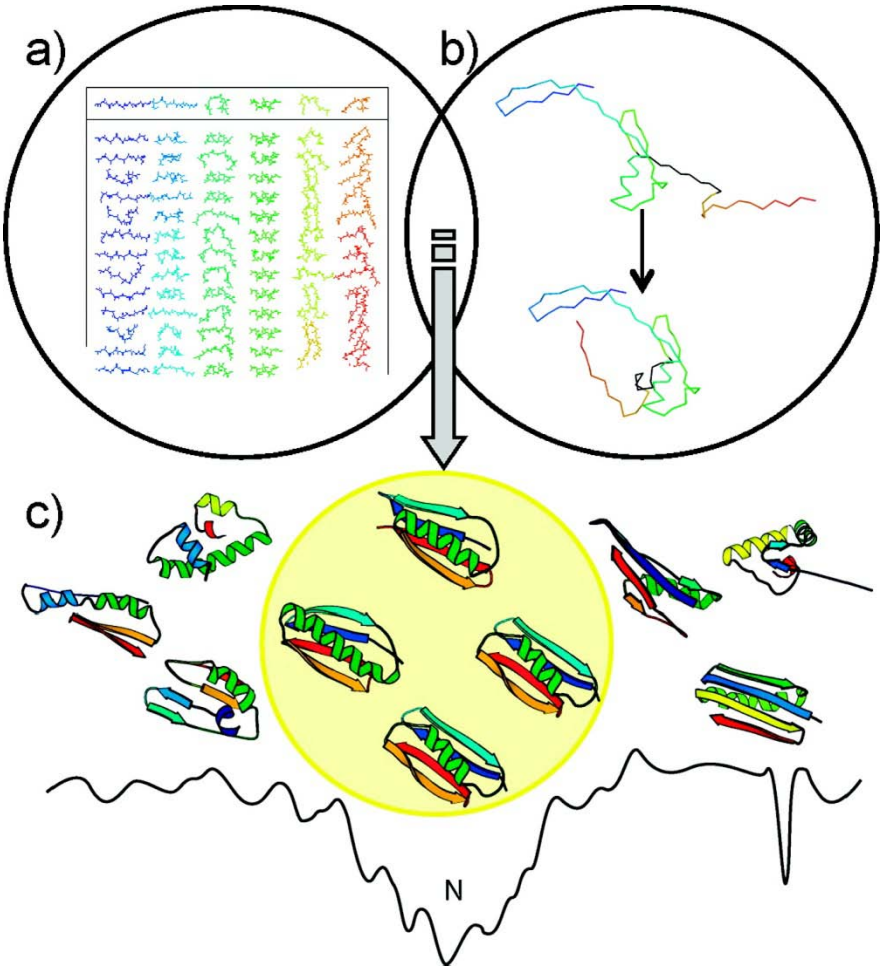


Figure 1. **De novo Folding Algorithm** Rosetta starts from *a*) fragment libraries with sequence dependent phi-psi angles which capture the local conformational space accessible to a sequence. *b*) Combining different fragments from the libraries folds the protein through optimization of non-local contacts. The low resolution energy function depicted in *c*) smoothes the rough energy surface resulting in a deep, broad minimum for the native conformation. Metropolis Monte Carlo minimization drives the structure towards the global minimum. Reprinted with permission from (Kaufmann, Lemmon et al.).

entire complex undergoes full atom refinement. Fold-and-Dock assumes that secondary structural elements of a homomer are symmetric, and inserts the same fragments into every subunit. As the interface between a homomer is highly buried, docking while folding allows this region to be protected and stabilized during the folding simulation, which greatly improved accuracy. Using this method, the structure of a K138A mutant of 1510-N membrane protease (PDB code: 3bbp) (Yokoyama, Hamamatsu et al. 2008) was predicted to within 1 RMSD of the crystal structure in a

blind prediction test. To further improve resolution, sparse NMR-derived chemical shift restraints were added yielding models with RMSD-values smaller 1. Typically, structure elucidation for homomers with NMR-derived restraints would have required extensive datasets of RDCs, NOEs, chemical shifts, and scalar couplings.

Comparative modeling

Comparative modeling in ROSETTA starts after the alignment of a target sequence to a template protein, using sequence-sequence or sequence-structure alignment tools as described by Raman et al (Raman, Vernon et al. 2009).The quality of the alignment determines the aggressiveness of the sampling in ROSETTA (Raman, Vernon et al. 2009). In case of high sequence homology (sequence identity larger than 50%), the protein backbone is only rebuilt in regions surrounding insertions and deletions in the sequence alignment (Rohl, Strauss et al. 2004; Raman, Vernon et al. 2009). Consequently, ROSETTA starts with the template structure and builds in missing loops using fragment insertion or randomization of phi, psi angles followed by one of the loop closure algorithms such as cyclic coordinate descent or kinematic closure (Canutescu and Dunbrack 2003; Coutsias, Seok et al. ; Mandell, Coutsias et al.). In the case of medium to low sequence identity between template and target, Raman et al. applied a more aggressive iterative stochastic rebuild and refine protocol that allowed the complete rebuilding of large regions of the protein, which in some cases included entire secondary structure elements (Raman, Vernon et al. 2009).

Mandell et al.(Mandell, Coutsias et al. 2009) recently developed a Loop Closure algorithm in ROSETTA that achieves RMSD-values better than 1. Their adaptation of Kinematic closure (KIC) first selects 3 C_α atoms as pivots. Next non-pivot (φ,ψ) torsion angles are sampled, leading to a chain break at the middle pivot. Finally KIC is used to find torsion angles for the pivot atoms that close the loop. For a dataset of 25 loops containing 12 residues each, ROSETTA achieved a median accuracy of 0.8 RMSD (see Figure 2). This demonstrates an improvement over both the standard ROSETTA

cyclic coordinate descent protocol and a state-of-the-art molecular dynamics protocol (median accuracies of 2.0 RMSD and 1.2 RMSD respectively).

Model relaxation and refinement

After constructing a protein backbone via *de novo* protein folding or comparative modeling, the model enters atom-detail refinement (Bradley, Misura et al. ; Misura, Chivian et al. ; Qian, Raman et al.). During the iterative relaxation protocol φ and ψ angles of the backbone are perturbed slightly while maintaining the overall global conformation of the protein. The side chains

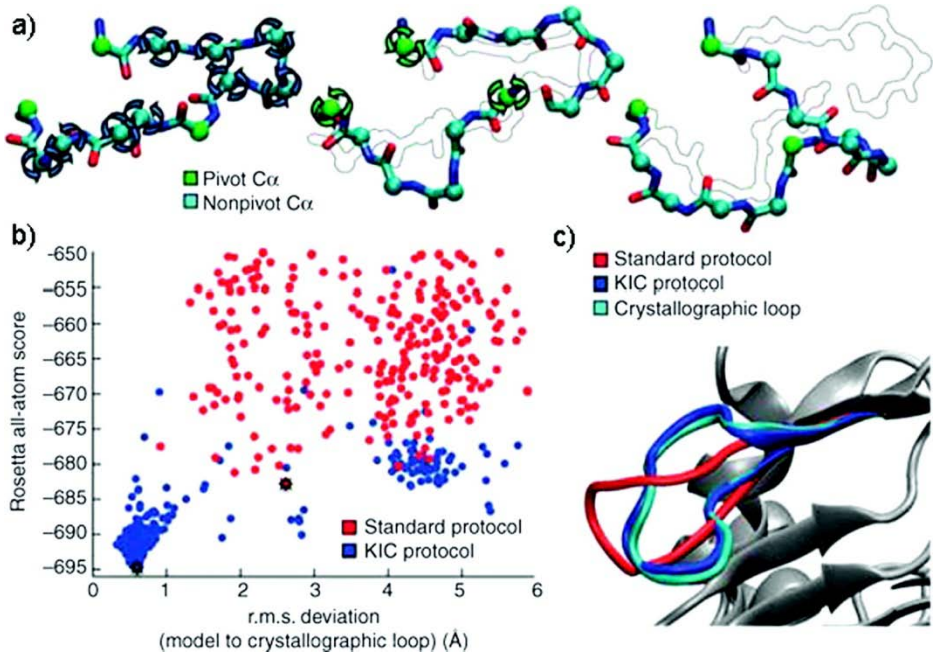


Figure 2. **Kinematic Loop Closure** *a*) The kinematic loop closure algorithm samples (φ, ψ) torsion angles at the cyan Cα spheres from a residue specific Ramachandran map. The (φ, ψ) torsion angles at green Cα spheres are determined analytically to close the loop. *b*) The energy versus RMSD plot shows that the improved sampling offered by the kinematic closure protocol results in a sub angstrom prediction for the loop conformation. *c*) The kinematic closure prediction better resembles the crystallographic conformation. Reprint with permission from (Mandell, Coutsias et al. 2009).

of the protein are adjusted using a simulated annealing Monte Carlo Metropolis search of the rotamer space. Finally, gradient minimization is applied to all torsional degrees of freedom (φ, ψ, ω, and χ). The repulsive portion of van der Waals potential is increased incrementally, moving the structure to the nearest energy minimum. Extensive use of the all-atom model refinement has proven integral to the success of ROSETTA in the recent Critical Assessment of Structure Prediction (CASP) experiments. The Relax protocol has been extremely valuable for performing rapid Monte-Carlo minimization of protein backbones. Recently, Qian et al (Qian, Raman et al. 2007) applied the refinement protocol to protein structures determined *de novo*, via comparative modeling, or using NMR-derived restraints.

In this protocol, protein models derived from NMR constraints or comparative modeling were used as a basis for solving the crystallographic phase problem. The model was initially minimized using ROSETTA's all atom Monte Carlo Refinement protocol. The results of this refinement were used to identify regions likely to deviate from the native structure, as regions of high variability between refined models often correlate to areas of deviation from the native structure. These areas were re-sampled using the fragment assembly approach used by Rosetta's *de novo* structure prediction protocol. These re-sampled models are then subjected to all atom refinement. This cycle of refinement and rebuilding is preformed iteratively, each time using a small selection of the lowest energy structures from the previous round of refinement. The iterative refinement process is repeated until the system converges. These structures were then used to in molecular replacement to solve the crystallographic phase problem, and as a means of refining models derived from medium resolution NMR data. In a blind test, this *ab initio* phase solution method resulted in significant improvement in structural resolution over the original unrefined models (Qian, Raman et al. 2007).

ROSETTA in Practice

ROSETTA's performance in the CASP experiment

ROSETTA has displayed remarkable success in *de novo* structure prediction in the last several blind critical assessment of structure prediction (CASP) experiments as is evidenced by the method's ranking among the top structure prediction groups (Bonneau, Tsai et al. 2001; Bradley, Chivian et al. 2003; Chivian, Kim et al. 2003; Rohl, Strauss et al. 2004; Bradley, Malmstrom et al. 2005; Das, Qian et al. 2007; Raman, Vernon et al. 2009). During CASP sequences of proteins not yet reported in the PDB are distributed among participating laboratories. Within a given time limit predictions are collected and assessed based on the experimental structure that is typically available shortly after the CASP experiment (www.predictioncenter.org). Generally ROSETTA has superseded competing approaches at predicting the structure of small to moderately sized protein domains with less than 200 amino acids *de novo*. Shortly after the CASP6 (held in 2004) Bradley et al. showed that for a benchmark of small proteins ROSETTA *de novo* structure prediction found models at atomic detail accuracy in an encouraging 8 out of 16 cases (Bradley, Misura et al. 2005). In that same year, Misura et al. found that homology models built with ROSETTA can be more accurate than their templates (Misura, Chivian et al. 2006). During CASP 7 with the assistance of the ROSETTA@HOME distributed computer network, several moderately sized domains were predicted to atomic-detail accuracy within 2 of the experimental structure for the first time {Das, 2007}. Based on the performance of ROSETTA in improving models over the best template structures available (Raman, Vernon et al. 2009) (see Figure 3), Raman et al. suggest that the limitation of the ROSETTA structure prediction protocol remains in the sampling algorithms rather than the energy function. For this reason, prediction of larger domains becomes possible upon introduction of experimental data which restricts the conformational space.

ROSETTA leverages sparse data from NMR and EPR experiments

ROSETTA allows incorporation of many types of experimental restraints. ROSETTA's ability to deal with restraints derived from Nuclear Magnetic Resonance (NMR) spectroscopy is the most developed (Rohl 2005). NMR restraints have two entry points into the ROSETTA protein structure

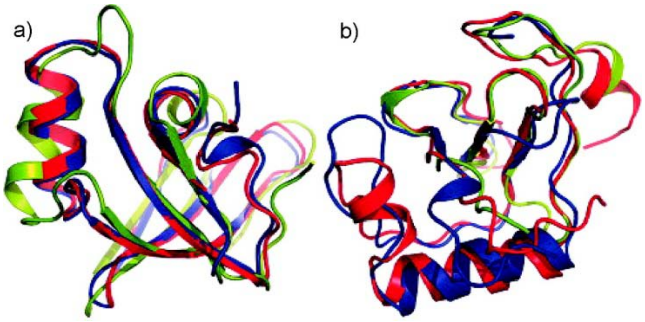


Figure 3. **Comparative Modeling CASP Performance** Raman and colleagues demonstrated that comparative models refined with Rosetta improved upon the best template structure available for several targets, for example a) T0492 and b) T0464. The native structure is shown in blue, the best submitted Rosetta model in red, and the best template structure in green. The Rosetta models for T0492 resulted in atomic level accuracy for side chains in the core. For T0464 a 25 residue insertion was predicted resulting in models significantly improved over the best templates available. One of the models was further improved in the model refinement category. Reprinted with permission from (Raman, Vernon et al. 2009)

prediction routine. Chemical shift assignments for backbone atoms can be converted to phi, psi backbone angle restraints (Cornilescu, Delaglio et al. 1999) and are used during the selection of the fragment libraries. Distance and orientation restraints (e.g. nuclear Overhauser effect (NOE) restraints and Residual Dipolar Couplings (RDCs) respectively) are incorporated into the scoring function used during folding. Bowers et al. showed that a sparse mixture of short and long range NOE restraints (1 restraint per residue) in addition to the backbone chemical shifts enables ROSETTA to build models at atomic-detail accuracy (Bowers, Strauss et al. 2000). Rohl and Baker (Rohl and Baker 2002) likewise demonstrated that limited RDC measurements (1 per residue) in combination with backbone chemical shifts identify the correct fold. Meiler and Baker presented a protocol that uses unassigned NMR restraints for rapid protein fold determination (Meiler and Baker 2003). More recently, Shen et al. showed the use of a modified fragment selection protocol

that ROSETTA can be used to generate structures of a quality comparable to those from traditional NMR structure determination methods (Shen, Lange et al. 2008). Furthermore, Shen found that ROSETTA sampling can compensate for the incomplete and incorrect NMR restraints (Shen, Vernon et al. 2009). A major point to note is that in each of these examples ROSETTA is used to complement structure restraints obtained early in the structure determination process. Consistently ROSETTA models are accurate at atomic detail that would only be apparent from either significantly more or higher resolution experimental information. For example, Rohl and Baker found that ROSETTA

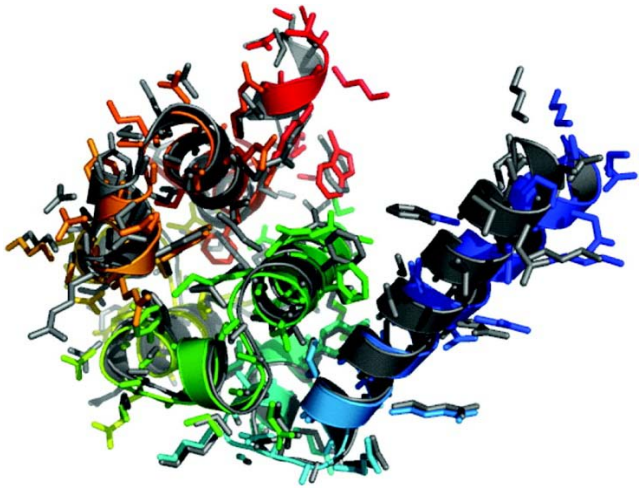


Figure 4. **De novo Protein Structure Prediction from Sparse EPR Data** Alexander et al. demonstrated that approximately one low resolution distance restraint for every four residues is sufficient to determine a model of T4-Lysozyme that is accurate at atomic-detail. The didstance restraints had been determined using SDSL-EPR experiments. The native T4-Lysozyme structure is show in grey, while the model with an RMSD of 1.0Å is shown in a rainbow coloring scheme. Side chain conformations in the core of the protein are accurately represented in the model. Reprinted with permission from (Alexander, Bortolus et al. 2008)

produced ubiquitin models less than 4 RMSD of the experimental structure using RDC restraints that were also consistent with models that have a RMSD greater than 20 (Rohl and Baker 2002). Beyond NMR restraints, any experimental data suitable to represent the distance between atoms can be used in the simulation. Through site-directed spin-labeling (SDSL) such distance restraints can be obtained from Electron Paramagnetic Resonance (EPR) spectroscopy (Alexander, Bortolus et al. 2008). Alexander et al. generated accurate atomic-detail models of T4-Lysozyme (see Figure 4) and

the heat shock protein alpha crystalline using SDSL-EPR data using as few as a single distance restraint for every four residues. Similar approaches can be used to model multimeric complexes from monomers as Hanson et al. showed for the Arrestin tetramer in solution (Hanson, Dawson et al. 2008).

ROSETTA models assist determining molecular structures from electron diffraction data

De novo predicted models have also been used to assist in phasing of X-ray diffraction data (see Figure 5) (Qian, Raman et al. 2007; Das and Baker 2008; Ramelot, Raman et al. 2009). Das and Baker found that for 15 of 30 benchmark cases ROSETTA *de novo* models successfully solved the phase problem by molecular replacement(Das and Baker 2009). Das and Baker suggest that approximately one in six X-ray diffraction data sets for proteins of 100 residues or less in length can be solved via molecular replacement using ROSETTA generated *de novo* models. In a subsequent study, Ramelot et al. showed that refinement of NMR ensembles using ROSETTA results in higher quality molecular replacement solutions to X-ray diffraction data than directly using the NMR ensembles(Ramelot, Raman et al.). DiMaio et al. extended ROSETTA to directly build models from electron density(DiMaio, Tyka et al. 2009). Both Lindert and DiMaio have obtained atomic accuracy models into cryo-electron microscopy density maps at resolutions of 4-7 using ROSETTA (DiMaio, Tyka et al. ; Lindert, Staritzbichler et al. ; Lindert, Stewart et al.). In both cases resulting models are of a higher resolution than the density from which they were built.

Protein structure prediction servers

Large parts of the ROSETTA protein structure prediction protocol including generation of fragments, *de novo* folding, and comparative modeling have been replicated in an automated server ROBETTA(Chivian, Kim et al. 2003; Kim, Chivian et al. 2004; Chivian, Kim et al. 2005). Chivian found that comparative models built with early versions of ROBETTA generally did not improve upon templates

from close homologs(Chivian, Kim et al. 2005), however recently ROBETTA performed well in fold recognition and produced models which serve as good starting points for further refinement. In the most recent CASP however ROBETTA produced several models with accuracy comparable to the

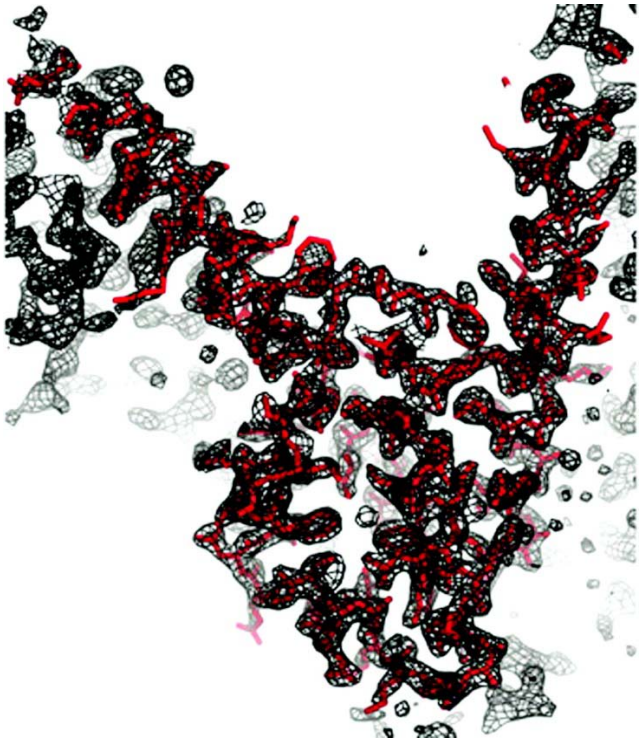


Figure 5. **The Crysallographic Phase Problem.** Qian et al. demonstrated that Rosetta-predicted protein models can be used in conjunction with automated molecular replacement algorithms to determine phases for electron density maps. The coordinates deposited in the PDB (2hh6) (shown in red) fit well into the isosurface of the electron density solved by molecular replacement using a Rosetta prediction for T0283 at CASP 7. Reprinted with permission from (Qian, Raman et al. 2007)

best human predictions (Raman, Vernon et al. 2009). For instance ROBETTA's top model for the server only target, T0513 domain 2, had an RMSD of 0.84. In general ROBETTA performance compared to other servers increases as the quality of template structures decreases(Raman, Vernon et al. 2009). ROBETTA is publicly accessible at robetta.bakerlab.org.

Protein-Protein Docking

While protein function is often determined by interactions with other proteins, most structures found in the Protein Databank contain single chains. Because of the difficulties in determining structures of protein complexes, computational methods for predicting protein/protein interactions are important. ROSETTADOCK provides tools for predicting the interaction of two proteins (Gray, Moughon et al. 2003). ROSETTADOCK employs first a low-resolution rigid-body docking. The second high-resolution refinement stage provides for side-chain conformational sampling and backbone relaxation.

The ROSETTADOCK algorithm begins with random reorientation of both proteins (Gray, Moughon et al. 2003). Next one protein slides into contact with the other. The following low resolution docking conformational search involves 500 Monte Carlo rigid body movements. These moves rotate and translate one protein around the surface of the other with movements chosen from a Gaussian distribution centered at 5° degrees and 0.7 Å. Each conformation is scored using the low-resolution energy function based on residue pair interaction statistics, residue environment statistics, and van der Waals attractive and repulsive terms. In this low resolution step, side-chains are represented by their centroids.

Next, 50 cycles of high resolution refinement at atomic detail are performed. Each cycle consists of a 0.1 Å random rigid-body translation, Monte Carlo based side-chain rotamer sampling (packing), and gradient-based rigid-body minimization to find a local energy minimum. Finally backbone flexibility is introduced around the protein interface. ROSETTADOCK is available as a web server (rosettadock.graylab.jhu.edu) but is limited to complexes for which the approximate binding orientation is known. The server produces 1,000 structures and returns details for the 10 lowest energy models (Lyskov and Gray 2008).

ROSETTADOCK successfully recovered the native structures of 42 out of 54 benchmark targets from which side-chains had been removed (Gray, Moughon et al. 2003). Starting with randomly placed proteins, ROSETTADOCK predicts more than 50% of the interface contacts for 23 out of 32 benchmark

targets (Gray, Moughon et al. 2003). These results have improved with the addition of backbone flexibility (Wang, Bradley et al. 2007), and conformational sampling (Chaudhury and Gray 2008).

ROSETTADOCK has been used to predict the structures of anthrax protective antigen (Sivasubramanian, Maynard et al. 2008) and epidermal growth factor (Sivasubramanian, Chao et al. 2006) bound to monoclonal antibodies. Both docking studies led to predicted interface structures consistent with known mutant binding properties and were used to select residues for site directed mutagenesis. The antibody modeling protocol has been made accessible through a web server (antibody.graylab.jhu.edu).

ROSETTADOCK was benchmarked in the Critical Assessment of PRediction of Interactions (CAPRI) experiment (Figure 6), where it achieved full-atom RMSDs of better than 1.6 Å for most targets

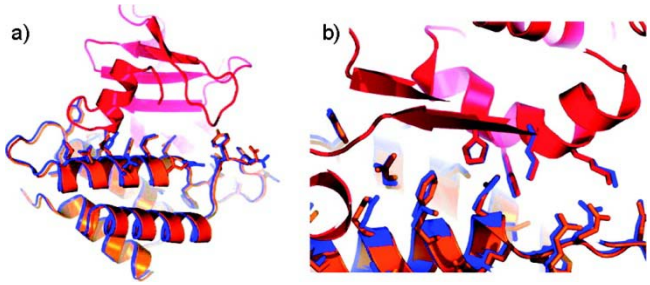


Figure 6. **Protein Interface Prediction** High-resolution CAPRI prediction of Colicin D /Immunity protein D interface. Both rigid-body orientation and side-chain conformation were modeled. The crystal structure is shown in red and orange, the Rosetta model is shown in blue. a) Whole protein complex. b) The interface shows the side-chains of the catalytic residue His611 and additional positively charged residues that are thought to bind to the RNA, as well as their matching negatively charged residues in the immunity protein. Reprinted with permission from (Schueler-Furman, Wang et al. 2005)

(Schueler-Furman, Wang et al. 2005). Its success has been attributed to advances such as the inclusion of gradient-based energy minimization of side-chain torsion angles (Schueler-Furman, Wang et al. 2005), incorporation of biochemical data (Chaudhury, Sircar et al. 2007), and coupling of docking with loop modeling (Chaudhury, Sircar et al. 2007).

Sircar and Gray recently reported on an extension of the ROSETTADOCK algorithm that allows for accurate modeling of antibody-antigen complexes in the absence of an antibody crystal structure (Sircar and Gray 2010). SNUGDOCK simultaneously samples the rigid body antibody-antigen positions, the

orientation of antibody light and heavy chains and the conformations of the six complementary determining loops. Additionally antibody conformational ensembles can be provided to mimic conformational selection. As in ROSETTADOCK side chain rotamers are sampled during high resolution refinement.

SNUGDOCK was compared with ROSETTADOCK in a blind prediction of human MCP-1 binding 11k2 antibody (PDB ID 2bdn) (Reid, Rushe et al. 2006). While the lowest energy structure produced by ROSETTADOCK is incorrect, the model produced by SNUGDOCK meets the CAPRI acceptable criterion of having more than 30% of the residue-residue contacts predicted correctly (39%). When combined with ensemble sampling, five of the ten lowest energy models meet the CAPRI medium quality criterion of correctly predicting more than 50% residue-residue contacts. Similar results were seen in a benchmark of 15 antibody/antigen complexes.

Protein-Ligand Docking

Ligand docking seeks to predict the interaction between a protein and a small molecule. Most ligand docking applications struggle to correctly predict conformational selection or induced-fit effects (Taylor, Jewsbury et al. 2002) resulting from ligand and protein flexibility. As applications originally designed for protein/ligand docking, flexibility is often a feature added as an afterthought. On the other hand ROSETTA was originally developed for *de novo* structure prediction. As such it was designed from its inception to efficiently model flexibility.

ROSETTALIGAND is an application for docking a small molecule in the binding pocket of a protein that considers both ligand and protein flexibility (Meiler and Baker 2006). The ROSETTALIGAND algorithm is a modification of the ROSETTADOCK algorithm. First, a ligand conformer is chosen randomly from a user provided ligand conformational ensemble. Second, the ligand is moved to a user defined putative binding site. A low-resolution shape-complementarity search translates and rotates the ligand

optimizing attractive and repulsive score terms. In the high-resolution phase cycles of Monte Carlo minimization perturb the ligand pose and optimize amino acid side-chain rotamers and ligand conformers. Lastly all torsion degrees of freedom in ligand and protein undergo gradient minimization and the model is output.

In a benchmark by Meiler and Baker (Meiler and Baker 2006), ROSETTALIGAND successfully recovered the native structure of 80/100 protein/ligand complexes with RMSD better than 2.0 Å. When docking ligands into experimental protein structures determined with different binding partners (cross-docking), ROSETTALIGAND recovered the native structure in 14 of 20 cases. Comparing binding energy predictions with 229 experimentally determined binding energies from the Ligand-Protein Database (lpdb.chem.lsa.umich.edu) (Roche, Kiyama et al. 2001), ROSETTALIGAND achieved an overall correlation coefficient of 0.63, which is comparable to the best scoring functions available for protein-ligand interfaces (Ferrara, Gohlke et al. 2004).

Recently, backbone flexibility was added to the docking algorithm which led to improved predictions, including lower RMSDs among top scoring ligands (Davis and Baker 2009). Backbone flexibility allows prediction of induced-fit effects that occur upon ligand binding. When tested in a blind study on a set of lead-like compounds ROSETTALIGAND performance was comparable to other commercially available docking programs (Davis, Raha et al. 2009). The authors caution however that current docking programs fail 70% of the time on at least one of the receptor in the test set.

References

Alexander, N., M. Bortolus, et al. (2008). "De novo high-resolution protein structure determination from sparse spin-labeling EPR data." Structure **16**(2): 181-195.

Berman, H. M., T. Battistuz, et al. (2002). "The Protein Data Bank." Acta Crystallogr D Biol Crystallogr **58**(Pt 6 No 1): 899-907.

Bernstein, F. C., T. F. Koetzle, et al. (1977). "The Protein Data Bank: a computer-based archival file for macromolecular structures." J Mol Biol **112**(3): 535-542.

Bonneau, R., C. E. Strauss, et al. (2002). "De novo prediction of three-dimensional structures for major protein families." J Mol Biol **322**(1): 65-78.

Bonneau, R., J. Tsai, et al. (2001). "Rosetta in CASP4: progress in ab initio protein structure prediction." Proteins **Suppl 5**: 119-126.

Bowers, P. M., C. E. Strauss, et al. (2000). "De novo protein structure determination using sparse NMR data." J Biomol NMR **18**(4): 311-318.

Bradley, P., D. Chivian, et al. (2003). "Rosetta predictions in CASP5: successes, failures, and prospects for complete automation." Proteins **53 Suppl 6**: 457-468.

Bradley, P., L. Malmstrom, et al. (2005). "Free modeling with Rosetta in CASP6." Proteins **61 Suppl 7**: 128-134.

Bradley, P., K. M. Misura, et al. (2005). "Toward high-resolution de novo structure prediction for small proteins." Science **309**(5742): 1868-1871.

Bystroff, C., K. T. Simons, et al. (1996). "Local sequence-structure correlations in proteins." Curr Opin Biotechnol **7**(4): 417-421.

Canutescu, A. A. and R. L. Dunbrack (2003). "Cyclic coordinate descent: A robotics algorithm for protein loop closure." Protein Science **12**(5): 963-972.

Chaudhury, S. and J. J. Gray (2008). "Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles." J Mol Biol **381**(4): 1068-1087.

Chaudhury, S., A. Sircar, et al. (2007). "Incorporating biochemical information and backbone flexibility in RosettaDock for CAPRI rounds 6-12." Proteins **69**(4): 793-800.

Chivian, D., D. E. Kim, et al. (2003). "Automated prediction of CASP-5 structures using the Robetta server." Proteins **53 Suppl 6**: 524-533.

Chivian, D., D. E. Kim, et al. (2005). "Prediction of CASP6 structures using automated Robetta protocols." Proteins **61 Suppl 7**: 157-166.

Cornilescu, G., F. Delaglio, et al. (1999). "Protein backbone angle restraints from searching a database for chemical shift and sequence homology." J Biomol NMR **13**(3): 289-302.

Coutsias, E. A., C. Seok, et al. (2004). "A kinematic view of loop closure." J Comput Chem **25**(4): 510-528.

Das, R., I. André, et al. (2009). "Simultaneous prediction of protein folding and docking at high resolution." Proc Natl Acad Sci USA **106**(45): 18978-18983.

Das, R. and D. Baker (2008). "Macromolecular modeling with rosetta." Annu Rev Biochem **77**: 363-382.

Das, R. and D. Baker (2009). "Prospects for de novo phasing with de novo protein models." Acta Crystallogr D Biol Crystallogr **65**(Pt 2): 169-175.

Das, R., B. Qian, et al. (2007). "Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home." Proteins **69 Suppl 8**: 118-128.

Davis, I. W. and D. Baker (2009). "RosettaLigand docking with full ligand and receptor flexibility." J Mol Biol **385**(2): 381-392.

Davis, I. W., K. Raha, et al. (2009). "Blind docking of pharmaceutically relevant compounds using RosettaLigand." Protein Sci **18**(9): 1998-2002.

DiMaio, F., M. D. Tyka, et al. (2009). "Refinement of protein structures into low-resolution density maps using rosetta." J Mol Biol **392**(1): 181-190.

Dunbrack, R. L. and M. Karplus (1993). "Backbone-Dependent Rotamer Library for Proteins - Application to Side-Chain Prediction." Journal of Molecular Biology **230**(2): 543-574.

Ferrara, P., H. Gohlke, et al. (2004). "Assessing scoring functions for protein-ligand interactions." J Med Chem **47**(12): 3032-3047.

Gray, J. J., S. Moughon, et al. (2003). "Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations." J Mol Biol **331**(1): 281-299.

Hanson, S. M., E. S. Dawson, et al. (2008). "A model for the solution structure of the rod arrestin tetramer." Structure **16**(6): 924-934.

Kaufmann, K. W., G. H. Lemmon, et al. (2010). "Practically useful: what the Rosetta protein modeling suite can do for you." Biochemistry **49**: 2987-2998.

Kim, D. E., D. Chivian, et al. (2004). "Protein structure prediction and analysis using the Robetta server." Nucleic Acids Res **32**(Web Server issue): W526-531.

Kortemme, T., A. V. Morozov, et al. (2003). "An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes." J Mol Biol **326**(4): 1239-1259.

Kuhlman, B. and D. Baker (2000). "Native protein sequences are close to optimal for their structures." Proc Natl Acad Sci U S A **97**(19): 10383-10388.

Kuhlman, B., G. Dantas, et al. (2003). "Design of a novel globular protein fold with atomic-level accuracy." Science **302**(5649): 1364-1368.

Lazaridis, T. and M. Karplus (1999). "Effective energy function for proteins in solution." Proteins **35**(2): 133-152.

Leaver-Fay, A., B. Kuhlman, et al. (2005). "Rotamer-Pair Energy Calculations Using a Trie Data Structure." Lecture Notes in Computer Science **3692**: 389-400.

Levinthal, C. (1968). "Are there pathways for protein folding." Journal de Chimie Physique et de Physico-Chimie Biologique **65**: 44-45.

Lindert, S., R. Staritzbichler, et al. (2009). "EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps." Structure **17**(7): 990-1003.

Lindert, S., P. L. Stewart, et al. (2009). "Hybrid approaches: applying computational methods in cryo-electron microscopy." Curr Opin Struct Biol **19**(2): 218-225.

Lyskov, S. and J. J. Gray (2008). "The RosettaDock server for local protein-protein docking." Nucleic Acids Res **36**(Web Server issue): W233-238.

Mandell, D. J., E. A. Coutsias, et al. (2009). "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling." Nat Methods **6**(8): 551-552.

Mandell, D. J., E. a. Coutsias, et al. (2009). "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling." Nature methods **6**: 551-552.

Mandell, D. J. and T. Kortemme (2009). "Computer-aided design of functional protein interactions." Nat Chem Biol **5**(11): 797-807.

Meiler, J. and D. Baker (2003). "Rapid protein fold determination using unassigned NMR data." Proc Natl Acad Sci U S A **100**(26): 15404-15409.

Meiler, J. and D. Baker (2006). "ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility." Proteins **65**(3): 538-548.

Misura, K. M., D. Chivian, et al. (2006). "Physically realistic homology models built with ROSETTA can be more accurate than their templates." Proc Natl Acad Sci U S A **103**(14): 5361-5366.

Qian, B., S. Raman, et al. (2007). "High-resolution structure prediction and the crystallographic phase problem." Nature **450**(7167): 259-264.

Raman, S., R. Vernon, et al. (2009). "Structure prediction for CASP8 with all-atom refinement using Rosetta." Proteins **77 Suppl 9**: 89-99.

Raman, S., R. Vernon, et al. (2009). "Structure prediction for CASP8 with all-atom refinement using Rosetta." Proteins.

Ramelot, T. A., S. Raman, et al. (2009). "Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study." Proteins **75**(1): 147-167.

Reid, C., M. Rushe, et al. (2006). "Structure activity relationships of monocyte chemoattractant proteins in complex with a blocking antibody." Protein Eng Des Sel **19**(7): 317-324.

Roche, O., R. Kiyama, et al. (2001). "Ligand-protein database: linking protein-ligand complex structures to binding data." J Med Chem **44**(22): 3592-3598.

Rohl, C. A. (2005). "Protein structure estimation from minimal restraints using Rosetta." Methods Enzymol **394**: 244-260.

Rohl, C. A. and D. Baker (2002). "De novo determination of protein backbone structure from residual dipolar couplings using Rosetta." J Am Chem Soc **124**(11): 2723-2729.

Rohl, C. A., C. E. Strauss, et al. (2004). "Modeling structurally variable regions in homologous proteins with rosetta." Proteins **55**(3): 656-677.

Rohl, C. A., C. E. Strauss, et al. (2004). "Protein structure prediction using Rosetta." Methods Enzymol **383**: 66-93.

Rohl, C. A., C. E. M. Strauss, et al. (2004). "Protein structure prediction using rosetta." Numerical Computer Methods, Pt D **383**: 66-+.

Schueler-Furman, O., C. Wang, et al. (2005). "Progress in protein-protein docking: Atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility." Proteins-Structure Function and Bioinformatics **60**(2): 187-194.

Schueler-Furman, O., C. Wang, et al. (2005). "Progress in protein-protein docking: atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility." Proteins **60**: 187-194.

Shen, Y., O. Lange, et al. (2008). "Consistent blind protein structure generation from NMR chemical shift data." Proc Natl Acad Sci U S A **105**(12): 4685-4690.

Shen, Y., R. Vernon, et al. (2009). "De novo protein structure generation from incomplete chemical shift assignments." J Biomol NMR **43**(2): 63-78.

Simons, K. T., C. Kooperberg, et al. (1997). "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions." J Mol Biol **268**(1): 209-225.

Simons, K. T., I. Ruczinski, et al. (1999). "Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins." Proteins **34**(1): 82-95.

Sircar, A. and J. J. Gray (2010). "SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models." PLoS Comput Biol **6**(1): e1000644.

Sivasubramanian, A., G. Chao, et al. (2006). "Structural model of the mAb 806-EGFR complex using computational docking followed by computational and experimental mutagenesis." Structure **14**(3): 401-414.

Sivasubramanian, A., J. A. Maynard, et al. (2008). "Modeling the structure of mAb 14B7 bound to the anthrax protective antigen." Proteins **70**(1): 218-230.

Taylor, R. D., P. J. Jewsbury, et al. (2002). "A review of protein-small molecule docking methods." Journal of Computer-Aided Molecular Design **16**(3): 151-166.

Wang, C., P. Bradley, et al. (2007). "Protein-protein docking with backbone flexibility." J Mol Biol **373**(2): 503-519.

Yokoyama, H., S. Hamamatsu, et al. (2008). "Novel dimer structure of a membrane-bound protease with a catalytic Ser-Lys dyad and its linkage to stomatin." J Synchrotron Radiat **15**(Pt 3): 254-257.

CHAPTER II

SMALL MOLECULE ROTAMERS ENABLE SIMULTANEOUS OPTIMIZATION OF SMALL MOLECULE AND PROTEIN DEGREES OF FREEDOM IN ROSETTALIGAND DOCKING[1]

Introduction

Representing protein flexibility through side chain rotamers (Dunbrack and Karplus 1993) has been central to the success of protein structure prediction, protein docking, and protein design. This discretization of protein side chain conformations observed in the Protein Databank is, for example, used by the ROSETTA program in the *de novo* prediction of protein structure (Bradley, Misura et al. 2005). Furthermore, rotamers form critical components of successful protein docking and protein design strategies such as ROSETTADOCK (Gray, Moughon et al. 2003) (Schueler-Furman, Wang et al. 2005) and ROSETTADESIGN (Kuhlman, O'Neill et al. 2001; Dantas, Kuhlman et al. 2003; Kuhlman, Dantas et al. 2003). Finally, ROSETTA incorporates the rotamer probability when performing alanine scanning mutagenesis to identify key residues in protein-protein interfaces (hot spots) (Kortemme, Kim et al. 2004). The above success of rotamers for modeling protein side chain flexibility makes adapting the concept for small molecule flexibility attractive.

Leach first introduced using rotamers in modeling small molecule flexibility during docking (Leach 1994). He took small molecule conformations in local minima of a molecular dynamics forcefield as rotamers. However, Leach observed a failure of the energy function on docking of phosphocholine to the antibody McPc 603. We independently implemented a similar method using rigid ligands and full side chain flexibility in the ROSETTA (Meiler and Baker 2006) protein modeling suite. The ROSETTALIGAND energy function identified native conformations for 71 of 100 small molecule protein complexes in a self docking test and 14 of 20 small molecule protein complexes in a ”cross docking“ benchmark. In the cross

[1] Published as Kristian Kaufmann, Kirsten Glab, Ralf Mueller, and Jens Meiler “Small molecule rotamers enable simultaneous optimization of small molecule and protein degrees of freedom in RosettaLigand docking” *Proceedings from the German Conference on Bioinformatics* **2009** September 9-12 Dresden Germany 148-157

docking benchmark, a small conformational ensemble containing 10 conformations, one of which was close to the crystallized conformation, was used to simulate small molecule flexibility. In the present work it is our objective to simulate small molecule flexibility using small molecule rotamers generated from a crystal structure database of small molecules. This setup mirrors the amino acid side chain rotamer approach used in ROSETTA for small molecules and thus capitalizes on "knowledge based" rotamers and energy functions deemed responsible for the success of ROSETTA.

In an analogous manner to the amino acid side chain rotamers, we employ small molecule crystal structures from the Cambridge Structural Database (CSD) (Allen 2002) to construct small molecule rotamers. Unlike amino acid side chains in the Protein Data Bank (PDB), in the case of small molecules we lack multiple conformations of the same chemical configuration. Instead, torsion profiles are created from chemical similar groups. OMEGA, a highly regarded program for generating small molecule conformations, makes use of such torsion profiles extracted from the CSD. OMEGA generates conformational ensembles from overlapping fragments in a rule based manner using torsion profiles(Bostrom, Greenwood et al. 2003). Perola and Charifson, in a study of crystallized bioactive small molecules, found OMEGA to be the best available tool for generating ensembles containing the bioactive conformation (Perola and Charifson 2004).

Most current small molecule docking programs approach the docking problem from the perspective of the small molecule. In contrast, ROSETTALIGAND approaches small molecule docking from the perspective of protein modeling. We hypothesize that ROSETTALIGAND will more accurately represent small molecule protein interactions because of its focus on the protein point of view which allows the accurate simulation of protein flexibility and associated energetics. In our previous paper we demonstrated the utility of the ROSETTA energy function to discriminate native-like models (Meiler and Baker 2006). Here our objective is to demonstrate that the concept of rotamers in protein structure prediction can be extended to small molecules. We show that small molecule rotamers can be created using crystal structure data. In addition, these small molecule rotamer ensembles contain conformations similar to the bioactive conformation, in particular for small molecules with a number of torsions similar

to those in protein side chains (≤ 6 rotatable bonds). We show that these rotamer ensembles successfully simulate small molecule flexibility in small molecule docking benchmarks.

Methods

Creating Torsion Profiles from the Cambridge Structural Database

We use 28 atom types to describe atoms in small molecules defined by element type, hybridization state, and number of bonded hydrogens (Meiler, Maier et al. 2002). We measure non-hydrogen atom torsions for each atom type pairing for all structures in the Cambridge Structural Database (CSD) (Allen 2002), excluding torsions in ring systems. Histograms are constructed for every pair of the 28 atom types using bins with a width of 10°. Histograms with less than 100 data points are excluded as containing too little information. The distributions are made symmetric by summing counts of symmetry equivalent bins.

A knowledge based torsion energy to model the interactions between atoms separated by three bonds is calculated using the inverse Boltzmann equation (Eq. 1)

$$E_i = -\log P_{torsion} = -\log\left(\frac{N_i + 1}{N_{tot} \times P_{i,ran}}\right) \qquad (1)$$

where $P_{torsion}$ is the propensity of the torsion, N_i is the number of counts in a bin, N_{tot} is the total number of torsions observed for this type, $P_{i,ran}$ is the probability of selecting the bin from a uniform distribution. The propensity of the torsion is the probability of the torsion divided by the background probability the torsion value occurring by chance. The pseudo count of 1 is added to avoid zero probability bins, which result in infinite energies. The background probability is drawn from the uniform distribution since we assume that no other forces bias the torsions observed in the CSD. The weighting of the other internal energies will counterbalance the error introduced by this assumption. The minima in the energy profiles form the set of allowed dihedral angles for the rotamer ensemble generator.

Small Molecule Rotamer Ensemble Generation

The small molecule ensemble generation protocol (see Figure 1) creates an ensemble of acceptable energy rotamers. The protocol maximizes coverage of the conformational space accessible to the small molecule by maximizing pair-wise root mean squared deviation (RMSD) for all rotamers. Starting from a conformation with idealized bond lengths and angles, a set of dihedral angles is chosen from the minima of the appropriate torsion profiles. Rotamers containing overlapping atoms are discarded. If the energy is acceptable then the rotamer is provisionally accepted. Otherwise, a new set of dihedral angles are chosen. Using this protocol a list of 10,000 candidates is obtained for pruning. The pruning first selects the lowest energy rotamer from this list and makes it the first rotamer of the ensemble considered for docking.

The energy incorporates van der Waals interaction for atoms separated by four or more covalent bonds (Kuhlman, O'Neill et al. 2001), the knowledge based torsion energy described in the previous section, an intra-molecular hydrogen bonding term (Morozov and Kortemme 2005), a desolvation energy based on the Lazaridis-Karplus approximation (Lazaridis and Karplus 1999), and a coulomb electrostatics term (Kuhlman, O'Neill et al. 2001).

Next the protocol iteratively identifies the candidate rotamer that has the largest RMSD to all current members of the docking ensemble and adds it to this ensemble. The protocol continues until the desired number of 500 rotamer has been reached or all candidate rotamers are within a user defined cutoff of 0.2 Å RMSD of one member of the docking ensemble.

Flexible Small Molecule Docking

Given a protein structure and small molecule conformation the protocol (see Figure 1) first generates a conformational ensemble for the small molecule. Next iteratively conformations are chosen from the ensemble and placed at a random position and orientation within the manually defined binding site.

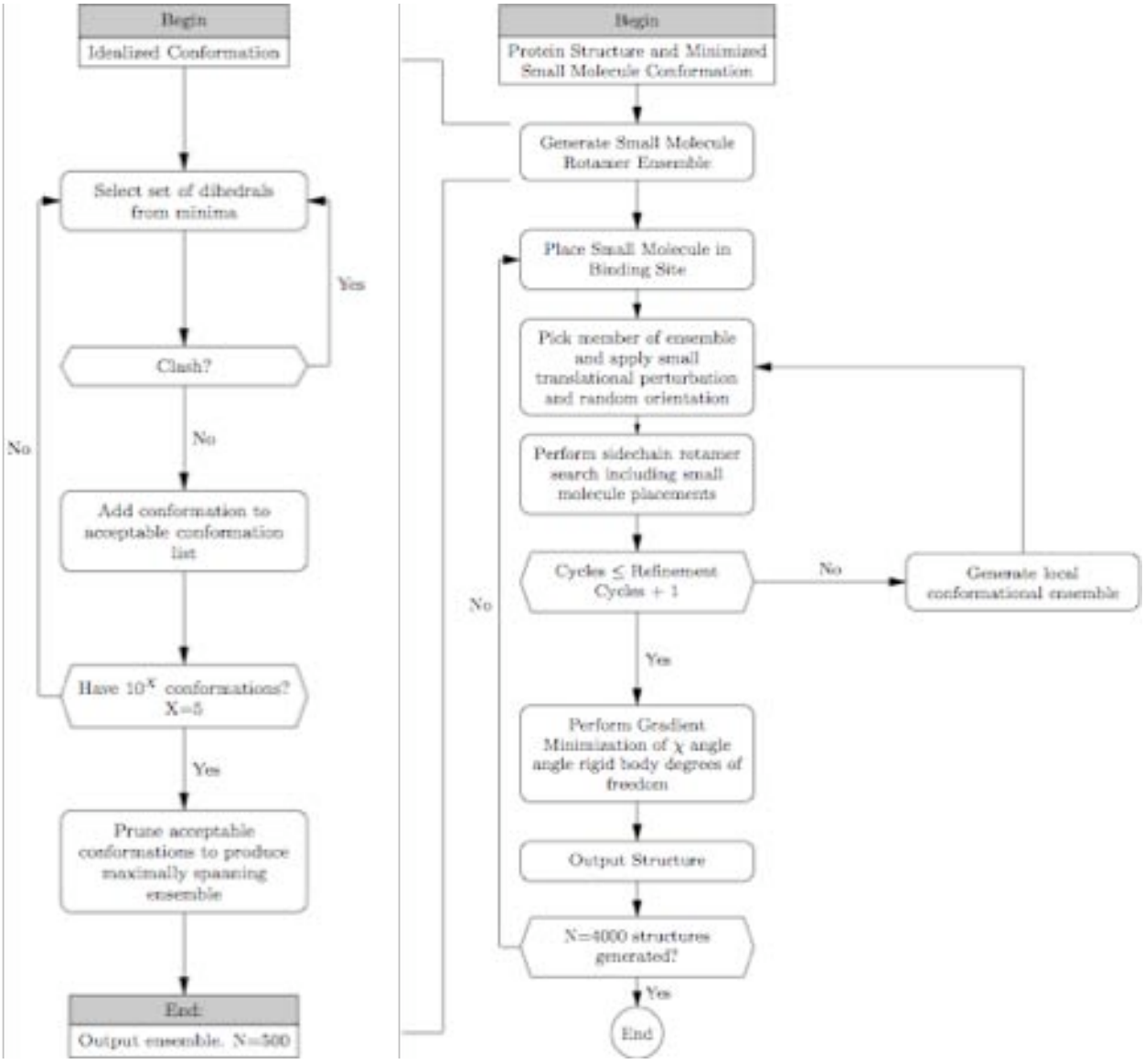


Figure 1. Small molecule docking protocol incorporates a rotamer ensemble generating protocol into a monte carlo search of protein side chain conformations. Reprinted with permission from (Kaufmann, Glab et al. 2008)

The first 100 non-clashing placements are incorporated as small molecule rotamers into the protein side-chain rotamer conformational optimization. After completion of this optimization a local ensemble of up to 100 rotamers is created for refinement by allowing small random changes sampled from a uniform distribution of [-5°, +5°] to all rotatable bonds of the optimized small molecule conformation. After four rounds of side chain optimization with this local discrete conformational ensemble, a gradient minimization of the amino acid side chain χ angles and the small molecule position and orientation take the structure to a local minimum. This structure is then written out. The sequence is repeated until 4,000 models have been generated.

Small Molecule Flexibility Benchmark Sets

Compounds for the ensemble generation test set were taken from the 2007 PDBBind database (Wang, Fang et al. 2004). All molecules with ≤ 6 rotatable non-hydrogen atom torsions were selected.

Two docking benchmarks were carried out. The self docking benchmark tests whether our protocol recovers the correct position, orientation, and conformation of a small molecule in the protein crystal structure solved with that same small molecule. Using the protein structure crystallized with the small molecule ensures the backbone of the protein is in the correct conformation for binding of this substance. The structures used in the self docking benchmark are listed in Table 1. The set contains 10 small molecules crystallized in 7 proteins. The cross docking benchmark is comprised of the same small molecules, but uses protein coordinates from other crystal structures. Hence, the cross docking benchmark assesses the capacity of the protocol to recover the placement of a small molecule in a real world situation where the protein was not crystallized with the small molecule of interest. The structures used are listed in Table 1. Meiler and Baker previously evaluated all structures in both docking benchmarks (Meiler and Baker 2006). The set was reduced to contain only small molecules with ≤ 6 rotatable non-hydrogen atom torsions.

Results and Discussion

Small Molecule Flexibility Benchmark Sets

The torsion profiles generated cover 103 common bond types (see supplement). The profiles obtained show similar characteristics to profiles seen in the AMBER (Wang, Wolf et al. 2004) forcefield (see Figure 2a).

Table 1. Crystal structures forming small molecule benchmark sets. Reprinted with permission from (Kaufmann, Glab et al. 2008).

Self docking protein structure	Small molecule	Number of torsions	Cross docking protein structures
1aq1 human Cyclin Dependent Kinase 2	Straurosporine	1	1dm2
1dm2 human Cyclin Dependent Kinase 2	Hymenialdisine	0	1aq1
1dbj IGG1-κ DB3 FAB	Aetiocholanolone	0	2dbl
2dbl IGG1-κ DB3 FAB	5-α-pregnane-3-β-ol-hemisuccinate	6	1dbj
1pph Trypsin	m-aminophenyl-3-alanine	5	1ppc
1p8d Liver X receptor β small molecule binding domain	24(S),25-epoxycholesterol	4	1pq6,1pqc
2ctc carboxypeptidase A	L-phenyl lactate	3	7cpa
2prg small molecule binding domain of peroxisome proliferator activated receptor γ	2,4-thiazolidiinedione, 5-[[4-[2-(methyl-2-pyridinylamino) ethoxy] phenyl]methyl]-(9cl)	5	1fm9
4tim Triosephosphate isomerase	2-phosphoglyceric acid	4	6tim
6tim Triosephosphate isomerase	SN-Glycerol-3-phosphate	4	4tim

However, some profiles exhibit minima not present in the AMBER forcefield. The aryl oxygen profile, shown in Figure 2b, displays additional minima at ± 90º. Klebe and Meitzner found that these additional minima arise from meta substituted compounds (Klebe and Mietzner 1994). The additional minima give the CSD torsion profiles an advantage, since they allow the ensemble generator to sample conformations that might otherwise be excluded.

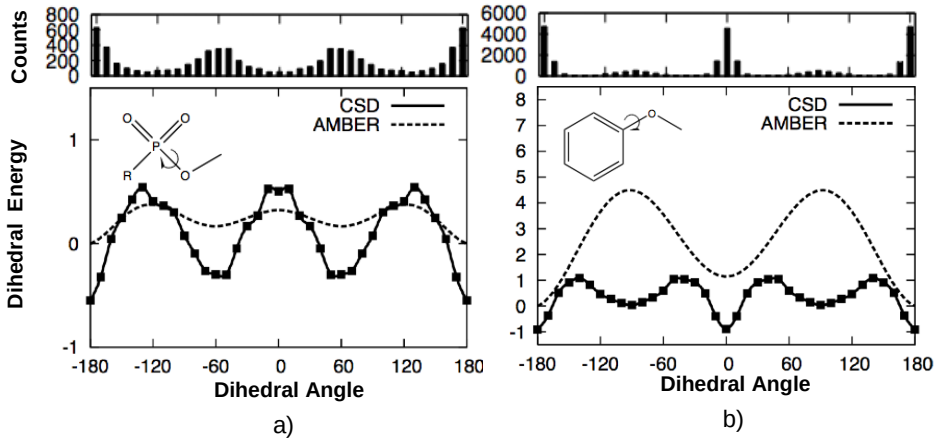


Figure 2. Torsion profiles for phosphate ether and aromatic carbon oxygen bond. Histograms for a) phosphate ether torsion and b) aromatic carbon oxygen torsion yield energy. Reprinted with permission from (Kaufmann, Glab et al. 2008)

Small Molecule Rotamer Ensemble Generation

The ensemble generator created ensembles for 628 small molecules with ≤ 6 rotatable bonds. The atomic coordinates of no two conformations within the ensemble were allowed to be closer 0.2 Å RMSD. Ten thousand conformations were generated while constructing the ensemble. The final ensembles

Table 2. Performance of rotamer ensemble generator evaluated by computing RMSD of closest and furthest conformation in the ensemble to the crystallized conformation of the small molecule Reprinted with permission from (Kaufmann, Glab et al. 2008)

Number of Torsions	Number of Molecules	Average RMSD of closest conformation	Average RMSD of furthest conformation
1	92	0.14 ± 0.16	1.12 ± 0.47
2	118	0.33 ± 0.26	1.74 ± 0.69
3	118	0.41 ± 0.22	2.13± 0.62
4	135	0.47 ± 0.21	2.45 ± 0.69
5	97	0.61 ± 0.30	2.83± 0.81
6	118	0.79 ± 0.32	3.07 ± 0.87
Overall Totals	628	0.46 ± 0.31	2.23± 0.94

contained up to 500 conformations. On the set of 628 molecules, the ensemble generator produced a rotamer with 0.46 ± 0.31 Å RMSD to the crystallized conformation. As expected, the accuracy decreases from 0.14 ± 0.16 Å RMSD to 0.79 ± 0.32 Å RMSD as the number of rotatable torsion angles increases from 1 to 6 (see Table 2). Improvement of these numbers might be possible by increasing the size of the

ensemble,and increasing the number of rotamers generated during construction of the ensemble. The additional cost of such increases may outweigh the benefits.

Flexible Small Molecule Docking

The small molecule docking results are summarized in Table 3 and Table 4. For the self docking, 9 of the 10 cases show a native-like model in the top 1 % by energy. In 7 of the 10 cases the top ranked model is native-like. For the cross docking benchmark 8 of 11 cases show a native-like structure in the top 1 % by energy.

Table 3. Summary of small molecule cross docking benchmark. Rank is determined by energy of best structure recapturing the binding mode. RMSD is of best structure recapturing the binding mode. Reprinted with permission from (Kaufmann, Glab et al. 2008)

Source structure for small molecule	Source structure for protein	Rank	RMSD
1aq1	1aq1	**1**	**0.56**
1dm2	1dm2	**1**	**0.31**
1dbj	1dbj	**1**	**1.36**
2dbl	2dbl	**1**	**1.45**
1p8d	1p8d	**1**	**1.63**
1pph	1pph	*6*	*1.49*
2prg	2prg	639	1.94
2ctc	2ctc	**3**	**0.82**
4tim	4tim	**1**	**1.87**
6tim	6tim	**1**	**1.77**

In only 2 of the 11 cases was the top ranked model a native-like model. In Figure 3a the RMSD energy plot demonstrates that the scoring function identifies the native binding mode as the most favorable (see Figure 3c). However, in other cases the RMSD energy plots appear like that of Figure 3b. Some models are present in the native binding mode (see Figure 3d), but low energy does not imply low RMSD.

The self docking results are comparable to those in Meiler and Baker (Meiler and Baker 2006). Meiler and Baker achieved a 71% success rate in a self docking benchmark of 100 small molecules. We see the same success rate on our reduced set despite the increased conformational space sampled for the small molecule. However in the cross docking benchmark our results fall short. One possible cause is the much increased conformational space sampled for small molecules in the present protocol.

The previous evaluation used an ensemble size of only ten in which one conformation was close to the crystallized conformation. Here, we create unbiased ensembles with up to 500 conformations. The increase in conformational diversity represents an increased challenge to the search process as well as the scoring function.

Table 3. Summary of small molecule cross docking benchmark. Rank is determined by energy of best structure recapturing the binding mode. RMSD is of best structure recapturing the binding mode. Reprinted with permission from (Kaufmann, Glab et al. 2008)

Source Structure for small molecule	Source Structure for protein	Rank	RMSD
1aq1	1dm2	4296	1.87
1dm2	1aq1	**1**	**0.56**
1dbj	2dbl	**1**	**1.80**
2dbl	1dbj	468	3.49
1p8d	1pq6	181	1.62
1p8d	1pqc	*10*	*1.28*
1pph	1ppc	*2*	*1.96*
2ctc	7cpa	*3*	*0.95*
2prg	1fm9	*16*	*2.02*
4tim	6tim	*2*	*1.90*
6tim	4tim	*5*	*1.77*

Conclusion

We have extended the amino acid concept of rotamers to include small molecules. When the number of torsions is in the same range as those seen in amino acids, small molecule rotamer ensembles contain conformations close to those seen in crystal structures of protein small molecule complexes. In such cases rotamer ensembles can efficiently simulate flexibility for small molecules.

However, as the number of rotamers grow (due to increased flexibility) and the precision of the protein structures decrease (due to inaccuracy in the protein backbone), the discriminatory power of the scoring function decreases. The components of the scoring function have not been optimized for the increased flexibility; doing so may yield increased discrimination. Improved fine grain sampling of protein backbone motion may also assist in the docking process.

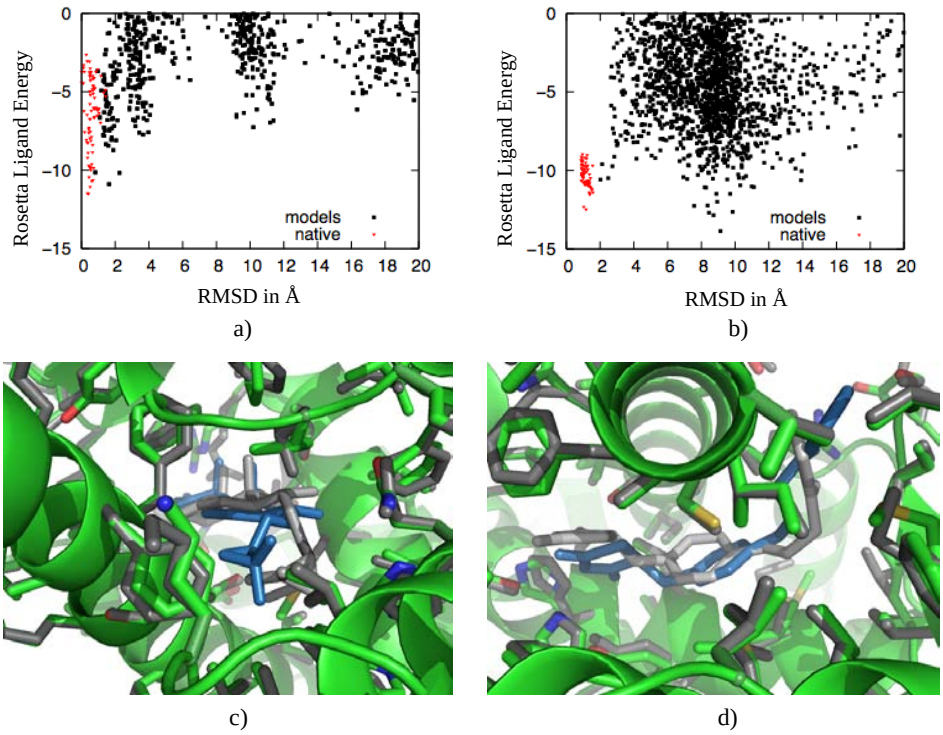


Figure 3. RMSD energy funnels show successful discrimination of native binding mode in a) for the case epoxycholesterol into Liver X receptor from 1P8D, and failure to rank the native binding mode as the lowest energy structure in b) for thiazolidiinedione from 2PRG in the 1FM9 structure of the peroxisome proliferator activated receptor. Panel c) shows the best scoring model (in green and blue) overlaid on the 2PRG structure (shown in gray). Panel d) shows the best scoring model under 2 Å RMSD (in green and blue) from the native structure from 2PRG (in grey). Reprinted with permission from (Kaufmann, Glab et al. 2008)

Additionally, the method must be extended to larger small molecules. We intend on expanding our method by breaking small molecules into multiple residues. The residues would then be reassembled in the protein binding site to form the small molecule. Thereby, we decrease the conformational complexity and incorporate information from the protein environment.

References

Allen, F. H. (2002). "The Cambridge Structural Database: a quarter of a million crystal structures and rising." Acta Crystallogr B **58**(Pt 3 Pt 1): 380-388.

Bostrom, J., J. R. Greenwood, et al. (2003). "Assessing the performance of OMEGA with respect to retrieving bioactive conformations." J Mol Graph Model **21**(5): 449-462.

Bradley, P., K. M. Misura, et al. (2005). "Toward high-resolution de novo structure prediction for small proteins." Science **309**(5742): 1868-1871.

Dantas, G., B. Kuhlman, et al. (2003). "A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins." Journal of Molecular Biology **332**(2): 449-460.

Dunbrack, R. L. and M. Karplus (1993). "Backbone-Dependent Rotamer Library for Proteins - Application to Side-Chain Prediction." Journal of Molecular Biology **230**(2): 543-574.

Gray, J. J., S. Moughon, et al. (2003). "Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations." J Mol Biol **331**(1): 281-299.

Kaufmann, K., K. Glab, et al. (2008). Small Molecule Rotamers Enable Simultaneous Optimization of Small Molecule and Protein Degrees of Freedom in ROSETTALIGAND Docking. German Conference on Bioinformatics, Dresden, Gesellschaft fur Informatik.

Klebe, G. and T. Mietzner (1994). "A fast and efficient method to generate biologically relevant conformations." J Comput Aided Mol Des **8**(5): 583-606.

Kortemme, T., D. E. Kim, et al. (2004). "Computational alanine scanning of protein-protein interfaces." Sci STKE **2004**(219): pl2.

Kuhlman, B., G. Dantas, et al. (2003). "Design of a novel globular protein fold with atomic-level accuracy." Science **302**(5649): 1364-1368.

Kuhlman, B., J. W. O'Neill, et al. (2001). "Conversion of monomeric protein L to an obligate dimer by computational protein design." Proceedings of the National Academy of Sciences of the United States of America **98**(19): 10687-10691.

Lazaridis, T. and M. Karplus (1999). "Effective energy function for proteins in solution." Proteins **35**(2): 133-152.

Leach, A. R. (1994). "Ligand docking to proteins with discrete side-chain flexibility." J Mol Biol **235**(1): 345-356.

Meiler, J. and D. Baker (2006). "ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility." Proteins **65**(3): 538-548.

Meiler, J., W. Maier, et al. (2002). "Using neural networks for (13)c NMR chemical shift prediction-comparison with traditional methods." J Magn Reson **157**(2): 242-252.

Morozov, A. V. and T. Kortemme (2005). "Potential functions for hydrogen bonds in protein structure prediction and design." Adv Protein Chem **72**: 1-38.

Perola, E. and P. S. Charifson (2004). "Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding." J Med Chem **47**(10): 2499-2510.

Schueler-Furman, O., C. Wang, et al. (2005). "Progress in protein-protein docking: Atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility." Proteins-Structure Function and Bioinformatics **60**(2): 187-194.

Wang, J., R. M. Wolf, et al. (2004). "Development and testing of a general amber force field." J Comput Chem **25**(9): 1157-1174.

Wang, R., X. Fang, et al. (2004). "The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures." J Med Chem **47**(12): 2977-2980.

CHAPTER III

A PHYSICAL MODEL FOR PDZ-DOMAIN/PEPTIDE INTERACTIONS[1]

Introduction

Protein/peptide interactions play an important biological role in an array of cellular processes. One frequently used motif for such interactions is the well characterized PDZ (**P**SD-95, **D**iscs large, **Z**ona occludens 1) domain(Kim and Sheng 2004). Within *Homo sapiens*, *Drosophila melanogaster*, and *Caenorhabditis elegans* Schultz et al. have estimated the existence of 440 PDZ domains in 259 different proteins, 133 PDZ domains in 86 proteins, and 138 PDZ domains in 96 proteins, respectively(Schultz, Copley et al. 2000). PDZ domains perform critical roles in signaling cascades of bacteria, yeast, plants, and animals(Ponting 1997) by acting as intracellular scaffolding proteins(Pawson and Scott 1997; Kurschner and Yuzaki 1999). Pathogens disrupt host-signaling processes using linear peptide motifs to target PDZ binding sites.(Tonikian, Zhang et al. 2008) Developing inhibitors of these interactions is one avenue of therapeutic development.(Dev 2004) The wide-spread presence of the PDZ domain in nature and its integral role in numerous biological processes and diseases make it an ideal focus for studying the specificity of protein/peptide interactions.

PDZ domains bind peptides through strong backbone hydrogen bonds

PDZ domains are typically composed of 80-90 amino acids(Hung and Sheng 2002) and consist of a central bent six-stranded β-sheet surrounded by two α-helices. The peptide binding interface (Figure 1) lies at the edge of the β-sheet. The peptide binds in an extended, antiparallel conformation, using the unsatisfied hydrogen bonding capabilities of PDZ β-strand 2 (β2) to extend the β-sheet by one additional strand. The ligand also engages in side chain interactions with the second α-helix (α2) of the PDZ domain

[1] Published as Kristian Kaufmann, Nicole Shen, Laura Mizoue, Jens Meiler "A physical model of PDZ-domain/peptide interactions" *Journal of Molecular Modeling* **2011** 17 p.315

which lines the other side of the binding groove. The binding pocket contains a characteristic hydrophobic loop (β1:β2) which binds the peptide carboxy-terminus through the formation of three hydrogen bond interactions. Overall, the interface is characterized by strong backbone-backbone hydrogen bonding contacts within a hydrophobic environment(Nourry, Grant et al. 2003).

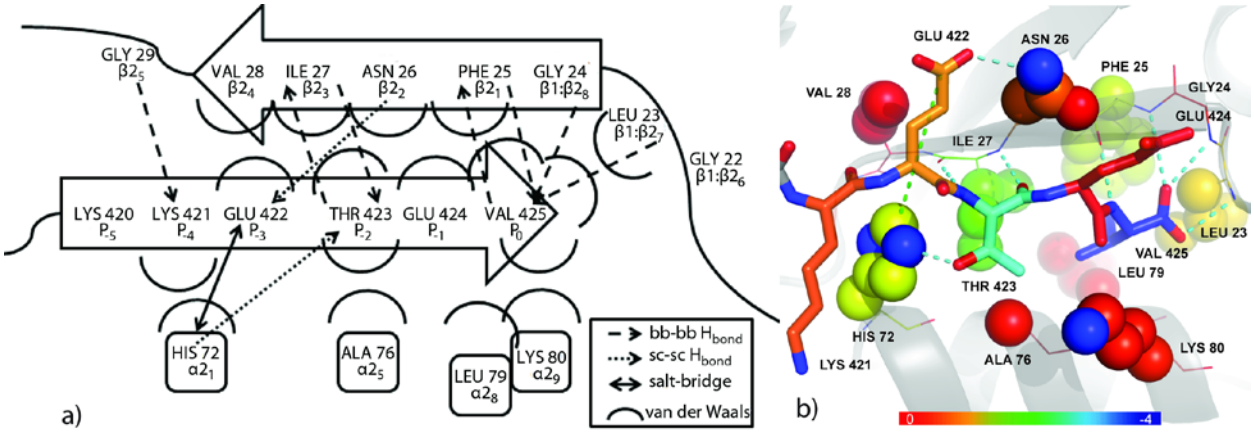


Figure 1 **Binding site of PSD-95 a class I domain.** PDZ domains strongly bind peptides through backbone-backbone hydrogen bonds; dashed lines in (a) indicate these interactions while dotted lines indicate protein-peptide side chain-side chain hydrogen bonds, and the arrow points from the hydrogen donating nitrogen to the oxygen acceptor. Color in (b) illustrates each residue's overall energetic involvement in binding the peptide, summing the weighted ROSETTA energy function of the individual attractive, solvation, repulsive, rotamer, pairwise, and hydrogen bonding energy contributions (ΔΔG values). In (b), the strong backbone-backbone hydrogen bonds (shown as blue dashed lines) between the PDZ3 protein and peptide residues V425 and T423 highlight the known PDZ3 protein preference for X-Thr/Ser-X-Val-COO(-) peptides [4]. This is further emphasized by E424 and K421's lack of strong backbone-backbone hydrogen bonds and orange and dark yellowish residue colors, which have overall 0.1 and -0.3 ΔΔG values, respectively. This is in contrast to T423 and V425's teal color and respective -3.0 and -3.1 ΔΔG values. Reprinted with permission from (Kaufmann, Shen et al. 2011)

In addition to hydrogen bonding interactions, important salt bridge (shown as green dashed lines) and van der Waals interactions (protein side chains contributing to van der Waals interactions shown with spheres) are involved in tightly binding the peptide between a beta sheet and alpha helix. H72 of the alpha helix seems to be the most involved in binding the peptide, forming a salt bridge with E422, interacting with T423 through side chain-side chain hydrogen bonding, and engaging in van der Waals interactions with K421. L79 and K80 similarly interact with V425 through van der Waals interactions, experiencing the attractive part of the Leonard Jones potential energy curve. This probably creates a favorable, hydrophobic environment for the non-polar, valine side chain and amplifies the strength of the hydrogen bonds formed between the peptide carboxy terminus and P25, G24, and L23.

PDZ domain specificity is governed by side chain interactions

Although the general binding mode of PDZ domains is the same, different proteins interact with different targets. While specificity has been studied extensively in the PDZ family, an unambiguous classification of the PDZ domain remains a challenge(Tonikian, Zhang et al.). Generally PDZ domains have been grouped into three classes (I, II, and III) depending on the characteristics of the β1:β2 loop(Tonikian, Zhang et al.) and position -2 (P_{-2}) of the ligand (see Figure 1). Class I domains have a G-L/Y-G-F β1:β2 loop that binds C-terminal peptide residues of sequence X-S/T-X-V/I/L (P_{-3}-P_0)(Nourry, Grant et al.). Additionally, the peptide hydroxyl group at P_{-2} makes an important hydrogen bonding contact with the histidine side chain of $\alpha 2_1$(Tonikian, Zhang et al.). Class II proteins have a similar β1:β2 loop sequence of X-L/V-G-F/I/L that binds peptide sequences having a hydrophobic amino acid at position P_{-2} (X-ø-X-ø)(Hung and Sheng). Class III domains are less widespread and have a G-L-G-F β1:β2 loop sequence that binds peptides having an acidic amino acid at P_{-2} (X-D/E-X-ø)(Doyle, Lee et al. 1996; Stricker, Christopherson et al. 1997).

PDZ class I, II, and III proteins and their peptides have variable sequence similarities (between 5% and 90%) but are structurally highly similar. Indeed, Stiffler et al. found only a weak correlation between sequence identity and PDZ domain specificity(Stiffler, Chen et al.). Instead, Stiffler developed a modified position specific scoring matrix based on the profiles of peptides which bind to a domain. Chen et al. later developed a method that incorporated structural information on protein/peptide residue pairs within close proximity of each other(Chen, Chang et al.). The model was capable of predicting PDZ domain specificity for multiple species from primary sequences and it was argued that including structural information via the protein/peptide residue position specific interaction matrix was sufficient to predict the specificity of PDZ domains.

PDZ domains display a diverse and finely tuned specificity profile

PDZ domain classification can be extended beyond the three naïve classes discussed here. Specificity within these classes depends upon other differences in the protein/peptide interface that result

in a diversified sequence profile. Tonikian et al. performed profiling of 91 point mutants of a model PDZ domain to create a specificity map. Using this map, 82 protein domains of the PDZ family were reclassified into 16 classes distinguished by specificity for peptide residues up to the P_{-6} position(Tonikian, Zhang et al.). While sequence-based analysis alone reveals diverse specificity profiles, the inclusion of structure-based information should provide a more general model for predicting PDZ specificity. Such a physical model would be a useful tool for PDZ domain classification, specificity prediction, and design.

The ROSETTA protein modeling software predicts specificity of protein/protein interfaces

In a series of experiments, Kortemme et al. demonstrated the power of the knowledge-based energy function of the modeling software ROSETTA to characterize and design protein/protein interfaces(Kortemme and Baker). A model for protein/protein binding was created using a data set of alanine mutants at protein/protein interfaces. The model was able to successfully predict the results of alanine scanning experiments on globular proteins (743 mutations) and 19 protein/protein interfaces (233 mutations) with low standard deviations of 0.8 kcal/mol and 1.1 kcal/mol, respectively(Kortemme and Baker). The model was applied to create new DNase-inhibitor protein pairs with altered specificities that functioned both *in vitro* and *in vivo*(Kortemme, Joachimiak et al. 2004). It was also used to fuse domains of two homing endonucleases creating a chimera that recognized a new DNA target and functioned as a highly specific artificial endonuclease(Chevalier, Kortemme et al. 2002).

While this model proved successful in modeling protein/protein interfaces, the derived parameterization is not optimal for protein/peptide interfaces as these are characterized by distinct features that require a tailored parameterization, such as smaller hydrophobic surface area and a greater dependence of hydrogen bonding interactions. Sood and Baker explored the use of ROSETTA to design elongated p53 and dystroglycan-based peptides that bind with increased affinity to Mdm2 oncoprotein and dystrophin, respectively. These studies included backbone flexibility and allowed side chain flexibility through repacking of a rotamer library but used the standard ROSETTA energy function with a

packing score derived from the change in solvent accessible surface area(Sood and Baker). Sood and Baker found that sampling of the backbone conformation improved recovery of sequence diversity in designed peptides and in cases where the algorithm fails, insufficient sampling of backbone degrees of freedom explains the error.

A ROSETTA parameterization tailored for PDZ domain/peptide interfaces

It is the objective of the present work to develop a model for predicting the specificity of PDZ domains using the protein structure prediction program ROSETTA. Saro et al. conducted isothermal titration calorimetry (ITC) measurements on a series of peptides binding the third PDZ domain (PDZ3) of postsynaptic density 95 protein (PSD-95), a class I domain. They recorded the thermodynamic properties ΔΔG, ΔΔH, and TΔS for a series of six-residue peptides of sequence (X-X-X-T-X-V), with different X amino acids influencing binding(Saro, Li et al. 2007). We parameterize ROSETTA to accurately predict these thermodynamic parameters.

Methods

Dataset for energy function parameterization

The dataset contains free energy (ΔΔG), enthalpy (ΔΔH), and entropy (TΔS) measurements for binding of 28 peptides to the PDZ3 domain of PSD-95 (Table 1)(Saro, Li et al.). The crystal structure of the PDZ3 domain of PSD-95 with the highest resolution (1.54 Å) from the PDB was used for structural modeling (PDBID 1TP5). The crystal structure was determined in complex with the peptide KKETWV.

Introduction of mutations and initial minimization of structural models

ROSETTADESIGN(Liu and Kuhlman 2006) protocols allow *in silico* mutation of amino acids. Briefly, the side chain of the amino acid in question is removed and replaced with a side chain of the

Table 1. Experimentally determined thermodynamic parameters by Saro et al. Binding energy changes do to point mutations on the native peptide, KKETEV were determined using ITC and represent the average of at least two independent experiments. Reprinted with permission from (Kaufmann, Shen et al. 2011)

Peptide	K_d (μM)	ΔG (kcal/mol)	ΔH (kcal/mol)	TΔS (kcal/mol)
KKETEV	1.9 ± 0.1	−7.8 ± 0.1	−6.2 ± 0.1	1.6 ± 0.1
KKETEA	91.0 ± 2.0	−5.5 ± 0.1	−4.6 ± 0.2	0.9 ± 0.2
KKETEL	7.9 ± 1.3	−7.0 ± 0.1	−4.1 ± 0.3	2.9 ± 0.2
KKETEI	7.7 ± 1.2	−7.0 ± 0.1	−4.3 ± 0.2	2.7 ± 0.1
KKETEM	21.0 ± 2.0	−6.4 ± 0.1	−6.8 ± 0.2	−0.4 ± 0.1
KKETEF	57.0 ± 2.0	−5.8 ± 0.1	−4.4 ± 0.4	1.4 ± 0.4
KKETET	105.0 ± 6.0	−5.4 ± 0.1	−5.9 ± 0.2	−0.5 ± 0.2
KKESEV	6.6 ± 0.9	−7.1 ± 0.1	−4.8 ± 0.1	2.3 ± 0.2
KKECEV	72.0 ± 7.0	−5.7 ± 0.1	−1.7 ± 0.1	4.0 ± 0.2
KKESEL	33.0 ± 2.0	−6.1 ± 0.1	−4.0 ± 0.1	2.1 ± 0.1
KKESEI	24.0 ± 6.0	−6.3 ± 0.2	−5.0 ± 0.2	1.3 ± 0.4
KKESEF	98.0 ± 16.0	−5.5 ± 0.1	−3.1 ± 0.1	2.4 ± 0.1
KKETGV	2.4 ± 0.0	−7.7 ± 0.1	−5.7 ± 0.2	2.0 ± 0.2
KKETAV	0.5 ± 0.1	−8.7± 0.1	−5.3 ± 0.4	3.4 ± 0.4
KKETVV	1.3 ± 0.2	−8.1 ± 0.1	−5.9 ± 0.1	2.2 ± 0.1
KKETLV	1.8 ± 0.3	−7.8 ± 0.1	−3.7 ± 0.4	4.1 ± 0.3
KKETPV	0.9 ± 0.2	−8.2 ± 0.1	−4.3 ± 0.1	3.9 ± 0.2
KKETWV	2.8 ± 0.4	−7.6 ± 0.1	−3.5 ± 0.2	4.1 ± 0.1
KKETDV	20.0 ± 2.0	−6.4 ± 0.1	−4.1 ± 0.3	2.3 ± 0.3
KKETKV	1.2 ± 0.0	−8.1 ± 0.1	−5.6 ± 0.6	2.5 ± 0.6
KKGTEV	80.0 ± 3.0	−5.6 ± 0.1	−2.7 ± 0.1	2.9 ± 0.1
KKATEV	21.0 ± 4.0	−6.4 ± 0.1	−2.4 ± 0.1	4.0 ± 0.2
KKQTEV	4.0 ± 0.0	−7.4 ± 0.1	−4.9 ± 0.3	2.5 ± 0.3
KKDTEV	85.0 ± 12.0	−5.6 ± 0.1	−3.9 ± 0.3	1.7 ± 0.2
KKKTEV	27.0 ± 4.0	−6.2 ± 0.1	−2.7 ± 0.3	3.5 ± 0.4
KKGTGV	273.0 ± 30.0	−4.9 ± 0.1	−2.6 ± 0.3	2.3 ± 0.2
KKATAV	8.3 ± 1.5	−6.9 ± 0.1	−3.0 ± 0.1	3.9 ± 0.2
YKETEV	1.2 ± 0.1	−8.1 ± 0.1	−6.9 ± 0.1	1.2 ± 0.2

target amino acid. The conformation of the introduced amino acid is chosen from a backbone-dependent rotamer library(Dunbrack and Karplus 1993) to minimize the ROSETTA energy function. First, the tryptophan at position P_{-1} of 1TP5 was reverted to a glutamate to match the base peptide KKETEV used

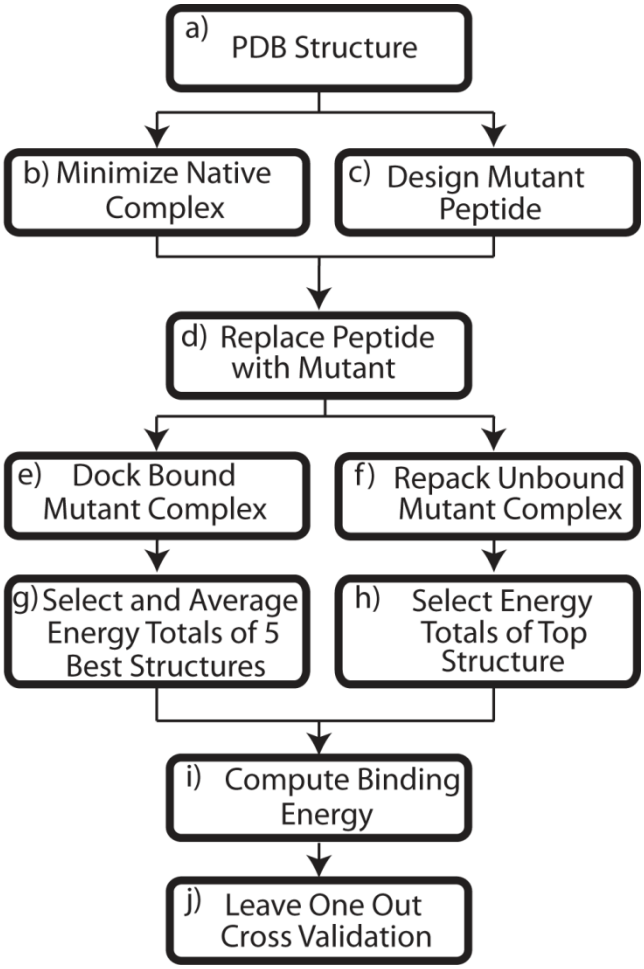


Figure 2. **Procedural flowchart**. Protein and peptide mutants of the PSD-95 PDZ3 domain, 1TP5, were each downloaded (a) and minimized (c) in Rosetta to remove initial clashes. Experimental mutants were reproduced in the computational analysis using design (b). Mutant proteins and peptides were combined with minimized peptides and proteins, respectively (d). These bound mutants were docked (e), yielding 100 decoys, of which the 5 structures with the lowest energy were selected and averaged over selected energy components (g), including attractive, solvation, repulsive, Dunbrack, residue pair electrostatics, hydrogen bonding, amino acid reference energies. The protein structure with the peptide far removed was repacked over selected residues determined from docking the structure (f). Because the 100 unbound structures had the same total Rosetta energy, a single structure's energy values were used rather than the average of five structures (h). The unbound energy values for each structure were subtracted from the corresponding, averaged bound energy values of the structure (i), producing $\Delta\Delta E$ values for each energy term, which were weighted and summed to produce the overall change in energy due to the protein binding the peptide, which was correlated to the experimentally determined binding energy, yielding the best correlation with particular weights (j). For the specificity analysis, 17 PDB files were downloaded (a) and separated into their fundamental protein, peptide components. All possible protein–peptide complexes were combined and minimized (c). Steps (d) through (i) as previously discussed were followed. Energy terms from (i) were weighted using weights determined from the mutational investigations (j). Reprinted with permission from (Kaufmann, Shen et al. 2011)

in the study by Saro et al.(Saro, Li et al.). Following this modification, the 28 PDZ domain/peptide complexes were built (Table 1, Figure 2b). All models underwent gradient minimization using ROSETTA to remove initial clashes (Figure 2c)(Bradley, Misura et al. 2005). The protocol involves eight rounds of gradient-based minimization of all torsional degrees of freedom which is alternated with side chain repacking using a rotamer library. The all-atom RMSD of the structure changed by 0.40 Å on average with a maximum of 0.51 Å observed for complex 6 containing the KKETEF mutant peptide.

ROSETTADOCK generation of structural models for protein/peptide complexes

To generate minimized models for energy evaluation, all bound structures underwent a small perturbation protocol applied to the transformational degrees of freedom in the protein/peptide complex using ROSETTADOCK (Figure 2e)(Gray, Moughon et al. 2003). This rigid body motion is complemented by a simultaneous optimization of side chain coordinates through a fast repacking protocol. The backbone coordinates of protein and peptide are held fixed in the process.

The protocol is setup in an iterative fashion. First a random small perturbation of up to 0.1 Å translation and up to 2° rotation is made to the rigid body degrees of freedom. Then the side chain conformations are allowed to change by substituting discrete rotamers from a library of conformations commonly seen in the PDB. If the substitution results in a lower total energy, ROSETTA keeps the new conformation of the protein. If the energy is higher, ROSETTA may still accept the substitution with a probability inversely proportional to the energy increase (Metropolis criterion). On average, around 50 of these iterations are completed in order to find the best combination of amino acid side chain conformations. The output model is the lowest energy complex observed throughout the entire trajectory. Lastly, a gradient-based minimization on the rigid body degrees of freedom moves the final model into the nearest local minimum in the ROSETTA energy landscape. A total of 100 bound models were generated for each complex. The 5 models with the lowest overall energy were selected for further analysis (Figure 2g).

Modeling apo structures in ROSETTA

The unbound (apo) structures were created by removing the peptide from the binding pocket and away from the protein by a distance sufficiently large to prevent any interaction (> 100 Å). The side chains, which were allowed to move during the docking protocol, were allowed to rearrange using repacking algorithms(LIU AND KUHLMAN 2006). One hundred models were generated for each of the mutants, and the total ROSETTA energy was used to select a single most favorable unbound conformation for each of the 28 complexes (Figure 2h).

Calculation and evaluation of binding free energy

The ROSETTA energy function contains six energy terms. Van der Waals energies are modeled using a Lennard-Jones 12-6 potential. The potential is split into an attractive (atr) and a repulsive (rep) component. ROSETTA introduces a solvation energy (sol) that imposes a penalty for polar atoms buried in the core of a protein accounting for the exposure preferences of polar and non-polar atoms(Lazaridis and Karplus 1999). Side chain conformational probabilities are reflected by an energy (dun) derived from rotamer probabilities(Dunbrack and Karplus 1993). Electrostatic interactions are mimicked by a knowledge-based pair-wise potential (pair) derived from statistics over the PDB. Hydrogen bonds (hbnd) are captured by an orientation dependent potential(Kortemme, Morozov et al. 2003). Note that in the past hydrogen bonds have been classified into three classes: long range-backbone-backbone (lr-bb), backbone-side chain (bb-sc), and side chain-side chain (sc-sc) hydrogen bonds(Kortemme and Baker 2002).

Within each structure, all residues were individually evaluated. To obtain the total energy of the model, the sum over all amino acids was computed and averaged over the top five bound structures (Figure 2g). For the unbound models energies from the single structure with lowest ROSETTA energy were directly used (Figure 2h). The binding free energy was computed for each of the above-mentioned terms $\Delta\Delta E_{binding}^{term}$ using:

$$\Delta\Delta E_{binding}^{term} = \frac{1}{5}\sum_{i=1}^{5} \Delta E_{i\ bound}^{term} - \Delta E_{unbound}^{term} \qquad (1)$$

$\Delta E_{i\ bound}^{term}$ is the ROSETTA energy one of the five complex models, respectively; $\Delta E_{unbound}^{term}$ is the ROSETTA energy of the single unbound model.

Multiple linear regression is used to parameterize an overall free energy function

To obtain an energy function optimized for the analysis of protein/peptide interactions (Figure2i), a multiple linear regression (MLR) analysis was used. Each of the $\Delta E_{binding}^{term}$ terms is affiliated with a weight w^{term}:

$$\begin{aligned}\Delta\Delta E_{binding} &= w^{atr}\Delta\Delta E_{binding}^{atr} + w^{rep}\Delta\Delta E_{binding}^{rep} + w^{sol}\Delta\Delta E_{binding}^{sol} \\ &+ w^{pair}\Delta\Delta E_{binding}^{pair} + w^{dun}\Delta\Delta E_{binding}^{dun} + w^{hbnd}\Delta\Delta E_{binding}^{hbnd} + bias\end{aligned} \qquad (2)$$

The bias is introduced to account for contributions to the binding free energy not represented in the ROSETTA energy function, such as the loss in entropy. The bias assumes that these contributions are constant, an obvious limitation of the present model.

The weights were determined by performing a Leave-One-Out (LOO) cross validation analysis. In a round-robin setup, 27 of the 28 mutants with known experimental binding affinities were used to determine an optimal weight set given these 27 data points. Afterwards, the binding free energy of the 28^{th} mutant was predicted and compared with the experiment to enter a correlation analysis. This experiment was repeated for all 28 mutants.

To determine whether an energy term contributes significantly to an optimal energy function for protein/peptide interfaces, energy terms were systematically removed. The subset of energy terms that resulted in the optimal correlation coefficient within the cross-validation experiment was used. The final

weight set reported consists of the average weights and standard deviations over of all 28 experiments. The protocol was implemented using the MATHEMATICA software package (Figure 2j).

Results

The physical model for protein/peptide interactions depends on van der Waals, solvation, and hydrogen bonding

The optimal weight set was determined by a Leave-One-Out (LOO) cross validation analysis as described in the Methods section. Of the six ROSETTA energy terms considered, only van der Waals attraction (atr), solvation (sol), and hydrogen bonding energies (hbnd), contributed to an energy function that optimally reproduced experimentally determined binding free energies:

$$\Delta\Delta E_{binding} = 0.47 \times \Delta\Delta E_{binding}^{atr} + 0.40 \times \Delta\Delta E_{binding}^{sol} + 1.34 \times \Delta\Delta E_{binding}^{hbnd} + 3.90 \qquad (3)$$

The correlation coefficient for the independent dataset is 0.66 (Figure 3).

Known characteristics of the PDZ binding domain are mirrored within the model

Figure 1 displays the per amino acid changes in free energy upon peptide binding for the PSD-95 PDZ3 in complex with the peptide KKETEV as determined by our model. Strong backbone-backbone hydrogen bonds between the class I domain and the peptide residues $V(P_0)$ and $T(P_{-2})$ agree with the anti-parallel β-strand binding motif of the PDZ domain which forms two backbone hydrogen bonds for every other amino acid. In this particular case, the C-terminal amino acid $V(P_0)$ engages in three hydrogen bonds. This alternative pattern is further highlighted by $E(P_{-1})$ and $E(P_{-3})$ which contribute only 0.1 and -0.3 kcal/mol to the binding free energy, respectively. In contrast, $V(P_0)$ and $T(P_{-2})$ contribute -3.1 and -3.0 kcal/mol, respectively.

In addition to hydrogen bonding interactions, important salt bridges and van der Waals interactions are involved in tightly binding the peptide. H($\alpha 2_1$) is the most important residue within α2 for binding the peptide as it forms a hydrogen bond with T(P_{-2}) through side chain hydrogen bonding, and

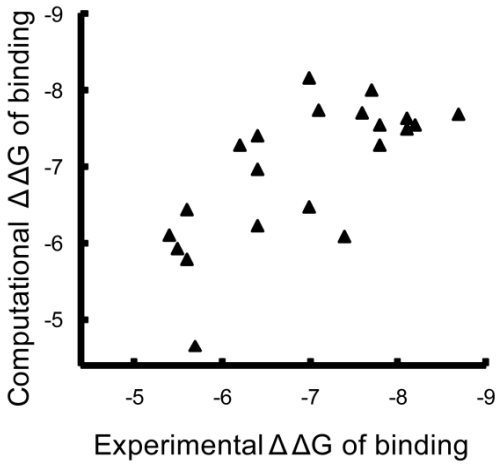


Figure 3. **Correlation of experimentally and computationally measured ΔΔG values over peptide mutants of the PDZ3 domain**. Experimentally calculated binding energies were determined using isothermal titration calorimetry (ITC). Computational binding energies were determined after a leave-one-out (LOO) cross validation analysis of the summed calculation of the various combinations of the weighted changes in the attractive (atr), solvation (sol), repulsive (rep), residue pair electrostatics (pair), dunbrack (dun), and hydrogen bonding (hb and hb_bb) energy terms over all the residues due to the protein binding the peptide. The individual binding term energy changes were calculated using Eq. 1. Different combinations of these terms were weighted and totaled according to Eq. 3. By calculating these weights when each mutant was left out and then applying the determined function, the LOO cross validation analysis measures the weighted energy function's predicting power. The predicted binding energies (y-axis) for the various peptide mutants correlated nicely with the experimentally observed binding energies (x-axis), having an r value equal to 0.66. The overall computational function $f(x) = 0.47 * E_{atr} + 0.40 * E_{sol} + 1.34E_{hbnd} + 3.90$ indicates the importance of the attractive, solvation, and side chain hydrogen bonding energy terms. Reprinted with permission from (Kaufmann, Shen et al. 2011)

engages in van der Waals interactions with K(P_{-4}). L($\alpha 2_8$) and K($\alpha 2_9$) interact with V(P_{-0}) through van der Waals attractive interactions. This creates a favorable, hydrophobic environment for the non-polar valine side chain and amplifies the strength of the hydrogen bonds formed between the peptide carboxyl terminus and F($\beta 2_1$), G($\beta 1$:$\beta 2_8$), and L($\beta 1$:$\beta 2_7$).

Enthalpic and entropic contributions to the binding free energy map to different components of the ROSETTA energy function

The investigation was extended to other thermodynamic characteristics of protein/peptide binding including enthalpy and entropy (Table 2). The independent correlation observed for the binding enthalpy $\Delta\Delta H_{binding}$ is with 0.60 only slightly reduced from the value observed for the Gibbs binding free energy (0.66, Figure 3). In contrast, when correlating with respect to experimentally measured entropy changes the independent correlation drops to 0.17.

Table 2. **Weighted energy terms over thermodynamic binding properties**. Reprinted with permission from (Kaufmann, Shen et al. 2011)

	correlation	atr	rep	sol	hbnd	rotamer	pair
ΔG	0.66	0.47±0.04	0.00±0.00	0.40±0.06	1.34±0.07	-	-
ΔH	0.60	-	-	-	2.25±0.12	-	1.28±0.16
ΔS	0.17	-	-	-	0.74±0.10	0.36±0.03	-

Specificity prediction for twelve PDZ domains with available crystal structures

For a specificity analysis, a set of twelve PDZ protein/peptide complexes with available crystal structures was used (Table 3). All experimentally determined structures with resolutions of 2.30 Å or better were considered (PDBID 1BE9(Doyle, Lee et al.), 1N7F(Im, Park et al. 2003), 1OBY(Kang, Cooper et al. ; Kang, Cooper et al.), 1RZX(Peterson, Penkert et al. 2004), 1TP3, 1TP5, 1V1T(Grembecka, Cierpicki et al.), 1W9E(Grembecka, Cierpicki et al.), 1W9O(Grembecka, Cierpicki et al.), 1W9Q(von Ossowski, Oksanen et al. ; von Ossowski, Tossavainen et al.), 2I04(Zhang, Dasgupta et al. 2007), 2QT5(Long, Wei et al. 2008)). Structures used in the specificity analysis were initially separated into their protein and peptide components. Peptides were truncated to include five carboxy-terminal residues. All possible combinations between PDZ domains and peptides were created yielding a total of 144 complexes. Each complex was refined using the protocol described above (Figure 2).

The binding energies for each complex were then computed using the PDZ optimized weight set. The heat map in Figure 4a shows that the PDZ optimized weight set captures specificity within each PDZ

class. The complexes group into two blocks reflecting the two classes of PDZ domains. Figure 4b shows the receiver operating characteristics (ROC) curve

where a complex is regarded as a true complex if both peptide and protein come from the same PDZ class. The area under the curve is 78%, 28% better than a random predictor.

Discussion

Energy Function Weights from LOO Analysis are Stable

The deviations from a perfect correlation are attributed to imperfection in the ROSETTA energy function which is simplified to only contain pair-wise decomposable energetic terms(Kuhlman and Baker 2000). The small standard deviations observed for the individual weights (Table 4) demonstrate internal consistency as the analysis of all 28 complexes yielded very similar weight sets.

Table 3. **Specificity data set.** Bold letters indicate amino acids that were used for specificity prediction (P0-P-4). Reprinted with permission from (Kaufmann, Shen et al. 2011) Reprinted with permission from (Kaufmann, Shen et al. 2011)

PDB ID	PDZ Class	Peptide Sequence	Resolution (Å)	Domain
1TP5	1	K**KETWV**	1.54	PSD95-3
1BE9	1	**KQTSV**	1.82	PSD95-3
1TP3	1	K**KETPV**	1.99	PSD95-3
1RZX	1	V**KESLV**	2.10	Par-6B
2I04	1	RR**RETQV**	2.15	MAGI1-1
2QT5	1	NNLQ**DGTEV**	2.30	GRIP1-12
1N7F	2	ATV**RTYSC**	1.80	GRIP1-6
1W9E	2	**TNEFYF**	1.56	Syntenin-2
1W9Q	2	**TNEFAF**	1.70	Syntenin-2
1V1T	2	**TNEYKV**	1.80	Syntenin-2
1W9O	2	**TNEYYV**	2.25	Syntenin-2
1OBY	2	**TNEFYA**	1.85	Syntenin-2

computational model for protein/peptide binding. The weight set is optimized to predict the binding free energies PDZ domains. In particular, the hydrogen bonding weight is substantially increased relative to other weights. This result can be explained in part by the backbone hydrogen bonds between peptide and PDZ domain. These hydrogen bonds contribute significantly to the stability of the PDZ-domain/peptide interface. However, as these hydrogen bonds are present in all PDZ domain/peptide

complexes, they do not govern specificity but contribute an approximately equal amount to all interfaces studied.

Our results indicate that a high weight on side chain hydrogen bonds is particularly important for accurate specificity prediction. Interestingly, a holistic weighting with a single hydrogen bonding weight gave the best results. This is in contrast to the earlier reported optimal weight set for protein/protein

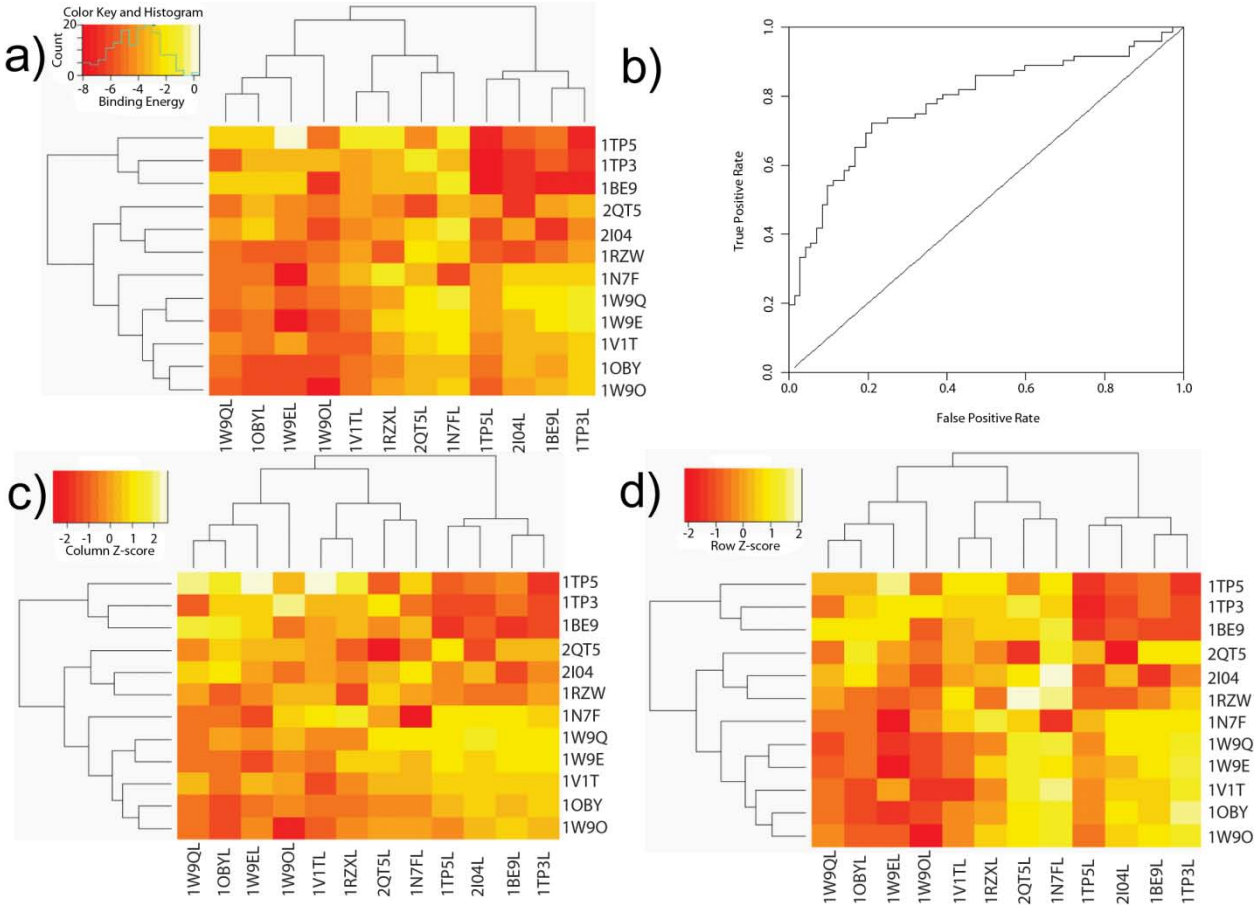


Figure 4 **Specificity based on computed binding energy.** Each column shows the computed binding energies of the peptide from a structure to each of the PDZ domains. Each row displays the computed binding energies of each peptide to a given PDZ structure. a Heat map with colors scaled according to the raw computed binding energy. b Receiver operating characteristics (ROC) curve for PDZ classification. c Heat map of the binding energies colored by the z-score computed to the peptide group (i.e., within column). dColoring scaled according to the z-score computed by the PDZ structure group (within row). Reprinted with permission from (Kaufmann, Shen et al. 2011)

interfaces where hydrogen bonds contributed differently depending on the level of solvent exposure(Kortemme, Kim et al. 2004). Beyond this aspect, changes in the weight set are small.

Energy Function Components Capture Enthalpic but not Entropic Contributions

The weighted energy terms vary significantly when correlated to the different thermodynamic binding properties. Enthalpy is best predicted from hydrogen bonding (hbnd) and electrostatic (pair) interactions; entropy correlates best with a combination of hydrogen bonding (hbnd) and rotamer probability. Overall we expected that ROSETTA derived energy terms correlate best with binding free energies. Their knowledge-based character can be well aligned with the definition of free energy in statistical thermodynamics. Hence, every one of the ROSETTA energy terms contains both entropic contributions and enthalpic contributions. However, the term can be dominated by one of the two if it is better represented by the simplified two-body equations used within ROSETTA. Our results demonstrate that entropic contributions are least accurately reflected and prevent ROSETTA from predicting to higher degrees of accuracy.

Computed Binding Energies Correctly Classify PDZ Domains

The correlation of the binding energies within each class is apparent, but the computed binding energies across all PDZ complexes do not accurately rank the complexes. However when holding either the protein or the peptide constant, the binding energies display a better correlation with specificity as seen in Figures 4c and 4d. This may reflect the need to sample a greater conformational space. In fact, Sood and Baker found a better recovery of peptides sequence profiles upon introducing backbone flexibility into their design protocol(Sood and Baker).

Summary

This study presents a physical model for PDZ domain/peptide interactions. Parameterization of the ROSETTA energy function was achieved by fitting a linear model to experimentally determined binding free energies for 28 PDZ domain/peptide complexes. The energy function is dominated by van

der Waals attractive, solvation, and hydrogen bonding interactions. It reproduces well-known determinants of PDZ domain/peptide interactions such as an alternating pattern of backbone hydrogen bonding to the second strand of the PDZ domain (β2) and side chain interactions with the second helix (α2). While the Gibbs free energy correlates well with experimental values (R=0.66), correlation of enthalpy (R=0.60) and particularly entropy (R=0.17) is reduced. This reduction is attributed to the knowledge-based nature of ROSETTA energy functions which aligns well with the definition of free energy in statistical mechanics. The resulting weight set was able to classify a given PDZ/peptide complex 28% better than a random predictor.

Table 4. **Weight set optimized for protein/peptide interfaces compared to a weight set optimized for protein/protein interfaces (Kortemme and Baker 2002)** and to the default weight set. sc=side chain bb=backbone. atr=attractive component of van der Waals energy, rep=repulsive componen component of van der Waals energy, sol=implicit solvation energy, hbnd=hydrogen bonding, rotamer=knowledge based energy for conformation for a side chain. Reprinted with permission from (Kaufmann, Shen et al. 2011)

	atr	rep	sol	hbnd			rotamer
Protein/Peptide	0.47±0.04	0.00	0.40±0.06			1.34±0.07	0.00
Protein/Protein	0.44	0.07	0.32	sc-bb		0.49	0.28
				sc-sc	exposed	0.16	
				sc-sc	intermediate	0.44	
				sc-sc	buried	0.94	
ROSETTA default	0.42	0.10	0.37			0.24	0.06

References

Bradley, P., K. M. Misura, et al. (2005). "Toward high-resolution de novo structure prediction for small proteins." Science **309**(5742): 1868-1871.

Chen, J. R., B. H. Chang, et al. (2008). "Predicting PDZ domain-peptide interactions from primary sequences." Nat Biotechnol **26**(9): 1041-1045.

Chevalier, B. S., T. Kortemme, et al. (2002). "Design, activity, and structure of a highly specific artificial endonuclease." Mol Cell **10**(4): 895-905.

Dev, K. K. (2004). "Making protein interactions druggable: targeting PDZ domains." Nat Rev Drug Discov **3**(12): 1047-1056.

Doyle, D. A., A. Lee, et al. (1996). "Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ." Cell **85**(7): 1067-1076.

Dunbrack, R. L., Jr. and M. Karplus (1993). "Backbone-dependent rotamer library for proteins. Application to side-chain prediction." J Mol Biol **230**(2): 543-574.

Gray, J. J., S. Moughon, et al. (2003). "Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations." J Mol Biol **331**(1): 281-299.

Grembecka, J., T. Cierpicki, et al. (2006). "The binding of the PDZ tandem of syntenin to target proteins." Biochemistry **45**(11): 3674-3683.

Hung, A. Y. and M. Sheng (2002). "PDZ domains: structural modules for protein complex assembly." J Biol Chem **277**(8): 5699-5702.

Im, Y. J., S. H. Park, et al. (2003). "Crystal structure of GRIP1 PDZ6-peptide complex reveals the structural basis for class II PDZ target recognition and PDZ domain-mediated multimerization." J Biol Chem **278**(10): 8501-8507.

Kang, B. S., D. R. Cooper, et al. (2003). "Molecular roots of degenerate specificity in syntenin's PDZ2 domain: reassessment of the PDZ recognition paradigm." Structure (Camb) **11**(7): 845-853.

Kang, B. S., D. R. Cooper, et al. (2003). "PDZ tandem of human syntenin: crystal structure and functional properties." Structure **11**(4): 459-468.

Kaufmann, K., N. Shen, et al. (2011). "A physical model for PDZ-domain/peptide interactions." J Mol Model **17**(2): 315-324.

Kim, E. and M. Sheng (2004). "PDZ domain proteins of synapses." Nat Rev Neurosci **5**(10): 771-781.

Kortemme, T. and D. Baker (2002). "A simple physical model for binding energy hot spots in protein-protein complexes." Proc Natl Acad Sci U S A **99**(22): 14116-14121.

Kortemme, T. and D. Baker (2002). "A simple physical model for binding energy hot spots in protein-protein complexes." Proceedings of the National Academy of Sciences of the United States of America **99**: 14116-14121.

Kortemme, T., L. A. Joachimiak, et al. (2004). "Computational redesign of protein-protein interaction specificity." Nat Struct Mol Biol **11**(4): 371-379.

Kortemme, T., D. E. Kim, et al. (2004). "Computational alanine scanning of protein-protein interfaces." Sci STKE **2004**(219): pl2.

Kortemme, T., A. V. Morozov, et al. (2003). "An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes." J Mol Biol **326**(4): 1239-1259.

Kuhlman, B. and D. Baker (2000). "Native protein sequences are close to optimal for their structures." Proc. Natl. Acad. Sci. U. S. A. **97**(19): 10383-10388.

Kurschner, C. and M. Yuzaki (1999). "Neuronal interleukin-16 (NIL-16): a dual function PDZ domain protein." J Neurosci **19**(18): 7770-7780.

Lazaridis, T. and M. Karplus (1999). "Effective Energy Function for Proteins in Solution." PROTEINS: Structure, Function, and Genetics **35**: 133–152.

Liu, Y. and B. Kuhlman (2006). "RosettaDesign server for protein design." Nucleic Acids Res **34**(Web Server issue): W235-238.

Long, J., Z. Wei, et al. (2008). "Supramodular nature of GRIP1 revealed by the structure of its PDZ12 tandem in complex with the carboxyl tail of Fras1." J Mol Biol **375**(5): 1457-1468.

Nourry, C., S. G. Grant, et al. (2003). "PDZ domain proteins: plug and play!" Sci STKE **2003**(179): RE7.

Pawson, T. and J. D. Scott (1997). "Signaling through scaffold, anchoring, and adaptor proteins." Science **278**(5346): 2075-2080.

Peterson, F. C., R. R. Penkert, et al. (2004). "Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB-PDZ transition." Mol Cell **13**(5): 665-676.

Ponting, C. P. (1997). "Evidence for PDZ domains in bacteria, yeast, and plants." Protein Sci **6**(2): 464-468.

Saro, D., T. Li, et al. (2007). "A thermodynamic ligand binding study of the third PDZ domain (PDZ3) from the mammalian neuronal protein PSD-95." Biochemistry **46**(21): 6340-6352.

Schultz, J., R. R. Copley, et al. (2000). "SMART: a web-based tool for the study of genetically mobile domains." Nucleic Acids Res **28**(1): 231-234.

Sood, V. D. and D. Baker (2006). "Recapitulation and design of protein binding peptide structures and sequences." J Mol Biol **357**(3): 917-927.

Stiffler, M. A., J. R. Chen, et al. (2007). "PDZ domain binding selectivity is optimized across the mouse proteome." Science **317**(5836): 364-369.

Stricker, N. L., K. S. Christopherson, et al. (1997). "PDZ domain of neuronal nitric oxide synthase recognizes novel C-terminal peptide sequences." Nat Biotechnol **15**(4): 336-342.

Tonikian, R., Y. Zhang, et al. (2008). "A specificity map for the PDZ domain family." PLoS Biol **6**(9): e239.

von Ossowski, I., E. Oksanen, et al. (2006). "Crystal structure of the second PDZ domain of SAP97 in complex with a GluR-A C-terminal peptide." FEBS J **273**(22): 5219-5229.

von Ossowski, L., H. Tossavainen, et al. (2006). "Peptide binding and NMR analysis of the interaction between SAP97 PDZ2 and GluR-A: potential involvement of a disulfide bond." Biochemistry **45**(17): 5567-5575.

Zhang, Y., J. Dasgupta, et al. (2007). "Structures of a human papillomavirus (HPV) E6 polypeptide bound to MAGUK proteins: mechanisms of targeting tumor suppressors by a high-risk HPV oncoprotein." J Virol **81**(7): 3618-3626.

CHAPTER IV

ROSETTALIGAND: SMALL MOLECULE DOCKING INTO COMPARATIVE MODELS.

Introduction

Structure based comparative models of proteins in complex with small molecules advance science by creating hypotheses that can be tested experimentally. The process of modeling a protein in complex with a small molecule, often termed small molecule docking, has a long history stretching back more than 25 years(Kuntz, Blaney et al. 1982). There are two basic problems in small molecule docking, searching the space of possible arrangements of the atoms at the small molecule protein interface (sampling) and evaluating free energy for that configuration (scoring). Sampling requires accounting for both the position of the small molecule relative to the protein as well as the internal flexibility of the small molecule and the protein. This easily leads to thousands of degrees of freedom. In order to detect the correct configuration the method must accurately rank the free energy of the correct arrangement relative to alternative arrangements. Other authors have provided guides to small molecule docking and evaluation of the current best practices and software for this purpose(Taylor, Jewsbury et al. 2002; Rester 2006; Sousa, Fernandes et al. 2006; Warren, Andrews et al. 2006; Davis, Raha et al. 2009).

Until recently, small molecule docking programs have been validated mostly on experimental structures available for the protein rather than models of the protein(Verdonk, Mortenson et al. 2008). However, for the vast majority of protein sequences no experimental structure is available. For this reason, we and others turn our attention to evaluating small molecule docking into models of proteins(McGovern and Shoichet 2003; Kairys, Fernandes et al. 2006; Brylinski and Skolnick 2008; Fan, Irwin et al. 2009). Naively, one would expect comparative models to perform better than their templates in small molecule docking as sequence deviations between template and target protein have been

rectified. In particular, recent results from the Critical Assessment of Structure Prediction Techniques (CASP) indicate that comparative modeling methods can add information to models that is not present in templates. However, Kairys et al. found that docking into the experimental templates performed as well as docking into the homology models based on templates with sequence identities ranging from 30% to 90% and heavy atom RMSDs in the binding site ranging from 1-4 Å(Kairys, Fernandes et al. 2006). On the other hand, McGovern and Shoichet found that docking into a set of comparative models covering ten enzymes from ModBase is more successful than docking in just the experimental structure of the protein with no ligand bound (apo) in a virtual screening scenario(McGovern and Shoichet 2003). Ferrara and Jacoby found that in a virtual screen for insulin-like growth factor 1 receptor kinase ligands, homology models varied in enrichment capacity from random to as good as the experimental structure of the protein determined by X-ray crystallography(Ferrara and Jacoby 2007).

Although, stunning progress has been made in *de novo* protein structure prediction over the past years(Das, Qian et al. 2007), comparative models at atomic detail accuracy have been reported regularly at recent CASP experiments(Bradley, Misura et al. 2005; Das, Qian et al. 2007; Zhang 2009). Hence, comparative modeling remains the method of choice if a template with a structure similar to the target protein can be identified in the protein data bank (PDB)(Saxena, Wong et al. 2009). Template-based modeling focuses on modifying the known structure to reflect the sequence of the protein of interest. Generally, good quality models result if the sequence identity between the target and template protein is better than 30%(Chothia and Lesk 1986). However, results from the latest CASP experiment indicate that template detection methods are able to identify suitable templates with even lower sequence homology(Raman, Vernon et al. 2009; Tress, Ezkurdia et al. 2009). Only recently has high accuracy refinement of comparative models led to improvements upon the starting models(MacCallum, Hua et al. 2009). Comparative models accurate at atomic detail also open the possibility of obtaining highly accurate models of protein-small molecule complexes from comparative models of proteins.

In the following experiments, we establish a baseline for the performance of ROSETTALIGAND on small molecule docking into comparative models. We show that ROSETTALIGAND can correctly identify

binding modes in two sets of models. The first test composed of nine models submitted during the CASP experiment allows us to verify the method on comparative models built in a blind experiment by the best comparative modeling techniques available. We also construct a test set of 21 complexes from seven proteins with models from at least two different templates. This test set expands the test to more diverse chemotypes, examines the effect of template choice, and explores the limits of sampling and scoring methods used in ROSETTALIGAND.

Results and Discussion

Using two sets of comparative models we show ROSETTALIGAND is capable of sampling and identifying native-like complexes. In the first set, models for nine targets from the 8th CASP experiment which contained organic ligands were used to assess the ability of ROSETTALIGAND to dock small molecules into comparative models constructed blindly by a variety of best-practices comparative modeling protocols. The second set of seven proteins with in complex with three different ligands each expands the chemotype diversity of ligands and assesses the impact of the choice of the template as comparative models were constructed from two to five templates each.

Two factors are critical to the success of a small molecule docking study. First the energy function must guide the modeling method towards native-like complexes. Second, the sampling methods must allow the method to produce native-like models.

Native complexes occupy minima in the ROSETTALIGAND Energy Function

To assess the energy function we compare native-like complexes found by minimizing the ligand complex while constraining the complex in a native-like binding mode. We ran minimization on complexes from using both the comparative models and the crystal structures. Native-like complexes from the crystal structure score as good as or better than non-native complexes in 23 of the 30 cases

tested. In five of the seven failures a non-native conformation scored better by less than 2 ROSETTA Energy Units (REUs). In seven cases native-like conformations score at least 5 REUs better than any non-native conformation. In an additional 8 cases the native conformation scored better than 2 REUs over the best scoring non-native model.

Native-like complexes from comparative models score better than non-native binding modes in 18 of 23 cases. Three cases show the native-like binding having a REU score more than 5 less than the best scoring non-native binding mode, while a further seven score more than 2 REU less the best scoring non-native binding mode.

These trends indicate that the energy function is able to discriminate native-like complexes. Although the errors structure of the protein structure inherent in comparative models does decrease the apparent depth of the native binding mode energy well.

ROSETTALIGAND Samples Native-like Conformations

Having looked at the energy functions ability to discriminate native-like complexes we now turn our attention to the sampling problem. The first concern is whether Rosetta samples native-like conformations in a standard docking run. Indeed ROSETTALIGAND samples native-like conformations (ligand root mean squared deviation L-RMSD < 2 Å) in all cases (see Table 1). In the worst case (3D8B) conformations with L-RMSD < 2 Å were sampled but included in the cluster with a L-RMSD of 2.04 Å.

Energy Function Discriminates Native-like Complexes in Models Built Using Comparative Models.

Furthermore sampling in ROSETTALIGAND is dense enough that the sampled native-like complexes fall into native energy well. For 11 of 30 cases the best energy model had a native-like binding mode with binding energy funnels like those seen in Figure 1A. A further 10 cases saw one of the top 10 best energy models contain a native-like binding-mode and binding energy funnels like that in Figure 1B. A similar though slightly lower success rate is seen if one does not pool the models from all templates (Tables 2-8). On a by template basis 16 of 69 or 23% of the cases the best energy model was in a native-

like binding mode, while 36 of 69 or 52% had native-like binding modes among the ten best energy clusters.

Table 1. Minimum I-RMSD of models and L-RMSD of native-like binding modes. I-RMSD is calculated over all heavy atoms within 5 Å of the small molecule in X-ray crystal structure. L-RMSD are calculated over heavy atoms in the small molecule. Cluster Rank is the rank order of the cluster from lowest binding energy to highest binding energy. Error describes the spatial orientation to the native binding mode if the rank 1 cluster is non-native. I=inverted binding mode, W=wrong conformation of ligand, T=Translation, R=rotation, C=cofactors present in native which may influence binding mode.

	Native Energy		I-RMSD		Best Non Native				Model			
	Cry. Str.	Model	Min.	Ave.	Energy	Rank	RMSD	Error	Energy	Rank	L-RMSD	I-RMSD
3D8B	-17.31	-14.94	0.74	1.28	-15.00	1	8.06	TR 5A	-11.38	27	2.14	2.09
3DLZ	**-17.46**	**-17.62**	**2.96**	**7.78**	**-13.78**	**4**	**6.21**		**-17.62**	**1**	**1.46**	**26.38**
3DA0	-14.16	-15.00	0.34	1.56	-17.88	1	2.84	TR 1.5A	-13.51	35	1.84	2.66
3DA1	**-28.65**	**-33.72**	**0.69**	**1.43**	**-24.59**	**2**	**14.52**		**-31.37**	**1**	**0.52**	**1.63**
3DKP	-12.3	-13.73	0.76	1.49	-16.09	1	8.59	TR 6A	-10.73	56	1.30	1.63
3DLS	**-12.13**	**-12.80**	**1.32**	**2.18**	**-13.92**	**2**	**4.95**		**-14.22**	**1**	**1.95**	**2.02**
3DLC	*-30.06*	*-24.83*	*0.90*	*4.05*	*-22.33*	*1*	*3.27*	*W*	*-20.58*	*3*	*1.40*	*2.71*
3DME	*-40.93*	*-29.51*	*2.43*	*3.20*	*-23.81*	*1*	*2.61*	*W*	*-21.54*	*4*	*1.04*	*3.4*
3DOU	**-27.92**	**-24.27**	**0.43**	**1.86**	**-17.87**	**3**	**2.66**		**-19.88**	**1**	**0.79**	**2.64**
1Y1M	**-13.72**	**-13.66**	**1.43**	**2.53**	**-9.68**	**14**	**2.87**		**-13.66**	**1**	**0.67**	**1.49**
1PB9	*-11.86*	*-9.93*	*1.09*	*2.20*	*-8.51*	*1*	*2.88*	*TR 3A*	*-7.91*	*3*	*0.86*	*2.65*
1PBQ	*-15.07*	*-15.78*	*1.82*	*2.50*	*-17.57*	*1*	*4.17*	*TR 2A*	*-15.78*	*8*	*1.83*	*3.48*
2QWB	*-10.04*	*-14.77*	*1.33*	*2.36*	*-11.78*	*1*	*4.22*	*I*	*-11.58*	*3*	*1.51*	*2.52*
2QWD	*-14.95*	*-14.32*	*1.31*	*2.30*	*-13.56*	*1*	*3.44*	*TR 1A*	*-12.99*	*2*	*1.46*	*1.72*
2QWE	-15.17	-17.80	1.24	2.22	-13.41	1	6.19	TR 0.5A	-9.93	38	1.41	1.47
1FD0	-29.62	-17.97	2.54	3.41	-21.27	1	4.45	TR 3A	-17.52	16	1.38	3.21
1FCX	*-26.15*	*-18.13*	*2.56*	*3.39*	*-20.57*	*1*	*5.98*	*T 5A*	*-18.13*	*10*	*1.28*	*3.27*
1FCZ	-26.14	-17.73	2.54	3.38	-20.86	1	3.19	W	-16.62	26	1.60	3.04
1VFN	**-11.53**	**-11.87**	**2.39**	**3.01**	**-11.87**	**1**	**6.95**	**T 5A**	**-11.87**	**1**	**1.10**	**2.39**
1B8O	*-16.18*	*-14.73*	*2.39*	*3.26*	*-15.58*	*1*	*4.6*	*T 3A C*	*-14.56*	*3*	*1.20*	*2.07*
1V48	**-19.30**	**-19.20**	**1.64**	**2.38**	**-15.21**	**3**	**3.43**		**-16.67**	**1**	**1.76**	**1.78**
2FAI	-15.38	-14.21	2.12	3.68	-14.08	1	4.08	TR 3A	-12.83	19	1.37	2.12
2AYR	-23.20	-19.12	2.68	4.33	-21.07	1	8.29	IW	-17.26	58	1.81	2.74
2B1V	*-14.83*	*-13.43*	*1.97*	*3.58*	*-14.88*	*1*	*3.31*	*TR 2A*	*-13.27*	*4*	*1.90*	*2.35*
1NJA	**-14.48**	**-14.5**	**1.75**	**3.08**	**-13.33**	**2**	**6.94**		**-14.5**	**1**	**0.66**	**3.52**
1NJE	**-14.90**	**-16.25**	**1.82**	**3.03**	**-12.32**	**3**	**6.01**		**-16.25**	**1**	**1.66**	**2.35**
1TSY	-16.62	-13.81	1.67	3.26	-11.81	1	6.35	TR 5A	-7.89	77	1.91	2.11
1O3P	**-15.88**	**-19.24**	**1.70**	**2.45**	**-17.93**	**2**	**4.84**		**-19.24**	**1**	**0.88**	**1.97**
1F5K	**-10.74**	**-11.11**	**1.62**	**2.23**	**-10.18**	**2**	**5.59**		**-11.11**	**1**	**0.54**	**1.62**
1SQA	*-19.01*	*-19.50*	*2.51*	*3.09*	*-19.69*	*1*	*2.68*	*WC*	*-17.16*	*5*	*1.66*	*2.84*

Examining the Top Ranked Binding Mode indicate needed improvements in interface refinement, ligand conformations, and modeling of cofactors

Encouragingly, in 11 of 30 cases the lowest energy binding mode from a docking run is the correct solution as is seen in Figure 2a for the target 1O3P. However, for 12 cases the top ranked binding mode shows translation/rotation errors similar to those seen for 2FAI and 1B8O (Figures 2b and 2c), this along with the noted decrease native energy well depth in comparative models indicates improvements are needed in the interface refinement process. In five cases the rank 1 binding mode maintains many of

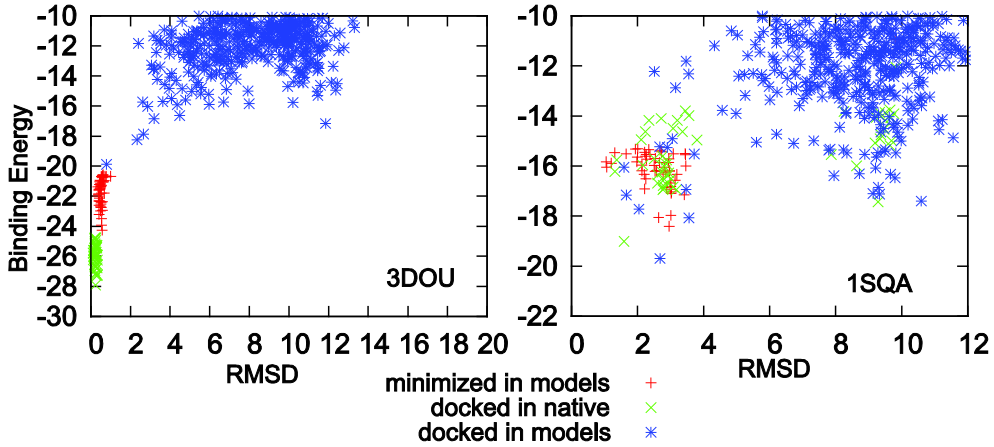


Figure 1 L-RMSD v. Binding Energy Plots. 3DOU shows comparative models can fall into the energy well of the native binding mode. However this is not the case for all proteins as is shown in the plot for 1SQA. Plots for all 30 targets can be found in the Appendix. With the exception of 1SQA, 1FD0, 1FCX, 1FCZ, and 3DLS each of the other 25 targets show overlap between the native binding mode minimized in the native PDB structure and the native binding mode minimized in the comparative models while remaining under the 2 Å radius. This indicates that the scoring function recognizes native-like binding modes as at least local minima, and 11 cases aas global minima.

the correct interactions, but adopts a non-native conformation for the ligand as is seen in Figure 2d for 1FCZ. Further improvements in ligand conformational sampling or the energetics of ligand conformations may decrease these errors (Kaufmann, Glab et al. 2008; Davis and Baker 2009). However, improvements in the accuracy of the protein side chain and backbone will also be necessary as can be seen from the degradation of the energy well depth when comparing for crystal structures to comparative models (see Table 1). In two cases (1SQA and 1B8O) the crystallized structure contained cofactors in the binding site

which overlap with rank 1 binding mode (see Figure 2c of 1B8O). Modeling the structures with these cofactors present might change the ranking of the binding modes.

Careful Template Selection Improves Docking

If the template contains a ligand and especially if that ligand is an analog of the ligand to be docked, ROSETTALIGAND has a higher success rate compared to the unliganded templates (see Figure 3: ~70 % for templates with ligand that has a similar chemotype to the target ligand, ~50 % for templates with a non-similar ligand bound, ~20% for templates without ligands). Tables 2-8 lists the occupancy state of the binding site for each template-target combination for the second set of comparative models. Twenty of the 23 templates contained small molecules in the binding site. In 43 of the resulting 60 template-target combinations, the ligand in the template was an analog of the small molecule in the target.

Table 2 N-methyl-D-Asparatate Receptor 1 ligand docking broken down by template. I-RMSD is calculated over all heavy atoms within 5 Å of the small molecule in X-ray crystal structure. L-RMSD are calculated over heavy atoms in the small molecule. Cluster Rank is the rank order of the cluster from lowest binding energy to highest binding energy. I=Template contains identical ligand, A=Template contains analogous ligand, PA=Template contains partial analog, L=Template contains , "-"= Template does not contain a ligand

Targets	Templates	Seq.ID./ I-Seq.ID.	Crystal Structure		I-RMSD		Rank 1		Model Native Binding Mode			
			Energy	Ligand	Min	Avg.	Energy	L-RMSD	Energy	Rank	L-RMSD	I-RMSD
1Y1M	2RCA	33%/33%		A	1.57	2.35			-11.51	1	1.93	1.68
	2A5S	37%/58%		A	1.59	2.06			-9.91	1	1.88	1.67
	2I0B	36%/25%		A	1.58	2.46			-11.33	1	1.91	1.75
	2RC7	36%/33%		A	1.43	2.31			-13.66	1	0.67	1.49
	1P1N	34%/42%		A	1.57	2.83			-12.36	1	1.87	1.67
	Combined		-13.72		1.43	2.53			-13.66	1	0.67	1.49
1PB9	2RCA	33%/33%		A	1.09	1.87	-8.51	2.88	-7.20	3	0.64	2.2
	2A5S	37%/58%		A	1.10	1.88			-7.18	1	0.28	1.14
	2I0B	36%/25%		A	1.09	1.89	-7.3	6.64	-6.67	2	1.91	1.16
	2RC7	36%/33%		A	1.21	2.15			-7.91	1	0.86	2.65
	1P1N	34%/42%		A	1.24	2.55	-7.61	4.81	-7.38	2	0.46	1.35
	Combined		-11.86		1.09	2.20	-8.51	2.88	-7.91	3	0.86	2.65
1PBQ	2RCA	33%/33%		A	1.94	2.47	-14.92	4.46	-9.68	22	1.88	2.04
	2A5S	37%/58%		A	1.69	2.31	-14.46	4.65				
	2I0B	36%/25%		A	1.90	2.49	-14.93	3.59	-13.33	2	0.88	2.69
	2RC7	36%/33%		A	1.82	2.49	-15.24	3.65	-12.37	7	1.73	2.71
	1P1N	34%/42%		A	1.94	2.51	-17.57	4.17	-13.21	16	1.53	2.84
	Combined		-15.07		1.82	2.50	-17.57	4.17	-15.78	8	1.83	3.48

Table 3 Neuramidase. ligand docking broken down by template. I-RMSD is calculated over all heavy atoms within 5 Å of the small molecule in X-ray crystal structure. L-RMSD are calculated over heavy atoms in the small molecule. Cluster Rank is the rank order of the cluster from lowest binding energy to highest binding energy. I=Template contains identical ligand, A=Template contains analogous ligand, PA=Template contains partial analog, L=Template contains , "-"= Template does not contain a ligand

Targets	Templates	Seq.ID./ I-Seq.ID.	Crystal Structure		I-RMSD		Rank 1		Model Native Binding Mode			
			Energy	Ligand	Min	Avg.	Energy	L-RMSD	Energy	Rank	L-RMSD	I-RMSD
2QWB	2HTY	50%/89%			2.06	2.75	-11.78	4.22	-7.65	31	1.76	2.93
	1V0Z	69%/95%			1.33	2.03	-10.76	5.11	-8.65	17	1.26	1.54
	1INF	36%/95%		A	1.67	2.30			-11.58	1	1.51	2.52
	Combined		-10.04		1.33	2.36	-11.78	4.22	-11.58	3	1.51	2.52
2QWD	2HTY	50%/89%			1.89	2.66	-12.96	5.34				
	1V0Z	69%/95%			1.31	2.01	-13.56	3.44	-12.99	2	1.46	1.72
	1INF	36%/95%		A	1.67	2.24	-12.22	5.59	-11.88	2	1.62	1.81
	Combined		-14.95		1.31	2.30	-13.56	3.44	-12.99	2	1.46	1.72
2QWE	2HTY	50%/89%			1.82	2.58	-12.98	2.42				
	1V0Z	69%/95%			1.24	1.88	-13.41	6.19	-9.93	19	1.41	1.47
	1INF	36%/95%		A	1.52	2.19	-11.21	5.57				
	Combined		-15.17		1.24	2.22	-13.41	6.19	-9.93	38	1.41	1.47

Table 4 Retanoic acid Receptor Gamma ligand docking broken down by template. I-RMSD is calculated over all heavy atoms within 5 Å of the small molecule in X-ray crystal structure. L-RMSD are calculated over heavy atoms in the small molecule. Cluster Rank is the rank order of the cluster from lowest binding energy to highest binding energy. I=Template contains identical ligand, A=Template contains analogous ligand, PA=Template contains partial analog, L=Template contains , "-"= Template does not contain a ligand

Targets	Templates	Seq.ID./ I-Seq.ID.	Crystal Structure		I-RMSD		Rank 1		Model Native Binding Mode			
			Energy	Ligand	Min	Avg.	Energy	L-RMSD	Energy	Rank	L-RMSD	I-RMSD
1FD0	2ACL	36%/28%		L	2.59	3.49	-18.46	2.32				
	1NQ0	37%/20%		L	2.54	3.28	-19.97	2.90	-17.52	5	1.38	3.21
	1PQ6	38%/28%		L	2.71	3.34	-21.27	4.45				
	2H77	39%/24%		L	3.22	3.60	-17.40	7.75				
	Combined		-29.62		2.54	3.41	-21.27	4.45	-17.52	16	1.38	3.21
1FCX	2ACL	36%/28%		L	2.49	3.46	-20.57	5.98	-18.07	3	1.21	2.89
	1NQ0	37%/20%		L	2.56	3.3	-20.48	3.21	-18.13	7	1.28	3.27
	1PQ6	38%/28%		L	2.88	3.35	-20.43	4.17				
	2H77	39%/24%		L	3.12	3.51	-14.01	6.42				
	Combined		-26.15		2.56	3.39	-20.57	5.98	-18.13	10	1.28	3.27
1FCZ	2ACL	36%/28%		L	2.59	3.49	-19.09	3.68	-16.62	7	1.60	3.04
	1NQ0	37%/20%		L	2.54	3.31	-20.86	3.19	-15.92	14	1.97	3.00
	1PQ6	38%/28%		L	2.71	3.31	-20.08	5.60				
	2H77	39%/24%		L	3.22	3.52	-14.18	7.11				
	Combined		-26.14		2.54	3.38	-20.86	3.19	-16.62	26	1.60	3.04

Table 5 Purine Nucleoside Phosphorylase ligand docking broken down by template. I-RMSD is calculated over all heavy atoms within 5 Å of the small molecule in X-ray crystal structure. L-RMSD are calculated over heavy atoms in the small molecule. Cluster Rank is the rank order of the cluster from lowest binding energy to highest binding energy. I=Template contains identical ligand, A=Template contains analogous ligand, PA=Template contains partial analog, L=Template contains , "-"= Template does not contain a ligand

Targets	Templates	Seq.ID./ I-Seq.ID.	Crystal Structure		I-RMSD		Rank 1		Model Native Binding Mode			
			Energy	Ligand	Min	Avg.	Energy	L-RMSD	Energy	Rank	L-RMSD	I-RMSD
1VFN	2P4S	56%/91%		A	2.42	3.11	-11.84	3.99	-9.40	63	0.94	
	1G2O	38%/91%		A	2.39	2.71			-11.87	1	1.10	
	1TCU	49%/82%		L	2.77	3.20			-11.34	1	0.78	
	Combined		-11.53		2.39	3.01	-11.87	6.95	-11.87	1	1.10	
1B8O	2P4S	56%/91%		I	1.78	2.73	-15.58	4.60	-11.12	16	1.30	
	1G2O	38%/91%		I	1.77	2.28			-14.56	1	1.20	2.07
	1TCU	49%/82%		L	2.26	2.79	-13.27	6.89	-12.14	6	0.87	
	Combined		-16.18		1.77	2.60	-15.58	4.60	-14.56	3	1.20	2.07
1V48	2P4S	56%/91%		PA	1.64	2.34	-14.84	2.59	-13.54	6	1.54	
	1G2O	38%/91%		PA	1.74	2.15			-16.67	1	1.76	
	1TCU	49%/82%		PA	2.12	2.65	-15.36	2.07				
	Combined		-19.30		1.64	2.38			-16.67	1	1.76	

Table 6 Estrogen Receptor ligand docking broken down by template. I-RMSD is calculated over all heavy atoms within 5 Å of the small molecule in X-ray crystal structure. L-RMSD are calculated over heavy atoms in the small molecule. Cluster Rank is the rank order of the cluster from lowest binding energy to highest binding energy. I=Template contains identical ligand, A=Template contains analogous ligand, PA=Template contains partial analog, L=Template contains , "-"= Template does not contain a ligand

Targets	Templates	Seq.ID./ I-Seq.ID.	Crystal Structure		I-RMSD		Rank 1		Model Native Binding Mode			
			Energy	Ligand	Min	Avg.	Energy	L-RMSD	Energy	Rank	L-RMSD	I-RMSD
2FAI	1QKN	60%/89%		PA	2.70	3.26	-14.08	4.08	-11.96	19	1.66	2.9
	1S9P	36%/42%		PA	2.12	3.94	-13.80	2.98	-12.83	7	1.37	2.12
	3CS8	46%/11%		L	2.79	3.83	-13.50	5.16	-11.08	28	1.66	3.33
	Combined		-15.38		2.12	3.68	-14.08	4.08	-12.83	19	1.37	2.12
2AYR	1QKN	60%/89%		A	3.40	3.91	-21.07	8.29				
	1S9P	36%/42%		PA	2.68	4.17	-19.75	7.46	-17.26	22	1.81	2.74
	3CS8	46%/11%		L	3.90	4.9	-20.55	5.48				
	Combined		-23.2		2.68	4.33	-21.07	8.29	-17.26	58	1.81	2.74
2B1V	1QKN	60%/89%		PA	1.97	2.76	-14.18	4.02	-13.27	4	1.9	2.35
	1S9P	36%/42%		PA	2.07	3.97	-14.88	3.31	-13.25	2	1.51	2.14
	3CS8	46%/11%		L	2.97	4.02	-12.72	2.36				
	Combined		-14.83		1.97	3.58	-14.88	3.31	-13.27	4	1.9	2.35

Table 7 Thymidylate Synthase ligand docking broken down by template. I-RMSD is calculated over all heavy atoms within 5 Å of the small molecule in X-ray crystal structure. L-RMSD are calculated over heavy atoms in the small molecule. Cluster Rank is the rank order of the cluster from lowest binding energy to highest binding energy. I=Template contains identical ligand, A=Template contains analogous ligand, PA=Template contains partial analog, L=Template contains , "-"= Template does not contain a ligand

Targets	Templates	Seq.ID./ I-Seq.ID.	Crystal Structure		I-RMSD		Rank 1		Model Native Binding Mode			
			Energy	Ligand	Min	Avg.	Energy	L-RMSD	Energy	Rank	L-RMSD	I-RMSD
1NJA	1QZF	43%/93%		A	2.24	3.28			-14.50	1	0.66	3.52
	1J3I	42%/93%		A	2.04	2.92	-13.30	6.21				
	1KZJ	50%/86%		A	1.75	3.05	-13.33	6.94	-7.21	103	1.76	3.06
	Combined		-14.48		1.75	3.08			-14.50	1	0.66	3.52
1NJE	1QZF	43%/93%		A	2.77	3.43	-9.75	7.68	-7.03	27	1.06	3.74
	1J3I	42%/93%		A	1.82	2.66	-11.08	7.15	-9.94	4	1.23	1.82
	1KZJ	50%/86%		A	1.89	3.01			-16.25	1	1.66	2.35
	Combined		-14.9		1.82	3.03			-16.25	1	1.66	2.35
1TSY	1QZF	43%/93%		I	2.75	3.51	-10.94	8.09				
	1J3I	42%/93%		I	1.67	2.68	-11.81	6.35	-7.89	11	1.91	2.11
	1KZJ	50%/86%		I	2.59	3.60	-10.81	6.32	-7.28	83	1.24	3.83
	Combined		-16.62		1.67	3.26	-11.81	6.35	-7.89	77	1.91	2.11

Table 8 Uridine Kinase Type Plasminogen Activator Ligand Docking broken down by template. I-RMSD is calculated over all heavy atoms within 5 Å of the small molecule in X-ray crystal structure. L-RMSD are calculated over heavy atoms in the small molecule. Cluster Rank is the rank order of the cluster from lowest binding energy to highest binding energy. I=Template contains identical ligand, A=Template contains analogous ligand, PA=Template contains partial analog, L=Template contains , "-"= Template does not contain a ligand

Targets	Templates	Seq.ID./ I-Seq.ID.	Crystal Structure		I-RMSD		Rank 1		Model Native Binding Mode			
			Energy	Ligand	Min	Avg.	Energy	L-RMSD	Energy	Rank	L-RMSD	I-RMSD
1O3P	1RTF	45%/70%		A	1.70	2.05			-19.24	1	0.88	1.97
	1YBW	40%/65%			2.48	2.84	-17.93	4.84				
	Combined		-15.88		1.70	2.45			-19.24	1	0.88	1.97
1F5K	1RTF	45%/70%		I	1.62	1.82			-11.11	1	0.54	1.62
	1YBW	40%/65%			2.36	2.64	-10.01	3.93				
	Combined		-10.74		1.62	2.23			-11.11	1	0.54	1.62
1SQA	1RTF	45%/70%		A	2.51	2.77	-19.69	2.68	-17.16	4	1.66	2.84
	1YBW	40%/65%			2.76	3.40	-17.40	10.59				
	Combined		-19.01		2.51	3.09	-19.69	2.68	-17.16	5	1.66	2.84

Five templates contained the same small molecule as the target complex. Figure 4 shows the target structures overlaid on template structures. Ligands in template binding sites decrease the probability of backbone deviations occluding the native binding mode as is seen for 1SQA on 1YBW in Figure 4a. Templates with analogs in the binding site help pre-form the binding pocket, thus increasing

the probability of finding native-like binding modes. Figure 4b shows almost perfect agreement between the binding mode in the target 1PB9 and for the ligand found in template 2RC7. However, target 2B1V on template 1QKN (Figure 4c) and target 2QWE on template 1INF (Figure 4d) show that not all functional groups transfer directly between complexes.

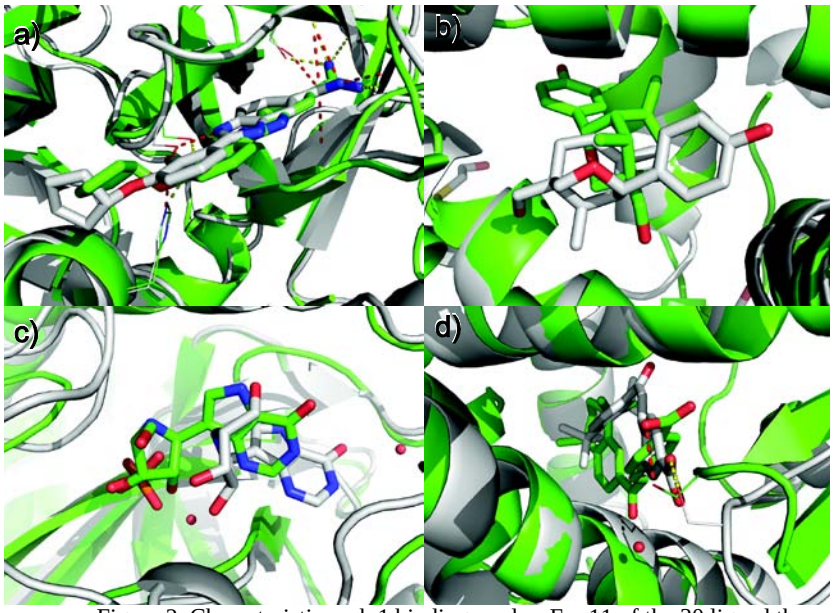


Figure 2. Characteristic rank 1 binding modes. For 11 of the 30 ligand the rank1 ligand is a native-like binding mode as for a) target1O3P. For 12 of the remaining case the rank 1 binding mode is rotated and/or translated compared to the native binding mode as is seen for b) target 2FAI and c) target 1B8O. Two targets (1SQA and 1B8O) contain cofactors in the target crystal structure which overlap with the rank 1 binding mode as seen for 1B8O in c). These targets might improve should cofactors be included. For the 6 remaining targets the rank 1 binding mode maintains many of the correct contacts but adopts a non-native conformation as seen in d) for target 1FCZ

Heuristics for template selection

Sequence identity of templates does not correlate with docking success. For the 21 small molecules docked into ROSETTA generated models, neither the overall sequence identities nor the sequence identities in the binding site serve as a good predictor of success. Figure 5 shows no discernible trend in success over the range of sequence identities from less than 30% to over 90%.

Docking into multiple templates can improve results. In the case of 4 of the 7 proteins, docking into a comparative model of a single template is sufficient for success (see Table 2-8). However a second

template is needed to identify the binding mode of all three small molecules for the three Neuramidase complexes (2QWE, 2QWD, 2QWB). Fan et al. also noted that multiple models improved results in the

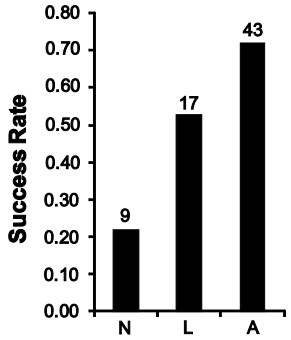


Figure 3 Docking success rate increases with occupancy of template binding site. N indicates no ligand present in binding site. L indicates small molecule with no chemical similiarity present in binding site. A indicates presence of an chemical analog in binding site. The number at top of column indicates total number of cases for each bin

context of virtual screening(Fan, Irwin et al. 2009).

Given that templates perform differently and that it is not always possible to use all available templates, template selection heuristics would be useful. The first naïve approach would be to take the template with the highest sequence similarity. However the results in this benchmark indicate that templates with sequence identities as low as 30 % perform as well as or better than templates of 60%. Additionally, sequence identity in the binding site does not correlate with success. One noticeable trend is that holo structures performed better than apo structures, particularly holo structures containing ligands similar to the target ligands. Templates containing ligands with function groups similar to the target ligand should be given preference. For the greatest gain, any chemical analogs found in the templates should be used to guide the modeling process as shown by Brylinski and Skolnick(Brylinski and Skolnick 2009).

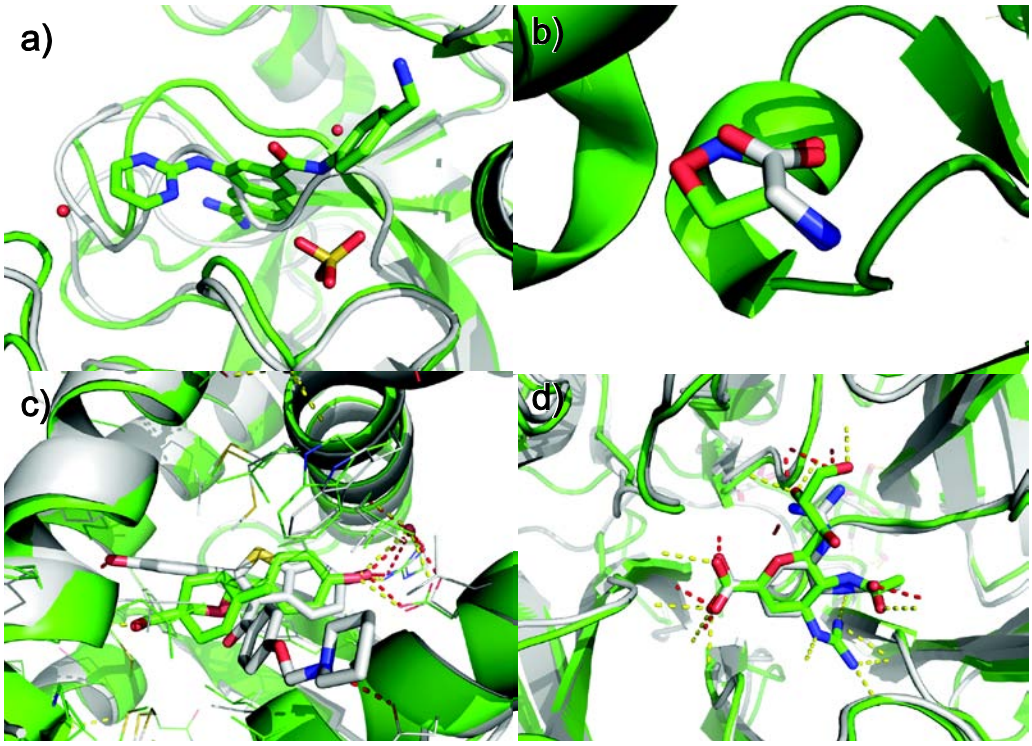


Figure 4 Careful template selection can improve docking results. a) Backbone differences in templates can preclude success by a docking program. Superimposition of 1SQA (green) on 1YBW (grey). Success in this case would require active remodeling of the binding site by the docking program. Selection of template with ligands similar to the target ligand pre-forms the binding site providing conserved binding motifs as seen for b) 1PB9 (green) on 2RC7 (grey), c) 2B1V (green) on 1QKN (grey), and d) 2QWE (green) on 1INF (grey). However a one to one correspondence is not guaranteed as in seen in both c) in which the phenyl group points to a different part of the pocket in the template as opposed to the target and d) in which the guanidinium head group occupies a different pocket in the binding site.

Conclusion

Modeling of small molecule protein binding sites is difficult. Davis et al. recently found that ROSETTALIGAND and other prominent docking software failed to generate a native-like binding mode on at least one protein 70% of the time. Thus docking to comparative models may seem like a fool's errand due the lack of accuracy in comparative models. However, the recent advances in comparative modeling techniques has improved the quality of comparative models(Misura, Chivian et al. 2006; Raman, Vernon et al. 2009; Zhang 2009). Indeed in some cases the comparative models have sub-angstrom accuracy at the protein small molecule interface. The fact that the native binding mode is sampled in all 30 cases is

encouraging. Further more, ROSETTALIGAND ranks the native-like binding modes from docking runs in the top 10 binding modes for 21 of 30 cases. The progressive degradation of the apparent native well energy well depth form crystal structures to comparative models indicate that docking to comparative models in ROSETTA would benefit greatly from improvements in sampling. Although some improves my require better gross protein modeling improvements the difference between energy well depth of minimized native binding modes in comparative models to docking runs into comparative models indicate that significant gains can be found in docking refinement improvements.

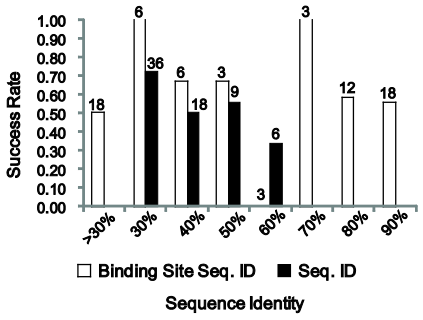


Figure 5 Docking success rate over a range of sequence identities. The number of targets in each sequence identity bin is found at the top of each bar.. Success rate does not correlate with sequence identity. Neither binding site sequence identity (residues with in 5 Å of the ligand) nor overall sequence identity show trends in docking success over the range of sequence identities covered.

The experiments described here-in point to three improvements that could be made in sampling. First, ligands in templates could be used to guide placement of functional groups. This could be accomplished by either using the ligand placements in templates as starting positions in a small perturbation Monte Carlo minimization protocol or by implementing constraints during the docking simulation. Second, cofactors could be included in the docking process. At present, due to limitations in the code, cofactors would remain fixed in place while docking occurs, although an iterative cycle could be employed to serial dock both ligands and cofactors. Upcoming changes to ROSETTALIGAND will allow simultaneous docking of multiple ligands. Third, the homology modeling process could be altered to

include ligands. Specifically, ligands found in templates could be retained in the structures during loop modeling and structure refinement. This may result in more accurate comparative models and thus allow ROSETTALIGAND to sample closer to the native binding mode.

Improvement of the scoring function is more complex. One glaring deficiency is the lack of a ligand internal energy. The internal energy of the ligand may prove particularly important for refinement of small molecule protein complexes. Examination of solvation and charge effects have also been observed to present problems for small molecule complexes(Nannemann, Kaufmann et al. 2010). Finally, using a dataset like similar to this one could be used to train an artificial neural network or support vector machine classifer to pick native-like binding modes in a manner similar to that employed by London and Schuler-Furman(London and Schueler-Furman 2007).

Docking to comparative models with ROSETTALIGAND can sample and identify native-like binding modes. Careful selection of templates and integration of biochemical data will increase the accuracy of the predicted interface. However, the native-like binding mode will be one of as many as 20 binding modes. Biochemical information will be required to prioritize the binding modes found by ROSETTALIGAND. Once a candidate binding mode is selected it should be carefully characterized using a series of mutations. The results of mutagenesis experiments and other biochemical experiments should then be integrated in the models.

Methods

The focus of this work was to assess the ability of ROSETTALIGAND to identify the binding mode of small molecules using comparative models. Two sources were chosen for the comparative models. The first set of comparative models was taken from the CASP experiment. The second source of models was prepared for a subset of systems in the PDBBind, a database of small molecule-protein structures with associated binding energies.

Preparation of CASP models

The models for the nine CASP targets containing organic ligands were downloaded from the CASP website (http://www.predictioncenter.org along with the corresponding crystal structures from the Protein Databank(Bernstein, Koetzle et al. 1977; Berman, Henrick et al. 2003) (www.rcsb.org). The top model submitted by each group was selected. Each model was structurally aligned to the crystal structure. First, a global alignment was performed using the PyMOL align command. This was followed by aligning all residues within 8 Å of the ligand. The ligand in the crystal structures was then transferred to the models. The procedure results in an optimally placed ligand and represents a theoretical limit for the quality models.

Building of Comparative Models

Building of comparative models requires the selection of a structural template, alignment of the sequence onto the structural template, followed by any refinement necessary to account for changes in the structure from the new sequence. In this study, potential structural templates were identified using a blast search of sequences in the PDB. At least one template was chosen for each 10% sequence identity bin ranging from 30% - 80%, if available. A multiple sequence alignment for the selected templates was constructed using the MUSTANG structural alignment program(Konagurthu, Whisstock et al. 2006). The sequence alignment used to construct the comparative model was then created using ClustalW's sequence to profile alignment options(Larkin, Blackshields et al. 2007). The sequence alignment was then mapped onto the template structures.

Any gaps or insertions were remodeled using the kinematic loop closure protocol in ROSETTA. The kinematic loop closure protocol has been previously described(Mandell, Coutsias et al. 2009). Briefly, each loop is chosen in a random order in a Metropolis Monte Carlo protocol. Six dihedral angle torsions are chosen from the residues in the loop. The remaining torsions are randomly sampled from Ramachandran probabilities of each amino-acid. The six torsions are solved analytically. The kinematic loop closure protocol is run several hundred times over varying sections of the loop with the new

conformation of the loop being accepted when it fulfills the Metropolis criteria. Once each of the loops has been built, a minimization of the protein structure is performed by iteratively performing Metropolis Monte Carlo repacking of the side chain conformations of the protein, followed by gradient minimization.

The loop_model application is called with the following options

```
-in
 -path
  #location of rosetta_database
  -database rosetta_database_dir/
-loops
 #allow ramachandran biased sampling of backbone torsions
 #under kinematic loop closer protocol
 -nonpivot_torsion_sampling
 #Allow expansion of loop definitions if necessary
 -random_grow_loops_by 4
 #build loops in random order
 -random_order
 #perform full atom kinematic loop closure
 -refine refine_kic
 #use reduced number of cycles for faster modeling
 -fast
 #perform iterative side chain repacking and gradient minimization
 #to refine models
 -relax seqrelax
 #start with input pdb
 -input_pdb target.pdb
 #input file defining residues in loops
 -loop_file target.loops
-in
 -file
  #Starting pdb file has side chains defined
  -fullatom
-out
 -file
  #output models should have side chains
```

```
   -fullatom
-out
 #output models should be placed in models/
 -path modelsI/
 #prefix for output model pdb file names
 -prefix model_name_prefix
 #Number of models to create
 -nstruct 100
#Starting pdb file has side chains defined
-fa_input
#Use expanded chi1 and chi2 expanded rotamer sets
-ex1 -ex2
```

After building approximately 4000 models using the above protocol, all models within 50 REU of the lowest energy models were selected for clustering. The Bio3D R package was used to align and calculate the RMSD matrix between the structures(Grant, Rodrigues et al. 2006) . The k-means clustering algorithm in R was used to find clusters of approximately 25 members. This clustering approach is meant to pick a maximally diverse subset of the structures for docking.

Docking to Models

The docking protocol has been described in detail previously(Davis and Baker 2009). The protocol begins by randomly placing the ligand center of mass in a 10 Å box. The protocol selects 1 from up to 1000 different orientations and conformations based on shape complementarities in a low resolution van der Waals grid. The side chains and ligand in the binding site then undergo six rounds of Metropolis Monte Carlo optimization of side chain and ligand conformations. Finally, a gradient minimization yields the structure of the complex. This protocol is repeated to generate 1000 models. The models were then ordered by the interface delta score. The interface delta score is the total energy of the complex with the ligand bound minus the total energy of the complex with the ligand separated from the binding site (i.e. 500 Å from the protein).

The ligand_dock application is called with the following options

-in

-path

#location of rosetta_database

-database *rosetta_database/*

-file

#model pdb of protein with ligand placed in the binding site

-s *protein_ligand.pdb*

#Parameter file defining topology of ligand

-extra_res_fa *ligand.params*

#Native PDB of complex used as reference in RMSD calculations

-native *native_complex.pdb*

-out

#Number of models to be built

-nstruct 1000

-file

#output file for models in atom tree difference format

-silent models_compressed.out

-docking

-ligand

#allow ligand torsion angles to change during minimization

-minimize_ligand

#place a harmonic constraint with force constant of 10 on ligand torsions

#during minimization

-harmonic_torsions 10

#allow backbone phi, psi to change during minimization

-minimize_backbone

#place harmonic constraint C_α to start position during minimization

#with harmonic constraint of 0.3

-harmonic_Calphas 0.3

#During docking process use soften van der Waals potential

-soft_rep

#Use electrostatics potential between ligand and protein as in ROSETTA 2.3

-old_estat

#Use 6 cycles of monte carlo minimization with side chain repacking

-protocol abbrev2

```
    #attempt up to 1000 different rotations in initial placement
    -improve_orientation 1000
    #Use a uniform distribution to translate ligand up to five angstroms
    #from starting position
    -uniform_trans 5
```

The models are extracted from the atom tree difference silent using the extract_atomtree_diffs application with the following options

```
-in
  -file
    #Input file in atom tree difference format
    -s models_compressed.out
    #tags for models to extract from input file
    -tags  model_0222
    # Parameter file defining topology of ligand
    -extra_res_fa ligand.params
  -path
    #location of rosetta_database
    -database rosetta_database_dir/
-out
  #directory in which to place models
  -path models_dir/
```

Identifying Binding Modes

In docking studies with non-native models, the discriminatory power of the ROSETTA docking energy function is decreased (London and Schueler-Furman 2007; Kaufmann, Dawson et al. 2009). The global minimum of the native energy funnel may be inaccessible because of limited sampling in either the protein or the ligand. As a result, local minima cannot be distinguished from the global minimum based on the ROSETTALIGAND energy function alone. Here, we use a clustering approach to identify the binding modes and then rank the binding modes by interface delta score. Clustering allows one to avoid considering models that contain the same binding mode.

To cluster models in best 5% by energy, the RMSD between the ligand heavy atoms for all pairs of models is computed. This matrix of RMSDs is then clustered in R(R Development Core Team) using complete linkage with a height of 3.00 RMSD. All clusters at this height are used regardless of the number of models in each cluster. The docking energy landscape is very rough. Consequently, the native binding mode may rarely be sampled. Thus, penalizing small clusters is counter-productive. Following hierarchical clustering in R the clusters are ranked by the energy of the best energy model in the cluster.

References

Berman, H., K. Henrick, et al. (2003). "Announcing the worldwide Protein Data Bank." Nat Struct Biol **10**(12): 980.

Bernstein, F. C., T. F. Koetzle, et al. (1977). "The Protein Data Bank. A computer-based archival file for macromolecular structures." Eur J Biochem **80**(2): 319-324.

Bradley, P., K. M. S. Misura, et al. (2005). "Toward high-resolution de novo structure prediction for small proteins." Science (New York, N.Y.) **309**: 1868-1871.

Brylinski, M. and J. Skolnick (2008). "Q-Dock: Low-resolution flexible ligand docking with pocket-specific threading restraints." Journal of computational chemistry **29**: 1574-1588.

Brylinski, M. and J. Skolnick (2009). "Q-Dock(LHM): Low-resolution refinement for ligand comparative modeling." Journal of computational chemistry.

Chothia, C. and a. M. Lesk (1986). "The relation between the divergence of sequence and structure in proteins." The EMBO journal **5**: 823-826.

Das, R., B. Qian, et al. (2007). "Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home." Proteins **69 Suppl 8**: 118-128.

Davis, I. W. and D. Baker (2009). "RosettaLigand docking with full ligand and receptor flexibility." Journal of molecular biology **385**: 381-392.

Davis, I. W., K. Raha, et al. (2009). "Blind docking of pharmaceutically relevant compounds using RosettaLigand." Protein Science **18**: 1998-2002.

Fan, H., J. J. Irwin, et al. (2009). "Molecular docking screens using comparative models of proteins." Journal of chemical information and modeling **49**: 2512-2527.

Ferrara, P. and E. Jacoby (2007). "Evaluation of the utility of homology models in high throughput docking." Journal of Molecular Modeling **13**: 897-905.

Grant, B. J., A. P. C. Rodrigues, et al. (2006). "Bio3d: an R package for the comparative analysis of protein structures." Bioinformatics (Oxford, England) **22**: 2695-2696.

Kairys, V., M. X. Fernandes, et al. (2006). "Screening drug-like compounds by docking to homology models: a systematic study." J. Chem. Inf. Model **46**: 365–379.

Kaufmann, K., K. Glab, et al. (2008). Small Molecule Rotamers Enable Simultaneous Optimization of Small Molecule and Protein Degrees of Freedom in ROSETTALIGAND Docking. German Conference on Bioinformatics, Dresden, Gesellschaft fur Informatik.

Kaufmann, K. W., E. S. Dawson, et al. (2009). "Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies." Proteins **74**(3): 630-642.

Konagurthu, A. S., J. C. Whisstock, et al. (2006). "MUSTANG: a multiple structural alignment algorithm." Proteins **64**(3): 559-574.

Kuntz, I. D., J. M. Blaney, et al. (1982). "A geometric approach to macromolecule-ligand interactions." Journal of molecular biology **161**: 269-288.

Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

London, N. and O. Schueler-Furman (2007). "Assessing the energy landscape of CAPRI targets by FunHunt." Proteins-Structure Function and Bioinformatics **69**(4): 809-815.

London, N. and O. Schueler-Furman (2007). "Assessing the energy landscape of CAPRI targets by FunHunt." Proteins **69**(4): 809-815.

MacCallum, J. L., L. Hua, et al. (2009). "Assessment of the protein-structure refinement category in CASP8." Proteins **77 Suppl 9**: 66-80.

Mandell, D. J., E. a. Coutsias, et al. (2009). "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling." Nature methods **6**: 551-552.

McGovern, S. L. and B. K. Shoichet (2003). "Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes." Journal of medicinal chemistry **46**: 2895-2907.

Misura, K. M., D. Chivian, et al. (2006). "Physically realistic homology models built with ROSETTA can be more accurate than their templates." Proc Natl Acad Sci U S A **103**(14): 5361-5366.

Nannemann, D. P., K. W. Kaufmann, et al. (2010). "Design and directed evolution of a dideoxy purine nucleoside phosphorylase." Protein Eng Des Sel **23**(8): 607-616.

R Development Core Team (2005). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Raman, S., R. Vernon, et al. (2009). "Structure prediction for CASP8 with all-atom refinement using Rosetta." Proteins **77 Suppl 9**: 89-99.

Raman, S., R. Vernon, et al. (2009). "Structure prediction for CASP8 with all-atom refinement using Rosetta." Proteins.

Rester, U. (2006). "Dock around the Clock – Current Status of Small Molecule Docking and Scoring." QSAR and Combinatorial Science **25**: 605 - 615.

Saxena, A., D. Wong, et al. (2009). "The basic concepts of molecular modeling." Methods in enzymology **467**: 307-334.

Sousa, S. F., P. A. Fernandes, et al. (2006). "Protein-ligand docking: current status and future challenges." Proteins **65**: 15-26.

Taylor, R. D., P. J. Jewsbury, et al. (2002). "A review of protein-small molecule docking methods." Journal of computer-aided molecular design **16**: 151-166.

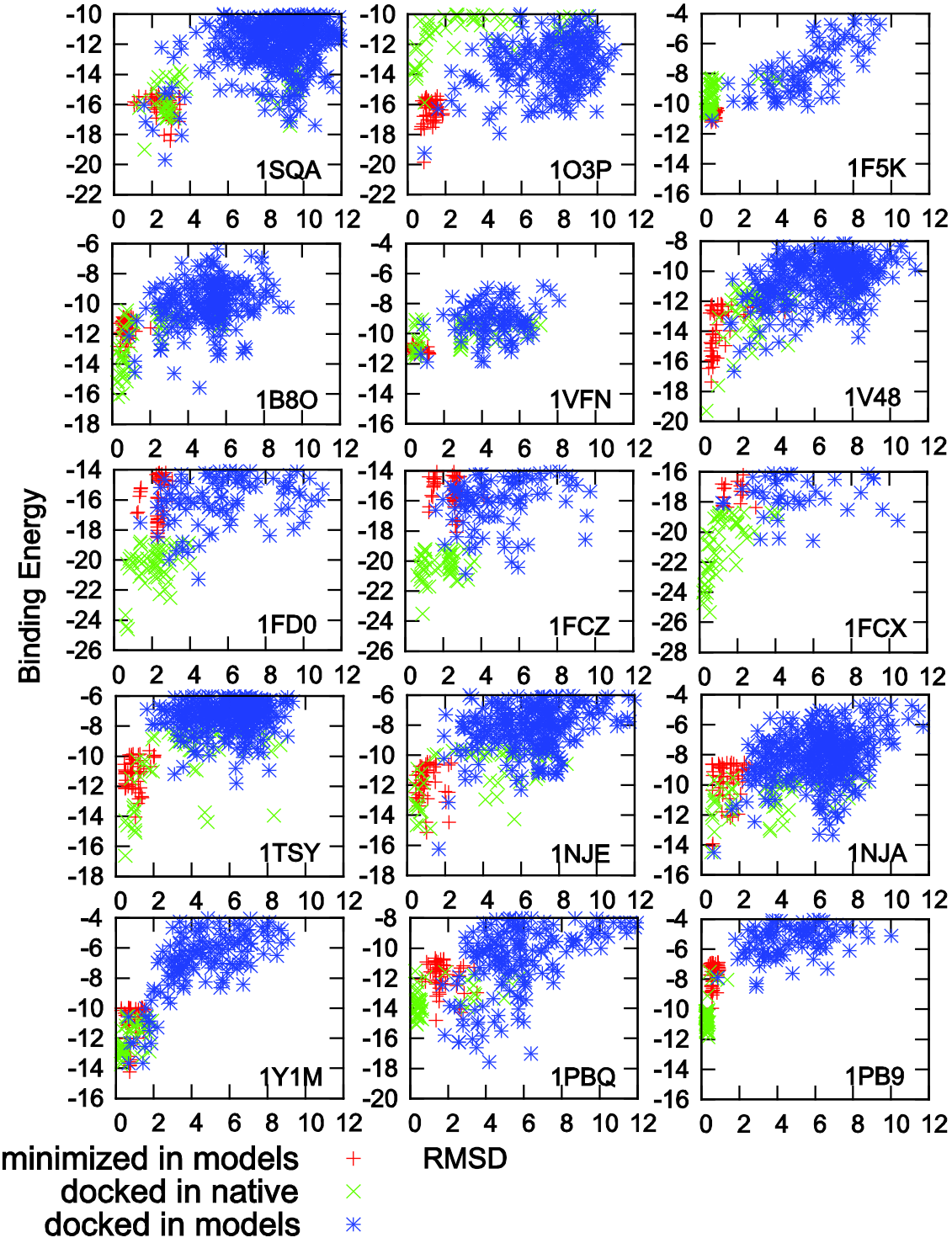
Tress, M. L., I. Ezkurdia, et al. (2009). "Target domain definition and classification in CASP8." Proteins **77 Suppl 9**: 10-17.

Verdonk, M. L., P. N. Mortenson, et al. (2008). "Protein-ligand docking against non-native protein conformers." Journal of chemical information and modeling **48**: 2214-2225.

Warren, G. L., C. W. Andrews, et al. (2006). "A critical assessment of docking programs and scoring functions." Journal of medicinal chemistry **49**: 5912-5931.

Zhang, Y. (2009). "I-TASSER: fully automated protein structure prediction in CASP8." Proteins **77 Suppl 9**: 100-113.

Appendix A1


Binding Energy
RMSD
1SQA
1O3P
1F5K
1B8O
1VFN
1V48
1FD0
1FCZ
1FCX
1TSY
1NJE
1NJA
1Y1M
1PBQ
1PB9
minimized in models
docked in native
docked in models

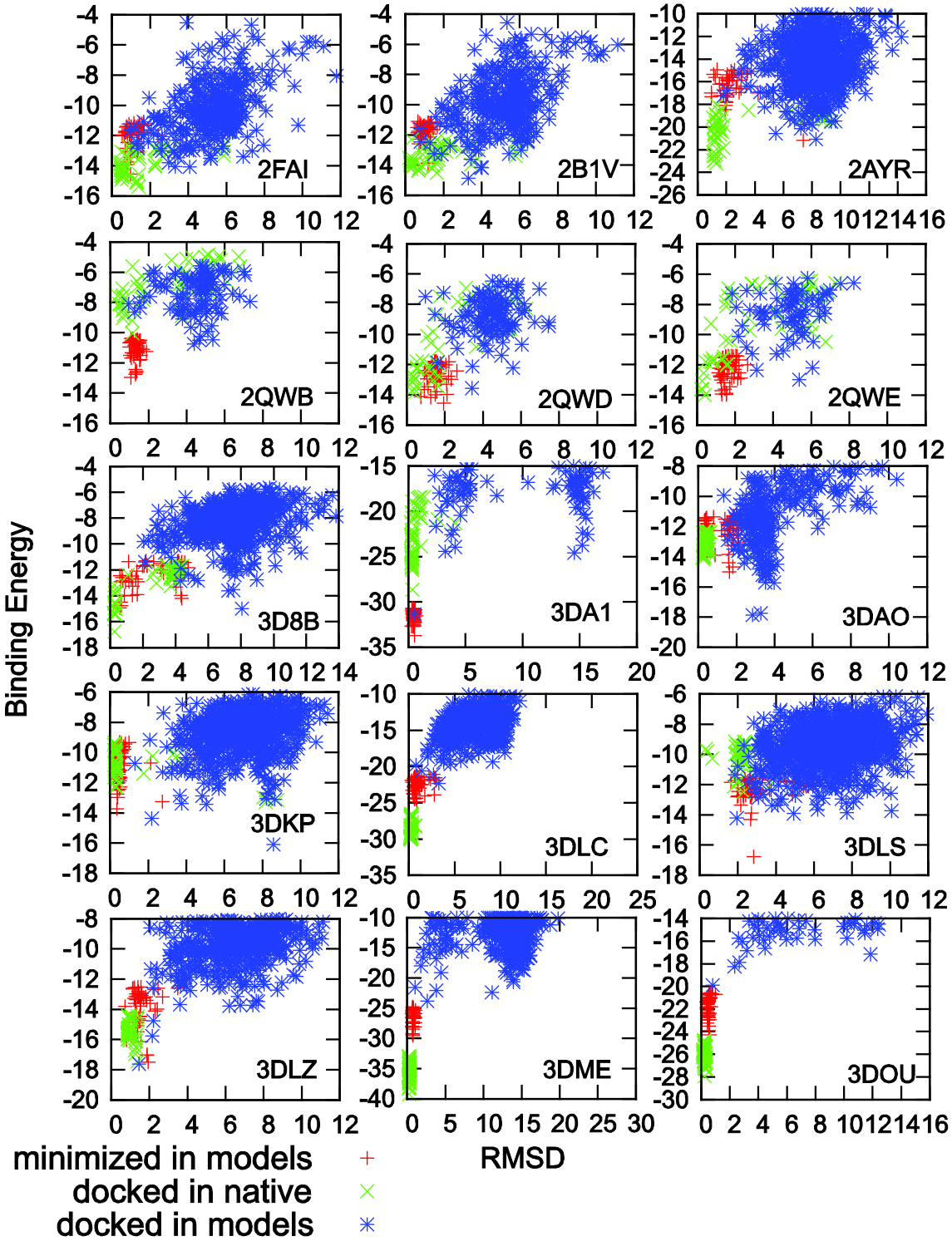


Figure A1 L-RMSD energy plots of complexes docked into multiple comparative models. Blue stars display docking clusters, green crosses show top models from docking into the target structure. Red crosses show models produce by Monte Carlo minimization of native-like binding modes in the comparative models.

CHAPTER V

STRUCTURAL DETERMINANTS OF SPECIES SELECTIVE SUBSTRATE RECOGNITION IN HUMAN AND *DROSOPHILA* SEROTONIN TRANSPORTERS REVEALED THROUGH COMPUTATIONAL DOCKING STUDIES[1]

Introduction

As members of the sodium and chloride dependent neurotransmitter transporter (SLC6) gene family, SERTs carry out the uptake of 5-HT across plasma membranes in the central nervous system, peripheral nervous system, placenta, platelets, and pulmonary system (Ramamoorthy, Bauman et al. 1993; Rothman and Baumann 2002). SERTs are targets of antidepressants and substances of abuse like cocaine and 3,4-methyldioxy-methamphetamine (MDMA, commonly known as "Ecstasy") (Roman, Saldana et al. 2004). Hydropathy analyses initially suggested that SERTs are integral membrane proteins with twelve α-helices (Blakely, Berson et al. 1991; Hoffman, Mezey et al. 1991; Ramamoorthy, Bauman et al. 1993). Site-directed mutagenesis and SCAM experiments on putative TMs and loops have supported this proposal (Chen, Liu-Chen et al. 1997; Chen, Liu-Chen et al. 1998; Henry, Adkins et al. 2003; Keller, Stephan et al. 2004; Zhang and Rudnick 2006).

Mutagenesis of key residues has provided insight into the structure and function of SERT. Shortening the ethylamine tail of tryptamine by one methylene group (dimethyl-tryptamine to gramine) causes a decrease in substrate uptake in rat SERT (rSERT). The addition of a methylene group, via a D98E mutation, restores uptake of gramine to levels expected for dimethyl-tryptamine, suggesting that residue D98 forms a direct (ion pair) interaction with the substrate(Barker, Moore et al. 1999). Chen et al.(1997) implicated I172 and I176 in substrate and inhibitor binding through protection of transporter function from inactivation by MTS in cysteine mutants of these residues. Several studies have identified amino acid sequence differences among SERT species that confer alternate specificities for substrates and

[1] Reprinted with permission from Kaufmann, K.W. Dawson E.S. Henry L.K., Field J.R. Blakely R.D. Meiler J. *Proteins* **2009** 74 630-642

inhibitors(Adkins, Barker et al. 2001; Rodriguez, Roman et al. 2003; Roman, Saldana et al. 2004; Henry, Field et al. 2006). Barker et al.(1998) used human SERT (hSERT) and *Drosophila* SERT (dSERT) chimeras to implicate Y95 in forming part of the recognition site for citalopram and mazindol, two biogenic amine reuptake inhibitors. Adkins et al.(2001) used the same approach to show the Y95F hSERT mutant exhibits dSERT-like recognition of N-isopropyl tryptamine. Henry et al.(2006) found the I172 residue in hSERT displays a marked functional divergence with respect to inhibitor but not substrate potencies when the residue is mutated to its dSERT identity (I172M). Although these advances have identified residues involved in 5-HT and antagonist recognition, interpretation of these data would benefit from a three-dimensional (3D) context provided by high-resolution transporter structures.

Comparative models of SERT have been reported that interpret the structure function implications of site directed mutagenesis data and substituted cysteine accessibility data using Na+/H+ antiporter cyro-EM densities and crystal structure as well as the Lac permease crystal structure (Ravna and Edvardsen 2001; Ravna, Sylte et al. 2003; Ravna, Jaronczyk et al. 2006). However, the low sequence homology and low functional correlation of these templates to SERTs limits predictive power of these models. The recently reported crystal structure for a bacterial Na+-dependent leucine transporter ($LeuT_{Aa}$), a bonafide member of the neurotransmitter sodium symporter (NSS) protein family, represents a critical break-through for the field (Yamashita, Singh et al. 2005; Henry, Defelice et al. 2006). The $LeuT_{Aa}$ structure confirms a predicted topology for NSS members consisting of twelve TM spanning α-helices. Unexpectedly, it features two five-helix bundles arranged in an inverted mirror symmetry. The final two helices, TMs 11 and 12, reside peripheral to the core transporter and may participate in homo-oligomerization (Just, Sitte et al. 2004). In the crystal structure of $LeuT_{Aa}$, the substrate leucine is located in a pocket formed by TMs 1, 3, 6, and 8. Notably, unwound regions in the centers of TMs 1 and 6 serve as contact points for the carboxyl and amine groups of leucine. Beuming et al. (2006) refined the primary sequence alignment of $LeuT_{Aa}$ using a large multisequence alignment of eukaryotic and prokaryotic NSS family members sequences, resulting in an alignment featuring an improved agreement with available biochemical data that underscores the utility of the $LeuT_{Aa}$ structure.

ROSETTA comparative modeling (Rohl, Strauss et al. 2004; Misura and Baker 2005) and docking (Meiler and Baker 2006) approaches are invoked for their power in building accurate models of membrane proteins from distant sequence homologs as recently demonstrated with voltage-gated K+ channels (Yarov-Yarovoy, Baker et al. 2006). Our approach involves comprehensive high-resolution docking of 5-HT into SERT comparative models based on the $LeuT_{Aa}$ structure. We refrain from using experimental data during model construction to permit rigorous testing of the predictive power of our models using data derived from site-directed mutagenesis, SCAM, and binding affinity experiments.

Methods

SERT Sequence Alignment

The alignment used in the comparative models of dSERT and hSERT on $LeuT_{Aa}$ was synthesized from the alignments of Beuming and colleagues (Beuming, Shi et al. 2006). The adjusted alignment published between $LeuT_{Aa}$ and rSERT was combined with the alignment of the eukaryotic NSS family provided by Beuming (Beuming, Shi et al. 2006).

The SERT sequences were divided into TM and binding site regions based on the $LeuT_{Aa}$ crystal structure and 5-HT docking results discussed below. TMs 1, 3, 6, and 8 form the core TMs that surround the leucine binding site. First shell residues are defined as any residues with a Cα atom within 7 Å of the leucine ligand in the $LeuT_{Aa}$ structure. We define the second shell binding site residues to be all residues with a Cα atom within 12 Å of the leucine ligand in the $LeuT_{Aa}$.

Docking leucine into the $LeuT_{Aa}$ (PDB ID 2A65) crystal structure.

The crystal structure 2A65 was obtained from the Research Collaboration for Structural Bioinformatics Protein Data Bank(Bernstein, Koetzle et al. 1977) website. All non-protein atoms were removed. Docking of leucine was performed using ROSETTALIGAND as described by Meiler and

Baker(Meiler and Baker 2006). In brief, twenty conformations, including the crystallized conformation, were generated for leucine. These rigid conformations of the ligand were placed in a random orientation and position inside a user-defined 8 Å cube around the native binding site. ROSETTALIGAND then simultaneously placed sidechain rotamers around the ligand and optimized the ligand pose in a Metropolis Monte Carlo simulated annealing algorithm to optimize the binding site structure and minimize the binding energy. The energy function used during the search contains terms for van der Waals attractive and repulsive forces, statistical energy derived from the Dunbrack probability of observing a sidechain conformation in the PDB, hydrogen bonding, electrostatic interactions between pairs of aminoacids, and solvation assessing the effects of both sidechain interactions and sidechain ligand interactions. Five hundred structures for each conformation of leucine were produced for a total of 10,000 structures for the $LeuT_{Aa}$-leucine complex. This experiment mimics the docking procedure applied for modeling of 5-HT interactions with SERT and provides a test of our computational methods on a closely related system.

SERT Comparative Model Construction

The backbone coordinates of the TM helices from the $LeuT_{Aa}$ crystal structure (PDBID 2A65) were retained in the comparative models of dSERT and hSERT. The loop regions were built in ROSETTA using Metropolis Monte Carlo fragment replacement(Rohl, Strauss et al. 2004) combined with Dunbrack cyclic descent loop closure (Canutescu and Dunbrack 2003). In short, ϕ-ψ angles of backbone segments of homologous sequence amino acid fragments from the PDB are introduced for the residues in the loops. After the fragment substitution slight changes in the ϕ-ψ angles are made to close breaks in the protein chain. Sidechains for all residues in the protein were built using ROSETTA's Metropolis Monte Carlo rotamer search algorithm(Kuhlman, Dantas et al. 2003). Subsequently the ten models generated for both dSERT and hSERT were iteratively subjected to 8 cycles of sidechain repacking and gradient minimization of ϕ, ψ, and χ angles in ROSETTA.

SERT Serotonin Docking

A conformational ensemble containing 100 conformations of 5-HT was generated using the mmff94 small molecule force field in MOE (Molecular Operating Environment, Chemical Computing Group, Montreal, Quebec, Canada). The ensemble contained representatives from the (±) gauche and the trans conformations of the ethylamine tail. Each conformation from the ensemble was placed into both hSERT and dSERT models for docking calculations using ROSETTALIGAND. ROSETTALIGAND placed 5-HT in a random orientation inside a 10Å cube centered at the same depth as leucine in the $LeuT_{Aa}$. ROSETTALIGAND then simultaneously placed sidechain rotamers around the ligand and optimized the ligand pose in a Metropolis Monte Carlo simulated annealing algorithm. The energy function used during the search contains terms for van der Waals attractive and repulsive forces, statistical energy derived from the probability of observing a sidechain conformation in the PDB, hydrogen bonding, electrostatic interactions between pairs of aminoacids, and solvation assessing the effects of both sidechain interactions and sidechain ligand interactions (Meiler and Baker 2006). Approximately 13,000 docked complexes each for hSERT and dSERT were generated.

Inaccuracies inherent in comparative models preclude identification of the native binding mode based solely on the score. When docking small molecules into crystal structures, the ROSETTALIGAND energy function reliably identifies the correct binding model (Meiler and Baker 2006). However when dealing with comparative models the energy funnel of the correct binding mode is shallower and local minima can have increased depth (see Fig. 1). Nonetheless, the correct binding mode can occupy a minimum in the energy landscape. As discussed by recent studies, docking to comparative models remains a difficult task; however they can prove useful in the design of experiments (DeWeese-Scott and Moult 2004; Kairys, Fernandes et al. 2006). Docked complexes occupying a physiologically relevant minimum in the energy landscape might then be identified through testing the predictive power of the models using available biochemical data as a filter.

The structures with the best protein ligand interaction energies were selected in a first filter. A second filter imposed a 3.6Å distance between the 5-HT amine tail and one of the D98 sidechain carboxyl

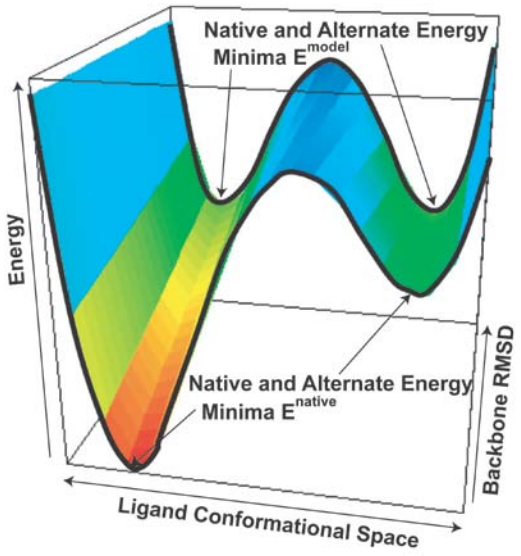


Figure 1. The docking energy landscape is shown as a function of backbone RMSD. The energy is indicated by color from low (red) to high (blue). As the error or RMSD in the backbone increases the native minimum in the energy landscape E^{native} is obscured. Alternate binding modes associated with higher energy can no longer be clearly distinguished from the native binding mode E^{model}. Comparative models by their nature have some error in their atom coordinates. In turn frequently multiple minima are observed when docking small molecules into comparative models. Additional experimental data are required to distinguish between these models. Reprinted with permission from (Kaufmann, Dawson et al. 2009)

oxygens. In a third-round filter binding modes were chosen that were present in both dSERT and hSERT based on the assumption that the 5-HT binding mode is conserved across the two species.

Serotonin Analog Docking

After identifying a common binding mode for 5-HT in the both hSERT and dSERT models, 5-HT analogs were placed into the ligand binding site while maintaining the putative binding mode of 5-HT. Each of the analogs then underwent Monte Carlo refinement and gradient energy minimization allowing small adjustments in ligand position and sidechain conformations. For each binding mode, the nine lowest total ROSETTA energy structures for each analog were selected. Out of the nine structures, the structure with the lowest total ROSETTA energy and with an indole ring less than 1 Å RMSD from the starting position was retained for binding energy calculations. The one exception to the RMSD constraint was the

7-benzyloxy-tryptamine analog, which was allowed to deviate further due to the large bulk of the substitution. The resulting lowest energy structures were visually inspected to verify that they retained the original binding mode.

The binding energy was calculated using,

$$\Delta E_{ligand\_binding} = \Delta E_{protein\_bound\_state} - \Delta E_{protein\_unbound\_state} \quad \text{(Eq. 1)}$$

where ΔEprotein_unbound_state is the energy of the protein in the unbound state, and ΔEprotein_bound_state is the energy of the protein in the bound state plus ligand protein interaction energy. The change in energy, ΔE, is given by

$$\Delta E = \Delta E_{atr} + \Delta E_{dun} + \Delta E_{hb} + \Delta E_{pair} + \Delta E_{sol} \quad \text{(Eq. 2)}$$

as was reported previously(Kortemme and Baker 2002; Meiler and Baker 2006). ΔEatr is the attractive portion of a van der Waals Lennard-Jones 12-6 potential energy term. ΔEdun is the energy derived from the Dunbrack rotamer probability. ΔEhb is the energy of hydrogen bonds involving sidechains. ΔEpair encodes for the energy due to electrostatic interaction between residues. ΔEsol is a Lazaridius-Karplus approximation of the solvation energy. The repulsive portion of the van der Waals energy was removed to decrease noise inherent in the sensitivity of this term. ΔE for each residue were summed to obtain the total ΔE for the protein binding energy. Amino acid residues with a ΔE < -1 were considered to be major contributors to the binding energy.

Model Refinement of Binding Mode with Bound Na+ Ion

Molecular models for the sodium (Na+) ion bound form of both hSERT and dSERT were generated and refined using the following protocol. The ROSETTALIGAND binding mode was taken as the starting point for model refinement using the AMBER forcefield (Wang, Cieplak et al. 2000). Briefly, the binding mode models for the hSERT and dSERT were aligned with the published structure of $LeuT_{Aa}$ (PDB ID: 2A65) and a single Na+ ion was added to both models by copying the coordinates of atom NA

572 (Na1 binding site). Models of the hSERT and dSERT sodium ion binding site were then refined with 50 steps of steepest descents and 450 steps of conjugate gradient energy minimization in AMBER9 (Bayly, Cieplak et al. 1993) followed by brief (1ns), low temperature (50K) molecular dynamics simulations in-vacuo using a distance-dependent dielectric constant and 12Å cutoff for non-bonded interactions. Partial charges for 5-HT were developed using the atom-centered point charge method of Bayley et al. (Bayly, Cieplak et al. 1993). All other molecular mechanics parameters for 5-HT and ions were taken from the standard AMBER force field. Two-dimensional schematics of the refined hSERT and dSERT ion binding sites were generated with ChemDraw 10.0 (Cambridge Soft) while 3D representations were rendered with PyMol (DeLano).

SVM Analysis for Tryptamine Analog Pharmacology

Support vector machines (SVM) (Vapnik 1998), a form of machine learning previously used by this group to study anti-cancer activity of epothilones (Bleckmann and Meiler 2003), were applied to derive a substitution sensitivity model for SERT substrates using uptake inhibition data from a previously published study of tryptamine analogs (Adkins, Barker et al. 2001). The freely available software, LIBSVM (Chang and Lin 2001), was applied to 26 tryptamine analogs to derive models for hSERT and dSERT sensitivity to substitution at positions around the indole ring and ethyl amine tail. The binary encoding scheme for each compound was configured to indicate the type of substituent at each of the following positions: R1/2, α2, X, R3, 4, 5, 6, 7 (see Fig. 2 and Appendix A Table AX). A total of 24 binary inputs are required to uniquely describe the configuration of each of the 26 tryptamine analogs in these nine positions. The resulting input vector of length 24 for each compound is associated with a normalized floating point representation of the experimentally measured binding constant for [3H]-5-HT uptake inhibition (K_i) for training of the SVM.

Epsilon support vector regression was applied with a cost of 0.2 and a polynomial kernel function with gamma of 0.1. Optimal cost (c) and gamma (γ) parameters were empirically determined via a systematic search for best RMSD for predict log K_i from leave one out cross validation. Description of

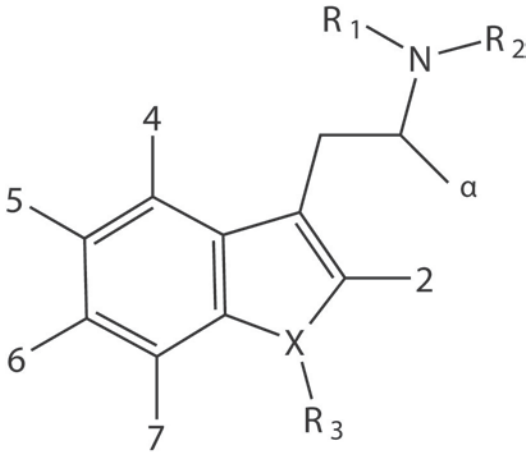


Figure 2. Tryptamine core used in fragment based substitution encoding for SVM sensitivity maps. Reprinted with permission from (Kaufmann, Dawson et al. 2009)

the theory and application of SVM can be found in the following references (Vapnik ; Chang and Lin). The sensitivity to each input was computed as the absolute partial derivative of the output (i.e SVM predicted binding constant) with respect to that input. The average sensitivity to substitution was computed by taking the mean of the sensitivities for all inputs coding for substitution at a position on the tryptamine core. The rationale of this approach is that large derivatives identify sensitive inputs that point to more critical regions for binding and vice versa. The average sensitivity to substitution at each position was displayed as a colored molecular surface using PyMOL (DeLano).

Results

Our strategy employs comparative modeling, ligand docking, and SAR methodology to address species selectivity for substrate recognition in hSERT and dSERT. Comparative modeling of a target sequence based on a known structural template requires identification of a related structural template, alignment of the target sequence to the structure, model construction, and assessment of the resulting structure (Baker and Sali 2001). Ligand docking programs seek to identify the lowest free energy

structure of the ligand protein complex (Ferrara, Gohlke et al. 2004). It is beneficial to categorize the available structural degrees of freedom into ligand internal degrees of freedom (ligand conformation), ligand translation and rotational degrees of freedom (pose), protein sidechain degrees of freedom (rotamer), and protein backbone degrees of freedom. Our approach optimizes all of these degrees of freedom during the course of the model development. In addition we use support vector machines (SVM) to condense data into substitution sensitivity maps (Vapnik 1998; Bleckmann and Meiler 2003). SVMs allow analysis of data sets containing noise and uneven distribution in the chemical space tested by offering an overview of the available data. The overview can then be interrogated in more depth.

Sequence Alignment Demonstrates High Similarity Between the LeuT and the SERT substrate binding sites

Sequence alignments offer insight into the structural similarity of two proteins. The sequence identities in Table 1, based on the alignment of hSERT and dSERT to the rSERT-$LeuT_{Aa}$ alignment in Fig. 3, reflect regions expected to have different degrees of involvement in the binding of substrates as defined in the Methods. The sequence identity increases from ~15% to greater than 50% as the focus narrows on the first shell of residues in the binding site. As the sequence identity increases, the confidence in the alignment and the resulting quality of the comparative models increases(Forrest, Tang et al. 2006).

Table 1. Relationship Between Sequence Identity and Expected Model Accruacy. Relationship between sequence identity of hSERT and dSERT to $LeuT_{Aa}$ in specific regions of the protein and the expected model accuracy. Core TMs are TMs 1,3,6, and 8. Second shell and 1st shell residues include all residues with C_α atoms within 12 and 7 Å respectively of an atom from the leucine ligand in the PDB structure 2A65. Reprinted with permission from (Kaufmann, Dawson et al. 2009)

		Overall	Loop regions	TMs	Core TMs	2nd Shell	1st Shell
Protein Sequence Identity	hSERT	17% 108/630	11% 40/362	25% 68/268	35% 38/108	40% 31/77	58% 11/19
	dSERT	18% 113/622	14% 51/354	23% 62/268	33% 36/108	36% 28/77	52% 10/19
Expected Backbone RMSD to true structure[39]		>5Å	>5Å	>=2.5Å	≈2Å		
Backbone RMSD to $LeuT_{Aa}$	hSERT			1.6-2.1	1.1-1.6	1.0-1.3	0.9-1.2
	dSERT			1.4-2.3	1.1-1.8	1.0-1.3	0.9-1.2

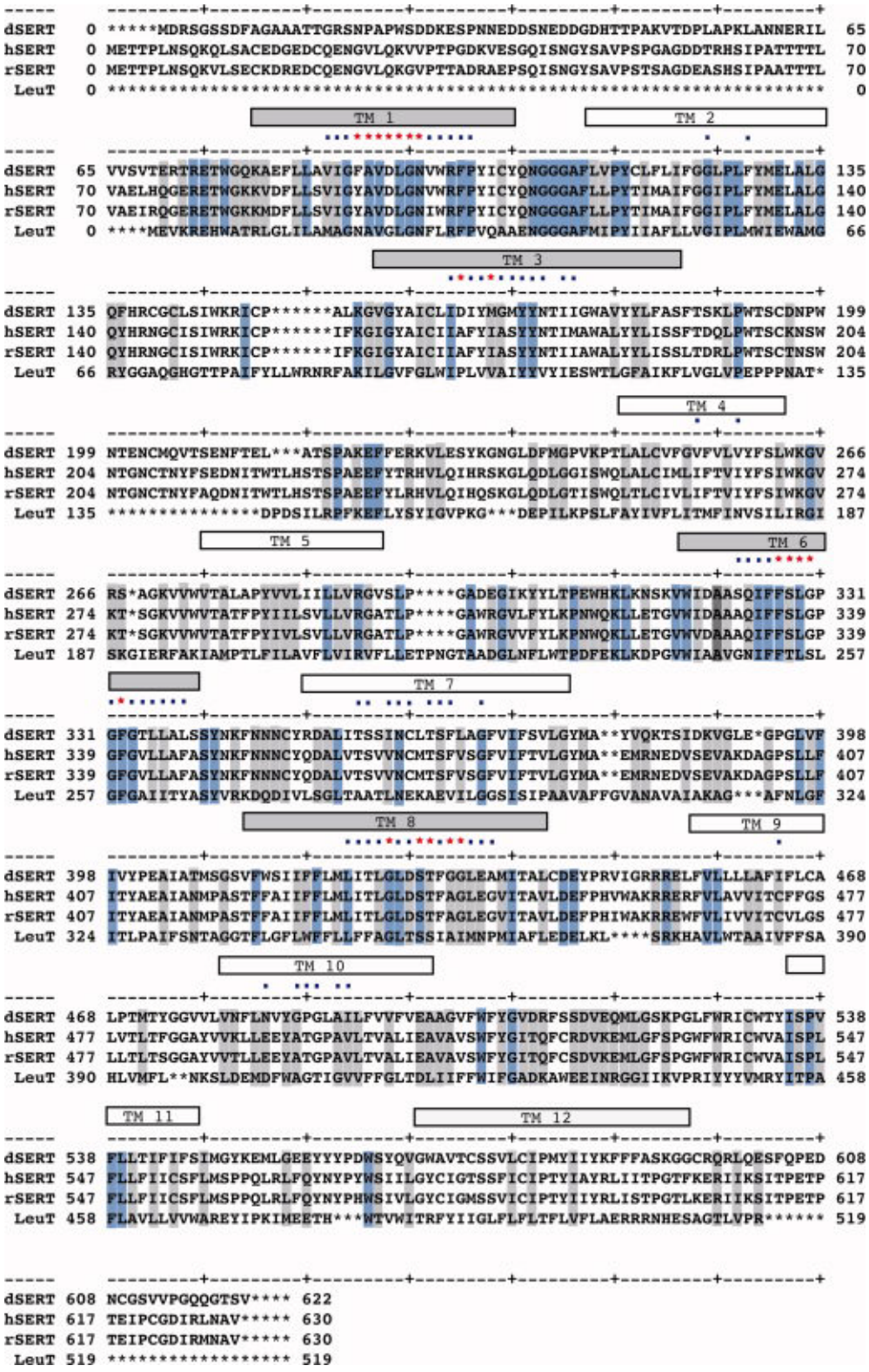


Figure 3. Sequence alignment between LeuTAa, hSERT, dSERT, and rSERT. Blue background denotes complete conservation of amino acid identity. Light gray background denotes similarity of amino acid identity across sequences. Rectangles above amino acids mark the transmembrane helices. Core transmembrane helices are shaded gray. Red stars denote amino acids in the first shell of the binding site. Blue squares highlight residue in the second shell of the binding site. Reprinted with permission from (Kaufmann, Dawson et al. 2009)

RosettaLigand Correctly Docks Leucine into the LeuT$_{Aa}$ Structure

The self-docking of leucine back into the LeuT$_{Aa}$ crystal structure serves to evaluate ROSETTALIGAND performance in docking to a NSS protein. The lowest energy structure recaptured the native binding mode (RMSD 0.81 Å). Plotting the predicted ligand binding energy versus the RMSD of the ligand to the native ligand coordinates yields the "energy funnel" seen in Fig. 4a. At the neck of the "energy funnel", the sidechain position of leucine is recovered along with the amine positioning with only one difference, a 90° deviation in the ψ angle (Fig. 4b). The absence of ions (e.g Na^+) in the docking process produced only a minor perturbation of the physiological leucine binding mode. Only R30 and I359 are in different rotamer states. RosettaLigand successfully docks leucine back into the LeuT$_{Aa}$ structure.

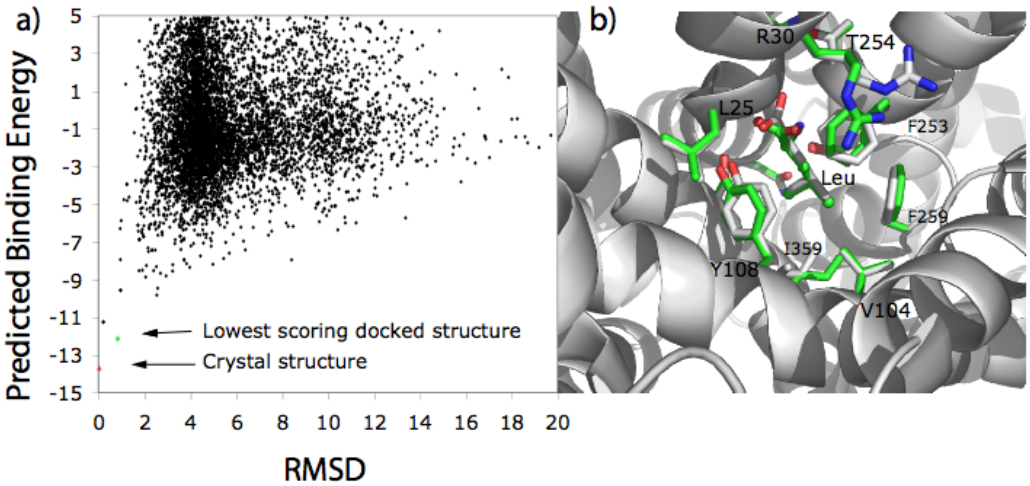


Figure 4. ROSETTALIGAND binding energies decrease (a) as the model approaches the native crystal structure. The lowest energy ROSETTA model has a deviation from the crystal structure of 0.81Å (b) Overlay of computational docked leucine in ball and sticks with green sidechains on the crystal structure with grey sidechains and leucine. This figure was prepared using PyMOL. (DeLano 2002). Reprinted with permission from (Kaufmann, Dawson et al. 2009)

SERT Comparative Models Extensively Sample Backbone and Sidechain Conformational Space

Side by side comparison of hSERT, dSERT, and LeuT$_{Aa}$ models highlight differences that may be responsible for differences in ligand recognition and transport. As can be seen in Fig. 5, many sidechains of the transporters retain not only their amino acid identity but also the χ angles, supporting the conserved functionality of these residues. Most of the diversity observed in the binding site is conserved across both

dSERT and hSERT and also occurs at the intracellular end of the binding site. The backbone RMSDs in the twenty SERT models range from as little as 0.9Å in the binding site up to 2.3Å in trans-membrane spans (see Table I). SCAM accessibility patterns in the regions comprising the binding site show a periodicity that agrees with available experimental data (Fig. 6).

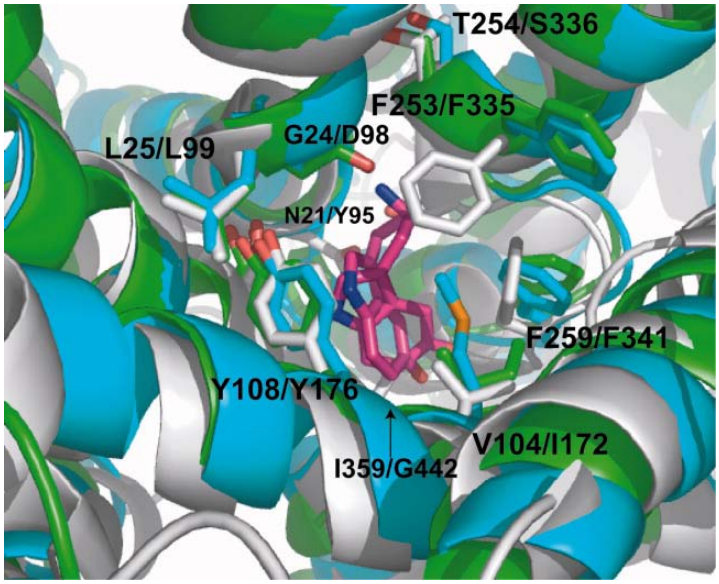


Figure 5. Overlay of hSERT comparative model in green and the dSERT model in cyan on LeuTAa crystal structure in gray. The conformational space sampled in this study remains close to that of the backbone captured in the LeuTAa structure. Gradient minimization retains most of the same side-chain interactions, due to the high sequence identity evident in the binding site. This figure was prepared using PyMOL. (DeLano). Reprinted with permission from (Kaufmann, Dawson et al. 2009)

Serotonin Docking Comprehensively Samples Translational and Rotational Degrees of Freedom in Protein-Ligand Complex and identifies 5 potential binding modes

Ligand docking searches for the most energetically favorable position of 5-HT in the binding pocket; thus identifying likely structural determinants for 5-HT recognition. Out of the top 100 lowest energy docked 5-HT complexes for each protein, 22 dSERT models and 24 hSERT models contained a D98 contact. Of those models six binding modes were present in both proteins. Five of the six binding modes place the amine in approximately the same location as seen as for leucine in the $LeuT_{Aa}$ structure. These five modes were carried forward for further analysis and are shown in Fig. 7. The first three

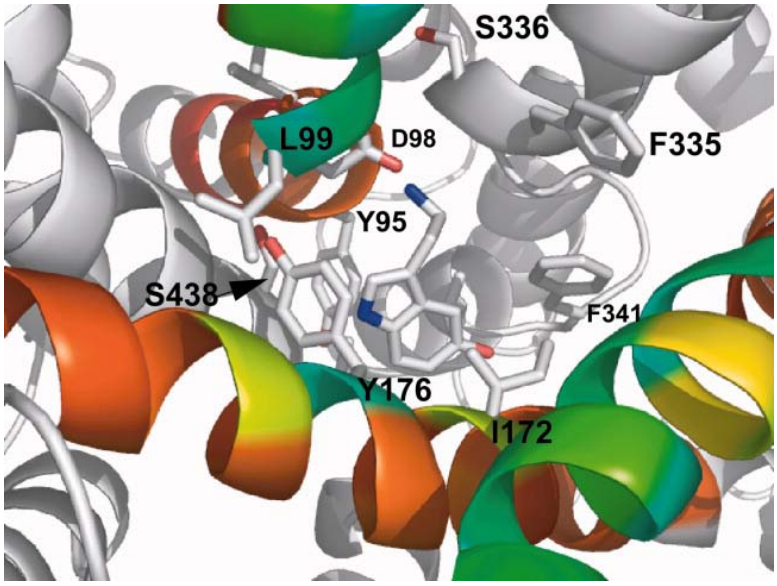


Figure 6. hSERT Down binding mode with substituted cysteine accessibility mapped onto TM 1, 3, and 10. Red to blue scale indicates no sensitivity to large sensitivity to MTS attack of a cysteine substituted at that residue. All three helices show patterns consistent with the helix orientations in the models. This figure was prepared using PyMOL. (DeLano). Reprinted with permission from (Kaufmann, Dawson et al. 2009)

binding modes Up_a, Up_b, and Up_c have the 5-hydroxyl group oriented in the general direction of the extracellular surface (Fig. 7a, b, c). In the first binding mode Up_a (Fig. 7a), the 5-hydroxy points towards F335, pushing the phenyl ring of F335 up against the TM 6 helix. The indole nitrogen neighbors T439 in TM 8 at the interface between TMs 3 and 8. For the second binding mode Up_b (Fig. 7b), the indole ring is rotated 180° relative to the orientation in Up_a. The indole nitrogen now faces F341. The 5-hydroxyl group is placed up against the ring of Y176 lining the upper side of the binding pocket. Up_c (Fig. 7c) has the indole ring rotated 90° relative to Up_a. It packs against the phenyl ring of Y176 in a π stacking interaction. The edge of the ring points towards the interface between TMs 8 and 3, with A173 and G442 opposite the indole nitrogen in that interface. In Up_c, the 5-hydroxyl group forms a steric contact with L99. The fourth binding mode (Side) has the 5-hydroxyl bond horizontal in the binding pocket pointing towards T439 and G442 in TM 8 at its interface with TM 3 (Fig. 7d). The indole ring lies sideways in the binding pocket with the side of the indole ring packing against I172. Additionally, the.

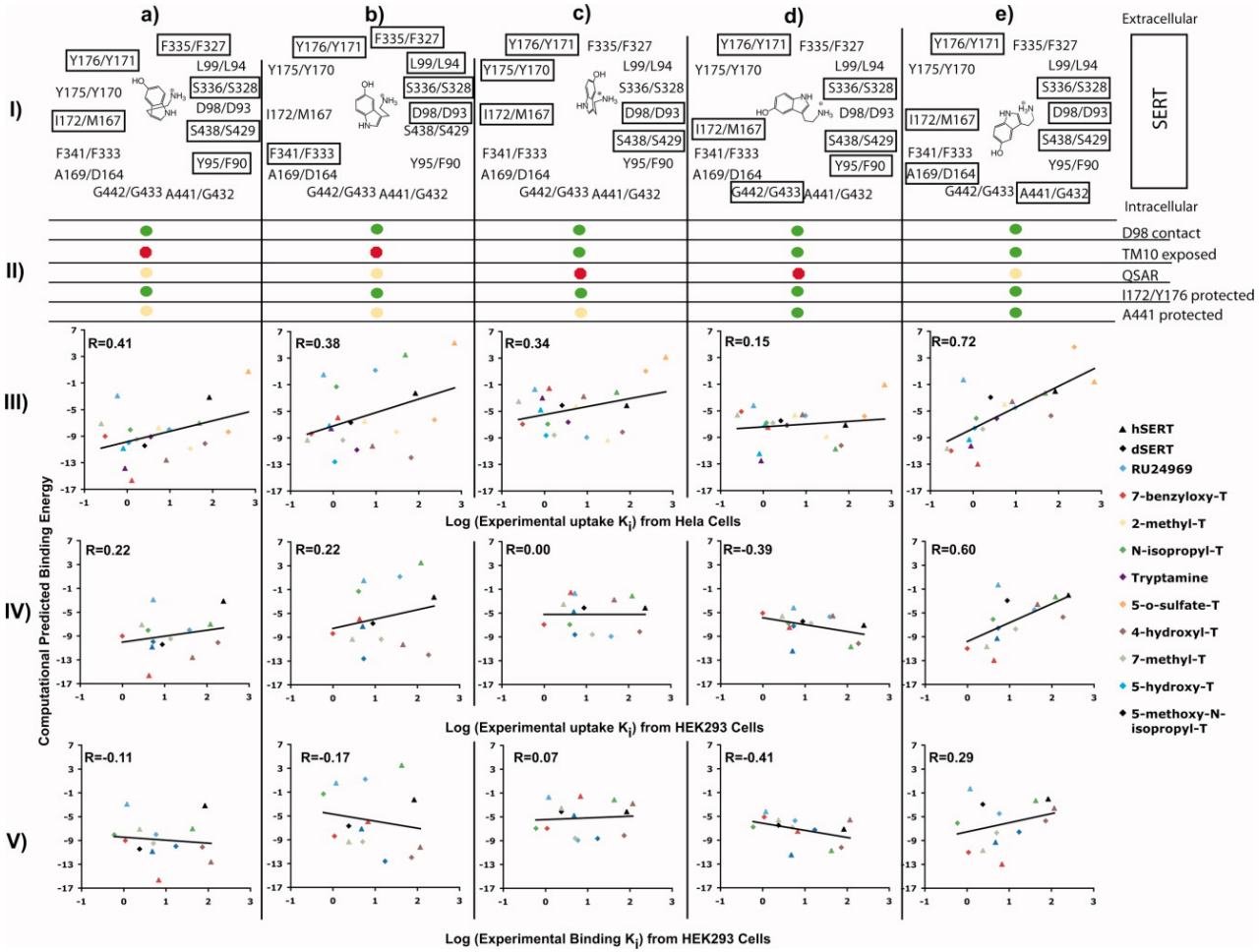


Figure 7. For each of the docked complexes (a) Up_a, (b) Up_b, (c) Up_c, (d) Side, (e) Down (I) shows a flattened representation of the binding site with residues contributing most to the computational binding energy outlined in rectangles with black borders. (II) shows agreement of each docking mode with biological data. Each mode contains a D98 contact. Up_a and Up_b display contacts with TM 10 that contradict the lack of protection from MTS inactivation. Up_c and Side binding modes do not match the SVM species difference maps. All the modes show interaction with I172 and Y176 explaining protection against MTS modification. The Side and Down modes pack closely to A441 in a manner which may explain protection of A441C by 5-HT from MTS modification. (III-V) Correlation plots for predicted log Ki (calculated on computational binding free energy of tryptamine analogs in these modes) and log Ki for uptake in Hela cells (III), for uptake in HEK293 cells (IV), and for binding in HEK293 cells (V). hSERT values are given in triangles and dSERT values in diamonds. All experimental transport and binding data taken from Adkins et al. (Adkins, Barker et al.). Reprinted with permission from (Kaufmann, Dawson et al. 2009)

indole nitrogen points toward F335 at top of the binding pocket. The Down binding mode (Fig. 7e) shows a 180° rotation of the indole ring relative to the position observed in Up_c. The indole nitrogen is in approximately the same position though pointed more towards T439 and N177. The 5-hydroxy is now pointed down towards A169 in TM 3 and G342 in TM 6. The residues contributing to the binding energy are boxed in a flattened representation of the binding pocket in each of the five binding modes as shown in Fig. 7I. The agreement of the biochemical data with each of the binding modes is shown in Fig. 7II.

SVM Derived Sensitivity Maps Highlight Species Differences in the SERT Substrate Tryptamine Pharmacology

Adkins et. al. (Adkins, Barker et al. 2001) reported the potencies of 27 tryptamine analogs to inhibit the uptake of [H]3-5-HT in the hSERT and the dSERT. Here we develop SVM sensitivity maps to visually display differences in the tryptamine pharmacology. The hSERT and dSERT show sensitivity to chemical identity at a variety of positions on the tryptamine core (Fig. 8a and 8b). The SVM maps trained

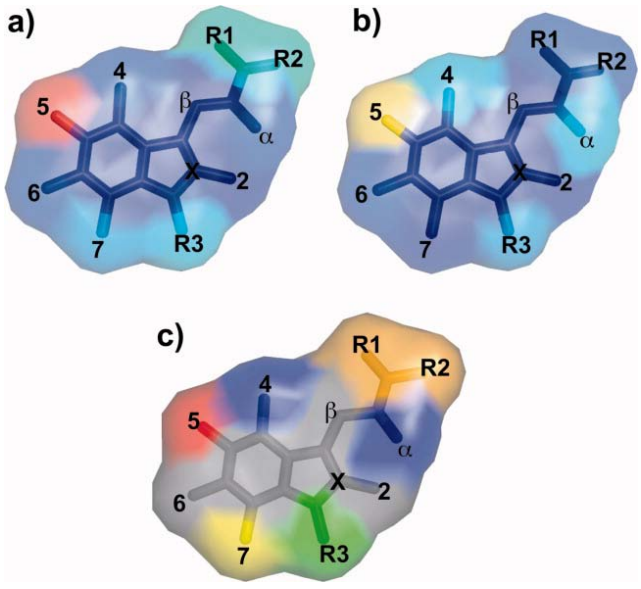


Figure 8. Sensitivities of positions to substitution predicted from support vector machine trained on SERT transporter substrate uptake Kis. Blue to red gradient indicates low to high sensitivity. (a) hSERT, (b) dSERT, (c) difference map (hSERT-dSERT) of the raw sensitivities. Blue shows higher sensitivity for dSERT. Green to red indicates moderate to higher sensitivity in hSERT. This figure was prepared using PyMOL. (DeLano). Reprinted with permission from (Kaufmann, Dawson et al. 2009)

on tryptamines assayed on the hSERT display strong sensitivity to substitution at the 5 position, and weaker sensitivity at the R3 indole position and R1 and R2 ethyl amine positions (Fig. 8a). The dSERT SVM maps also show strong sensitivity at the 5 position with a weaker sensitivity at the R3 indole position, the 4 position, and the α position to the ethyl amine (Fig. 8b). Strong differences in sensitivity between the hSERT and the dSERT SVM maps occur at the R1, R2, R3, α, 4, 5, and 7 positions (Fig. 8c). The hSERT SVM maps show higher sensitivity at the R3, 7, R1, R2, and 5 positions in order of increasing difference in sensitivity. The dSERT SVM maps show higher sensitivity at the α, and 4 position in increasing order of difference in sensitivity. Care is taken to avoid over-interpretation of the SVM maps by resorting to the original data when making use of the maps in the context of modeling

Serotonin Analog Docking Probes ROSETTALIGAND Identified Binding Sites through Binding Energy Prediction

It can be hypothesized that SERTs recognize tryptamine analogs in a conserved manner such that the indole ring occupies the same position in binding pocket. With this in mind, the native binding mode for 5-HT should explain the differences in the binding affinity seen for other tryptamine analogs. Representative deviations of the indole ring for a binding mode compared with 5-HT are shown in Fig. 9. In the Down mode, the substitution of the indole nitrogen causes Y176 to change rotamers. Substitutions at the 5-position interact with residues V343, G442, and A169 in this binding mode. Figure 7III-V show the correlations of the lowest predicted binding free energies of ligand binding and the log of the uptake and binding K_i values extracted from experimental competitive uptake and binding assays by Adkins(Adkins, Barker et al. 2001). The Down mode shows the highest correlation for all three datasets. The correlation coefficient of the Down binding mode to the log uptake K_i data from Hela cells is 0.72. The correlation coefficient to log uptake K_i data from HEK293 cells is 0.60. The coefficient falls to 0.29 when compared to log binding K_i data extracted from HEK293 competition binding assays. The first two datasets of uptake K_i's in HEK293 and Hela cells assess the ability of tryptamine analogs to competitively

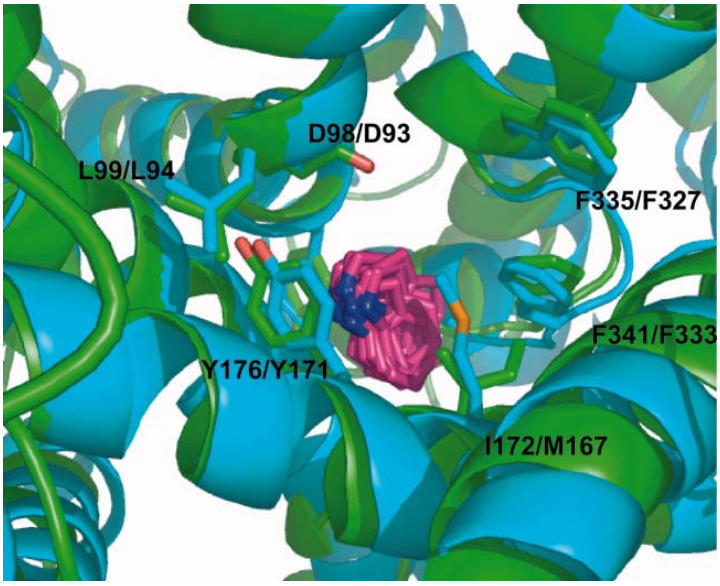


Figure 9. A superimposition of the indole ring of tryptamine derivatives in the Down binding mode is shown for hSERT and dSERT docking. It highlights the conserved manner in which tryptamine derivatives are recognized by SERTs. This figure was prepared using PyMOL. (DeLano). Reprinted with permission from (Kaufmann, Dawson et al. 2009)

inhibit uptake of tritiated serotonin across membranes with the SERT transporter. The third dataset of binding K_i's assesses the ability of tryptamine analogs to compete with a high affinity inhibitor to bind to the SERTs. The third category measures competitive binding events, a more close approximation to the binding energy measured in the present study. However, binding is thought to be an important step during uptake by transport, and the uptake studies examine the ability of chemical similar compounds to compete. Thus, uptake potency provides a relevant assessment of binding. In any case, the down binding mode remains the best correlated in the five binding modes (see Fig. 7e).

Model Minimization in Amber Force Field Confirms Hydrogen Bonding Contacts of 5-OH Group

We refined our final models using the AMBER force field employing a short molecular dynamics simulation as a minimization tool (Summa and Levitt).We leverage the ability of the molecular mechanics force field in AMBER to model ligand flexibility to optimize the models for the hSERT and dSERT 5-HT Down binding mode (Fig. 10). As this calculation is a local refinement with minimal movements, the ROSETTALIGAND conformations are not altered significantly. However, the geometry of hydrogen bonds

and other local interactions are improved. The conformation identified by ROSETTALIGAND proves to be stable after 1ns of molecular dynamics. The overall RMSD of the binding site in both models is < 1.0 Å indicating that, even though the sodium ion is not explicitly included in our model building and ligand docking to identify the 'down' binding mode, the conservation of the site may implicitly encode this information. The 5-OH substituent of 5-HT maintains a hydrogen bond to the dSERT D164 sidechain

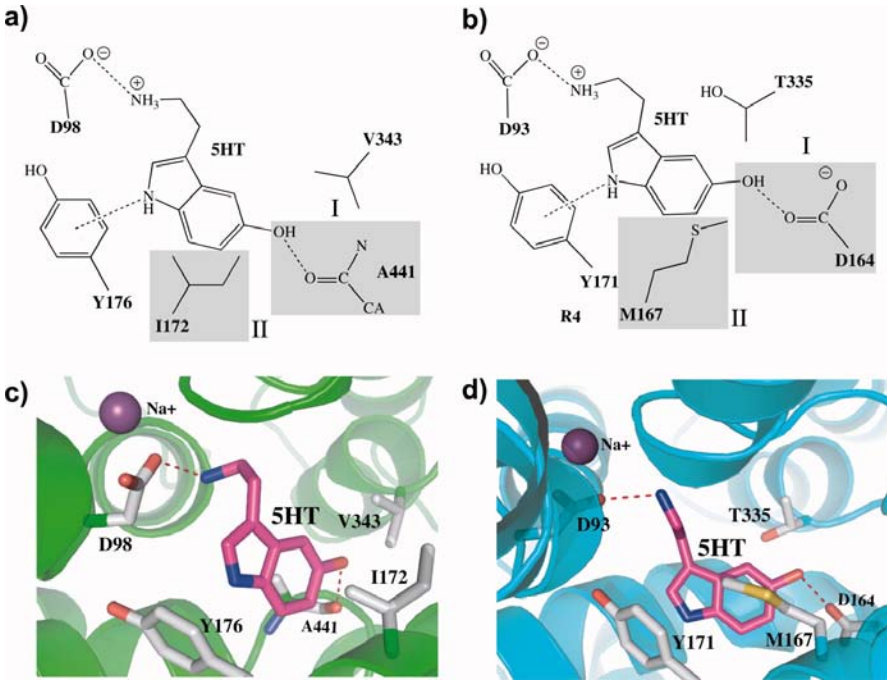


Figure 10. The Down binding mode in the hSERT and dSERT models. Dashed lines in (a) and (b) represent stable hydrogen bonding interactions observed during the 1 ns AMBER refinement of the best ROSETTALIGAND model [Fig.6(e)] of the substrate binding site. The dashed line from 5-HT to the aromatic ring of Y176 marks a T-type ring stacking interaction. The gray-shaded areas highlight major differences of the hSERT and dSERT models in the substrate binding site: (I) The A441/D164 hydrogen bonding interactions with the 5-OH position of 5-HT. (II) I172/M167 packing interactions with 5-HT indole ring. Panels (c) and (d) show 3D representations of the Down binding mode in hSERT and dSERT models. Reprinted with permission from (Kaufmann, Dawson et al. 2009)

carbonyl oxygen while in the hSERT the 5-OH of 5-HT forms transient hydrogen bonds to the backbone oxygens of residue A169 (dSERT D164) and A441 (dSERT G432).

Discussion

The present study examines two primary questions; “Can docking of 5-HT into comparative models of SERTs identify a physiologically relevant binding mode consistent with known mutagenesis, SCAM, and SAR data?” and “If so, what are the implications for SERT substrate recognition?” Computational docking on its own is unlikely to present a single correct solution due to the errors inherent in comparative models (Kairys, Fernandes et al. 2006). However, docking to comparative models may yield a physiologically relevant binding mode (DeWeese-Scott and Moult 2004). Functional conservation, sequence identity, and biochemical structural data all indicate promising potential for comparative models based on $LeuT_{Aa}$ structure. Chothia and Lesk found that functional conservation of proteins often implies a higher structural conservation than sequence identity would imply (Chothia and Lesk 1986). In a study of comparative modeling for membrane proteins, Forrest et al. reported that sequence identities above 30% in the transmembrane domains yield models with Cα-RMSD of ~2 Å (Forrest, Tang et al.) to the true structure. Biochemical structural information such as the SCAM profiles of TMs 1, 3, and 10 in SERTs are consistent with the $LeuT_{Aa}$ structure (Beuming, Shi et al.).

No single model resulting from this process is guaranteed to satisfy all the biochemical data available. However, in our study unbiased sampling of possible binding modes produced a single binding mode in line with all biochemical data. The collective satisfaction of these constraints indicates the physiological relevance of the Down binding mode shown in Fig. 7e and Fig. 10. For example, in the Down mode residues I172 and Y176 are protected from MTS modification and subsequent inactivation of transport. Only bulky or charged mutations at I172 have a significant effect on 5-HT transport (Henry, Field et al. 2006), indicating a purely steric impact of this position on the binding site as is indicated by the packing against the side of the indole ring. The hSERT G100A mutant is transport deficient, but maintains an unperturbed binding affinity (Kristensen, Larsen et al. 2004). Since the Down binding mode lies below G100, G100A would not significantly perturb this binding mode. TM 10 residues cannot be protected from MTS attack and inactivation by 5-HT binding (Keller, Stephan et al. 2004). The Down

binding mode predicts this since it leaves TM 10 amino acids, that are sensitive to MTS modification, solvent accessible. Finally, the A441C mutant is protected from MTS access by 5-HT (Androutsellis-Theotokis and Rudnick 2002) in line with the proximity of A441 to the 5-OH group. The sum of all these experimental data points support the Down binding mode as a physiologically relevant placement for 5-HT in the binding site.

SVM sensitivity maps reveal differences in the sensitivities of dSERT and hSERT to substitution at the R3, 4, 5, and α positions (see Fig. 8). The R3 indole nitrogen displays sensitivity to bulky substituents in hSERT (Adkins, Barker et al.). An isopropyl substitution causes a significant decrease in transport, whereas a methyl substituent in the same position causes little difference in uptake. These data indicate the indole nitrogen likely faces a sterically restricted area in hSERT. The Down binding mode places the indole nitrogen R3 substituents proximal to Y176/Y171. Y176 has been shown to be important for transport (Chen, Sachpatzidis et al. 1997), thus it is not surprising the substitutions perturbing this residue are detrimental to transport. Adkins identified a mutant hSERT, Y95F, which minimizes this effect (Adkins, Barker et al. 2001). Since no direct contact between R3 substituents and Y95 is seen in our models, we hypothesize an indirect effect: the tryptamine N-isopropyl substitution causes a shift in the indole ring towards the bottom of the pocket where Y95 is located in hSERT (F90 in dSERT). Mutation at position 95 allows for a structural rearrangement that accommodates additional bulk at the indole nitrogen position. If this is the case, then bulk reducing mutations at neighboring residues, such as V343, L344, and A441, could have a similar effect and serve to test our hypothesis. In contrast to hSERT, the intracellular base of the binding site in dSERT exhibits a more polarizable nature (e.g. hydrophobic to polarizable I172/M167, V343/T335 hydrophobic to polar, and A169/D164 hydrophobic to charged see Fig. 9). The hydrogen bond seen between the 5-OH of 5-HT and the sidechain of D164 reinforces this view. Furthermore, sensitivity to substitution at positions 4 and 5 as shown in the SVM sensitivity maps agree with the Down binding mode by placing hydroxyl groups near V343/T335 and A169/D164 in the hSERT/dSERT (Fig. 8c and Fig. 10).

The Down binding mode merits experimental investigation given agreement with the above biochemical data. The difference in polarity in this region in combination with the Down mode placing the 5-OH in this region implies that dSERT and hSERT should exhibit a differential preference for polarity surrounding the 4 and 5 position of the tryptamine ring. Further studies with species switching mutations of the above residues will ascertain the role of these residues in substrate specificity for 4 and 5 position tryptamine derivatives. Since the sparseness in the dataset for substitutions at α, R3, and 4 limits the further analysis of determinants of sensitivities to substitution at these positions, uptake and binding assays experiments with additional substrates modified at these positions should be useful in the context of our models.

The Down binding mode places the indole ring such that the 6 and 7 positions of the tryptamine core point towards the interface between TM 8 and TM 3. The amino acid identities of residues at this interface do not change significantly in hSERT and dSERT. However, future experiments with site directed mutants in this region may verify the orientation of indole ring of the Down binding mode. One prediction is that an hSERT T439A mutant would display differential recognition of polarity switching substitutions at the 7 position on the tryptamine core. Additional hSERT mutants, such as G442S, A173S, and A169S, would impact recognition of 6 position substituted tryptamines with varied hydrogen bonding capabilities. Assessing the function of these mutants in both hSERT and dSERT backgrounds could validate the assumption of a conserved mode for tryptamines in SERTs. Should the assumption prove incorrect, this constraint on the binding mode selection could be changed to find modes consistent with new experimental findings.

Despite the advances made with the current models, much still remains unknown. The $LeuT_{Aa}$ structure captures but one state in a multistep transport process. Structures of other states in the transport process are needed to fully understand species selectivity for substrates. Additionally, the $LeuT_{Aa}$ structure lacks a chloride in the binding site known to be required for function of the SERT. Studies are forthcoming to elucidate mechanism of chloride coupling in transport.

Jorgensen et. Al (Jorgensen, Tagmose et al. 2007) independently performed a manual docking and molecular dynamics study with 5-HT in hSERT. Interestingly, the binding mode identified is similar to our Down mode. Celik et al. recently reported a study on hSERT using the paired mutant-ligand analog complementation approach (Celik, Sinning et al. 2008). They report an alternate binding mode using this approach. Our approach places a lower priority on their proposed binding mode as it seems less consistent with the cross species sensitivities reported in the SVM sensitivity maps. We expect hSERT and dSERT to show differences in the amino-acids in regions surrounding the 5 position and the N position. Of course hSERT and dSERT could bind in different modes, but this is unlikely. Our study applies a different approach of comparing multiple tryptamine derivatives in both hSERT and dSERT, thereby identifying structural determinants of substrate specificity in these transporters.

Conclusion

Docking of 5-HT into hSERT and dSERT identifies a single conserved binding mode, in which the predicted binding energy of tryptamine derivatives correlates with inhibition uptake constants (R=0.72). The Down binding mode curls the ethylamine tail under F335 and S336 and orients the 5-OH group towards A169 with the indole nitrogen facing the top of the binding site covered by Y176. This binding mode correctly predicts, qualitatively, the decreased modification by SCAM reagents of cysteines substituted at I172, Y176, A441, and the extracellular half of TM 10 due to binding of 5-HT. The mode posits that polarity differences caused by A169D and V343T changes could be responsible for species selectivity observed for hSERT and dSERT recognition of tryptamine derivatives. As additional mutations in SERTs are produced and characterized, particularly in the context of substituted tryptamines, our models should be capable of local refinement to even more precisely focus its utility.

References

Adkins, E. M., E. L. Barker, et al. (2001). "Interactions of tryptamine derivatives with serotonin transporter species variants implicate transmembrane domain I in substrate recognition." Mol Pharmacol **59**(3): 514-523.

Androutsellis-Theotokis, A. and G. Rudnick (2002). "Accessibility and conformational coupling in serotonin transporter predicted internal domains." J Neurosci **22**(19): 8370-8378.

Baker, D. and A. Sali (2001). "Protein structure prediction and structural genomics." Science **294**(5540): 93-96.

Barker, E. L., K. R. Moore, et al. (1999). "Transmembrane domain I contributes to the permeation pathway for serotonin and ions in the serotonin transporter." J Neurosci **19**(12): 4705-4717.

Barker, E. L., M. A. Perlman, et al. (1998). "High affinity recognition of serotonin transporter antagonists defined by species-scanning mutagenesis. An aromatic residue in transmembrane domain I dictates species-selective recognition of citalopram and mazindol." J Biol Chem **273**(31): 19459-19468.

Bayly, C. I., P. Cieplak, et al. (1993). "A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model." Journal of Physical Chemistry **97**(40): 10269-10280.

Bernstein, F. C., T. F. Koetzle, et al. (1977). "The Protein Data Bank: a computer-based archival file for macromolecular structures." J Mol Biol **112**(3): 535-542.

Beuming, T., L. Shi, et al. (2006). "A comprehensive structure-based alignment of prokaryotic and eukaryotic neurotransmitter/Na+ symporters (NSS) aids in the use of the LeuT structure to probe NSS structure and function." Mol Pharmacol **70**(5): 1630-1642.

Blakely, R. D., H. E. Berson, et al. (1991). "Cloning and expression of a functional serotonin transporter from rat brain." Nature **354**(6348): 66-70.

Bleckmann, A. and J. Meiler (2003). "Epothilones: Quantitative structure activity relations studied by support vector machines and artificial neural networks." Qsar & Combinatorial Science **22**(7): 722-728.

Canutescu, A. A. and R. L. Dunbrack (2003). "Cyclic coordinate descent: A robotics algorithm for protein loop closure." Protein Science **12**(5): 963-972.

Celik, L., S. Sinning, et al. (2008). "Binding of Serotonin to the Human Serotonin Transporter. Molecular Modeling and Experimental Validation." J Am Chem Soc.

Chang, C.-C. and C.-J. Lin (2001). LIBSVM: a library for support vector machines.

Chen, J. G., S. Liu-Chen, et al. (1997). "External cysteine residues in the serotonin transporter." Biochemistry **36**: 1479-1486.

Chen, J. G., S. Liu-Chen, et al. (1998). "Determination of external loop topology in the serotonin transporter by site-directed chemical labeling." J Biol Chem **273**(20): 12675-12681.

Chen, J. G., A. Sachpatzidis, et al. (1997). "The third transmembrane domain of the serotonin transporter contains residues associated with substrate and cocaine binding." J Biol Chem **272**(45): 28321-28327.

Chothia, C. and a. M. Lesk (1986). "The relation between the divergence of sequence and structure in proteins." The EMBO journal **5**: 823-826.

DeLano, W. L. (2002). The PyMOL Molecular Graphics System. San Carlos, CA, USA, DeLano Scientific.

DeWeese-Scott, C. and J. Moult (2004). "Molecular modeling of protein function regions." Proteins **55**(4): 942-961.

Ferrara, P., H. Gohlke, et al. (2004). "Assessing scoring functions for protein-ligand interactions." J Med Chem **47**(12): 3032-3047.

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): 508-517.

Henry, L. K., E. M. Adkins, et al. (2003). "Serotonin and cocaine-sensitive inactivation of human serotonin transporters by methanethiosulfonates targeted to transmembrane domain I." J Biol Chem **278**(39): 37052-37063.

Henry, L. K., L. J. Defelice, et al. (2006). "Getting the message across: a recent transporter structure shows the way." Neuron **49**(6): 791-796.

Henry, L. K., J. R. Field, et al. (2006). "Tyr-95 and Ile-172 in transmembrane segments 1 and 3 of human serotonin transporters interact to establish high affinity recognition of antidepressants." J Biol Chem **281**(4): 2012-2023.

Hoffman, B. J., E. Mezey, et al. (1991). "Cloning of a serotonin transporter affected by antidepressants." Science **254**(5031): 579-580.

Jorgensen, A. M., L. Tagmose, et al. (2007). "Molecular Dynamics Simulations of Na(+)/Cl(-)-Dependent Neurotransmitter Transporters in a Membrane-Aqueous System." ChemMedChem.

Just, H., H. H. Sitte, et al. (2004). "Identification of an additional interaction domain in transmembrane domains 11 and 12 that supports oligomer formation in the human serotonin transporter." The Journal of biological chemistry **279**: 6650-6657.

Kairys, V., M. X. Fernandes, et al. (2006). "Screening drug-like compounds by docking to homology models: a systematic study." J. Chem. Inf. Model **46**: 365–379.

Kaufmann, K. W., E. S. Dawson, et al. (2009). "Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies." Proteins **74**(3): 630-642.

Keller, P. C., 2nd, M. Stephan, et al. (2004). "Cysteine-scanning mutagenesis of the fifth external loop of serotonin transporter." Biochemistry **43**(26): 8510-8516.

Kortemme, T. and D. Baker (2002). "A simple physical model for binding energy hot spots in protein-protein complexes." Proceedings of the National Academy of Sciences of the United States of America **99**: 14116-14121.

Kristensen, A. S., M. B. Larsen, et al. (2004). "Mutational scanning of the human serotonin transporter reveals fast translocating serotonin transporter mutants." Eur J Neurosci **19**(6): 1513-1523.

Kuhlman, B., G. Dantas, et al. (2003). "Design of a novel globular protein fold with atomic-level accuracy." Science (New York, N.Y.) **302**: 1364-1368.

Meiler, J. and D. Baker (2006). "ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility." Proteins **65**(3): 538-548.

Meiler, J. and D. Baker (2006). "ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility." Proteins **65**: 538-548.

Misura, K. M. S. and D. Baker (2005). "Progress and challenges in high-resolution refinement of protein structure models." Proteins-Structure Function and Bioinformatics **59**(1): 15-29.

Ramamoorthy, S., A. L. Bauman, et al. (1993). "Antidepressant- and cocaine-sensitive human serotonin transporter: molecular cloning, expression, and chromosomal localization." Proc Natl Acad Sci U S A **90**(6): 2542-2546.

Ravna, A. W. and O. Edvardsen (2001). "A putative three-dimensional arrangement of the human serotonin transporter transmembrane helices: a tool to aid experimental studies." J Mol Graph Model **20**(2): 133-144.

Ravna, A. W., M. Jaronczyk, et al. (2006). "A homology model of SERT based on the LeuT(Aa) template." Bioorg Med Chem Lett **16**(21): 5594-5597.

Ravna, A. W., I. Sylte, et al. (2003). "Molecular mechanism of citalopram and cocaine interactions with neurotransmitter transporters." J Pharmacol Exp Ther **307**(1): 34-41.

Rodriguez, G. J., D. L. Roman, et al. (2003). "Distinct recognition of substrates by the human and Drosophila serotonin transporters." J Pharmacol Exp Ther **306**(1): 338-346.

Rohl, C. A., C. E. Strauss, et al. (2004). "Modeling structurally variable regions in homologous proteins with rosetta." Proteins **55**(3): 656-677.

Rohl, C. A., C. E. Strauss, et al. (2004). "Protein structure prediction using Rosetta." Methods Enzymol **383**: 66-93.

Roman, D. L., S. N. Saldana, et al. (2004). "Distinct molecular recognition of psychostimulants by human and Drosophila serotonin transporters." J Pharmacol Exp Ther **308**(2): 679-687.

Rothman, R. B. and M. H. Baumann (2002). "Therapeutic and adverse actions of serotonin transporter substrates." Pharmacol Ther **95**(1): 73-88.

Summa, C. M. and M. Levitt (2007). "Near-native structure refinement using in vacuo energy minimization." Proc Natl Acad Sci U S A **104**(9): 3177-3182.

Vapnik, V. N. (1998). Statistical learning theory. New York, Wiley.

Wang, J., P. Cieplak, et al. (2000). "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?" Journal of Computational Chemistry **21**(12): 1049-1074.

Yamashita, A., S. K. Singh, et al. (2005). "Crystal structure of a bacterial homologue of Na+/Cl--dependent neurotransmitter transporters." Nature **437**(7056): 215-223.

Yarov-Yarovoy, V., D. Baker, et al. (2006). "Voltage sensor conformations in the open and closed states in ROSETTA structural models of K(+) channels." Proc Natl Acad Sci U S A **103**(19): 7292-7297.

Zhang, Y. W. and G. Rudnick (2006). "The cytoplasmic substrate permeation pathway of serotonin transporter." J Biol Chem **281**(47): 36213-36220.

Appendix A

Support Vector Machine Sensitivity Map Encoding Scheme

The encoding scheme generates a 24 bit binary number indicating the substituents on the base tryptamine core (see Fig. 2) recognized by SERTs. Table A1 shows the encoding scheme for each tryptamine analog

.

Table A1 SVM Sensitivity Map Encoding Scheme																											
			X	R3			2	4	5					6				7						α	amine		
	hSERT scaled logKi	dSERT scaled logKi	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
5-hydroxy-tryptamine	0.181	0.215	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5-hydroxy-7-methoxy-tryptamine	0.43	0.582	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1
5,7-dihydroxytryptamine	0.492	0.495	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1
3-(beta-Aminoethyl)-5-hydroxybenzothiophene	0.341	0.367	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4-hydroxy-tryptamine	0.453	0.699	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7-hydroxytryptamine	0.43	0.492	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1
tryptamine	0.192	0.355	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5,6,7-trihydroxytryptamine	0.938	1	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0	1
1-methyltryptamine	0.302	0.351	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5-Methoxytryptamine	0.572	0.529	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
N,N-dimethyltryptamine	0.162	0.337	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
6-Methoxytryptamine	0.252	0.215	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1
5-Methoxy-N,N-dimethyl-tryptamine	0.439	0.457	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
6-Fluorotryptamine	0.144	0.225	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1
5-Methyltryptamine	0.391	0.543	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Serotonin o-sulfate	0.939	0.816	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1-Methylserotonin	0.325	0.415	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7-Benzyloxytryptamine	0.235	0.066	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1
7-Methyltryptamine	0.041	0.266	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
5-Hydroxytryptophol	0.774	0.868	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2-Methyl-5hydroxy-tryptamine	0.405	0.606	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
alpha-Methyltryptamine	0.215	0.138	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
N-isopropyl-tryptamine	0.663	0.225	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5-Methoxy-N-isopropyl-tryptamine	0.726	0.316	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
5-Carboxamidotryptamine	0.628	0.57	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5,6-Dihydroxytryptamine	0.595	0.603	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1

element 1 encode 0 if indole nitrogen is N else encode 1	element 13 encode 1 if 6 position is OH else encode 0
element 2 encode 0 if substituent of R3 postion is H else encode 1	element 14 encode 1 if 6 position is methoxyl group else encode 0
element 3 encode 1 in R3 substituent is methyl else encode 0	element 15 encode 1 if 6 position is F else encode 0
element 4 encode 1 in R3 substituent is isopropyl else encode 0	element 16 encode 0 if 7 position is H else encode 1
element 5 encode 0 if 2 position is H else encode 1	element 17 encode 1 if 7 position is OH else encode 0
element 6 encode 0 if 4 position is H else encode 1	element 18 encode 1 if 7 position is methyl else encode 0
element 7 encode 0 if 5 position is H else encode 1	element 19 encode 1 if 7 position is methoxyl group else encode 0
element 8 encode 1 if 5 position is OH else encode 0	element 20 encode 1 if 7 position is benzoxyl else encode 0
element 9 encode 1 if 5 position is amide else encode 0	element 21 encode 0 if alpha position is H else encode 1
element 10 encode 1 if 5 position is sulphonate else encode 0	element 22 encode 1 if amine has diethyl substituents else encode 0
element 11 encode 1 if 5 position is methoxyl group else encode 0	element 23 encode 1 if amine has been substituted to an OH else encode 0
element 12 encode 0 if 6 position is H else encode 1	element 24 encode 1 if amine is a primary amine else encode 0