# Computational design of protein-small molecule interfaces

CrossMark

Brittany Allison [a], Steven Combs [a], Sam DeLuca [b], Gordon Lemmon [b], Laura Mizoue [c,f], Jens Meiler [a,b,d,e,f,g,]*

[a] Department of Chemistry, 7330 Stevenson Center, Station B 351822, Nashville, TN 37235, USA
[b] Chemical and Physical Biology Program, 340 Light Hall, Nashville, TN 37232, USA
[c] Department of Biochemistry, 607 Light Hall, Nashville, TN 37232, USA
[d] Department of Pharmacology, 476 Robinson Research Building, 2220 Pierce Avenue, Nashville, TN 37232, USA
[e] Department of Biomedical Informatics, 400 Eskind Biomedical Library, 2209 Garland Ave, Nashville, TN 37232, USA
[f] Center for Structural Biology, 465 21st Avenue South, Nashville, TN 37232, USA
[g] Institute for Chemical Biology, 896 Preston Research Building, Nashville, TN 37232, USA

## ABSTRACT

The computational design of proteins that bind small molecule ligands is one of the unsolved challenges in protein engineering. It is complicated by the relatively small size of the ligand which limits the number of intermolecular interactions. Furthermore, near-perfect geometries between interacting partners are required to achieve high binding affinities. For apolar, rigid small molecules the interactions are dominated by short-range van der Waals forces. As the number of polar groups in the ligand increases, hydrogen bonds, salt bridges, cation–π, and π–π interactions gain importance. These partial covalent interactions are longer ranged, and additionally, their strength depends on the environment (e.g. solvent exposure). To assess the current state of protein-small molecule interface design, we benchmark the popular computer algorithm Rosetta on a diverse set of 43 protein–ligand complexes. On average, we achieve sequence recoveries in the binding site of 59% when the ligand is allowed limited reorientation, and 48% when the ligand is allowed full reorientation. When simulating the redesign of a protein binding site, sequence recovery among residues that contribute most to binding was 52% when slight ligand reorientation was allowed, and 27% when full ligand reorientation was allowed. As expected, sequence recovery correlates with ligand displacement.

© 2013 Elsevier Inc. All rights reserved.

## 1. Introduction

Engineering protein-small molecule interactions is key for advancement of several grand challenges in computational biology. Protein-small molecule interactions are the basis for enzymatic catalysis, receptor–small molecule signaling, and transporter selectivity and are thus essential for carrying out biological processes and maintaining overall homeostasis in the body. Designed proteins that bind small molecule targets can act as therapeutics by sequestering ligands, stimulating or extinguishing signaling pathways, delivering other molecules to sites of action, and serving as *in vivo* diagnostics (Golan et al., 2008). For example, small molecule depletion has been suggested as a strategy for treatment of prostate cancer (Knudsen and Scher, 2009), cocaine abuse (Zhu et al., 2006), and bacterial infection (Clifton et al., 2009). Proteins that bind small molecules also have applications

in environmental chemistry and food chemistry as biosensors (Baeumner, 2003). Thus, the ability to engineer highly precise and specific interactions at protein interfaces can serve in many capacities.

Computational design of protein-small molecule interfaces continues to present challenges. Although the creation of new enzymes is a landmark achievement in protein design (Baker et al., 2008a,b, 2010; Zanghellini et al., 2006), the success rate is low and the designed proteins are poor catalysts compared to naturally-occurring enzymes. To help pinpoint the causes, a systematic study was conducted introducing mutations into the active site of three designed retro-aldolases (RA34, RA45, and RA95) derived from the TIM-barrel scaffold IGPS. In RA34 and RA95, mutations that increase substrate binding affinity and thereby enzymatic activity involve increases in side chain volume and hydrophobicity, including G233F/I/V/Y in RA34 (Wang et al., 2012) and T51Y, T83K, S110H, M180F and R182M in RA95 (Althoff et al., 2012). In contrast, many improvements to the RA45 design arose from large-to-small mutations including W8A/T/V, F133L, V159C, and R182V/I (Althoff et al., 2012). In all cases, key functional groups that engage the ligand are introduced or removed. These observations indicate that neither the hydrophobic packing nor the

positioning of substrate within the binding pocket were optimal in the initial designs. Similarly, a previously reported successful computational design of a protein-small molecule interface (Allert et al., 2004) did not withstand close examination (Hayden, 2009; Schreier et al., 2009).

Rosetta, a protein modeling software suite for protein structure prediction and design (Schueler-Furman et al., 2005), has been successfully used to tackle a number of interface design problems. Some of these successes include creating novel enzymes (Baker et al., 2008a,b, 2010), altering the specificity of protein-peptide (Sood and Baker, 2006), protein–DNA (Ashworth et al., 2010) and protein–protein interfaces (Kortemme et al., 2003), and designing proteins that bind a selected surface of a virus (Fleishman et al., 2011). Rosetta seeks to find the lowest energy conformation for a design by combining discrete side chain conformation (rotamer) optimization with Monte Carlo minimization (Schueler-Furman et al., 2005). This includes sampling random perturbations of the backbone torsion angles, rigid body degrees of freedom, and rotamer conformations, followed by an all-over local minimization to resolve clashes (Schueler-Furman et al., 2005). These methods enable much faster and larger exploration of sequence and conformational space compared to experimental methods such as phage display (Weng and DeLisi, 2002).

The energy function that Rosetta uses to discriminate between native-like and non-native-like atom arrangements includes a van der Waals-like attractive and repulsive potential, solvation term, hydrogen bonding potential, electrostatics potential, rotamer probability, and ($\varphi$, $\psi$) angle probabilities in the protein backbone (Meiler and Baker, 2006). The total energy of the system is computed as a weighted sum of all interactions with weights optimized through a series of benchmarks. All energy functions are pairwise decomposable (i.e. they depend on no more than two interacting partners). This design of the energy function maximizes algorithm speed since interaction energies can be pre-computed and stored. However, it also limits the accuracy of the energy function, particularly electrostatic and partial covalent interactions which vary greatly in strength depending on the environment of the interacting partners. Experimental characterization of some of the best scoring designs is used to validate and improve the computational protocols. In this way, both design successes and failures help test and expand our understanding of the fundamental forces involved in molecular recognition.

RosettaLigand is an application within Rosetta that was originally developed to dock small molecules into a protein with full protein and ligand flexibility (Baker and Davis, 2009; Meiler and Baker, 2006). In these studies, we expand RosettaLigand to include amino acid optimization (design) at the protein-small molecule interface. Using the full-atom energy function and Monte Carlo minimization procedure, RosettaLigand optimizes the small molecule position and protein side chain rotamers simultaneously (Meiler and Baker, 2006). RosettaLigand allows for protein backbone flexibility, side chain rotamer searching, and full ligand flexibility, all of which are necessary for accurately modeling the interface (Baker and Davis, 2009; Meiler and Baker, 2006). Fig.1 details each step of the ligand docking protocol. For each model, RosettaLigand calculates an 'interface energy' as the total score of the protein–ligand complex minus the total score of the apo-protein (Meiler et al., 2009). The accuracy of models in terms of ligand placement is determined by computing the root-mean-square distance (RMSD) over all ligand atoms between model and co-crystal structure. RosettaLigand is the foundation (Zanghellini et al., 2006) of a number of the successfully design enzymes (Jiang et al., 2008; Rothlisberger et al., 2008; Siegel et al., 2010) with the before-mentioned caveat that the computationally predicted residues are often sub-optimal even in the first shell surrounding the ligand. In order to understand its capabilities and limitations, the present work systematically
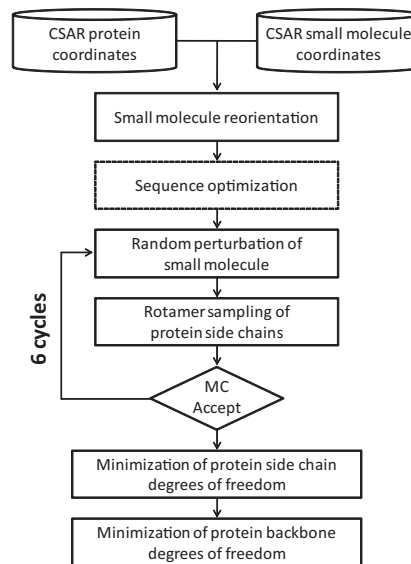


**Fig.1.** Flowchart of small molecule docking with design. The RosettaLigand protocol was modified to include interface design (dotted line box). From the input coordinates, the small molecule is allowed to rotate and translate before sequence optimization of nearby residues. After 6 cycles of small molecule perturbation, side chain rotamer sampling, and Monte Carlo (MC) minimization, a final gradient-based minimization of the protein is performed to resolve any clashes.

assesses RosettaLigand's ability to design protein-small molecule interfaces. This analysis is an important, and so far omitted, benchmark to identify design challenges that can currently be solved and to work towards improvements needed to achieve consistent success.

Recovering native protein-small molecule interfaces in sequence and conformation is a benchmark for designing novel interfaces. Creating new interfaces or even modifying existing ones requires computational tools that sample and select native-like interactions. In this study, we examine how RosettaLigand performs in sequence recovery within protein-small molecule interfaces while allowing for small molecule reorientation and side chain conformational changes. The benchmark consists of two parts. Part one tests overall sequence recovery when all residues within the protein-small molecule interface are allowed to change identity. Part two simulates a protein-smallmolecule design more closely by mutating up to five residues that contribute most to the interaction with the small molecule to alanine. This effectively removes the binding site's memory of the native ligand. In the design experiment a scoring bonus is given to the starting sequence. These experiments test RosettaLigand's ability to distinguish between native and non-native binding interaction and whether RosettaLigand can identify key mutations needed to bind the small molecule while limiting the total number of mutations. The results illustrate the types of ligands that Rosetta handles best and provide insights into weaknesses where continued method development is required.

## 2. Results and discussion

The setup of the experiments allows us to determine overall protein–ligand interface sequence recovery as well as an optimal strategy for re-designing proteins to recognize different small molecules using a minimal set of mutations. For this purpose separate measures for sequence recovery among the residues critical for ligand binding are determined. We investigate how sequence recovery varies with ligand size, binding affinity, and RosettaLigand interface energy. We appreciate that sequence recovery measures