This is to certify that the
thesis entitled

OPTIMIZING SIDE-CHAIN INTERACTIONS IN
PROTEIN-LIGAND INTERFACES

presented by

SAMEER ARORA

has been accepted towards fulfillment
of the requirements for the

M.S. degree in Computer Science

Major Professor's Signature

7/8/05

Date

OPTIMIZING SIDE-CHAIN INTERACTIONS IN PROTEIN-LIGAND INTERFACES

By

Sameer Arora

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Department of Computer Science

2005

ABSTRACT

OPTIMIZING SIDE-CHAIN INTERACTIONS IN PROTEIN-LIGAND INTERFACES

By

Sameer Arora

Proteins bind to other proteins or small molecules to perform essential cellular functions. Protein side-chain flexibility is crucial for binding and molecular recognition. Hence modeling side-chain flexibility in protein-ligand docking algorithms to predict the optimal inter- and intra-molecular interactions is extremely desirable. However, modeling side-chain flexibility in docking and screening is computationally expensive due to the numerous side chains and their many degrees of freedom.

Our research indicates that direct, intra-protein hydrogen bonds and hydrophobic interactions are preserved to a significant extent upon ligand binding. This provides guidance to restrict the number of side-chain candidates for conformational sampling in docking. While these bond-preservation tendencies limit side-chain movements, large side-chain motions are also observed in the protein-ligand interface. Subsequently, the extent of these large side-chain motions from ligand-free to ligand-bound crystal-structure conformations are characterized, as is the suitability of using backbone-dependent rotamers for sampling these larger motions.

The ability to accurately identify which side chains move significantly upon ligand binding as well as their optimal conformations is crucial for docking and screening. A new scoring function, having good linear correlation with experimentally determined protein-ligand binding affinities, is presented for scoring dockings and side-chain interactions by SLIDE. Using the new scoring function as a cost measure, a mean-field based algorithm, exploiting rotamer-based side-chain flexibility modeling, is proposed and tested for optimizing interactions in protein-ligand interfaces.

To my Mummyji and Papaji. It is your endless love, sacrifices, committment and support, that I am here, forever indebted.

ACKNOWLEDGEMENTS

First, I would like to express my deep gratitude for my advisor, Dr. Leslie A. Kuhn, for all hers support, guidance and encouragement throughout my graduate education. Her boundless energy and infectious enthusiasm for science helped me embark and persist on this research work. I am grateful for her mentorship, and feel lucky to have been associated with such a person, whose optimism, patience and dynamism continue to set high standards for me.

I want to thank my co-advisor, Dr. Bill Punch. His advice and guidance strengthened my decision and faith in pursuing this interdisciplinary research. And his support and patience helped me continue forward with a calm and positive mind.

I extend my thanks to Dr. Phil Duxbury. My discussions with him always left me with valuable insight and a physicist's perspective on our research problem.

I would also like to thank the past and present members of the Kuhn lab, Litian He, Chetan Sukuru, Rajesh Korde, Dr. Brandon Hespenheide, and Sandeep Namilikonda for creating an enjoyable and stimulating research environment. I am especially thankful to Dr. Mária Závodszky, who helped me grow, besides teaching me many things over the years, including Biochemistry and SLIDE. She was always ready to help, discuss ideas and answer countless (quick) questions.

Finally, I would like to thank my friends (especially Kantha Kumar) and family, without whose support and encouragement, it would be hard to imagine myself here.

Table of Contents

List of Tables

List of Algorithms

List of Figures

Images in this thesis/dissertation are presented in color.

Chapter 1

Flexibility in Protein-Ligand Docking

1.1 Introduction: Protein-Ligand Docking

Like people interact and cooperate with one another to perform many functions for the sustenance and growth of life at the social level, proteins too interact and cooperate with each other as well as with other molecules for the sustenance and growth of life at the cellular level. Proteins perform various vital functions in a cell - they provide cellular structure, bind and transport other proteins or organic compounds, and catalyze or inhibit reactions. Underlying all stable bindings between a protein and another molecule lies the mechanism for the two molecules to recognize each other and achieve a bound state more stable than their individual unbound states. Understanding the mechanism of protein binding is key understanding their function as well providing valuable insights for discovering novel compounds that can bind to specific proteins for designing therapeutic drugs.

Molecular docking is a term used to describe computational techniques that attempt to find the "best" mode binding between two molecules. Protein-ligand docking

aims at finding the optimal binding between a protein and small molecule to a specific site of the protein. In protein-ligand docking, the atomic structures of the two molecules are given as input in the most general form, no additional data is provided. However, in practice, additional biochemical information can be given, specifically, information about the location of the binding site.

Docking methods can be categorized in multiple ways. "Redocking" attempts to reconstruct a complex using bound structures of the receptor and the ligand. More challenging is the "unbound" docking, which attempts to reconstruct a complex using unbound structures of the receptor and the ligand. In this case, some degree of conformational change in the protein and the ligand must be modeled or accommodated. Another categorization may be based on flexibility: "rigid" docking keep the structures rigid during the docking, while flexible docking techniques allow flexibility in the receptor or ligand or both.

Typically, there are three key ingredients in docking algorithms:

1. representation of the molecular system
2. conformational and orientational space search ("sampling")
3. ranking of potential solutions ("scoring")

The present work focuses on modeling protein side-chain motion in docking, and the subsequent sections of this Introduction cover how flexibility has been modeled by others in a variety of protein modeling methods.

1.2 Overview of Protein-Structure Prediction Methods

Over several years, various techniques have been developed to predict protein structures from their amino acid sequences. These prediction techniques attempt to

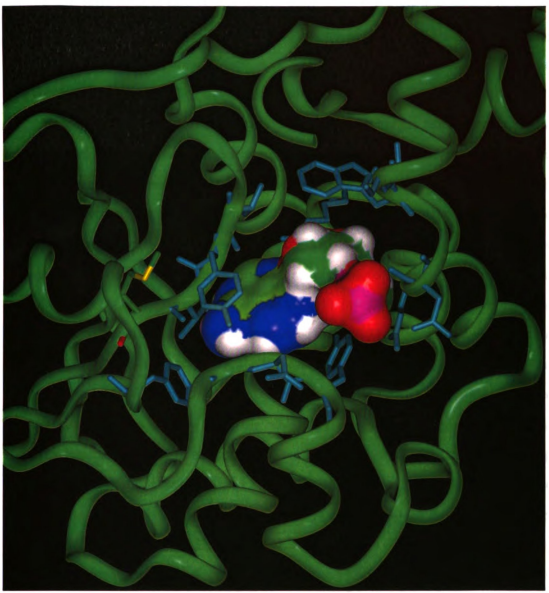


Figure 1.1: *This image is presented in color.* Example of a protein-ligand crystal complex. Crystal structure of α-momorcharin complexed with formycin 5′-monophosphate. Green ribbon represents the protein backbone while the ligand is displayed with Connolly solvent-accessible molecular surface, colored by atom type. Only binding-site side chains are displayed (in tubes) with interfacial side-chains colored cyan and side chains beyond the interface colored by atom type.

define the protein structure in the native state from the known sequence of amino acids, and in some cases aim to capture dynamics of processes like protein folding and protein-ligand or protein-protein docking. While each of these methods is used for structure prediction, they have been used in context of dockings as well.

Molecular Dynamics (MD) Simulation is one such technique that serves as an important tool in protein structural dynamics and refinement. It uses torsional degrees of freedom in proteins and ligands and key physical properties at the atomic level to solve Newton's equations of motion. Forces on individual atoms are represented as sums of potential terms (e.g. electrostatic and van der Waals) during the entire simulation. Each simulation run is divided into time steps, which are typically scaled small enough (1-2 femtoseconds) to ensure that the physical interactions are modeled accurately. While smaller time steps, combined with an accurate force field, lead to a better quality of simulation results, the large number of steps required makes the simulation extremely slow for docking and screening purposes, as each time step involves thousands of degrees of freedom in the molecule(s), requires evaluation of computationally expensive potential energy terms, and must also account for the surrounding solvent. The computational intensity also limits the amount of conformational space sampled for both protein and ligands. To optimize side-chain placement in docking, sometimes less computationally demanding energy minimization is used, but this involves very local motions. Carlson and colleagues[2] have used MD to determine relatively rigid regions from multiple crystal structures and focus on these immobile regions as templates for drug design.

Monte Carlo (MC) sampling is a stochastic sampling method applied to a number of diverse domains, including economics, physics, traffic regulation and others. In molecular simulations, an MC simulation is often guided by simulated annealing. In simulated annealing, the "temperature" of the system is raised and then system is allowed to cool down gradually. Here, temperature implies any parameter of the system

that controls the magnitude of sampled motions. Like high an input of energy helps electrons jump to higher energy orbital in atoms, high "temperature" in MC simulations helps torsional rotations overcome high energetic barriers. Once the systems has been allowed to sample a variety of states, temperature is dropped in a step-wise manner to allow the system to relax into a low energy state. This relaxation may be guided by a force field or could result from a random move. New conformations are accepted if the energy drops. the simulation stops when the temperature falls below a threshold. Since there is some randomness to the generation of the final molecular complex, and since there are usually many protein-ligand conformations with similar energy, an ensemble of many structures is usually calculated to observe which binding modes appear most often. This means that this computationally intensive program must be repeated many times.

The Dead-End Elimination (DEE) approach relies on a potential energy function to evaluate discrete side-chain conformations. A DEE run searches for a combination of side-chain conformations that result in global energy minimum for the protein or protein-ligand complex[6]. Typically, a set of side-chains is identified for conformational sampling. The rest of the atoms provide a representation of the immediate environment and additional constraints for the sampling. The DEE implementations typically use rotamer libraries to generate side-chain conformations. The relatively high efficiency of this relies on a mathematical expression, the DEE criterion, which rules out infeasible rotamers that cannot be part of the global minimum conformations. One example is that if a certain χ angle value of a particular side chain is observed to cause unresolvable collisions with neighboring residues, then this χ value is ruled out in all future moves. This criterion helps reduce the number of rotamers to be considered for side chains in the modeling set, resulting in a number that is sufficiently small to analyze by means of "brute force" combinatorial analysis. The DEE approach is useful for homology modeling, where backbone structure is usually

known from a related protein. For docking and screening purposes, DEE scales well as long as the number of side-chains sampled is small. Schaffer et al [40] have successfully applied DEE for side-chain optimization after docking HIV-1 protease mutant complexes using Monte Carlo/simulated annealing sampling.

The latest version of Dunbrack's SCRWL algorithm[1] for side-chain modeling uses results from graph theory to solve the combinatorial problem encountered in side-chain prediction. In this method, side chains are represented as vertices in an undirected graph. Any two residues that have rotamers with nonzero interaction energies are considered to have an edge in the graph. The resulting graph can be partitioned into connected subgraphs with no edges between them. These subgraphs can in turn be broken into biconnected components, which are graphs that cannot be disconnected by removal of a single vertex. The combinatorial problem is reduced to finding the minimum energy of these small biconnected components representing sets of mutually acceptable rotamers for different side chains and combining the results to identify the global minimum energy conformation. While SCRWL's main usage is in homology modeling, *ab initio* protein structure prediction, and protein design applications, it can be used for side-chain modeling in the presence of ligands too. While such an approach may help develop induced fit, the ligand largely remains rigid in the process.

In mean-field optimization methods[38, 26], the part of the protein that requires modeling, like the side chains, is oversampled by replacing single copies by multiple different copies. The copies of each side chain do not interact with each other but "see" the other side chains as an average. The multiple copies essentially correspond to a distribution function of one side chain, represented by a number of discrete structures with assigned, often equal, weights. Mean-field optimization has been used both for side-chain modeling in homology modeling [38] as well as modeling flexible side chains during docking in SLIDE, a ligand screening and docking software developed in our

laboratory [42, 41]. Jackson et al[19] also attempt side-chain optimization in protein-protein interfaces through iterative cycles of mean-field optimization and rigid-body energy minimization. Rotamers from the rotamer library of Tuffrey et al [46] are used for discrete sampling of side-chain conformational space while the main-chain atoms have fixed coordinates. The self-consistent mean field approach is used to determine the most probable set of rotamers from an ensemble of rotamers. In this closed system, the potential mean force on each rotamer is based on the internal energy of the rotamer itself, rotamer-backbone interactions, interactions between the rotamer and rotamers of other residues, and the rotamer's interactions with the surrounding solvent. Each rotamer's interactions are calculated prior to the optimization, thus saving time during the mean-field optimization step.

Replacing a system of side chains by a multiple copies per side chain forming a mean-field system has two major advantages. First, while mean-field optimization complexity is lower than the exponential complexity of searching the entire conformational space, the global energy minimum of the new mean-field system is the same[38]. Secondly, the barriers separating the minima in the mean-field calculation are lower than in the original system, making annealing easier. However, the result of optimization does depend on initial conditions, like the probability distributions for the multiple copies of the side chains.

1.3 Overview of Current Side-Chain Modeling Approaches in Docking Tools

Accounting for side-chain reorientation during docking is similar to predicting side-chain conformations in homology modeling. The search for candidate solutions in a docking problem is addressed by two essentially different approaches: (1) a gradual guided search through solution space, or (2) a full solution space search. The first

either scans only part of the solution space in a partially random and partially criteria-guided manner. This approach consists mainly of Monte Carlo (MC), simulated annealing, molecular dynamics (MD), and evolutionary algorithms such as genetic algorithms (GA). In contrast, the second scans the entire solution space in a predefined systematic manner such as using rotamer-library or geometric-hashing based searches of conformations.

The classical algorithm implemented for computational docking is that of DOCK, the first ligand docking tool [7, 5, 9, 44]. DOCK operates by generating a set of spheres to describe the volume, or negative image, of the binding site and uses the centers of these spheres as sites for matching to ligand atoms. Sets of receptor spheres are matched to sets of ligand atoms to generate a ligand orientation, which can then be scored according to their complementarity with the protein. The early internal DOCK scoring function, GRID [32], is a grid-based scoring function in the method of Goodford and colleagues [13]. Later implementations have used more robust scoring functions which are also grid-based. It is possible to use the ligand docking method of DOCK with an externally supplied scoring function. The initial implementation of DOCK used only steric fit and electrostatics as a determinant for ligand docking, but later versions implemented chemical type matching to better model chemical complementarity between ligand and receptor groups, including hydrogen bonding interactions. It should be noted that DOCK uses only rigid-body translations and rotations, including no internal molecular flexibility within the docking algorithm. Recently, Shoichet and colleagues have used conformational ensembles as input to DOCK[30].

Another popular docking algorithm is AutoDock[33], which employs a Monte-Carlo simulated annealing method to sample binding orientations and ligand conformations, by randomized rotation of torsional angles in the ligand. AutoDock is inexpensive, easy to use, achieves reasonable dockings, and has been applied success-

fully to predict ligand binding modes in several cases[33].

Another stochastic approach based on the use of a genetic algorithms (GA) was developed by Jones and colleagues[21]. Their approach uses a simple GA operating on rotational angles in the protein, rotational angles in the ligand, and on hydrogen bonds between protein and ligand. The fitness or scoring function encompasses terms for the hydrogen bond energy between protein and ligand, for the van der Waals energy between protein and ligand, and for the internal van der Waals energy of contacts within the ligand. Improvements to this algorithm resulted in the development of the GOLD algorithm[22], with changes in the representation of angles and hydrogen bonds and inclusion of a more robust scoring function. The GOLD algorithm achieved "acceptable" dockings for 71 of 100 test cases. Similar to GA approaches is the evolutionary programming approach AGDOCK [11], developed at Agouron Pharmaceuticals, which is able to correctly dock (within 1.5 Å of correct position), methotrexate into dihydrofolate reductase (DHFR) and a proprietary ligand, AG-1343, into HIV protease.

To address side-chain and limited main-chain flexibility of the receptor structure in docking, FlexE [3], was created. FlexE docks ligands into an ensemble of structures of the receptor instead of a single receptor binding site. The binding site ensembles can be from different crystallographic structures, as in [3], from a homology model with uncertain side-chain positions, from a series molecular dynamics time steps, or from another source. The key component of this algorithm is that multiple conformations of the protein can be used as a docking target simultaneously. In this algorithm, the receptor structures are merged, with regions of similar conformations reduced to a single structure and regions with dissimilar conformation constituting alternate positions. While this algorithm may handle some backbone movement in addition to side-chain rotations, the authors claim it is not able to work with large domain movements and limit their test set to protein ensembles with similar backbone traces.

1.4 Side-Chain Modeling in SLIDE

The docking and screening software SLIDE (Screening for Ligands by Induced-fit Docking, Efficiently) was developed in our laboratory for screening and docking small molecules into specified binding sites of target proteins [43, 41, 42]. Figure 1.2 pictorially describes the algorithm used in SLIDE for screening and docking small molecules. Surfaces of both the protein and the molecule being screened are represented by sets of strategically located interaction points, each point representing the binding site's or the ligand's polar atom or hydrophobic characteristics. The protein's interaction points are called template points, while the potential ligand's interaction points are called interaction centers. Each of these points is assigned one of the descriptor types : hydrogen-bond *donor* if the protein or ligand atom can donate a proton in an hydrogen bond, *acceptor* if the atom can accept a hydrogen bond, *donor/acceptor* if the atom can donate and/or accept a hydrogen bond, and *hydrophobic* if the atom or point represents a hydrophobic site in the ligand or protein. Unlike the potential ligand's interaction-centers, template points represent the binding-site's "negative" image, or ideal ligand atom types to bind at the position. Screening and docking is based on matching triplets of ligand's interaction centers onto triangles of template points based on geometry as well as descriptor types. Since there can be $O(n!)$ template triangles from n template points, SLIDE uses a multi-level geometric hashing for matching interaction-point triangles to pre-computed template triangles stored in the hash table, indexed using the triangle's descriptor types and geometry. Geometric hashing was originally applied in object matching and identification in computer vision [49]. Combined with an adequate molecular surface representation, geometric hashing yields a state-of-the-art toolkit for docking[10, 35]. This indexing empowers SLIDE to exhaustively explore orientational space for the ligand with respect to the protein speedily and efficiently. The part of the ligand within the triangle of interaction centers is treated as rigid, and called an anchor fragment. It serves as the

anchor for docking the ligand into the binding site. Any remaining fragments of the ligand are considered flexible. Single bonds in the flexible parts of both the protein and ligand candidates are rotated as needed to remove inter-atomic collisions and to generate a shape-complementary interface, before the complex is scored by the number of intermolecular hydrogen bonds and the hydrophobic complementarity of the contact surfaces.

SLIDE was the first method to balance protein and ligand flexibility in docking while developing induced fit between the protein and ligand surfaces during the docking step. It has identified and correctly docked diverse, known ligands into the ligand-free conformation of the binding site for a variety of proteins, e.g., subtilisin, cyclodextrin glycosyltransferase, uracil DNA glycosylase, rhizopuspepsin, HIV protease, estrogen receptor, and Asn tRNA synthetase [43, 41, 42].

1.5 Motivation and Focus in this Thesis Work

Various groups have studied protein side-chain flexibility modeling in docking. Docking techniques using MD or MC simulations can model side-chain flexibility with good accuracy, however they are slow. Besides, MD is poor at crossing energetic barriers at reasonable temperatures, hence limited in its sampling conformational space. Genetic algorithms encoding flexibility tend to be too slow for use in screening many molecules by docking. Exhaustive search space sampling methods using discrete rotamer libraries are fast and efficient if the number of side chains are limited. Methods like SCRWL using rotamer libraries have good accuracies for predicting $\chi_{1,2}$ χ angles, however such predictions have been confined to homology modeling or *ab initio* protein design rather than docking. Because rotamer libraries specify rotamers as average conformations representing clusters of similar side-chain conformations, a realistic use of libraries requires expanding the rotamers by sampling around the

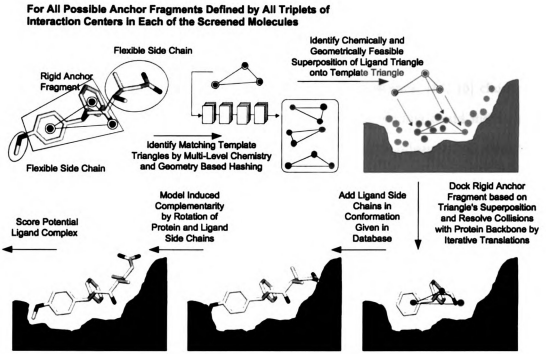


Figure 1.2: *This image is presented in color.* Screening and docking algorithm implemented in SLIDE [39]. SLIDEs docking of potential ligands into the binding site is based on mapping triplets of ligand interaction centers (hydrogen-bond donors, acceptors donor/acceptors, or hydrophobic atom centers) onto triangles of template points located above the protein surface. Feasible template triangles for each possible triplet in a screened molecule are directly accessed via a multi-level hash table, and the corresponding mapping is used to dock the rigid anchor fragment of the potential ligand. Single bonds in the flexible parts of both molecules are rotated to generate a shape-complementary interface, before the complex is scored by the number of intermolecular hydrogen bonds and hydrophobic complementarity of the contact surfaces. In all steps, the ligand triplets or dockings that do not meet a particular threshold are discarded.

dihedral angles within specified deviations[24]. DEE-based side-chain conformational search techniques are efficient but can miss global energy minima. The fuzzy-end elimination theorem[28, 25] corrects this problem but the search space becomes huge.

Some groups have studied specific side-chain flexibility aspects, such as the number of interfacial side chains undergoing large conformational changes or whether the side chains remain rotameric after ligand binding. For instance, [14, 15] claim that interfacial side chains are not always rotameric as ligand binding could induce non-rotamericity in the binding-site residues. Najmanovich and colleagues[34] show that for 85% of the binding pockets they studied, 3 or fewer side chains underwent a dihedral rotation of more than 30°. Knowledge about such proclivity might help restrict the conformational sampling in an exhaustive search without loosing the ability to model substantial side-chain conformational changes; and hence is very valuable for to improve docking algorithms.

This research was motivated to develop techniques that accurately model side chain positioning during docking to optimize the interactions in the protein-ligand interfaces. Optimal interactions, which may be defined as interactions observed in crystal complexes, play significant role in molecular recognition. Ability to predict optimal interactions in SLIDE would require developing an algorithm that not only samples larger side-chain conformational space than sampled by the current induced-fit collision resolution paradigm, but can also correctly identify those side chains for which bigger motions need to be sampled.

1.6 Organization of the Thesis

This thesis addresses the questions of which side chains are observed to undergo large rotations upon ligand binding (and why), how to sample such motions and the extent to which rotameric sampling and protein-ligand scoring facilitate or limit our

ability to model these motions accurately. Chapter 2 presents ana analysis of intra-protein noncovalent bond preservation upon ligand binding, conclusions from which help identify the side chains which undergo large motions during docking. Chapter 3 analyzes the rotamer-based sampling for these side chains. Of the many conformations available for a collection of side chains, a docking tool should be able to identify conformations that optimize interactions and rank them highly. Chapter 4 presents developing a new scoring function to predict the affinity of a protein-ligand complex with state-of-the-art accuracy and applies it to select from high-probability side-chain rotamers. Chapter 5 presents the side-chain interaction optimization algorithm in SLIDE and subsequent results. Chapter 6 concludes the thesis with a brief summary and a discussion on future directions of this work.

Chapter 2

Hydrogen Bond Preservation in Protein Binding Sites

2.1 Motivation

Predicting the structure of the complex of a small ligand with a protein is still a complex task, two major problems being the definition of an appropriate scoring function to discriminate good binding modes of the ligand among all possible binding modes and the huge size of the binding-mode search space itself. Assuming that one knows the correct scoring function, a successful search procedure should consider the three factors that give rise to the size of search space:

- sampling the relative orientations of ligand and receptor,
- sampling the low-energy ligand conformations, and
- sampling the low-energy protein conformations

SLIDE can sample different relative positions of ligand. Moreover it also addresses small-scale receptor and ligand flexibility through resolving steric overlaps and developing a shape complementarity between the two. However, while enhancing

shape complementarity, it does not yet take into account how chemical complementarity can also be enhanced during the docking. Enhancing chemical complementarity in the binding site through exhaustive conformational search for both protein and ligand leads to a very large conformational search space. If there are 'n' configurations for each of 'm' side chains in the protein-ligand interface, then there are n^m possible configurations for the protein alone to be assessed by scoring.

Such expensive spatial sampling is infeasible for high throughput screening algorithms. To perform such affinity enhancing sampling, we first focus on what can be learned from known changes in a receptor's non-covalent bond network upon ligand binding in order to limit the choices to those that are reasonable. This can help us identify circumstances in which side chains do not undergo much rearrangement. Hence, the knowledge of the circumstances under which side-chain motions occur, and their extent of motion, can provide guidance for developing improved docking prediction algorithms.

2.2 Methods

We selected 30 non-homologous protein crystal structures, both ligand free (apo-protein) and ligand bound (holo-protein) from the PDB macromolecular structural database.

Table 2.1: List of protein-structure pairs studied for hydrogen bond preservation upon ligand binding.

PDB Code[1]	Protein / Ligand Complex	Res(Å)[2]	RMSD[3]
1ahc/1ahb	α-momorcharin / formycin 5′-monophosphate	2.0/2.2	0.23
1apm/1atp	cAMP-dependent protein kinase / MnATP	2.0/2.2	0.35
1ajz/1aj2	dihydropteroate synthase / diphosphate	2.0/2.0	0.29
1ca2/1bcd	carbonic anhydrase II / trifluoromethane sulphonamide	2.0/1.9	0.20
1cgf/1hfc	fibroblast collagenase / HAP[4]	2.1/1.56	0.39
1gmq/1gmr	ribonuclease/2′-guanosine-monophosphate	1.8/1.77	0.25
1kem/1kel	catalytic antibody/hapten	2.2/1.90	0.89
2hvm/1llo	hevamine a/allosamidin	1.8/1.85	0.12
1swa/1swd	apo-core-streptavidin / biotin	2.0/1.90	0.97
2ptn/1tps	trypsin / a90720a	1.55/1.9	0.3
1xib/1xid	d-xylose isomerase / l-ascorbic acid	1.6/1.7	0.19
1ydc/1ydb	carbonic anhydrase ii / acetazolamide	1.95/1.9	0.13
1tli/3tmn	thermolysin / val-trp	2.05/1.7	0.22
6taa/7taa	family 13 alpha amylase / acarbose	2.1/1.98	0.30
1gta/1gtb	glutathione S-transferase / praziquantel	2.4/2.6	0.20
1hel/1mlc	hen egg-white lysozyme / monoclonal antibody Fab D44.1	1.7/2.1	0.49
1lib/1lic	adipocyte lipid-binding protein / hexadecanesulfonic acid	1.7/1.6	0.32
1nsb/1nsc	neuraminidase / N-acetyl neuraminic acid	2.2/1.7	0.12
1poa/1pob	phospholipase A2 / transition-state analogue	1.5/2.0	0.72
1syc/1syd	staphylococcal nuclease / 2′-deoxy-3′-5′-diphosphothymidine	1.8/1.7	0.41
1udg/1udh	uracil-DNA glycosylase / uracil	1.75/1.75	0.19
2act/1aec	actinidin / E64[5]	1.7/1.86	0.11
2apr/3apr	acid proteinase / reduced peptide inhibitor	1.8/1.8	0.13
2cla/3cla	chloramphenicol acetyltransferase / chloramphenicol	2.35/1.75	0.41
2ctv/5cna	concanavalin A / α-methyl-D-mannopyranoside	1.95/2.0	0.42
2sga/5sga	proteinase A / tetrapeptide Ace-Pro-Ala-Pro-Tyr	1.5/1.8	0.08
2wrp/1tro	Trp repressor / synthetic operator	1.65/1.9	2.18
3cox/1coy	cholesterol oxidase / 3-β-hydroxy-5-androsten-17-one	1.8/1.8	0.24
3dni/2dnj	deoxyribonuclease I / DNA	2.0/2.0	0.37
3enl/5enl	enolase / 2-phospho-D-glyceric acid	2.25/2.2	0.21
3grs/1gra	glutathione reductase / glutathione disulfide	1.54/2.0	0.12
5cpa/6cpa	carboxypeptidase A / phosphonate	1.54/2.0	0.36

[1]Ligand-free/ligand-bound
[2]Resolution of the crystallographic structures in Ångstroms.
[3]Main-chain RMS positional deviation from superposition of the ligand-bound and free structures.
[4]HAP is (N-(2-hydroxamatemethylene-4-methyl-pentoyl)phenylalanyl) methylamine.
[5]E64 is [N-(l-3-trans-carboxyoxirane-2-carbonyl)-l-leucyl]-amido(4-guanido)butane.

These structures are a set of diverse, non-homologous structures with resolution better than 2.2 Å . Positional root-mean-square deviation after superimposition of ligand-free and ligand-bound structures was less than 1.0 Å (mostly smaller than 0.5 Å) . The structures had no missing residues in the binding site.

Hydrogen bonds were identified between donor and acceptor groups according to the following geometric criteria [45, 31], shown graphically in Figure 2.1:

1. Donor-Acceptor distance, d $\leq$ 3.6Å
2. Hydrogen-Acceptor distance, r $\leq$ 2.6Å
3. Donor-Hydrogen-Acceptor angle, 90° $\leq \theta \leq$ 180°

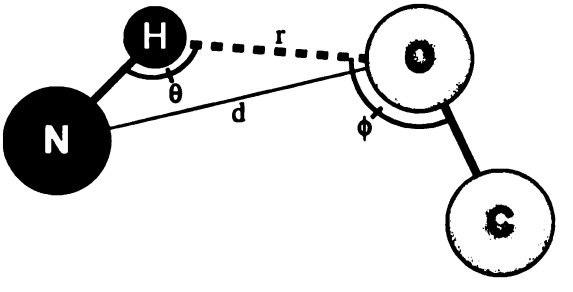


Figure 2.1: *This image is presented in color.* Geometric parameters used to identify hydrogen bonds and measure their energy. The hydrogen bond is depicted as a dashed line between the hydrogen and the acceptor oxygen. r is the hydrogen-acceptor distance, d is the donor-acceptor distance, θ is the donor-hydrogen-acceptor angle and ϕ is the hydrogen-acceptor-base atom angle, where the carbon is the base atom in this example.

The energy of each hydrogen bond was measured using a modified Mayo potential [16, 4]. The function evaluates the favorability of the observed hydrogen-bond

length relative to the optimal, equilibrium length for that pair of atoms based on their electron orbital hybridization, as well as the favorability of the angles between the donor and acceptor groups. The modification avoids non-physical H-bonds with angles near 90 deg (e.g., between C=O(i) and NH(i+3), rather than the important C=O(i)↔NH(i+4) interactions in the middle of α-helices). The energies of hydrogen bonds, E_{HB} were calculated using equations 2.1.

$$E_{HB} = V_0 \left\{ 5 \left(\frac{R_0}{R} \right)^{12} - 6 \left(\frac{R_0}{R} \right)^{10} \right\} F(\theta, \phi, \varphi) \tag{2.1}$$

with

$V_0 = 8$ kcal/mol	$R_0 = 2.80$ Å
sp^3 donor - sp^3 acceptor	$F = cos^2\theta e^{-(\pi-\theta)^6} cos^2(\phi - 109.5)$
sp^3 donor - sp^2 acceptor	$F = cos^2\theta e^{-(\pi-\theta)^6} cos^2\phi$
sp^2 donor - sp^3 acceptor	$F = cos^4\theta(e^{-2(\pi-\theta)^6})$
sp^2 donor - sp^2 acceptor	$F = cos^2\theta e^{-(\pi-\theta)^6} cos^2(\max[\phi, \varphi])$

R is the distance between the donor and acceptor atoms. The θ angle is the donor–hydrogen–acceptor angle, and ϕ is the hydrogen–acceptor-base atom angle, where the base atom is the atom bonded to the acceptor (e.g., carbonyl carbon for a carbonyl oxygen acceptor atom). The angle φ is an out-of-plane angle that arises when both the donor and acceptor have sp^2 hybridization.

First, both ligand free and ligand bound structures were pre-processed for FIRST [20] analysis to identify the hydrogen bond and hydrophobic interaction interactions within the protein. To correctly identify only intra-protein interactions, ligand was removed from the ligand-bound structure after adding hydrogen atoms by WHATIF [47] program. Adding hydrogen atoms before removing the ligand is needed to correctly assign hydrogen positions in the presence of ligand. Any other non-ligand non-water

hetero atoms, like SO_4, PO_4 or co-factors which were far from the binding site were also removed. Then FIRST analysis was run on both ligand-free and ligand-bound structures with hydrogen bond minimum energy cutoff of -0.1 kcal/mol as recommended in [20]. This energy cutoff helps in excluding large number of very hydrogen bonds. A hydrophobic interaction is determined between any two hydrophobic atoms whose inter-atomic distance is within the sum of their van der Waals radii plus 0.5 Å. FIRST analysis resulted in defining the network of non-covalent interactions, including hydrogen bonds, hydrophobic interactions and salt bridges.

The binding site was identified from the ligand bound structure. Any residue having any of its atoms within 4.0 Å of any ligand atom or any interfacial water atom is considered as part of the binding site. Interfacial waters were determined to be any water atoms within hydrogen-bonding distance (3.5 Å) of both ligand and protein structure, or any other water within hydrogen-bonding distance of any water atom that meets the previous criterion. Corresponding to residues of a binding site from ligand-bound structure, residues from the ligand-free structure were identified as part of binding site before ligand binding. These two sets of binding site residues, without ligand or any other non-water hetero-atoms, were used for experiments.

Using the binding site residues and list of interactions generated by FIRST , all those interactions were identified which had any atom in the binding site, for both ligand-free and ligand-bound structures. Using this list, intra-protein hydrophobic interactions, and direct, one-water-mediated hydrogen bonds and two-water hydrogen bonds are determined. These bonds are then categorized according to either atom categories according to the rules specified in table Table 2.2

Table 2.2: Rules for categorization of atom types. These categories are used to classify hydrogen bonds in the bond preservation analysis using FIRST.

Residue Type	Atom Name	Atom Category	Category Description
ALA	C*	ALPH	Aliphatic carbon
ARG	NE	N_Pos	Positive nitrogen
ARG	CZ	ARMT	Aromatic carbon
ARG	C*	ALPH	Aliphatic carbon
ASN	OD	Keto	Neutral keto group
ASN	ND	N_Ntrl	Neutral nitrogen
ASN	C*	ALPH	Aliphatic carbon
ASP	OD	O_Neg	Negative oxygen
ASP	C*	ALPH	Aliphatic carbon
CYS	C*	ALPH	Aliphatic carbon
GLN	OE	Keto	Neutral keto group
GLN	NE	N_Ntrl	Neutral nitrogen
GLN	C*	ALPH	Aliphatic carbon
GLU	OE	O_Neg	Negative oxygen
GLU	C*	ALPH	Aliphatic carbon
HIS	ND	His_ntrl_Pos	Histidine neutral or positive nitrogen
HIS	NE	His_ntrl_Pos	Histidine neutral or positive nitrogen
HIS	C*	ARMT	Aromatic carbon
ILE	C*	ALPH	Aliphatic carbon
LEU	C*	ALPH	Aliphatic carbon
LYS	NZ	N_Pos	Positive nitrogen
LYS	NZ	N_Pos	Positive nitrogen
LYS	C*	ALPH	Aliphatic carbon
MET	C*	ALPH	Aliphatic carbon
PHE	CB	ALPH	Aliphatic carbon
PHE	C*	ARMT	Aromatic carbon
PO_4	O	Ion	Ion
PRO	C*	ALPH	Aliphatic carbon
SER	OG	O_Ntrl_Hxyl	Neutral hydroxyl oxygen
SO_4	O	Ion	Ion
THR	OG	O_Ntrl_Hxyl	Neutral hydroxyl oxygen
THR	C*	ALPH	Aliphatic carbon
TRP	NE	N_Ntrl	Neutral nitrogen
TRP	CB	ALPH	Aliphatic carbon
TRP	C*	ARMT	Aromatic carbon
TYR	OH	O_Ntrl_Hxyl	Neutral hydroxyl oxygen
TYR	CB	ALPH	Aliphatic carbon
TYR	C*	ARMT	Aromatic carbon
VAL	C*	ALPH	Aliphatic carbon
*	NH1	N_Pos	Positive nitrogen
*	NH2	N_Pos	Positive nitrogen
*	OT1	O_Neg	Negative oxygen
*	OT2	O_Neg	Negative oxygen
*	OXT	O_Neg	Negative oxygen
*	O	MCHN_O	Main chain oxygen
*	N	MCHN_N	Main chain oxygen

2.3 Results and Discussion

Figure 2.2 displays preservation percentages of direct hydrogen bonds between different atom-category pairs or in short, "Bond Category". For each bond category, the label at far right indicates the bond preservation occurrences both in percentage terms and in terms of actual cases found. The red and green bars for each category also show the average displacement that was experienced by the bond centroid before and after ligation, bond centroid being the mid-point between the bonding pair of atoms.

Figure 2.3 shows the bond preservation percentages (in absolute counts) by residue type including both side-chain and main-chain bonds.

Two key trends stand out. 70% or more of the direct hydrogen bonds between two protein atoms are preserved for most atoms. Taking bond centroid displacement as an approximate measure of how much atoms move in preserved hydrogen bonds, it's clear that almost always their displacement is less than 0.5 Å. Both these observations may prove to be quite useful for docking algorithms. They help restrict the number of sidechains which might require rearrangements upon ligand binding. The displacement observations also measure the extent to which bonded interfacial sidechains can move while still preserving their hydrogen bond interactions. In fact, those groups that maintain intra-protein hydrogen bonds upon ligand binding also do not move significantly upon ligand binding.

Pair of interfacial receptor atoms may also participate in water mediated hydrogen bonds apart from direct hydrogen bonds . There is a higher chance of displacing the water upon ligand binding, hence a smaller percentage of water-mediated hydrogen bonds are preserved upon ligation, as presented in Figures 2.4 and 2.5.

Hydrophobic interactions are key to both protein folding as well as protein-ligand docking. Exposing hydrophobic patches to solvent has a high energy cost via unfavorable entropy. Figure 2.6 displays the percentages of hydrophobic interactions

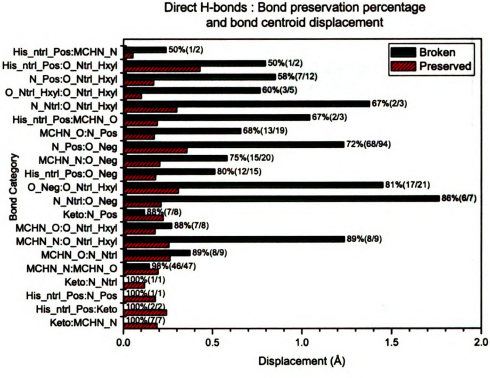


Figure 2.2: *This image is presented in color.* Percentages and displacements, by atom category, of preserved direct intra-protein hydrogen bonds in protein-ligand interfaces.

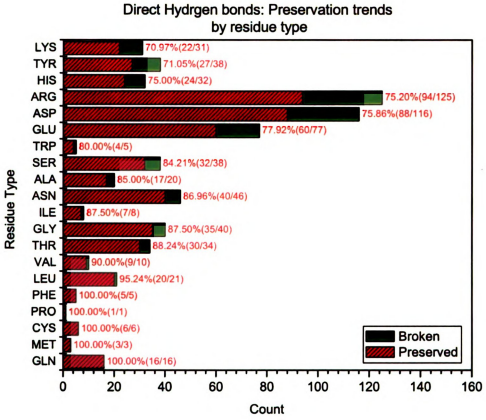


Figure 2.3: *This image is presented in color.* Percentages, by residue type, of preserved direct intra-protein hydrogen bonds in protein-ligand interfaces.

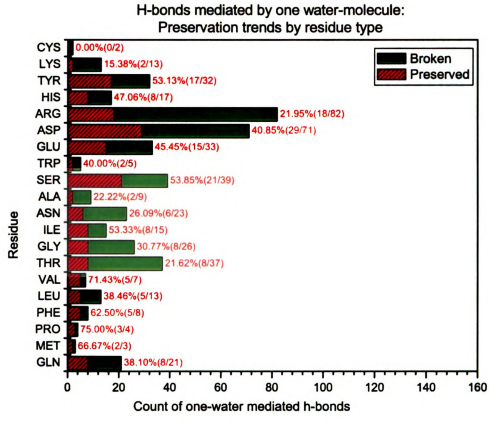


Figure 2.4: *This image is presented in color.* Percentages, by residue type, of preserved interfacial intra-protein hydrogen bonds mediated by one water molecule.

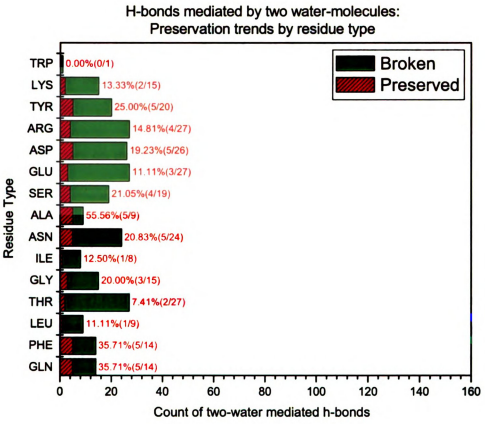


Figure 2.5: *This image is presented in color.* Percentages, by residue type, of preserved interfacial intra-protein hydrogen bonds mediated by two water molecules.

that stay preserved between the same pair of residues. Hydrophobic interactions are less specific, unlike hydrogen bonds. Hence, while hydrogen bondsare counted preserved if they persist between same pair of atoms upon ligand binding, hydrophobic interactions are evaluated preserved if the persist between the same residues. The criteria of evaluating hydrophobic interactions are :-

1. inter-atom distance $\leq$ sum of van der Waals radii of the atoms + 0.5
2. Both the participating atoms are connected to hydrogen, carbon or sulphur atoms only.

This focuses on more hydrophobic interactions rather than simple C-C van der Waals interactions.

2.4 Conclusions

The presented results provide powerful ways to guide spatial sampling during the docking process. Hydrogen-bond conservation probabilities can be utilized to limit the choices of rotatable bonds that can be rotated to resolve steric-clashes, deterministically or probabilistically, that are part of a sidechain participating in a hydrogen bond . Such limiting not only brings down the combinatorial effort involved in resolving clashes, but can also boost the chemical complementarity of the final docking by keeping the number of unsatisfied atoms buried in the interface low, thus improving the overall quality of dockings. Strong preservation trends of hydrophobic interactions can similarly guide docking algorithms to make wiser choices for developing induced fit.

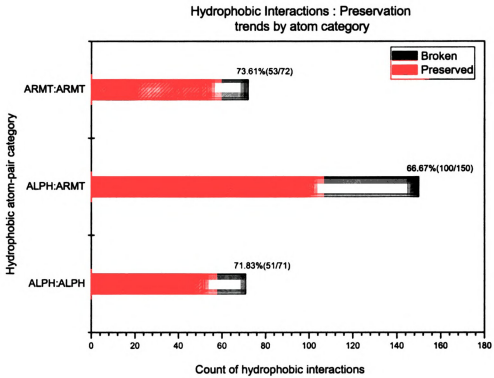


Figure 2.6: *This image is presented in color.* Percentages, by atom category, of preserved interfacial intra-protein hydrophobic interactions.

Chapter 3

Sampling Side-Chain Positions in Protein-Ligand Interfaces in SLIDE

The extent of intra-protein hydrogen bond preservation and the average displacement of the centroid of preserved and broken hydrogen bonds in Figure 2.2 indicate the tendency of most interfacial side chains to minimally adjust upon ligand binding. This chapter presents a detailed analysis of side-chain displacements as observed from ligand-free to ligand-bound conformations and how such displacements might be modeled in the docking tool SLIDE. Section 3.1 characterizes in detail the side-chain displacements observed in the 30 structures specified in Table 2.1. Displacement characterization is presented as the extent of rotations experienced by the different χ angles of side chains, as well as the extent of positional deviation (RMSD) the interfacial residues undergo upon ligation. Thereafter, the induced-fit model currently encoded in SLIDE is explained, as well as how it can be augmented.

3.1 Side-Chain Displacements upon Ligand Binding

Ligand binding can change the protein conformation by moving not only side chains, but also the protein backbone. A simultaneous conformational search for both backbone and side chains is infeasible in screening software due to the exponential complexity of the conformations to sample and test. However, sampling side-chain conformations only is relatively easier, though still a computationally daunting task; protein side-chain motion is still ignored in most docking software or modeled in a biased way based on known crystal complexes with different ligands.

To study the type of side-chain displacements through dihedral side-chain rotations, all residues with at least one rotatable bondbetween heavy atoms were considered. Hence alanine, glycine and proline were not considered. Rotation of the NE-CZ bond of arginine was not considered because the CD, NE, CZ, NH1 and NH2 atoms form a planar, partial double-bonded structure, severely restricting rotation. All the ligand-bound side chains analyzed were within 4.0 Å of the ligand in the 30 ligand-bound structures, defining the set of interfacial side chains; corresponding side chains from the ligand-free structures were also considered.

Because angular rotations can be compensatory and even side chains with significant χ-angular differences can be almost superimposible in Cartesian coordinates[29], the amplitude of side chain motions was also studied. Figure 3.1 shows the root-mean-square positional deviation (RMSD) experienced by interfacial side chains upon ligand binding across 30 structures. Up to 70% of the side-chain deviations were less than 0.5 Å, while another 15% moved within 0.5 Å to 1.0 Å, while the remaining 15% experienced somewhat bigger motion between 1.0 to 5.5 Å. This observation is in good agreement with the bond-preservation observations, where 70% of the intra-protein hydrogen bonds were preserved and, on average, the hydrogen bond preserving side

chains moved less than 0.5 Å. Observations of the prominent tendency for intra-protein hydrogen bond preservation is further supported by [50], where analysis of 63 ligand-bound and ligand-free structure pairs shows that 85% of side-chain motions in the interface are small rotations that typically do not lead to another rotameric conformation.

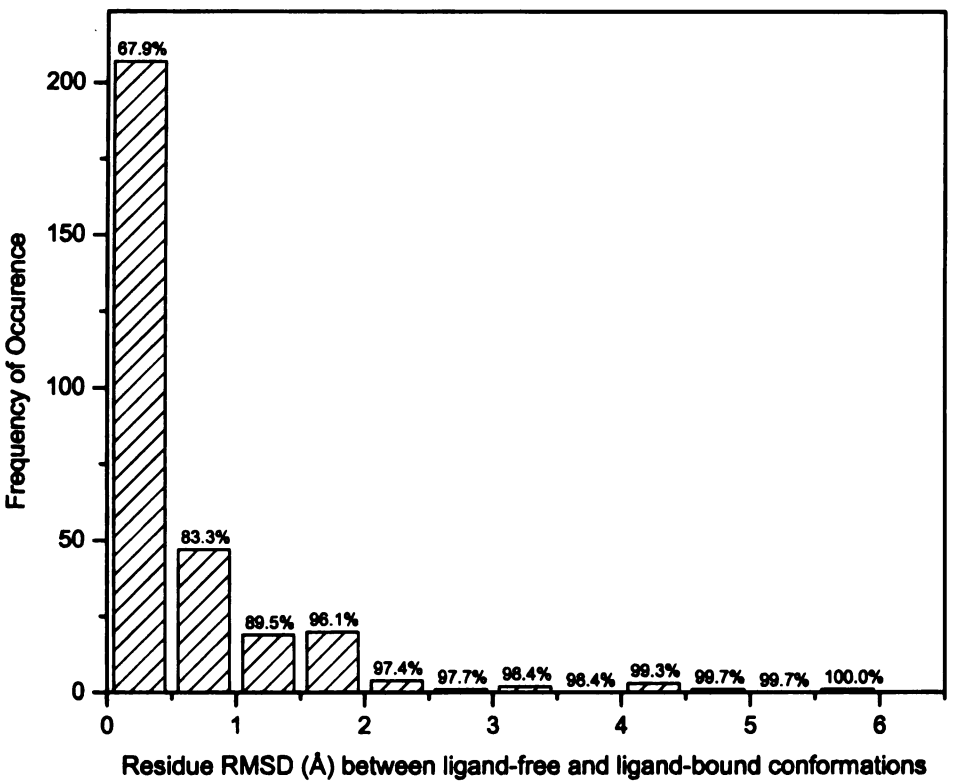


Figure 3.1: Frequency distribution of RMSD of ligand-bound interfacial side-chain positional shifts upon ligand binding in 30 structures. Percentages atop each bar are cumulative.

While RMSD measures the positional deviation of the entire side chain, changes

in each of the χ angle of all the interfacial side chains were also studied. These changes in χ angles are presented in Figures 3.2 and 3.3. As would be intuitive, higher χ angles, which were further from the main chain, experience not only larger changes when compared to lower χ angles, but the changes have a broader angular distributions, too. Possible explanation is that terminal side-chain bonds have more space to sample small as well as large rotations with a lower probability of causing steric overlaps, as compared to rotatable bonds closer to the backbone.

3.2 Current Rotation Paradigm in SLIDE

Here we present SLIDE's current paradigm for modeling protein side-chain and ligand flexibility, and assess the extent to which it models the kinds of side-chain motions observed in Figures 3.2 and 3.3.

As described earlier in Section 1.4, after matching a triangle from any three ligand-interaction points to a template-point triangle using geometric hashing and the chemical-complementarity criteria, SLIDE transforms the ligand triangle onto the template triangle using least-square fitting. Any ligand atoms within the perimeter of the ligand triangle are also transformed, thus docking what is now considered to be a rigid anchor fragment of the ligand, into the binding site. Transforming anchor-fragment atoms not representing the interaction centers can lead to inter-molecular bumps with the protein. SLIDE discards any ligand whose anchor fragment cannot escape clashing with protein backbone, after attempting small translations of the anchor fragment away from the clashing protein atoms.

If the anchor fragment has no steric clashes with the protein, the rest of the ligand is then transformed into the reference frame of the ligand anchor fragment to complete the ligand structure in the binding site. Reconstructing the ligand in the binding site can lead to new steric clashes between the protein and the ligand. Since

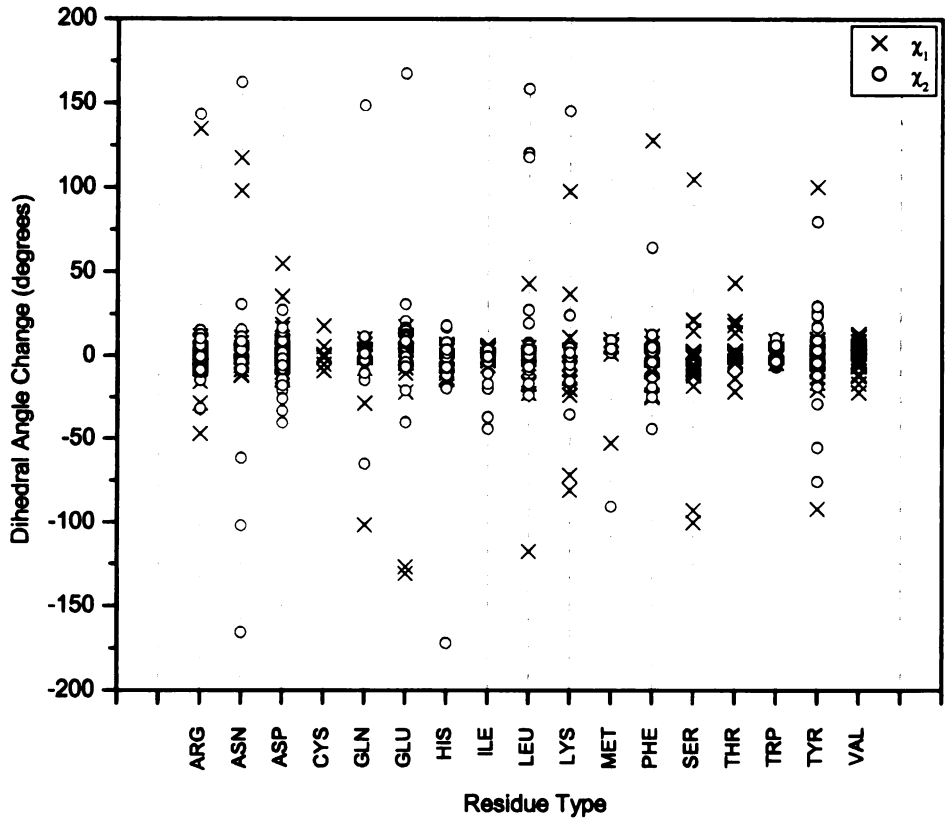


Figure 3.2: Change in χ_1 and χ_2 dihedral angles of all interfacial side chains within 4.0 Å of the ligand, across 30 structures listed in Table 2.1. The change is calculated as the difference of χ angle of interfacial side chains between their ligand-free and ligand-bound conformations. The standard deviation for χ_1 changes is 25.8 and for χ_2 changes is 34.7. Note that most changes in $\chi_{1,2}$ fall within the range of $-20°$ to $20°$.

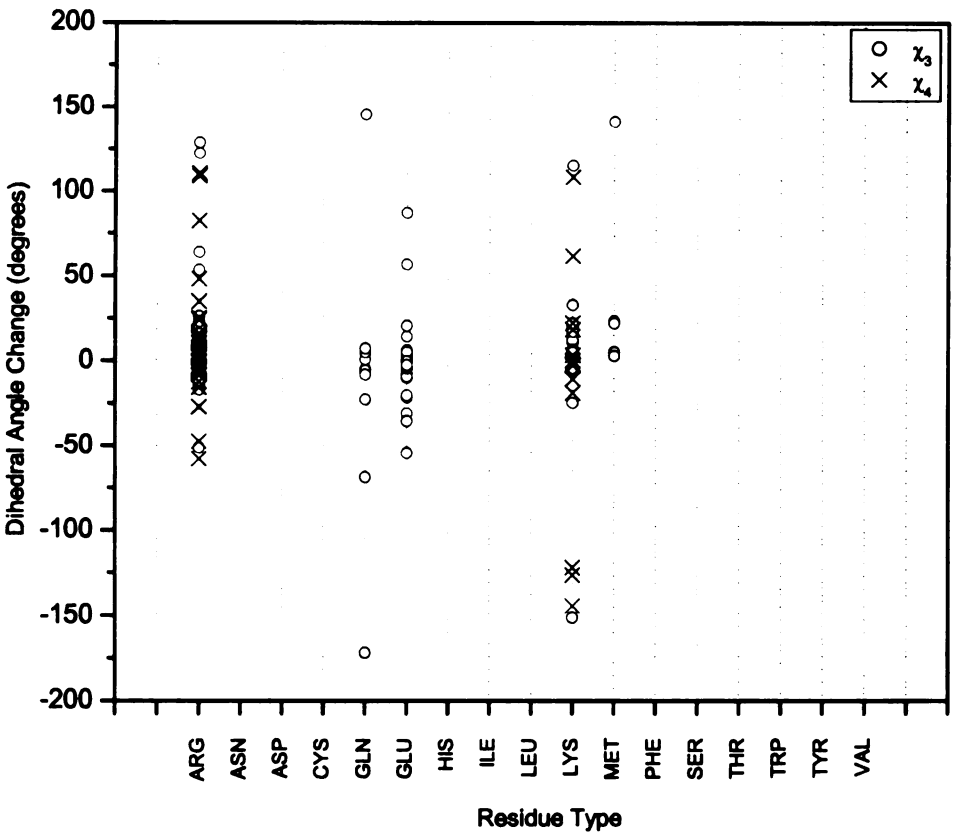


Figure 3.3: Change in χ_3 and χ_4 dihedral angles of all interfacial side chains within 4.0 Å of the ligand, across 30 structures listed in Table 2.1. The change is calculated as the difference of χ angle of interfacial side chains between their ligand-free and ligand-bound conformations. The standard deviation for χ_3 changes is 42.9 while for χ_4 changes is 46.4. Note the somewhat broader range of change in χ_3 and χ_4 compared to changes in χ_1 and χ_2 dihedral angles in Figure 3.2.

these steric clashes involve the flexible portion of the ligand, it is possible to resolve each steric clash independently by rotating either the ligand's flexible side chain, or a flexible group in the protein with which it overlaps. In this context, a ligand side chain is any group connected to the anchor fragment by a rotatable bond. When the target (protein) atom clashing with the ligand belongs to a flexible protein side chain, that side chain's rotation also presents another option to resolve the steric clash.

Geometrically, any collision between two atoms, which are modeled as va der Waals spheres can be resolved in multiple ways, given that each atom as 3 degrees of translational freedom. However, the colliding atoms are themselves bonded to other atoms by single or double bonds. For each single-bond rotation that can help resolve a steric clash, overlapping atom displaced by the bond rotation is considered mobile, while the other overlapping atom is considered fixed. When both the overlapping atoms are independently connected to single bonds, than their overlap resolution is evaluated with each of them being considered mobile and fixed, one at a time.

Single bond rotations allow moving the rotating atoms in a circular trajectory in planes perpendicular to the rotation axis, which is along the single bond. To resolve a collision through a specific bond rotation, the rotatable single bond is transformed along the Y-axis, with the atom of the bond further from the backbone forming the new origin. The same transformation is applied to all the atoms in the side-chain beyond, including both the atoms having the steric clash. The transformations are calculated in a way such that, on application, they would move the fixed atom into the X-Y plane, as shown in Figure 3.4. This later helps calculations for determining the rotatable bond's rotation angle for collision resolution.

While there are potentially infinite positions on the trajectory where mobile atom can be placed to resolve a clash, most of these are usually infeasible due to the high probability of causing new collisions. SLIDE's paradigm to choose a position to resolve collisions during docking is based on developing "induced fit" between the

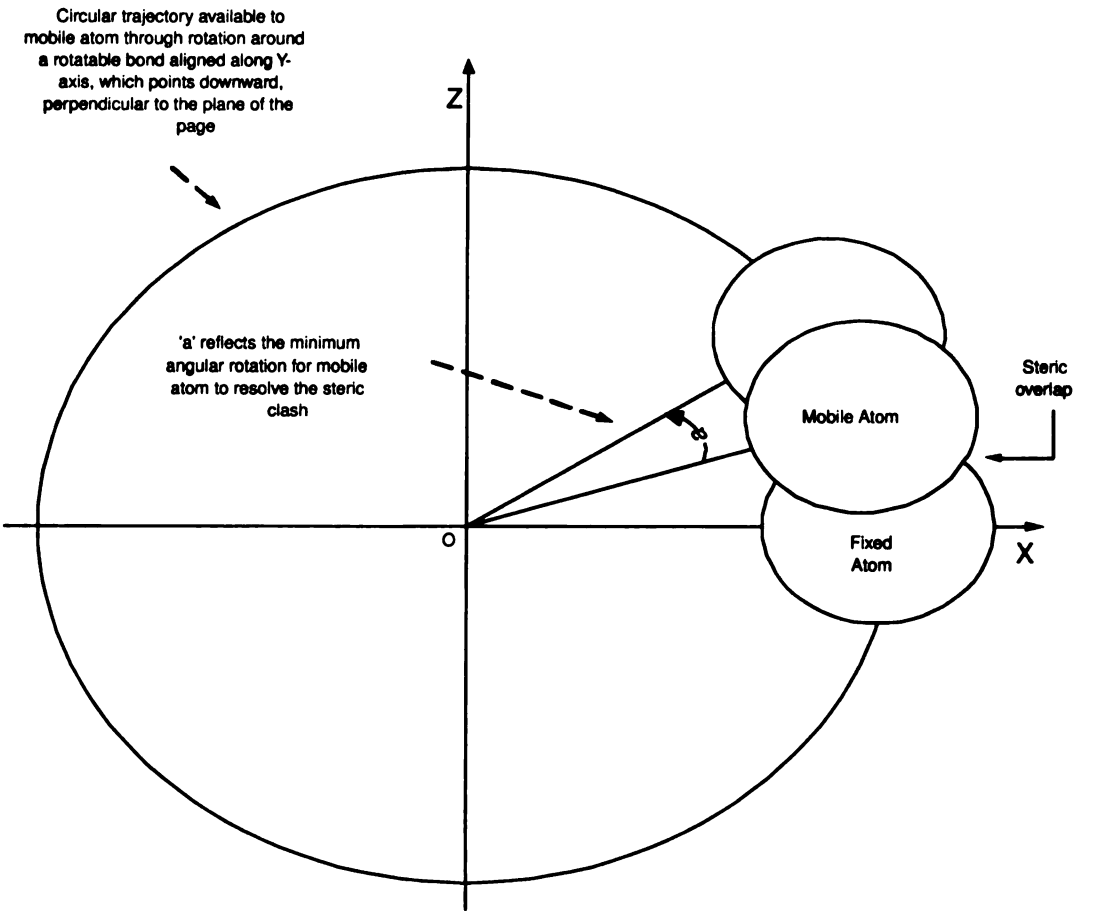


Figure 3.4: Induced fit development mechanism implemented in SLIDE performs directed rotations to resolve atomic collisions. Directed rotations are performed around a rotatable bond not adjacent to the mobile atom. The rotations are performed in a plane parallel to X-Z plane, by aligning rotatable bond along the Y-axis. The hinge-atom of the bond is new origin for the rotation. The fixed atom is transformed into the positive quadrant of the X-Y plane, while the mobile atom is rotated through the minimum angle of rotation, 'a' to resolve the steric overlap.

ligand and the protein binding site[23]. To mimic induced fit, SLIDE chooses the smallest angle through which the mobile atom can be rotated to resolve the collision.

However, choosing the smallest rotation angle for a bond rotation is not sufficient since there may be multiple single bonds in a side chain in residues like lysine, arginine which could be rotated to resolve a collision. SLIDE's approach for deciding which rotations to apply for resolving a set of collisions is based on mean-field theory[26, 19], implemented to optimize bond rotations[42]. The key feature of this approach is a probability matrix $P(i, j)$, which describes the probability that a particular collision i will be resolved by a rotation of bond j. First, all intermolecular collisions in the complex are identified. They form one dimension of the matrix. Only complexes with up to 20 collisions between atom pairs undergo side-chain collision resolution in SLIDE; ligand dockings with more collisions are discarded. All rotatable bonds that can be used to resolve at least one of the collisions build the other dimension of the matrix. Only those rotations are considered that do not result in an intramolecular collision based on the current conformation. Note that there is no differentiation between ligand and protein side chains in this matrix. All rotations that can resolve a particular collision are initialized with equal probability values. For each probability entry $P(i, j)$, a cost value $E'(i, j)$ is computed that reflects the cost of rotating bond j to resolve collision i. This cost is simply the product of the number of displaced non-hydrogen atoms and the absolute value of the rotation angle. This makes large rotation angles or the rotation of large side chains more costly, as they are more likely to cause steric problems elsewhere.

During the iterative cycles of the mean-field optimization process, the probability matrix P is updated to converge to high probabilities for those rotations that provide the lowest-cost conformational change of both molecules to resolve a maximal number of the observed collisions (Figure ??). In each cycle a mean cost, $E(i, j)$, is computed

```
Find all side-chain collisions i and all rotatable bonds j;
While there are between 1 and 20 side-chain collisions do:
        Compute probability matrix P and cost matrix E';
        For 10 cycles do:
                Compute mean cost E(i,j);
                update probability matrix P;
        Do feasible highest probability rotations;
        Find all remaining side-chain collisions i;
```

Figure 3.5: The algorithm for resolving side-chain collisions using the mean-field optimization technique. When there are still collisions exceeding the threshold after 10 iterations of the outer loop, this ligand orientation is discarded.

for each rotation, as follows:

$$E(i,j) = E'(i,j) + \sum_{h \neq i,\ k} dep[(i,j),(h,k)] \cdot P(h,k) \cdot E'(h,k)$$

The value of $dep[(i,j),(h,k)]$ is set to -1.0 if $j = k$, i.e., both entries refer to the same bond and both rotations are in the same direction. Hence, two collisions can be resolved at once by a rotation of this bond, and the -1.0 value results in a lower mean cost $E(i,j)$. The value of $dep[(i,j),(h,k)]$ is set to 1.0, with a resulting increase in $E(i,j)$, in two cases. The first case is when both entries refer to the same bond ($j = k$), but the corresponding rotation directions are opposed to each other. In the other case, bond j lies on the path from bond k to the anchor fragment or the main chain, respectively. Here, the mean cost of rotating bond j is increased, since if rotation (i,j) were applied, then bond k would be moved, and this would invalidate the assumptions made in the current iteration regarding rotations involving this bond.

At the end of each cycle, the entries in the probability matrix are updated based

on the mean costs $E(i,j)$ using the Boltzmann principle:

$$P(i,j) = \frac{e^{-E(i,j)/\mu}}{\sum_k e^{-E(i,k)/\mu}}$$

where μ is the average value of all computed mean costs. Convergence of the values in the probability matrix is usually observed in fewer than ten cycles, and those rotations with the highest probability are chosen to resolve the collisions. At this point, it is again necessary to check for negative correlations between bonds. Although this was already considered during the computation of the mean cost, two correlated bonds can receive high probabilities if they are the only bonds to resolve particular collisions, or if alternative rotations are much more expensive. During the mean-field optimization process, it is not possible to anticipate complex dependencies, e.g., which ligand rotations influence protein bonds related to other collisions. Rotations are only accepted if they do not cause any intramolecular collisions. Since it is likely that not all collisions can be resolved in one application of the mean-field optimization technique, up to 10 iterations of this process are executed, as outlined in Figure ??. Ligand conformations are discarded if they have more than 20 collisions at any time during the optimization or have collisions exceeding the threshold after 10 iterations.

3.2.1 Motions Modeling in SLIDE

Mean-field optimization helps choose bond rotations to minimize the conformational cost of developing shape complementarity. Figures 3.6 and 3.7 present a displacement comparison of side chains that were moved by SLIDE for collision resolution with displacements of corresponding side chains from their ligand-free to ligand-bound conformations. (Hence, these figures present a subset of interfacial side chains which were found to have moved in Figure 3.1).

To examine how well SLIDE models known side-chain motions, the distribution

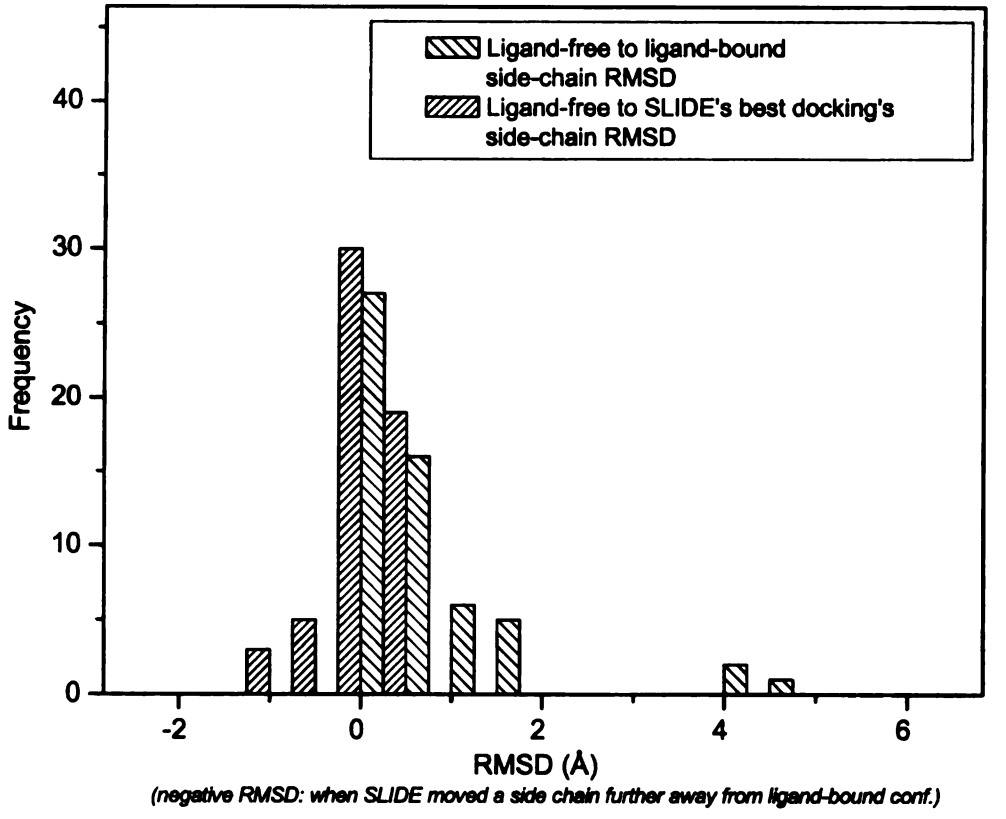


Figure 3.6: Frequency distributions of RMSD between ligand-free and ligand-bound side-chain conformation, as well as RMSD between ligand-free side-chain conformation and side-chain conformation in the best docking generated by SLIDE. The best docking was determined using the RMSD of the docked ligand relative to the crystal structure orientation. Positive values indicate the RMSD by which SLIDE's conformation of side chain got closer to crystal complex conformation, while negative values indicates that SLIDE moved the side chain further from crystal complex conformation. Observations are derived from SLIDE's best dockings of 24 structures, comparing only those interfacial side chains that were moved by SLIDE.

of root-mean-square deviations (RMSDs) between ligand-free and docked side-chain conformations found in the best dockings by SLIDE were compared with the distribution of RMSDs between ligand-free and ligand-bound side-chain conformations from crystal structures (Figure 3.6). This distribution data was derived from 24 of the 30 structures specified in Table 2.1. The remaining 6 structures did not permit docking without an increase in inter- atomic interpenetration parameters, so they were excluded from this analysis. The RMSD value of a docked side-chain conformation relative to its ligand-free conformation is displayed negative when it is greater than the RMSD value between the ligand-free and ligand-bound conformation in crystal structures. Hence, this figure compares side-chain displacement distributions between modeled and observed structures, both in magnitude and direction. While SLIDE's distribution of RMSD magnitudes was quite similar to nature, as most of the motions were restricted between 0-1 Å, a few large displacements in ligand-bound conformations are also observed. These observations are in agreement with [50], where it has been found that in 85% of the cases, interfacial side chains rotate through less than 45° upon ligand binding.

While Figure 3.6 compares displacement distributions, Figure 3.7 compares displacements on individual side-chain basis. Presented as a scatter plot is the comparison of side chain RMSDs between ligand-free and ligand-bound conformations, and between the ligand-free conformation and the conformation in the best dockings for 24 structures by SLIDE. Comparing with Figure 3.1, it is clear that SLIDE moves very few side chains compared to nature. The correlation between the ligand-free to ligand-bound side chain RMSD and ligand-free to docked side chain RMSD is small (0.28), indicating the extent to which SLIDE misjudges the magnitudes of motion, as is clear by the degree of scatter. The color coding indicates the quality of motion, answering the question: " did SLIDE move the side chain closer to ligand-bound conformation or further away ? " SLIDE, while removing steric overlaps, tends to

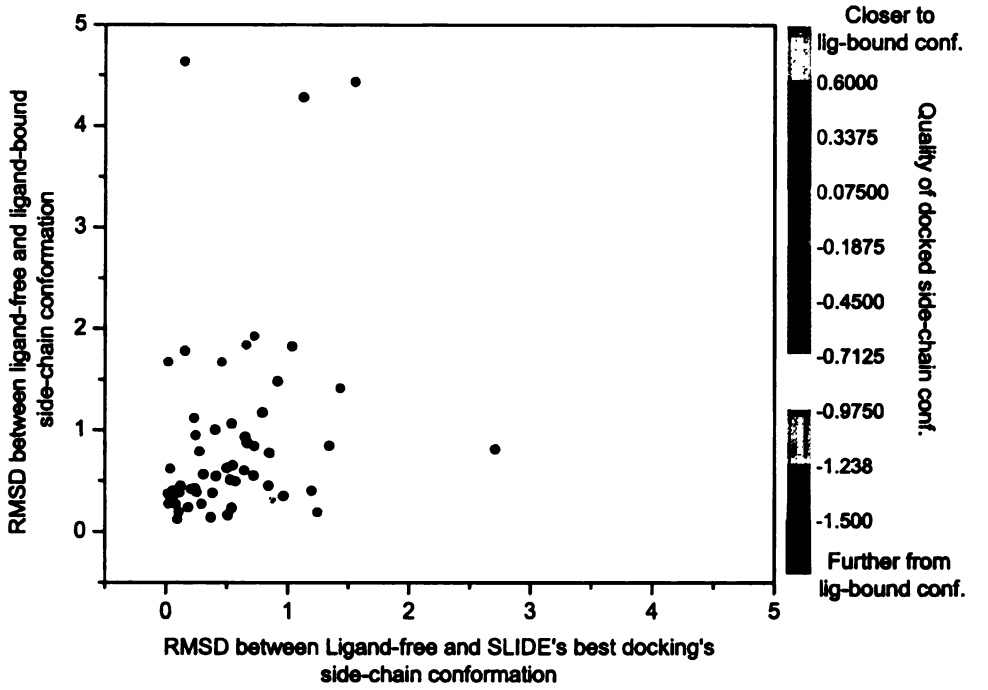


Figure 3.7: *This image is presented in color.* Comparison of side chain RMSDs between ligand-free to ligand-bound conformation, and between ligand-free conformation and conformation in the best docking. The best docking was determined using the RMSD of the docked ligand versus the crystal structure ligand orientation. The color scale indicates the quality of motion. Positive values indicate the RMSD by which the SLIDE-generated best docking's side-chain conformation became closer to crystal complex conformation, while negative values indicate that SLIDE moved the side chain further from the crystal complex conformation. Observations are derived from the best dockings of 24 structures, with only those interfacial side chains being compared which were moved by SLIDE. Correlation between the two RMSD distributions is 0.28.

move the side chains away from the ligand-bound conformation (colors ranging from cyan to red). For only 18 of 58 observations did SLIDE move side chain closer to the ligand-bound conformation (blue).

Having observed that most of the SLIDE motions are minimal to resolve collisions, that each collision can usually be resolved by either atom's rotation around a rotatable bond, and that often, SLIDE moves atoms further away from ligand-bound positions, it is postulated that both directions of rotation should be explored for collision resolution. While the angle of rotation in one direction will be bigger than the angle in the opposite direction, the ability to make better interactions in one direction versus the other should influence the direction of rotation. Furthermore, from the observed intra-protein hydrogen-bond preservation tendency presented in Chapter 2, any rotations which might disrupt direct hydrogen bonds between the binding-site side chains may be discouraged so that SLIDE's dockings will have a similar bias towards preserving most of the intra-protein hydrogen bonds.

3.3 Employing Hydrogen-Bond Preservation Bias in Mean-Field Optimization

As explained in previous section 3.2, the key feature of mean-field optimization process is the probability matrix $P(i, j)$, which describes the probability that a particular collision i will be resolved by a rotation of bond j. At the beginning of the optimization, each rotation that can resolve a collision is assigned equal probability. Subsequently, these probabilities are updated iteratively during the optimization process, depending upon the conformational cost of rotation and dependencies of resolving other collisions. To bias against those specific side-chain rotations that disrupt a hydrogen bond, initial probabilities can be reduced according to the hydrogen-bond preservation probabilities derived from the statistics shown in Figure 2.2. This should

effectively discourage rotations which can break a hydrogen bond. Nevertheless, despite low probability, a particular rotation that disrupts a hydrogen bond can still be effected if it is the only one that can resolve a particular collision, hence ensuring that final docking will be free from atomic overlaps. Biasing against breaking intra-protein hydrogen bonds would likely improve the chemical complementarity of the docking as well.

To measure the effect of incorporating hydrogen-bond preservation bias, a pilot study of 5 ligand-free and ligand-bound structure pairs was conducted. The templates representing the binding sites of the ligand-free structures included random sampling of potential interaction points as well as known interaction points with the ligand. Known points were included to ensure that at least one docking would result in placing the ligand close to its native position and orientation, so that side-chain motion would not be required to compensate for inaccurate ligand placement.

Figures 3.8, 3.9, and 3.10 present a comparison of SLIDE and the new SLIDE variant with a hydrogen-bond preservation bias based on the statistics from Figure 2.2. Data presented for both SLIDE versions include only the best dockings according to the ligand RMSD (relative to its crystal complex position), which must be less than 1.0 Å. The native ligand conformations from ligand-bound crystal complexes were used for docking into apo structures. Figure 3.8 presents the total number of intra-target (protein binding site) hydrogen bonds. The crystal complexes have more intra-protein hydrogen bonds upon ligand binding compared to the ligand-free structures. Both SLIDE versions preserve a similar number of hydrogen bonds, but fewer than what nature preserves in the crystal complexes. This indicates that new intra-target hydrogen-bonding opportunities are missed in both versions of SLIDE. Figure 3.9 presents intra-target hydrogen bonds that were lost upon ligation in the crystal complex and in the best dockings by SLIDE. Interestingly, while more hydrogen bonds were lost in the crystal complexes, Figure 3.8 indicates that these complexes form

new intra-target hydrogen bonds that were not present in the ligand-free structures, as shown in Figure 3.10. Even though the sample set of 5 structures is small, it nevertheless suggests that nature breaks and makes more intra-target hydrogen bonds upon ligation through optimal arrangement of side chains. SLIDE, on the other hand, due to its minimal rotation collision resolution model, tends to keep the binding site side-chain positions and their interactions more or less the same, breaking and making fewer intra-target hydrogen bonds compared to nature. Considering that each hydrogen bond has an energy of about -5 Kcal/mol, and that typical complex formation involves a favorable energy change of only $\sim$3 times this number, careful modeling of interfacial hydrogen bonds is likely to be crucial to correctly sample and assess the best dockings.

3.4 Sampling Large Side-Chain Motions

Comparison of the extent of displacement in Figures 3.6 and 3.7, as well as hydrogen-bond preservation results from best dockings by SLIDE presented in Figures 3.8, 3.9, and 3.10 indicate that aside from developing good shape complementarity through small side chain adjustments, optimal rearrangement of the intra-target hydrogen-bond network is important for favorable binding affinity between the molecules. This rearrangement may involve not only small but also large rotations, which SLIDE's induced-fit mechanism does not encourage.

To explore if there are any prominent reasons for rearrangements through large side-chain rotations, all interfacial residues that had undergone a side-chain dihedral rotation greater than 60° upon ligand binding were analyzed by molecular graphics to understand the reasons for these large rotations. Any binding sites having gaps between residues were excluded from this analysis using the molecular graphics tool InsightII, since those mobile residues could influence motions in unpredictable ways.

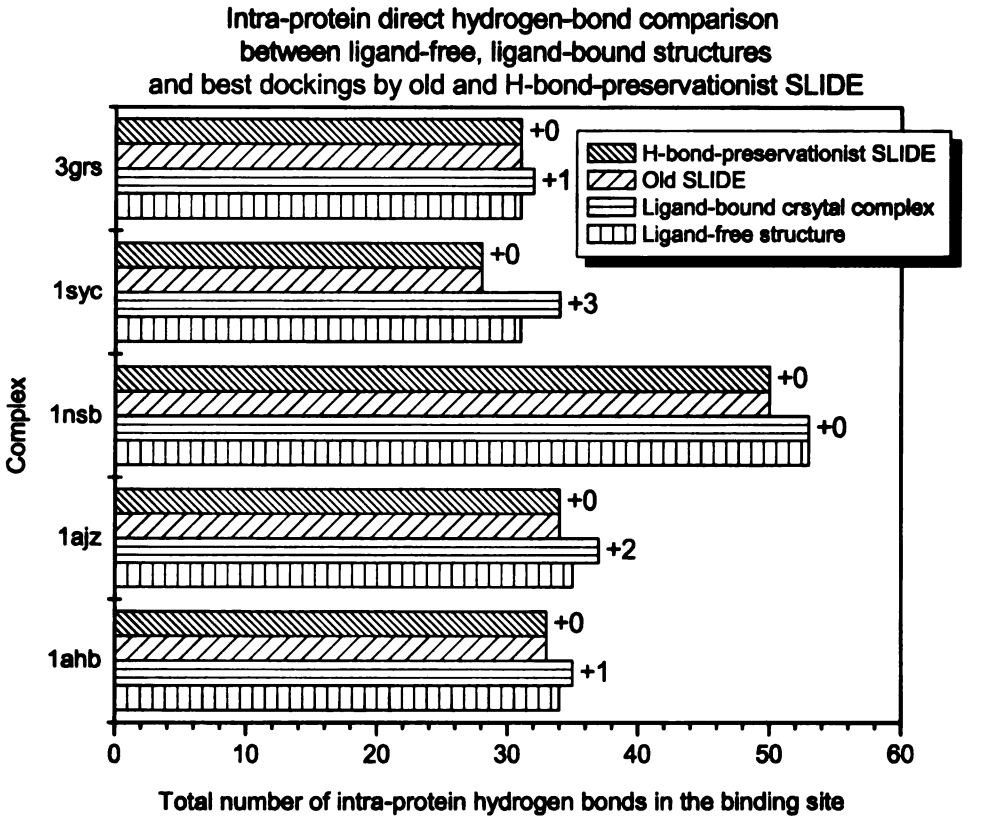


Figure 3.8: Comparison of the number of intra-protein hydrogen bonds before and after ligand binding in crystal complexes, in best dockings by SLIDE and the hydrogen-bond-preservationist version SLIDE.

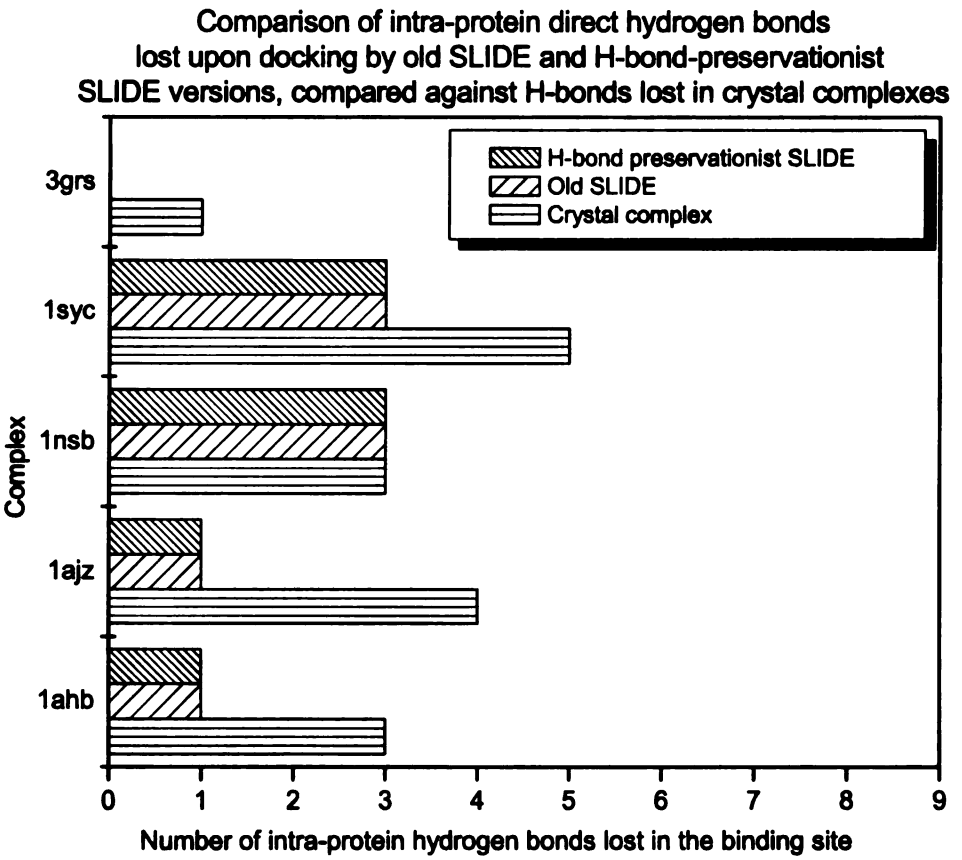


Figure 3.9: Comparison of the number of intra-protein hydrogen-bonds lost upon ligand binding in nature (between ligand-free and ligand-bound structures), in best dockings by SLIDE, and in the best dockings by hydrogen-bond-preservationist version SLIDE.

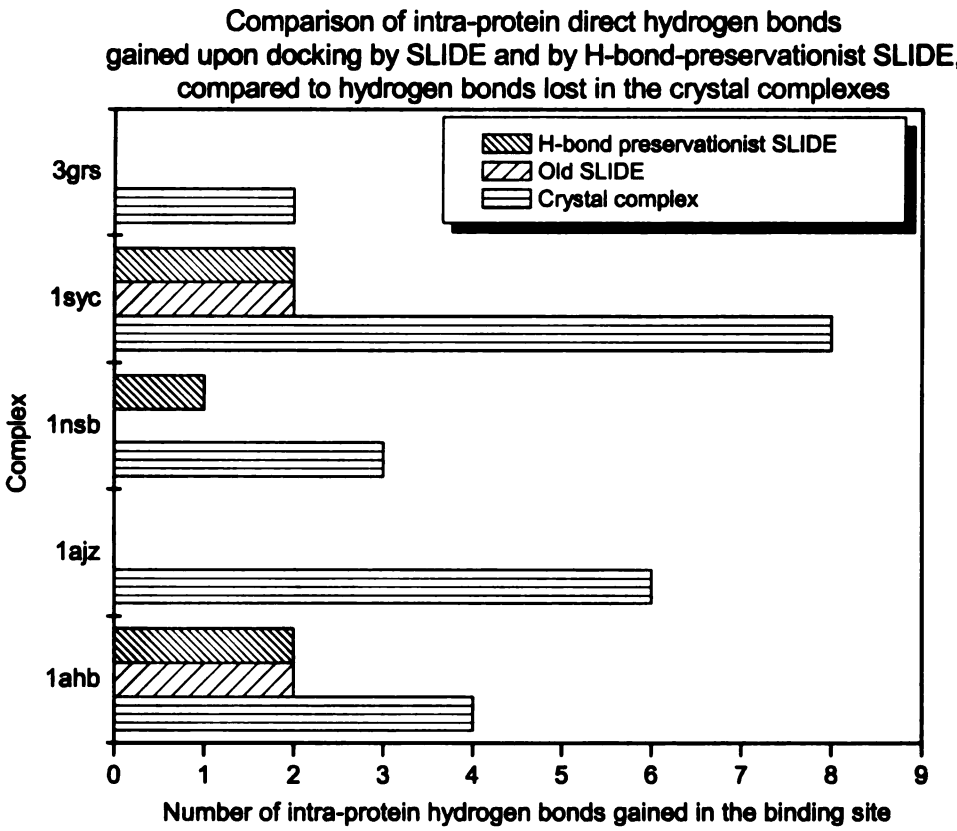


Figure 3.10: Comparison of the number of intra-protein hydrogen bonds gained upon ligand binding in crystal complexes, in the best dockings by SLIDE and by the hydrogen-bond-preservationist version of SLIDE.

Table 3.1 presents the conclusions from visual inspection of such large rotations. The process involved super-imposing the ligand-free binding site onto ligand-bound site, followed by analyzing steric and chemical factors that could encourage large rotations.

Several causes for rotations or displacements were found - main-chain movements of more than 1.0 Å (for residues like L33 in PDB entry 1kel L33, residues A45 and A49 in 1swd), side-chain movements caused by motion in adjacent residues, aromatic interaction with the ligand (1ahb, residue 70), π-cation interaction with the ligand (1gmr residue B40, 1syd residue 115), surface exposure or even crystal packing effects (1udh residue 87). Nevertheless, in two-thirds of the cases, side chains moved not to resolve any steric clashes upon ligation. Rather, side chains moved to satisfy polar atoms (in 10 cases, excluding cases with aromatic or π-cation interactions) which would have remained unsatisfied and buried if they had remained in ligand-free conformations.

Visual inspection and results trends indicate that, in 50% of the cases, the reason to satisfying the hydrogen-bond potential of a polar buried atom encouraged large side-chain rotations in the interface upon ligand-binding. Exhaustive modeling of possible large side-chain rotations in SLIDE around single bonds would mean sampling 300° of dihedral-angular space for each single bond. Given that a side chain can have from zero to four rotatable bonds, the combinatorial sampling, depending upon the fineness of sampling and the number of side chains to model, would be prohibitively time-consuming to do in docking.

3.4.1 Rotamer Libraries

Fortunately, for most of the bond-rotation angles, protein side chains adopt primarily a staggered dihedral angle such that covalently bonded, sp^3-hybridized atoms tend to stagger, rather than eclipse, their substituent atoms. As the number of solved protein structures has increased over the years, statistical distributions of the

Table 3.1: Visual analysis of reasons behind large dihedral rotations (> 60°) in interfacial side chains.

Structure	Residue#	Residue Type	Did side chain clash with the ligand ?	Would side-chain H-bond potential have remained if side chain had not moved ?
1aj2	221	LYS	NO	YES
1atp	E120	MET	NO	YES
1atp	E53	SER	NO	YES
1cgu	257	GLU	NO	YES
1gmr	B40	ARG	NO	YES
1pob	A52	ASN	NO	YES
1syd	43	GLU	NO	YES
1syd	115	TYR	NO	YES
1tps	192	GLN	NO	YES
7taa	296	HIS	NO	YES
1gmr	B38	GLN	NO	NO
1kel	L33	ASN	NO	NO
1bib	183	LYS	NO	YES
1lic	58	LYS	NO	NO
1ahb	70	TYR	YES	YES
1gmr	B65	ARG	YES	YES
1kel	H56	LYS	YES	YES
1pob	A30	ARG	YES	NO
1aj2	115	ASN	YES	NO
1lic	57	PHE	YES	NO
1pob	A2	LEU	YES	NO
1coy	122	MET	YES	NO

side chain bond-rotational angles for each residue type have been characterized[24]. These bond-rotation angles are labelled χ angles by convention, as explained in Figure 3.11. Side-chain χ angles have been found to occur in tight clusters around certain values. This is both because of the hybridization of the bonded atoms, as well as to prevent collisions with the main-chain atoms of the residue and its neighbors. Each single bond in the side chain can sample its dihedral degrees of freedom subject to these constraints. The resulting side chain conformation is called 'rotational isomer' of the side chain. Abbreviated as 'rotamer', this favored orientation is represented as a set of values, one for each dihedral angle degree of freedom. Since bond angles and bond lengths in proteins have rather small variances, they are usually not included in the definition. A library of such favored sets of side-chain χ-angle values for each residue type is called a rotamer library. Rotamer libraries usually also contain information about the frequency of each rotamer, as some conformations are more likely than others due to side chain stereochemistry, besides the information about the variance around the dihedral angle mean or mode.

Using rotamers to sample side chain conformation can help avoid exhaustive spatial sampling and make tractable the identification of a suitable combination of dihedral angles. With known geometries derived from experimentally solved crystal structures, not only can time be saved during sampling, it is also more likely to arrive at a side chain conformation that is natural and low in energy.

Using rotamer library for side chain modeling does have certain drawbacks. Firstly, rotamers are typically averaged conformations representing a cluster of conformations; they may not be real conformation themselves. Besides, granularity of dihedral space sampling of side chains within the rotamer library can itself be a limit, especially because rotamers largely reflect favored side-chain orientations outside of interfaces. Another aspect is that even though rotamers are thought to represent local energy minima on a potential energy landscape, not all rotamers may be local

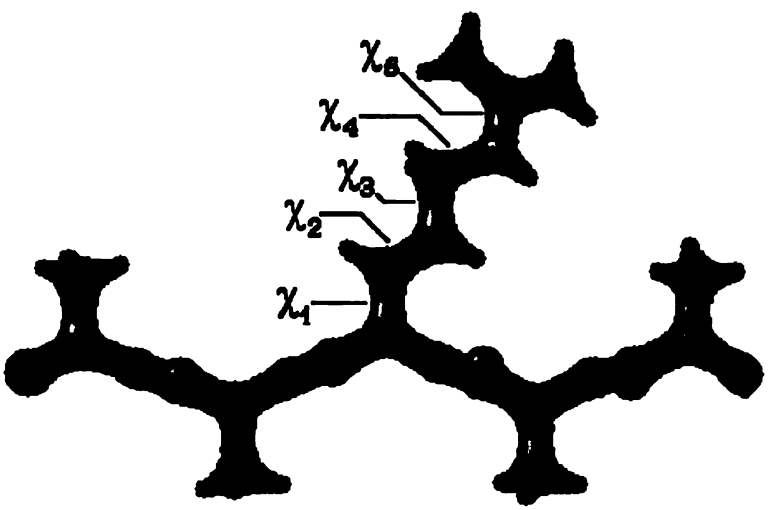


Figure 3.11: Illustration of bond-rotation angles associated with single bonds in an arginine side chain. The number of single bonds in a side chain, ranging from 0 to 5, depends on the residue type. A single bond is free to sample dihedral rotations and its associated dihedral angle is called a χ angle. The χ angles are labelled χ_1, χ_2, χ_3 and so on, according to the level of the single bond from the residue's backbone.

extrema since distributions of side chain dihedral angles are fairly broad ($\pm 20°$ or $30°$ relative to the average value presented).

Nevertheless, rotameric conformations can definitely be considered candidate conformations to sample, and the extent to which they approach optimal conformation can be assessed. For sampling larger conformational space for side chains, rotamer libraries offer time savings while providing ready-to-use, pre-calculated rotameric side chain conformations.

Rotamer libraries can be backbone-independent, backbone-dependent, or even secondary-structure-dependent. Backbone-dependent rotamers have dihedral angles and/or frequencies that are binned according to the local backbone conformation. Backbone-independent rotamers are calculated using all available side chains of a residue type. Backbone-dependent rotamers have the advantage that no rotameric side chain can have steric clashes with its own backbone.

To see if one can find rotamers from a library similar to ligand-bound conformations of few side chains which rotated through more than $60°$ on ligand-binding, the Dunbrack May 2002 backbone dependent rotamer library was searched (downloaded from http://dunbrack.fccc.edu/bbdep). Rotamer searches were done in incremental fashion - first comparing only χ_1, then χ_1 and χ_2 and so on up to χ_4, depending upon the number of χ angles in the side chain. A rotamer was considered similar to a side chain if each of the side-chain χ angles were within the range $\chi \pm \sigma$, where σ is the standard deviation of respective χ angle specified for the rotamer in the rotamer library.

As is clear from search results presented in Figure 3.12, for side chains with χ_3 and χ_4 angles, fewer suitably close rotamers were found. Furthermore, for a few of the side chains, like Tyr70 in 1ahc and Tyr115 in 1syc, no rotamer was found. This result, in effect, places an upper bound on how close, in χ-angular space, to a ligand-bound side chain conformation can rotamers from this particular library reach.

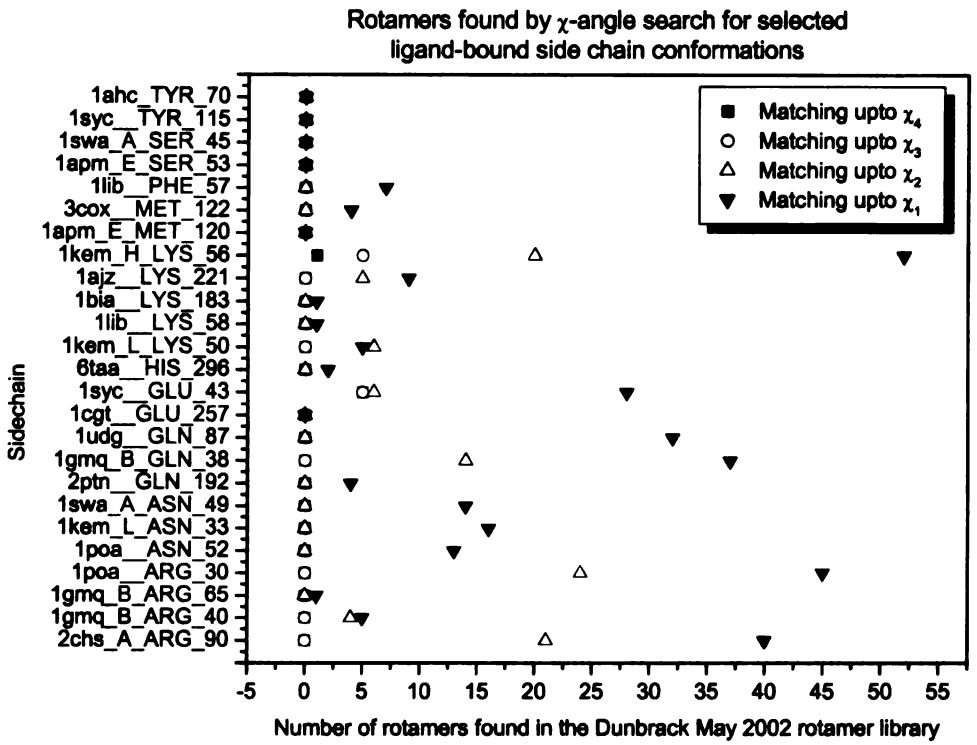


Figure 3.12: Number of rotamers, from Dunbrack May 2002 backbone-dependent rotamer library (http://dunbrack.fccc.edu/bbdep), approximating dihedral angles of target conformations for 25 interfacial side chains undergoing large rotation upon ligand binding. Selected side chains across structures are those that had rotated more than 60° from the ligand-free to ligand-bound conformation. Side chain names (Y-axis) include the PDB code, chain ID (if present), residue type and residue number. Backbone ϕ, ψ angles were used to locate the bins in which rotamers were searched, according the χ angles of the side chain in question. In the rotamer library, the ϕ, ψ resolution is 10°. The neighboring 8 bins, using $\phi \pm 10$, $\psi \pm 10$, were also searched. Rotamer searches were done in incremental fashion - first comparing only χ_1, then χ_1 and χ_2, and so on, up to χ_4, depending upon the number of χ angles in the side chain. A rotamer was considered similar to a side chain if each of the side-chain χ angles were within the range $\chi \pm \sigma$, where σ is the standard deviation of respective χ angle specified for the rotamer in the rotamer library. Rotamers searched consisted of only those rotamers which had the probability of at least 0.05 times the probability of the highest probable rotamer within the same ϕ, ψ bin. This helps exclude rotamers that were very rare, since many of them could be poorly resolved side-chain conformations in PDB.

Another worthwhile aspect to explore is to determine how close rotameric side chains, which are the 3D side-chain orientations generated from the rotamer library's ϕ, ψ and χ angles, can approach ligand-bound conformations in terms of RMSD. The reason for this is that single-bond rotations through adjacent χ-angle changes can either compensate for each other or augment each other in terms of the resulting side-chain motion. The closest rotamers found, in 3D Cartesian space, for ligand-bound conformations are presented Figure 3.13. Rotameric conformations within 1 Å of the ligand-bound conformations were found for about half of the 22 side chain cases presented here. This again effectively defines the limit on how close rotameric modeling approach can approach ligand-bound side-chain conformations using Dunbrack's backbone-dependent rotamer library.

In Figure 3.13, the relationship between B-values for the large-motion side chains and the probability of their having a rotameric conformation is presented. The motivation for this is that poor resolution of high B-value residues could be reflected in the unlikelihood of finding these conformations in the rotamer library. On a color-scale representing the maximum B-factor value found for any atom of the side chain, black to yellow indicate that the atom coordinates are low in mobility, while light-green to grey indicate that the side chain had at least one atom whose position was highly mobile and thus likely to be poorly resolved in the crystal complex. So, while rotamers close to ligand-bound conformations were found (RMSD $<$ 0.7 Å) for 1bib-Lys183 and 1gmr-ArgB40, high B-factor values indicate that these side chains do not have a well-defined conformation. Hence reliable conclusions may not be drawn about the effectiveness of a rotamer library for its ability or inability to find a close rotamer for side chains that consist of atoms having high B-factor values.

However, for most (13 out of 19) side chains having low B-factor values, a rotamer within an RMSD rang of 0.28 Å to 0.83 Å was found in each case. For 5 side chains having low B-factor values for which no close rotamers were found, three (1ahb-

Tyr70, 1cgu-Glu257 and 1poa-A52) moved to interact with the ligand, while 1poa-A2 moved due to excessive steric clashes with the ligand, indicating that their favored conformations are influenced by the ligand as well as the side-chain conformational energetics.

3.5 Conclusions

In this chapter, the bond-preservation bias discovered in 2 chapter was implemented in SLIDE. However, on further experiments it was found that SLIDE models small motions well enough during docking to preserve most of the hydrogen-bond interactions and, thus, additional bias towards preserving hydrogen bonds is not warranted. Relatively large side-chain rotations in ligand-bound conformations and causes for them were further investigated. Trying to satisfy buried polar groups was determined as a predominant reason. Investigating how large rotations can be modeled in SLIDE, the effectiveness of using Dunbrack backbone-dependent rotamer library to sample larger conformational space was analyzed. For χ_1 and χ_2 angles, the rotamer libraries always contained entries reflecting the conformations of side chains that underwent large side-chain motions upon ligand binding. However, for long side chains there were often no rotamers that also reflected their final ligand-bound χ_3 and χ_4 angles. In Cartesian RMSD space, 13 out of 22 ligand-bound side-chain conformer were represented in the rotamer library to within 1.0 Å RMSD and 8 of the remainder to within 2.0 Å RMSD. For all cases where side-chain mobility values were low, rotamers close to the ligand-bound conformations were found. At other times, side chains were found to be non-rotameric either because of ligand-induced strain[14, 15] in the interface or because they moved to enhance interactions with the ligand.

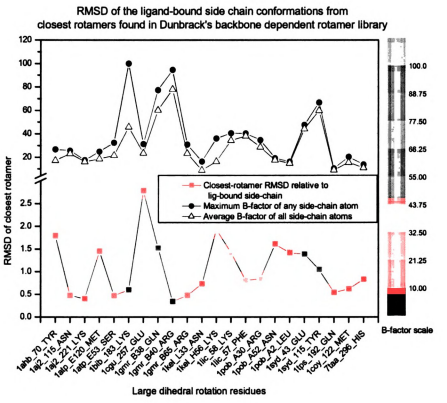


Figure 3.13: *This image is presented in color.* RMSD of rotamers from Dunbrack's May 2002 backbone-dependent rotamer library found to be closest to selected ligand-bound side-chain conformations in Cartesian space. Selected side chains across structures are those that had dihedral rotations more than 60° from the ligand-free to ligand-bound conformation. Side-chain observation names include the PDB code, chain ID (if present), residue type, and the residue number. Backbone ϕ, ψ angles were used to locate the bins in which rotamers were searched, using a ϕ, ψ resolution of 10°. The neighboring 8 bins, using $\phi \pm 10$, $\psi \pm 10$ were also searched. Each rotamer in these bins was converted to a side-chain orientation in Cartesian space and transformed into the reference frame of the selected side chain. The rotamer library consisted of only those rotamers which had a probability of at least 0.05 times the probability of the highest probable rotamer within the same ϕ, ψ bin. This helped exclude rotamers that were rare, since they could be poorly resolved side-chain conformations in the PDB. Maximum and average B-factor values over all the side-chain atoms are also shown in the top half of the figure. So, while close rotamers (RMSD < 0.7 Å) were found for 1bib-Lys183 and 1gmr-ArgB40, high B-factor values convey that some target side-chain positions are in fact poorly defined. In general, there is no trend between the B-value mobility of a side-chain and the probability that this conformation is present in the rotamer library.

Chapter 4

Scoring SLIDE Dockings

4.1 Introduction

Computer-aided structure-based methods for virtual screening are aimed at predicting the binding mode of a ligand in the binding site of a protein or molecular target, and estimating the resulting binding affinity. A conformational search algorithm can produce an immense number of solutions from which the right solution(s) must be selected by an energy-based or other scoring procedure. Thus, it is extremely valuable to predict binding affinity accurately so the user may select good conformations and orientations from poor ones with confidence.

For ranking protein-ligand dockings, a class of programs, namely 'scoring functions' has been actively researched for many years. Since docking algorithms search for potentially good binding modes between a specified target and, and in some cases, thousands of small molecules, they must have the ability to discriminate not only between promising and poor binders, but also between different binding modes of the same ligand. Theoretically, free-energy simulations can be a reliable check for discriminating good from bad dockings, but this approach is too expensive for screening large libraries of small molecules and remains a techniques performed by relatively

few experts.

Scoring functions can be broadly categorized into three categories: force-field based, knowledge-based and empirical. Force-field based scoring functions apply classical molecular-mechanics based energy functions to approximate the binding free energy of the protein-ligand using a sum of van der Waals, electrostatic, bond length, and bond angle terms. Solvation is usually taken into account using a distance dependent dielectric or solvent-accessible surface term, although explicit modeling of discrete water molecules is also done. Non-polar contributions are usually assumed to be proportional to the solvent-accessible surface area. A drawback is that energy landscapes associated with the force-field potentials are usually rugged, and therefore minimization is required prior to any energy evaluation. Furthermore, small inaccuracies in positioning atoms during docking can lead to large discrepancies in the energy score.

Knowledge-based scoring functions represent the binding affinity as a sum of protein-ligand pairwise atomic interactions. Using the protein-ligand complexes deposited in the Protein Data Bank (PDB), knowledge-based potentials are derived from the observed preferences for particular atom-pair interactions to occur at given distances pr distance ranges between ligands and proteins. However, structures in PDB do not provide a thermodynamic ensemble at equilibrium. Hence knowledge-based potentials should be considered as a statistical preference rather than an actual potential.

Recently, many empirical functions have emerged as alternatives to force-field-based and knowledge-based scoring functions. Unlike force fields, the weights of terms in empirical scoring functions are directly calibrated using a set of protein-ligand complexes with experimentally determined structures and binding affinities through multivariate regression analysis. These scoring functions may also include derived terms that include interactions that would otherwise involve several terms in

a force-field based function, e.g. terms for hydrogen bonds and hydrophobic interactions. Empirical scoring functions have several appealing features. They can be calibrated against a set of diverse protein-ligand complexes. Secondly, each term in an empirical function has an intuitive physical meaning. The resulting regression coefficients before each term conveys the relative contribution of each interaction in the ligand-binding process. Thirdly, empirical scoring functions are extremely fast with reasonable accuracy, which makes them acceptable for structure-based drug design applications like virtual database screening and *de novo* ligand generation. Finally, most force-field-based energy functions have been parameterized solely to fit peptidyl data, and are just beginning to be parameterized to recognize and accurately represent atom and bond types that occur in non-protein ligand molecules.

4.2 SLIDE Scoring Function

The current SLIDE scoring function is an empirical scoring function trained to score a collision-free complex using a linear combination of hydrophobic and hydrogen-bond interactions terms:

$$\text{SCORE(P, L)} = A \cdot \text{HPHOB(P, L)} + B \cdot \text{HBONDS(P, L)}$$

The weights A and B have been fit to optimize SCORE(P, L) to match the affinities of 89 high-resolution complexes listed in reference [8]. For more information about the terms and their tuning in current version of SLIDE, please see reference [42].

While the current SLIDE scoring function is based on hydrogen-bonding and hydrophobic terms, more detailed terms to represent both stabilizing and destabilizing thermodynamic interactions can be explored for developing new scoring functions for SLIDE. Inclusion of such terms is also desirable to help discriminate among different

binding orientations generated while sampling rotamers for buried unsatisfied side chains and to detect interactions that could be improved through further optimization.

4.3 Methods

Empirical scoring functions are typically trained and tested against series of experimentally determined protein-ligand crystal complexes whose binding affinities are known. Unfortunately, no experimental data exists for suboptimal orientations of ligands with proteins. For designing a new empirical scoring function for SLIDE to be used during and after docking, we developed and tested a series of scoring functions resulting from linear combinations of individual terms listed in Table 4.1. The choice of which terms to combine was driven by low-correlation among the terms, as well their ability to represent different interaction types.

Table 4.2: Scoring function variants evaluated to predict binding affinities. 269 crystal complexes, listed in Table 4.3, with experimentally known affinities were used for training and testing the scoring functions. Terms are as defined in Table 4.1. Linear regression constant α, and coefficients β through η were determined to best fit the binding affinity values.

Scoring Function	Expression
f_1	$\alpha + \beta A$
f_2	$\alpha + \beta B$
f_3	$\alpha + \beta C$
f_4	$\alpha + \beta D$
f_5	$\alpha + \beta E$
f_6	$\alpha + \beta F$
f_7	$\alpha + \beta G$
f_8	$\alpha + \beta H$
f_9	$\alpha + \beta I$
f_{10}	$\alpha + \beta J$
f_{11}	$\alpha + \beta K$
f_{12}	$\alpha + \beta L$
f_{13}	$\alpha + \beta M$
f_{14}	$\alpha + \beta N$
f_{15}	$\alpha + \beta O$
f_{16}	$\alpha + \beta P$
f_{17}	$\alpha + \beta Q$
f_{18}	$\alpha + \beta R$
f_{19}	$\alpha + \beta S$

Continued on next page

Table 4.2 – continued from previous page

Scoring Function	Expression
f_{20}	$\alpha + \beta T$
f_{21}	$\alpha + \beta U$
f_{22}	$\alpha + \beta V$
f_{23}	$\alpha + \beta J + \gamma B$
f_{24}	$\alpha + \beta(J + K) + \gamma B$
f_{25}	$\alpha + \beta(H + I) + \gamma B$
f_{26}	$\alpha + \beta D$
f_{27}	$\alpha + \beta D + \gamma I$
f_{28}	$\alpha + \beta D + \gamma(I + K)$
f_{29}	$\alpha + \beta I + \gamma(G + H) + \delta(J + K + L + M) + \epsilon P + \zeta R$
f_{30}	$\alpha + \beta I + \gamma(G + H) + \delta(J + K + L + M) + \epsilon P$
f_{31}	$\alpha + \beta I + \gamma(G + H) + \delta(P)$
f_{32}	$\alpha + \beta B + \gamma(I + G + H) + \delta P + \epsilon R$
f_{33}	$\alpha + \beta B + \gamma(I + G + H) + \delta R + \epsilon$
f_{34}	$\alpha + \beta B + \gamma I + \delta R + \epsilon$
f_{35}	$\alpha + \beta B + \gamma I$
f_{36}	$\alpha + \beta O + \gamma(I + G + H) + \delta(J + K + L + M)$
f_{37}	$\alpha + \beta O + \gamma(I + G + H) + \delta(J + K + L + M) + \epsilon R$
f_{38}	$\alpha + \beta O + \gamma(I + G + H) + \delta(J + K + L + M) + \epsilon P$
f_{39}	$\alpha + \beta S + \gamma I + \delta(G + H) + \epsilon(J + K + L + M) + \zeta P + \eta R$
f_{40}	$\alpha + \beta S + \gamma I + \delta(G + H) + \epsilon(J + K + L + M) + \zeta P$
f_{41}	$\alpha + \beta S + \gamma I + \delta(G + H) + \epsilon P$
f_{42}	$\alpha + \beta S + \gamma(I + G + H) + \delta P + \epsilon R$
f_{43}	$\alpha + \beta S + \gamma(I + G + H) + \delta P$
f_{44}	$\alpha + \beta S + \gamma(I + G + H) + \delta R$
f_{45}	$\alpha + \beta S + \gamma I + \delta R$
f_{46}	$\alpha + \beta S + \gamma I$
f_{47}	$\alpha + \beta S + \gamma(I + G + H) + \delta(J + K + L + M)$
f_{48}	$\alpha + \beta S + \gamma(I + G + H) + \delta(J + K + L + M) + \epsilon R$
f_{49}	$\alpha + \beta S + \gamma(I + G + H) + \delta(J + K + L + M) + \epsilon P$
f_{50}	$\alpha + \beta B + \gamma I + \delta G + \epsilon H$
f_{51}	$\alpha + \beta S + \gamma I + \delta G + \epsilon H$
f_{52}	$\alpha + \beta B + \gamma I + \delta(G + H)$
f_{53}	$\alpha + \beta S + \gamma I + \delta(G + H)$
f_{54}	$\alpha + \beta T + \gamma I + \delta(G + H)$
f_{55}	$\alpha + \beta T + \gamma I + \delta(G + H) + \epsilon P$
f_{56}	$\alpha + \beta T + \gamma(I + G + H) + \delta(J + K + L + M) + \epsilon R$
f_{57}	$\alpha + \beta S + \gamma G + \delta(H + I) + \epsilon L + \zeta J$
f_{58}	$\alpha + \beta T + \gamma G + \delta(H + I) + \epsilon L + \zeta J$
f_{59}	$\alpha + \beta S + \gamma G + \delta(H + I) + \epsilon L + \zeta J + \eta V$
f_{60}	$\alpha + \beta U + \gamma G + \delta(H + I) + \epsilon L + \zeta J$
f_{61}	$\alpha + \beta U + \gamma G + \delta(H + I) + \epsilon L + \zeta J + \eta V$

4.3.1 Calculation of Terms for Scoring Function

The overall free energy change in a protein-ligand binding can be represented as

Table 4.1: Potential terms to capture hydrophobic and polar interactions as well as molecular size and entropy while scoring a ligand-protein complex.

Ligand-size terms	
A	number of ligand atoms
O	number of interfacial ligand atoms
N	percentage of ligand atoms that are interfacial
Hydrophobic terms	
B	sum of atomic hydrophobicity values[2] for P-L[1] hydrophobic contacts
C	difference of atomic hydrophobicity values[2] for P-L hydrophobic contacts
S	number of P-L hydrophobic contacts (without considering hydrophobicity values)
U	hydrophobic term from current SLIDE score [42]
Polar terms	
E	sum of atomic hydrophilicity values for P-L hydrophilic contacts
F	difference of atomic hydrophilicity values for P-L hydrophilic contacts
G	number of favorable protein-bound metal interactions with ligand
H	number of P-L salt bridges
I	number of P-L hydrogen bonds
Mismatch terms	
P	number of exposed hydrophobic ligand atoms
D	difference of atomic hydrophobicity values for hydrophobic-hydrophilic contacts
J	number of interfacial, unsatisfied polar atoms
K	number of interfacial, unsatisfied charged atoms
L	number of P-L repulsive polar interactions
M	number of P-L repulsive charge interactions
W	number of hydrophobic-hydrophilic P-L contacts
Entropic terms	
Q	number of rotatable bonds in the ligand
R	number of rotatable interfacial bonds in both ligand and protein
V	number of rotatable interfacial bonds in ligand

[1] Protein-ligand.
[2] from Kuhn et al [27].

$$\Delta G_{bind} = \Delta G_{protein-ligand\ h-bond} + \Delta G_{salt-bridge} + \Delta G_{hydrophobic} + \Delta G_{unsatisfied\ buried\ polar\ atoms} + \Delta G_{entropy} + \alpha \quad (4.1)$$

Here, ΔG_{h-bond} accounts for the hydrogen bonding between the ligand and the protein and $\Delta G_{hydrophobic}$ accounts for the protein-ligand hydrophobic interactions. $\Delta G_{unsatisfied\ buried\ polar\ atoms}$ penalizes the score if dockings have unsatisfied polar atoms buried in the interface, while $\Delta G_{entropy}$ coarsely approximates the rotational entropy by counting the number of interfacial bonds, in either the ligand only or both the ligand and the protein. α is the regression constant which implicitly includes other effects for which no explicit scoring function term is included such as solvation and rotational/translational main-chain entropy changes.

The complexes being scored for training or testing are assumed to be free from any van der Waals overlaps among atoms. This is true for PDB structures in the data set. When dockings are scored from within SLIDE, van der Waals overlaps are resolved by directed rotations before scoring a docking. Interactions are determined in the following order:

1. **Intermolecular repulsive polar contacts** If any ligand and protein atom-pair consists of both atoms as hydrogen-bond acceptors or donors, and their interatomic distance is between 2.5 Å and 3.5 Å, the interaction is considered as repulsive.

2. **Protein-ligand metal-bonds** If the interaction is not a repulsive polar contact, and the protein-ligand atom-pair includes an acceptor or a donor or a donor/acceptor, that will match either donor or acceptor, ligand-atom that is within of 2.6 Å for Co, Cu, Fe, Mg, Mn, Ni, or Zn metal ions or within 2.9 Å

for Ca, Na, or K metal ions bound to the protein, the interaction is considered as a favorable metal interaction.

3. **Protein-ligand salt bridges** If an interaction is not characterized as meeting the criteria in (1) or (2), and the atoms are complementary (acceptor matched to donor, or a donor/acceptor that will match either donor or acceptor), interatomic distance is between 2.5 Å and 4.5 Å, and the charges of the atoms are non-zero and complementary, then the interaction is counted as a salt bridge.

4. **Protein-ligand hydrogen bond** If none of the above, the atom-pair is evaluated for an hydrogen bond . To calculate intermolecular hydrogen bonds, the position of the shared hydrogen in each intermolecular hydrogen bond is computed, if not provided in the protein or ligand structure. Hydrogen atoms and partial charges were added to the ligand structures by utilizing molcharge option in the *AM1BCC* [18]. If a potential hydrogen bond needed to be explored when no hydrogen atom was present in the protein structure, the position of the hydrogen atom was analytically determined. This position is well-defined by bond lengths and angles for all but the terminal hydrogen atoms in lysine and hydroxyl side chains; for these side chains, the optimal hydrogen-atom position is chosen, with respect to maximizing hydrogen bonds, on the circle of possible positions defined by covalent bonding constraints. All hydrogen bonds with a donor-acceptor distance up to 3.5 Å and a donor-hydrogen-acceptor angle larger than 120° contribute equally to the score. If water molecules are included in the interface, all water-mediated hydrogen bonds between protein and ligand are also counted [42].

5. **Hydrophobic interactions** A protein-ligand atom-pair is evaluated for making a hydrophobic interaction if it has been found not to participate in any of the previous interactions. The interatomic distance for the atom pair must

be within 4.5 Å to qualify for a hydrophobic interaction. Calculations for the hydrophobicity measure, term U, is explained in detail in [42]. Other hydrophobic terms are either counts of hydrophobic contacts, or sums or differences in hydrophobicity values.

6. **Intra-molecular polar interactions** At this point intra-molecular salt bridges, direct hydrogen bonds, and water-mediated hydrogen bonds are evaluated, in this order, within the protein and the ligand for atoms that did not contribute to any of the previous protein-ligand interactions.

7. **Unsatisfied buried polar atoms** If any interfacial polar or charged protein or ligand atom remains unpaired in any of the above interactions, it is counted as unsatisfied.

8. **Entropic terms** Counts of single (rotatable) bonds in the entire ligand and in the interfacial residues of the ligand and protein are used to generate terms that partially measure loss of degrees of freedom upon ligand binding.

When evaluating score of a docking from within SLIDE, interactions that are lost due to docking are also considered. For instance, unsatisfied protein atoms which previously had intra-protein hydrogen bonds that were lost due to side-chain rotations, or hydrogen bonds to waters that were displaced upon ligand docking, are considered in the penalty terms J and K.

4.3.2 Preparation of Test Set

The data set used in this study is constructed from 269 protein-ligand complexes specified in Table 4.3. The resolution range of the crystal complexes was from 0.95 Å to 3.16 Å with 85% of the complexes having resolution $\leq$ 2.5 Å. This set was assembled from training sets used by other empirical scoring function studies[8, 37, 48]. All of

these complexes have crystal structures and experimentally measured affinity values. Coordinates of all the complexes were downloaded from the Protein Data Bank. No energy minimization was performed on structures. Each complex was split into a protein molecule, which was saved in PDB file format, and a ligand molecule, which was saved in Mol2 file format. Metal ions, if present in the protein-ligand interface, were kept in the protein file. Water molecules were excluded and will be considered in future work. Any other organic or inorganic cofactors were kept with the protein.

Values were generated for each of the terms mentioned in Table 4.1 for each of the complexes in Table 4.3. Before combining various terms, the linear correlation coefficients between the terms were calculated. Combining highly correlated terms was avoided when defining new combinations as candidate scoring functions.

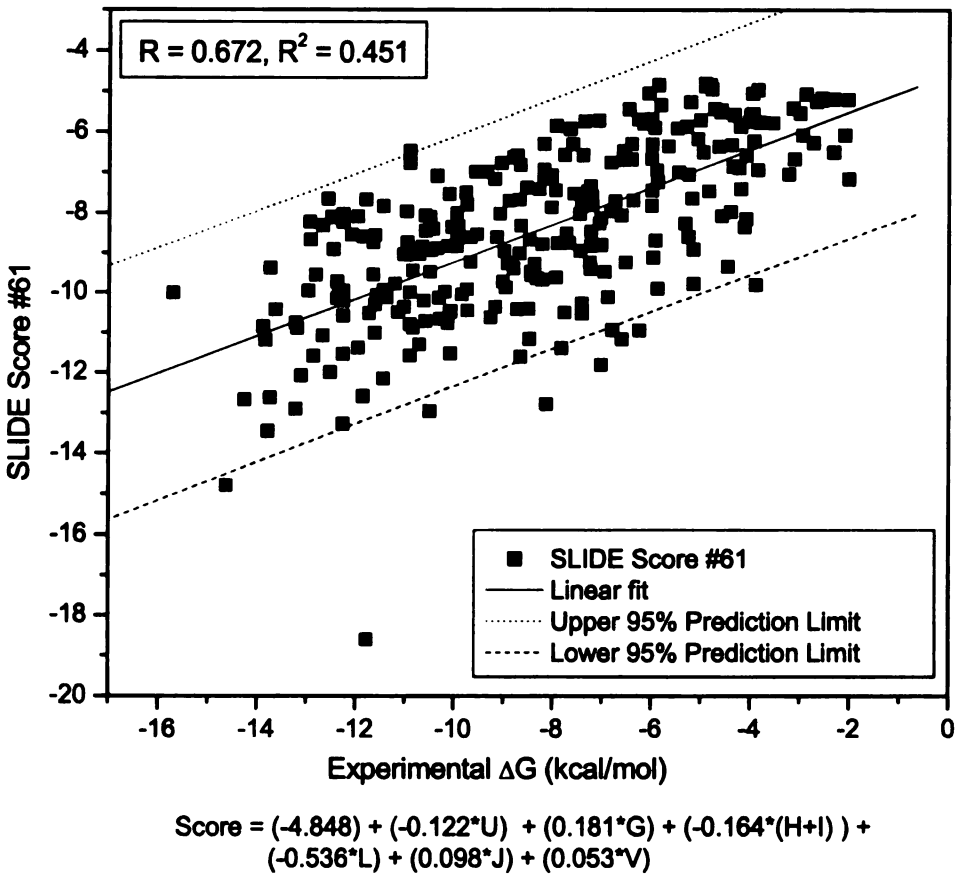
4.3.3 Training and Testing

Since scoring functions evaluated were linear expressions, training for a scoring function involved performing *linear multiple regression* between the binding affinity and scoring function terms to determine the weights of each terms and the constant, α, and to minimize the error between the predicted and actual affinity values. Raw terms specified in Table 4.1 were calculated for each of the complex specified in Table 4.3. Then the set of 23 terms for 269 complexes was randomly divided into halves. One half was used for training to derive regression coefficients (term weights) and the constant, then these weights were used with the values of terms to predict binding affinities for the complexes in second half. To avoid over-fitting or training on any specific subset of the complexes, random repartitioning of the 269 complexes into halves was performed 10 times, and the training and testing sets were also interchanged for each repartition. For each scoring function, its weights, constants and the resulting correlation coefficient with experimentally determined binding affinity values were determined across these 20 training and testing sets. These values are presented in

Table 4.3: PDB codes of crystal complexes used for training and testing the scoring functions

1a46	1bmn	1fq8	1rus	2xis	6tim
1a4w	1bxo	1g6n	1sre	2ypi	6tmn
1a5g	1bxq	1hbv	1tet	3cla	7abp
1a94	1bzm	1hdt	1tha	3cpa	7acn
1aaq	1cbx	1hew	1tlp	3csc	7cat
1abe	1cil	1hih	1tmn	3er3	7cpp
1abf	1cim	1hos	1tmt	3er5	7dfr
1ac4	1cin	1hpv	1tng	3fx2	7est
1ac8	1cla	1hpx	1tnh	3pgm	7hvp
1acj	1cnw	1hri	1tni	3ptb	7tim
1adb	1cnx	1hsg	1tnj	3tmn	7upj
1add	1cny	1hsl	1tnk	3ts1	8abp
1adf	1cps	1htf	1tnl	4cla	8acn
1ae8	1csc	1htg	1tpp	4dfr	8atc
1aeb	1ctt	1hvi	1ulb	4er1	8cpa
1aed	1d3d	1hvj	1uvs	4er2	8cpp
1aee	1d3p	1hvk	1xig	4er4	8hvp
1aef	1d3q	1hvl	1xli	4fab	8icd
1aeg	1d3t	1hvr	2ak3	4gr1	8xia
1aeh	1dbb	1hvs	2cgr	4hvp	9aat
1aej	1dbj	1l83	2cpp	4mdh	9abp
1aek	1dbk	1l87	2csc	4phv	9hvp
1aem	1dbm	1ldm	2dbl	4sga	9icd
1aen	1dif	1lgr	2dri	4tim	9rub
1aeo	1dih	1mbi	2er0	4tln	
1aeq	1dog	1mcf	2er6	4tmn	
1aes	1dr1	1mcj	2er7	4ts1	
1aet	1drf	1mcs	2er9	4xia	
1aeu	1dwb	1mdq	2gbp	5abp	
1aev	1dwc	1mfc	2ifb	5acn	
1af2	1dwd	1mfe	2ldb	5cna	
1ajv	1eed	1mnc	2mcp	5cpp	
1ajx	1ela	1nnb	2msb	5enl	
1anf	1elc	1nsc	2phh	5er2	
1apt	1ent	1nsd	2pk4	5hvp	
1apu	1epo	1okl	2qwc	5icd	
1apv	1epp	1pgp	2qwd	5ldh	
1apw	1etr	1phf	2qwe	5p21	
1avn	1ets	1phg	2qwf	5sga	
1b5g	1ett	1ppc	2qwg	5tim	
1ba8	1fbc	1pph	2r04	5tln	
1bai	1fbf	1ppk	2rnt	5tmn	
1bap	1fbp	1ppl	2tmn	5xia	
1bbz	1fkb	1ppm	2tsc	6abp	
1bcu	1fkf	1rbp	2upj	6apr	
1bhf	1fmo	1rgk	2wea	6cpa	
1bid	1fq4	1rne	2web	6enl	
1bll	1fq5	1rnt	2wec	6gst	

Table 4.4.



Score = (-4.848) + (-0.122*U) + (0.181*G) + (-0.164*(H+I)) + (-0.536*L) + (0.098*J) + (0.053*V)

Figure 4.1: Correlation between experimental binding affinity and score values determined by SLIDE scoring function 61.

4.4 Results

Scoring functions 58, 60 and 61 have the highest correlations with binding affinity affinities. Scoring function 61 was chosen since the standard deviation of its weights

Table 4.4: Predicted and experimental binding-affinity linear correlation coefficients and average weights derived from linear multiple regression. 20 rounds of training and testing were performed to best fit the experimental affinities.

Scoring func.	Avg Corr.	α	β	γ	δ	ϵ	ζ	η
1	0.511	-5.972	-0.085					
2	0.522	-5.506	0.259					
3	0.485	-5.57	-1.172					
4	0.474	-5.212	-0.109					
5	0.339	-6.389	-0.254					
6	0.324	-6.769	-0.549					
7	0.041	-8.51	0.404					
8	0.001	-8.429	-0.007					
9	0.329	-6.917	-0.274					
10	0.252	-7.217	-0.194					
11	0.221	-7.777	-0.457					
12	0.286	-7.583	-0.765					
13	0.104	-8.268	-0.192					
14	0	-8.857	0.431					
15	0.56	-5.485	-0.111					
16	0.102	-8.293	-0.132					
17	0.415	-6.811	-0.125					
18	0.432	-5.716	-0.1					
19	0.51	-5.435	-0.053					
20	0.593	-5.804	0.752					
21	0.542	-6.192	-0.097					
22	0.404	-6.705	-0.163					
23	0.596	-5.509	-0.1	0.242	0.16			
24	0.606	-5.529	-0.122	0.246	0.141			
25	0.579	-4.535	-0.043	-0.098	0.171			
26	0.528	-4.75	-0.734	-0.064	0			
27	0.526	-4.689	-0.828	-0.039	-0.104			
28	0.525	-4.74	-0.791	-0.045	-0.06			
29	0.56	-4.323	0.285	-0.206	-0.148	-0.004	-0.001	0.027
30	0.561	-4.36	0.267	-0.161	-0.119	0.017	0.009	
31	0.567	-4.335	0.255	-0.16	-0.114	0.01		
32	0.565	-4.315	0.292	-0.168	0.002	0.022		
33	0.565	-4.319	0.293	-0.169	0.023			
34	0.551	-4.784	0.219	-0.145	-0.014			
35	0.55	-4.884	0.231	-0.169	0			
36	0.571	-5.631	-0.144	-0.025	0.124			
37	0.58	-5.798	-0.187	-0.105	0.103	0.08		
38	0.573	-5.613	-0.149	-0.025	0.125	0.091		
39	0.56	-4.055	-0.057	-0.206	-0.145	-0.006	-0.018	0.017
40	0.562	-4.085	-0.054	-0.177	-0.126	0.008	-0.01	
41	0.568	-4.084	-0.053	-0.175	-0.123	-0.011		
42	0.565	-4.044	-0.058	-0.165	-0.014	0.012		
43	0.569	-4.053	-0.055	-0.148	-0.009			
44	0.565	-4.049	-0.058	-0.164	0.011			
45	0.552	-4.567	-0.044	-0.143	-0.021			
46	0.547	-4.722	-0.047	-0.184	0			
47	0.564	-4.071	-0.057	-0.149	0.011			
48	0.562	-4.045	-0.058	-0.163	0.003	0.009		
49	0.563	-4.061	-0.056	-0.15	0.011	-0.01		
50	0.562	-4.336	0.255	-0.16	-0.177	-0.106		
51	0.563	-4.089	-0.053	-0.175	-0.169	-0.117		
52	0.567	-4.344	0.254	-0.16	-0.112			
53	0.568	-4.094	-0.053	-0.175	-0.122			
54	0.604	-5.178	0.727	-0.082	-0.085			
55	0.605	-5.175	0.743	-0.081	-0.089	0.049		
56	0.611	-5.551	0.977	-0.155	0.031	0.057		
57	0.596	-4.184	-0.064	0.081	-0.128	-0.465	0.149	
58	0.628	-5.58	0.839	0.126	-0.055	-0.399	0.154	
59	0.594	-4.163	-0.064	0.087	-0.13	-0.466	0.147	0.002
60	0.626	-5.074	-0.111	0.19	-0.128	-0.505	0.129	
61	0.627	-4.848	-0.122	0.181	-0.164	-0.536	0.098	0.053

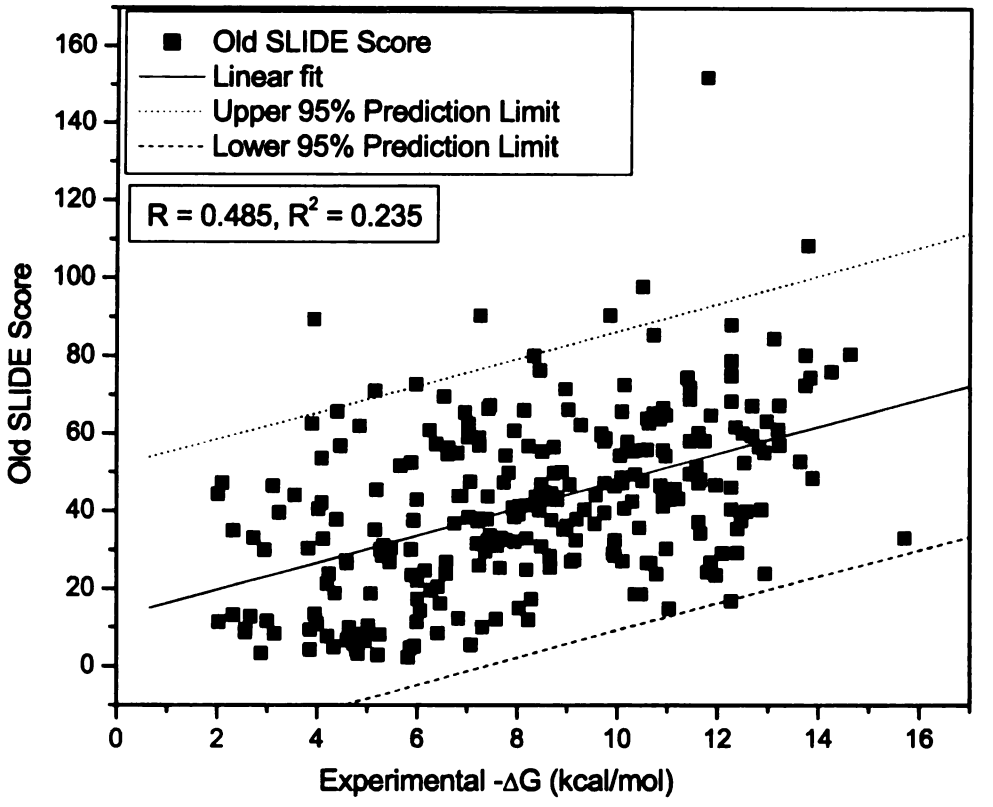


Figure 4.2: Correlation between experimental binding affinity and score values determined by the previous SLIDE scoring function.

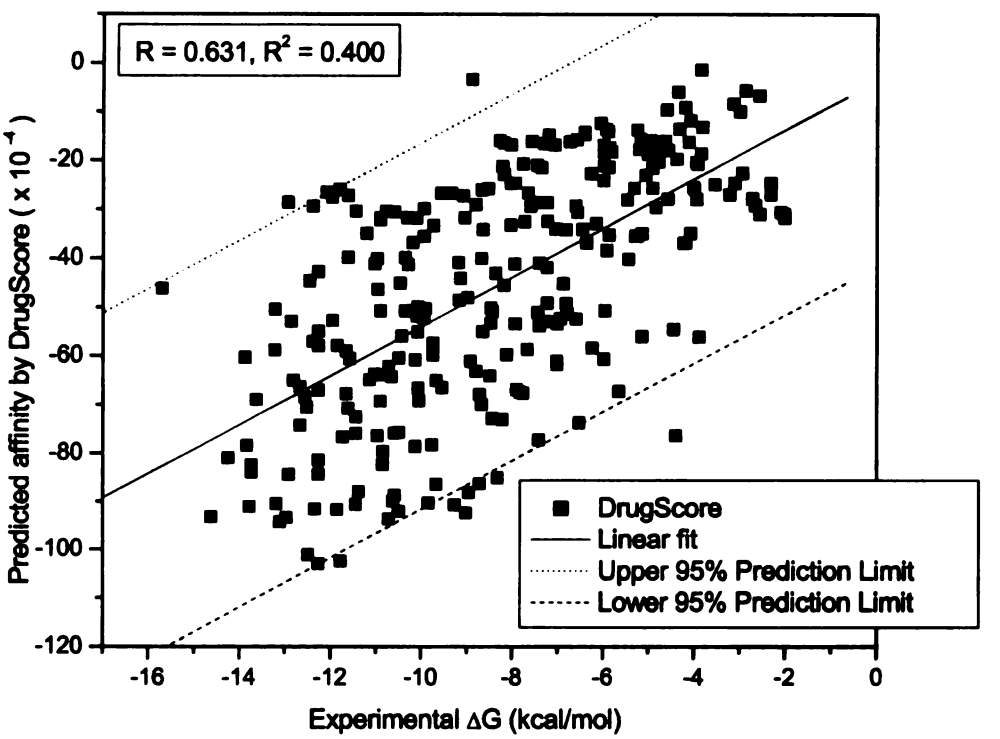


Figure 4.3: Correlation between affinity values known experimentally and those determined by DrugScore [12].

and the regression constant was least among the three (data not shown), indicating that these weights do not vary much as training set varies. Apart from representing hydrogen bonds and hydrophobic interactions, function 61 also includes terms for metal interactions, the entropic cost of ligand binding as well as penalizing the score for unsatisfied polar or charged atoms buried in the interface.

Figures 4.1 and 4.2 present the correlation of experimentally known affinities against scores predicted by scoring function 61 and old-SLIDE scoring function respectively. Not only scoring function 61 perform much better than old scoring function, but it performs comparable to DrugScore [12], whose correlation with known affinities of same 269 complexes is presented in Figure 4.3.

4.5 Conclusions

When sampling larger rotations to model large side-chain displacements in SLIDE is desirable, a suitable scoring function representing favorable as well unfavorable interactions is needed to discriminate among the candidate side-chain conformations. Scoring function 61 developed by our team not only predicts affinities more accurately than our previous scoring function, but it also performs somewhat better than the popular knowledge-based scoring function DrugScore and is considerably (10-fold) faster. Thus, this scoring function has been employed to evaluate how side-chain rotamers influence the affinity of the protein-ligand complex, and select among the rotamers, as described in the following chapter.

Chapter 5

Guiding Sampling by Score

While rotamers present a mechanism to model side chain flexibility through a range of motions, a scoring function is used to determine which specific rotamers enhance the interactions over the existing side chain conformation. Two requirements need to be met. One is to identify which side chains provide definite opportunities to enhance interactions through rotations after a steric overlap-free docking has been achieved by SLIDE. Second is to determine which orientations for these side chains collectively enhance the interactions the most. The work and results described in the previous two chapters provide tools to address these requirements. While buried unsatisfied groups are typically not observed in known protein-ligand complexes and provide definite interaction-enhancing opportunities, rotamers provide a reasonable method to sample larger rotations needed to satisfy hydrogen bonding groups. Using a scoring function that rewards favorable orientations as well as penalizes unsatisfactory or unfavorable orientations would help choose from several rotamers for a side chain.

If there are multiple rotamers for multiple side chains which improve interactions for unsatisfied groups buried in the interface, there are exponentially many possible configuration sets to choose from. If each single bond can sample 3 low-energy χ angles, *gauche*$^-$, *gauche*$^+$ and *trans*, the number of possible rotamers for a side chain

with 'c' χ angles is 3^c. Further, if there are 'r' residues to sample for rotamers, the total number of possible conformations for the 'r residues is 3^{rc}. For instance, if there were four side chains to be optimized with three χ angles each, there would be 3^{12} configurations to test, which is unfeasible except if many configurations can be ruled out by dead-end elimination or other pruning approaches.

5.1 Sampling Choices for Maximizing Score

Thus, generating and evaluating ensembles of side-chain configurations for enhancing dockings is expensive. To reduce the configurational search space, the number of rotamers sampled from the rotamer library is limited to only rotamers with backbone conformation similar to the side chain being sampled and having a backbone-dependent probability of at least 5% of that of the highest probability rotamer. Rotamers with probability lower than 5% of the threshold represent either poorly resolved side-chain conformations or rare geometries which are unlikely to occur frequently enough to be considered as backbone dependent rotamers (personal communication with Roland Dunbrack, developer of this rotamer library).

Since more than one rotamer could potentially enhance the interactions for a residue, measured by the change in the score, one could keep the rotamer that increases the score the most with respect to the initial steric overlap-free docked conformation of the ligand and protein achieved by SLIDE. However, this may lead to a set of rotamers which may not be compatible together due to van der Waals overlaps or which may even decrease the overall score due to repulsive contacts amongst the repositioned side chains. Moreover, the backbone-dependent probability for a rotamer may be low when compared to other acceptable rotamers. Alternatively, two or more rotamers may enhance the score equally, but one may be more likely of all.

To avoid local minima as well as choose high-probability rotameric conformations

for side chains, self-consistent mean-field optimization is used to maximize the chemical complementarity in the interface after the induced-fit docking has been achieved by SLIDE (see section 3.2). The mean-field optimization is based on multi-copy sampling. For each residue being optimized, these multiple copies are generated by sampling through high-probability rotamers with their backbone conformations same as the residue's. Prior to sampling, the induced-fit docked complex is recorded as the *base conformation*. This base conformation is restored before any rotamer is sampled for any residue. This ensures that every rotamer that is modeled for a side chain is evaluated in the same environment. More importantly, this also ensures that the rotamers for all side chains are evaluated in the same environment, thus avoiding any dependence on the order in which side chains are optimized. As each rotamer is modeled as a side chain for each residue, any resulting steric overlaps are resolved through the induced-fit mechanism. Any substituted rotamer causing unresolvable overlaps is discarded. From these overlap-free rotameric conformations for a side chain, maximum n are retained based on the magnitude of score improvement. In our experiments, n is 5, though it can be easily scaled to larger number of the rotamers used.

With these multiple copies, or rotameric conformations, for each residue being optimized, the probability and cost matrices for the mean-field optimization are constructed. Each element, $P(i,j)$, of probability matrix, P, describes the probability that rotamer j is optimal for residue i. To encourage final side-chain conformations to be highly likely rotamers, each element is assigned the respective backbone-dependent probability of the rotamer derived from the rotamer library. For each probability entry $P(i,j)$, an initial cost value $S'(i,j)$ is assigned the value of score improvement brought by rotamer j for residue i over the base conformation. In each iteration of

the optimization, the mean cost for each rotamer is computed as:

$$S(i,j) = S'(i,j) \cdot P(i,j)$$

At the end of each cycle, the entries in the probability matrix are updated based on the mean costs $S(i,j)$ using the Boltzmann principle:

$$P(i,j) = \frac{e^{-S(i,j)/\mu}}{\sum_k e^{-S(i,j)/\mu}}$$

where μ is the average value of all computed mean costs. In effect, probability for each rotamer for a residue is refined iteratively by Boltzmann weighting each rotamer by a term encoding both the score-improvement as well as the rotamericity bias introduced by the initial backbone-dependent probabilities. Convergence of the values in the probability matrix is usually observed in fewer than ten cycles, and those rotamers with the highest probability are chosen to model the side chains for the selected residues.

This class of mean-field methods, called self-consistent mean-field optimization, used to determine side-chain conformations to achieve the global energy minimum (in our case, the global score minimum) is inspired from [42, 26, 17]. However, the quality of solution depends critically on the choice of conformations for sampling - here, the rotamers - which serves as the basis for the probability distribution. Similarly, the probability function, which includes the score improvement as well as rotamer probabilities, also critically affects the quality of results.

In Figure 5.1, the flowchart of the algorithm used to select optimal combination of rotamers for a set side chains is presented. A protein-ligand docking free of van der Waals overlaps, achieved through induced-fit docking by SLIDE, defines the base conformation of the interface for which rotamers are sampled. After achieving the base configuration, residues consisting of buried unsatisfied polar groups are identified

by scoring function 61 (see previous chapter. Then, rotamer conformations having the same $\phi - \psi$ angles are sampled for each of these residues, keeping top 5 acceptable rotamers. Acceptable rotamers are those side-chain conformations that are free from steric collisions with other atoms, besides improving the score. Top 5 rotamers are selected based on steric compatibility and maximum score improvements over the base configuration. Iterative self-consistent mean-field optimization is then started, with backbone probabilities of the rotamers used as initial probabilities, and score improvement used as the cost. The goal is converge to high probabilities for rotamers which together maximize the cost. This optimization is very similar to the mean-field technique described in section 3.2, except for three differences. One, is that the score improvement here replaces force. Two, the initial probabilities are backbone dependent rather than equal probabilities, since keeping side chain conformations rotameric and appropriate for the backbone conformation is desirable. Finally, no inter-residue rotamer dependencies are modeled yet.

The minimization yields an optimal set of rotamers with respect to rotamer probabilities and score improvement, one per residue, which together maximize the overall score. These chosen rotamers are transformed into the binding site and the entire docked complex is again checked and overlap resolution is performed by SLIDE to remove any new van der Waals overlaps. Since induced-fit motions are small in nature, it is expected that very few key interactions would be lost during the induced-fit collision resolution. In fact, none were observed in our earlier studies of how overlap resolution influences scoring. If the collisions can be resolved and the final score is better than the pre-rotamer-search score, the new conformation is recorded. Otherwise, the pre-rotamer search configuration is restored and SLIDE moves to generate the next docking. In future, this can be modified to explore the next-best set of rotamers that are compatible while enhancing the score.

1. Generate an overlap-free docking from ligand-free protein structure and ligand. Record these protein, ligand docking conformations as *base conformations* and their score as *base score*.

2. Use scoring function to determine protein side chains with interfacial, unsatisfied polar atoms.

3. **if** *no unsatisfied residues exist* **then**
   Exit rotamer sampling. Write the docking files. Continue to generate and score next binding mode.
else
   Loop over the unsatisfied residues.
   if *currently considered residue is N- or C-terminal* **then**
      Skip this residue, since ϕ or ψ cannot be calculated. Continue sampling rotamers for next residue.
   else
      a. Determine ϕ, ψ for the residue. Determine the 10°- spaced $\phi - \psi$ bin in the rotamer library in which these values of ϕ, ψ fall.
      b. Loop over the rotamers in this $\phi - \psi$ bin.
      c. Restore protein, ligand to their *base conformations* before sampling a rotamer.
      d. Transform the rotamer into the binding site in place of original side-chain conformation.
      if *steric overlaps are found* **then**
         Try to resolve steric overlaps through SLIDE's induced-fit collision resolution.
         if *overlaps cannot be resolved* **then**
            Discard this rotamer, continue by sampling the next rotamer.
         else
            Evaluate score of the entire docking.
            if *score is worse compared to base score* **then**
               Discard this rotamer, continue by sampling the next rotamer.
            else
               record current conformation of side chain, keeping at most top 5 score-enhancing conformations for this residue.
            end
         end
      else
         Evaluate score of the entire docking.
         if *score is worse compared to base score* **then**
            Discard this rotamer, continue by sampling the next rotamer.
         else
            Record current conformation of side chain, keeping at most top 5 score-enhancing conformations for this residue.
         end
      end
   end
   if *more than one unsatisfied residue* **then**
      a. By performing mean-field optimization on recorded candidate rotameric conformations for the unsatisfied residues, identify the rotamers, one per residue, that together maximize the score.
      b. Transform these rotamers onto their residues in the interface, replacing their side chains.
      c. **if** *steric overlaps are found* **then**
         i. Try to resolve steric overlaps through SLIDE's induced-fit collision resolution.
         ii. If overlaps cannot be resolved, restore the protein, ligand *base conformations*, write the docking files and continue to generate next binding mode.
      else
         i. Evaluate score of the entire docking.
         ii. If score is worse compared *base score*, restore the protein, ligand *base conformations*.
      end
      d. Write docking files and continue to generate next binding mode.
   else
      Identify the best-scoring rotamer conformation the single unsatisfied residue.
   end
   Substitute the most score-enhancing set of rotamer(s). If steric overlaps are found, attempt to resolve them; if not possible, revert to pre-rotamer search conformation.
end

Algorithm 1: Pseudo-code for rotamer sampling for unsatisfied side chains and mean-field optimization for maximizing score. Flowchart of implementation is presented in Figure 5.1.

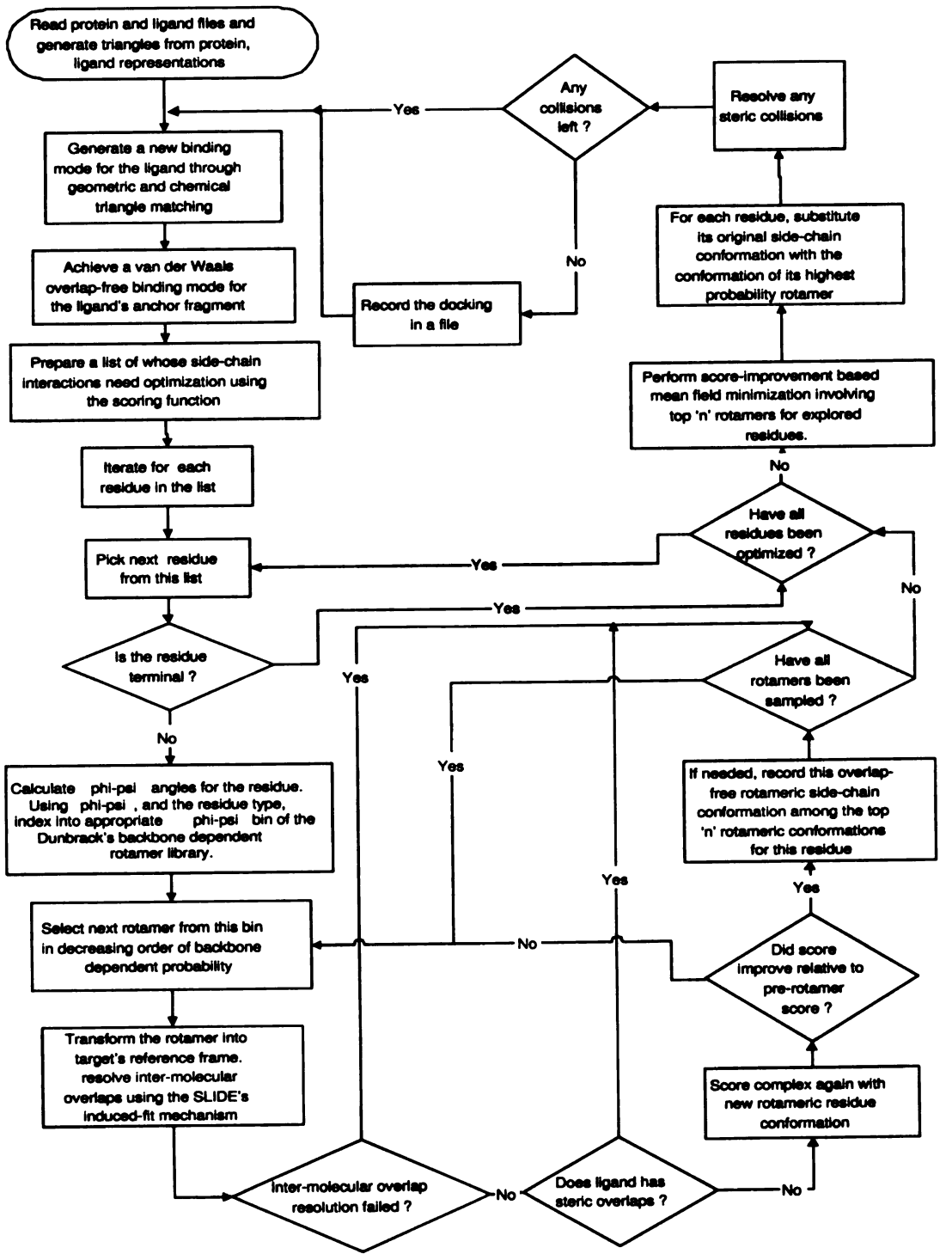


Figure 5.1: Flowchart of algorithm used to guide selection of a rotamer from a pool of rotamers. Pseudo-code is presented in algorithm 1.

5.2 Methods

Experiments involved 30 pairs of ligand-free and ligand-bound crystal structures specified in Table 2.1. To be able to compare a docked ligand's RMSD from its crystal complex conformation and orientations, the ligand-free protein backbone structure. The 30 proteins were chosen to have a backbone RMSD less than 1.0 Å, so that side-chain motion would not be required to compensate for incorrect backbone positions.

After superposition, the binding site in the ligand-free structure was initially identified as the set of all residues having any atom within 9.0 Å of the ligand. Corresponding to this set of ligand-free residues, ligand-bound residues were identified as the binding site of the ligand-bound protein structure.

Each ligand was then separated from the ligand-bound complex. Hydrogen atoms were added to the ligand structures by the *AM1BCC* [18] program to determine partial charges and add protons as needed to polar atoms of the ligand. A biased template representation of the active site generated by SLIDE to represent strategic points of potential interactions between protein and this ligand was generated using the ligand conformation from ligand-bound structure [51]. This ensures that most of the dockings would place ligand in its correct position, so that side-chain conformations of the docked complexes could then be compared with side-chain conformations in the ligand-bound crystal complex.

Scoring function #61 was implemented in both current version of SLIDE (referred to as old in figures) as well as the version with rotamer sampling (referred to as new). Default SLIDE parameters were used for docking most of the known ligands into respective apo protein binding sites[36]. However, the parameter SIDE_CHAIN_OVERLAP was relaxed from 0.3 to 0.5 Å, and parameter INTRA_OVERLAP was relaxed from 0.1 to 0.2 Å to allow ligand dockings for structures 1poa, 1apm, 3cox, 1lib. Water molecules were not included at any stage of the docking or analysis.

Of the 30 structures experimented, no successful dockings were generated by either SLIDE version due to excessive steric overlaps for structures 1lib, 3enl, 1udg, 1ca2, 1xib, and 2ctv. The predominant reason was that the ligand anchor fragment had unresolvable van der Waals overlaps with the protein main chain. Since dockings were not generated by either SLIDE version, these cases were not pursued further, and will require other refinements in ligand or main-chain flexibility modeling in SLIDE.

To observe the effects of rotamer sampling, dockings generated by different SLIDE versions need to be compared. Only those dockings that matched the same anchor fragment of the ligand to the same template-point triangle and resulted in an accurate docking were compared. Comparing such dockings, henceforth called *common dockings*, ensures fair evaluation of side-chain modeling performance across SLIDE versions. Since new SLIDE version, unlike the old version, optimizes side-chain conformations for chemical complementarity, the side-chain conformations in a common docking generated by new SLIDE may be different than those in the corresponding docking by old SLIDE, even though the ligand orientation in both the dockings are the same.

5.3 Results

This section presents results from the score-optimizing mean-field method implemented in SLIDE for sampling and substituting rotamers in the protein-ligand interface. To analyze the performance of the new SLIDE version, there were some specific measures of success:

- **Chemical Improvements**

    1. *Does the score improve for dockings generated by new SLIDE as compared to the same dockings by old SLIDE ?*

       This provides insight into the effectiveness of rotamer sampling and subse-

quent mean-field optimization to choose a chemically optimal rotamer set for residues with unsatisfied hydrogen-bond potential. In Figure 5.2, we present the distribution of improvement in score for all common dockings, generated by both old and new SLIDE versions for the 24 protein-ligand complexes. New SLIDE samples rotamers for unsatisfied side chains and optimizes rotamer choices for maximizing score improvement. The score improvement for each common docking is defined as (score of docking by new SLIDE - score of identical docking by old SLIDE). As shown, for more than 90% of almost 2500 common dockings, the score improves. The inset graph, provided to show the same distribution while excluding the most frequent score-improvement bin of 0 - 0.5, shows that bigger score improvements (greater than 0.5) are also significant in number. Note that the score improvement is an estimation of improvement in $\Delta G_{binding}$, since the new scoring function # 61 was developed to predict protein-ligand binding affinities. Through Figure 5.2, we summarize that by combining rotamer-sampling for unsatisfied polar groups and subsequent score-based mean-field maximization, we are able to enhance the score for most of the dockings generated by the new version of SLIDE.

2. *Does the score improve for the best docking by new SLIDE as compared to best docking by old SLIDE ?*

   While rotamer sampling and score maximization was effective for most of the dockings, the typical usage of SLIDE aims at identifying best dockings from a multitude of binding modes generated for each ligand. When docking known ligands, as in this research work, the best docking, involving the ligand-free conformation of the protein's active site docked with the known ligand, implies the docking that has the ligand position and conformation closest to that in the ligand-bound crystal structure. Focusing on the best

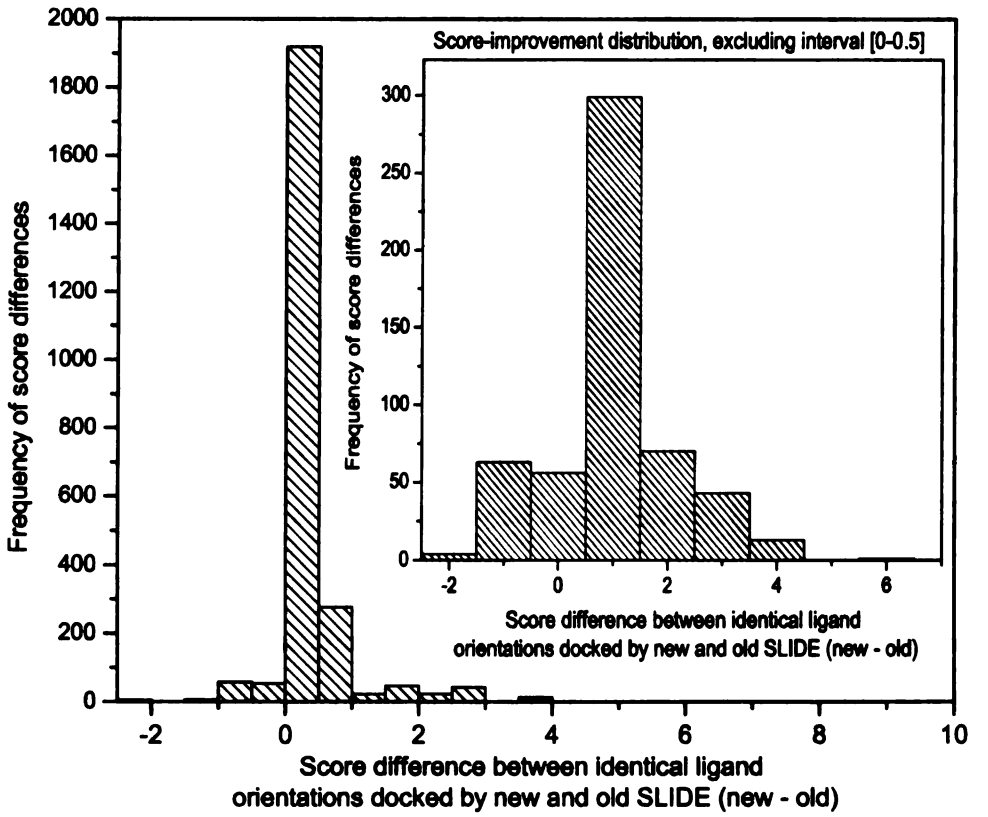


Figure 5.2: Distribution of change in score across all dockings with triangle matches in common, generated by both old and new SLIDE versions. Both old and new SLIDE use scoring function #61. The common triangle matches ensure ligand is in the same orientation in dockings by both old and new SLIDE versions. However, side-chain positions may differ, since the new SLIDE samples rotamers for unsatisfied side chains and optimizes rotamer choices for score improvement. Score improvement for each common docking is defined as (score of docking by new SLIDE - score of docking with same orientation by old SLIDE). Inset shows the same distribution with expanded y-axis scale, omitting the largest score-improvement bin.

docking is needed since the effectiveness of side-chain optimization can be evaluated fairly only with dockings that have the ligand in the correct position as well as conformation. In Figure 5.3, we present scores of the best dockings achieved by old and new SLIDE versions, comparing against the score evaluated for the crystal complex itself. In 50% of the best dockings generated by new SLIDE (12 out of 24 cases), the scores improved as compared to the best dockings generated by old SLIDE, and in no case was it significantly worse. In about one-third of the cases, the new SLIDE dockings scored better than the crystal complex, which indicates that the scoring function still requires improvement.

3. *Does the number of unsatisfied polar atoms decrease in best dockings generated by new SLIDE as compared to old SLIDE ?*

   Since the rotamer sampling was restricted to only those residues that had unsatisfied polar groups buried in the interface, we compared the number of such groups remaining in the best dockings by both old and new SLIDE in Figure 5.4. Restricting analysis to the best dockings, for the reasons cited above, we observe that as compared to best dockings generated by old SLIDE, the number of unsatisfied polar groups buried in the interface in new SLIDE decreased in close to 50% (11 out of 24 cases) of the cases. The decrease in number of unsatisfied polar groups varied from 3 to 1. Another encouraging fact is that the criterion for detecting candidate residues for rotamer sampling could easily be expanded to include repulsive contacts and exposed hydrophobic groups, as well as unsatisfied polar groups.

- **Conformational Improvements**

  While the score-guided rotamer sampling aims to maximize the side-chain interactions in the binding interface, another criterion to measure the effectiveness of rotamer sampling is how closely the sampled side-chain con-

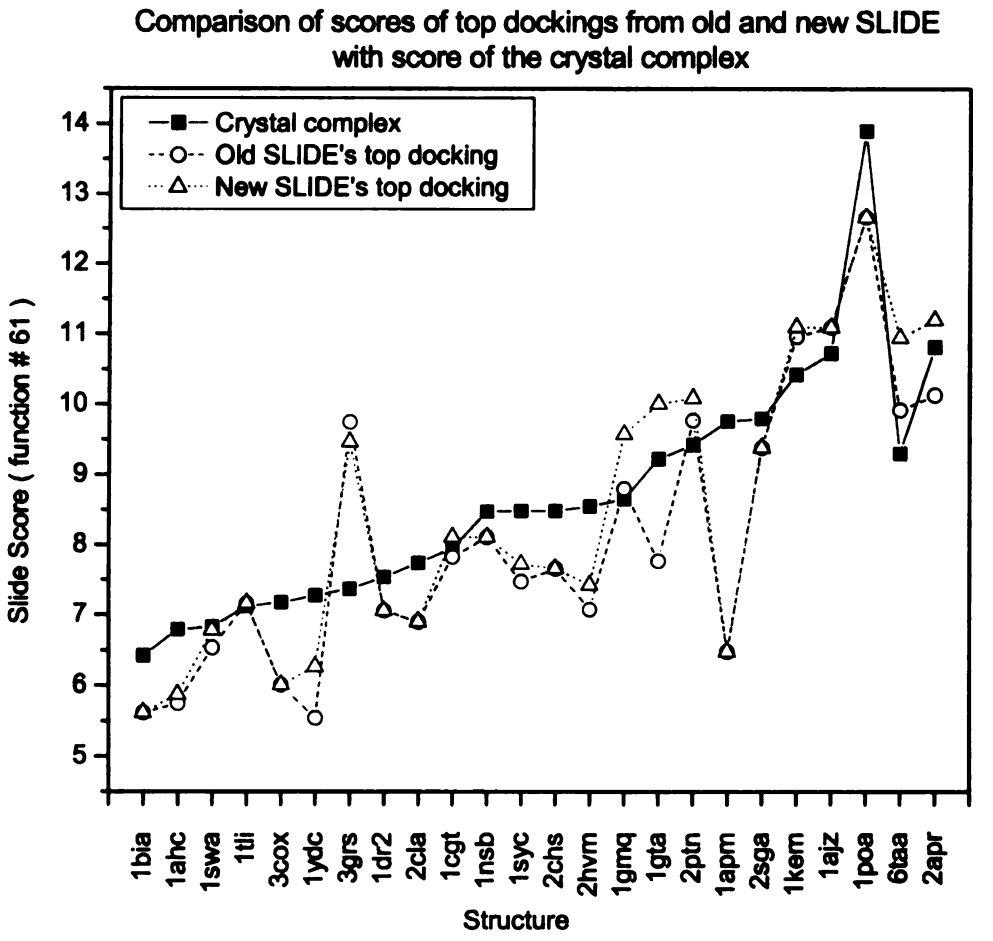


Figure 5.3: Comparison of scores of best dockings by old and new SLIDE versions for each protein, as well as scores evaluated for the corresponding crystal complexes. The best docking is identified using the lowest RMSD docked ligand conformation from the ligand's crystal-complex conformation. Scoring function #61 was used to score the dockings after they were recorded by old and new versions of SLIDE.

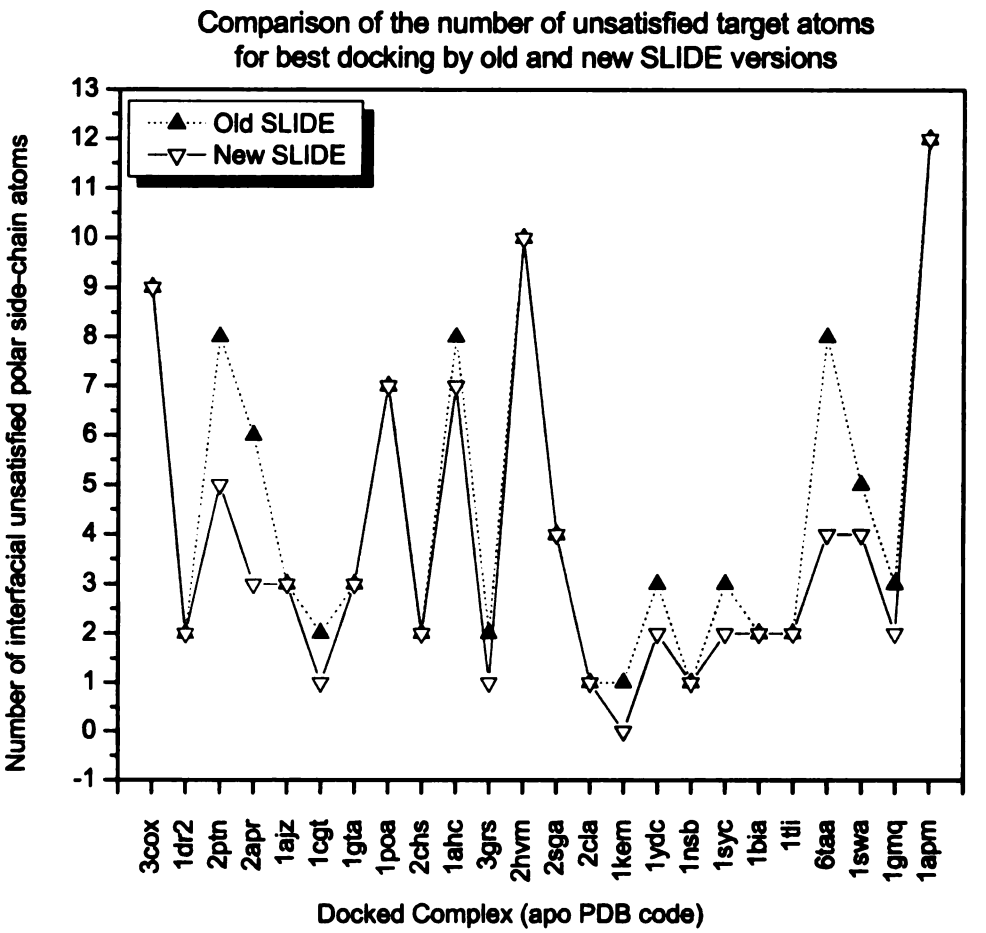


Figure 5.4: Comparison of the number of interfacial unsatisfied polar groups in the best dockings generated by old and new SLIDE versions. New SLIDE sampled rotamers for only those residues which had at least one interfacial polar atom unsatisfied.

formation approaches the to ligand-bound conformation upon docking. Is new SLIDE able to model motions, both in magnitude and direction, that are similar to the motions from ligand-free to ligand-bound conformations found in crystal complexes ? This study was organized to answer the following key questions:

1. *Does new SLIDE correctly detect the set of side chains that undergo large motions from ligand-free to ligand-bound conformations ?*

A large motion is defined here as side-chain's positional displacement, measured as RMSD between ligand-free and ligand-bound conformations, of at least 0.5 Å . While chapter 3 focused on large rotations, here focus is on large displacements since evaluation of docking performance is measured using positional deviation measures like RMSD.

In Figures 5.5, 5.6, and 5.7, all side chains are presented that were either moved by SLIDE or underwent a large motion from ligand-free to ligand-bound conformation, across best dockings. This set of side chains was selected to compare any side-chain motion by any SLIDE version relative to any large interfacial side-chain motions observed from the ligand-free and ligand-bound crystal structures. There were 93 individual side-chain cases. Figure 5.5 presents side chains moved to resolve collisions. As expected, new SLIDE moves almost the same set of side chains (52 in number) as old SLIDE does (55). Since collision resolution and induced-fit development are handled the same as before, therefore old and new SLIDE versions perform almost identically.The few differences that arise between motions by old and new SLIDE versions are the cases where a side chain that was rotated to resolve a steric overlap was also detected having unsatisfied hydrogen-bond potential by the new SLIDE, as in the

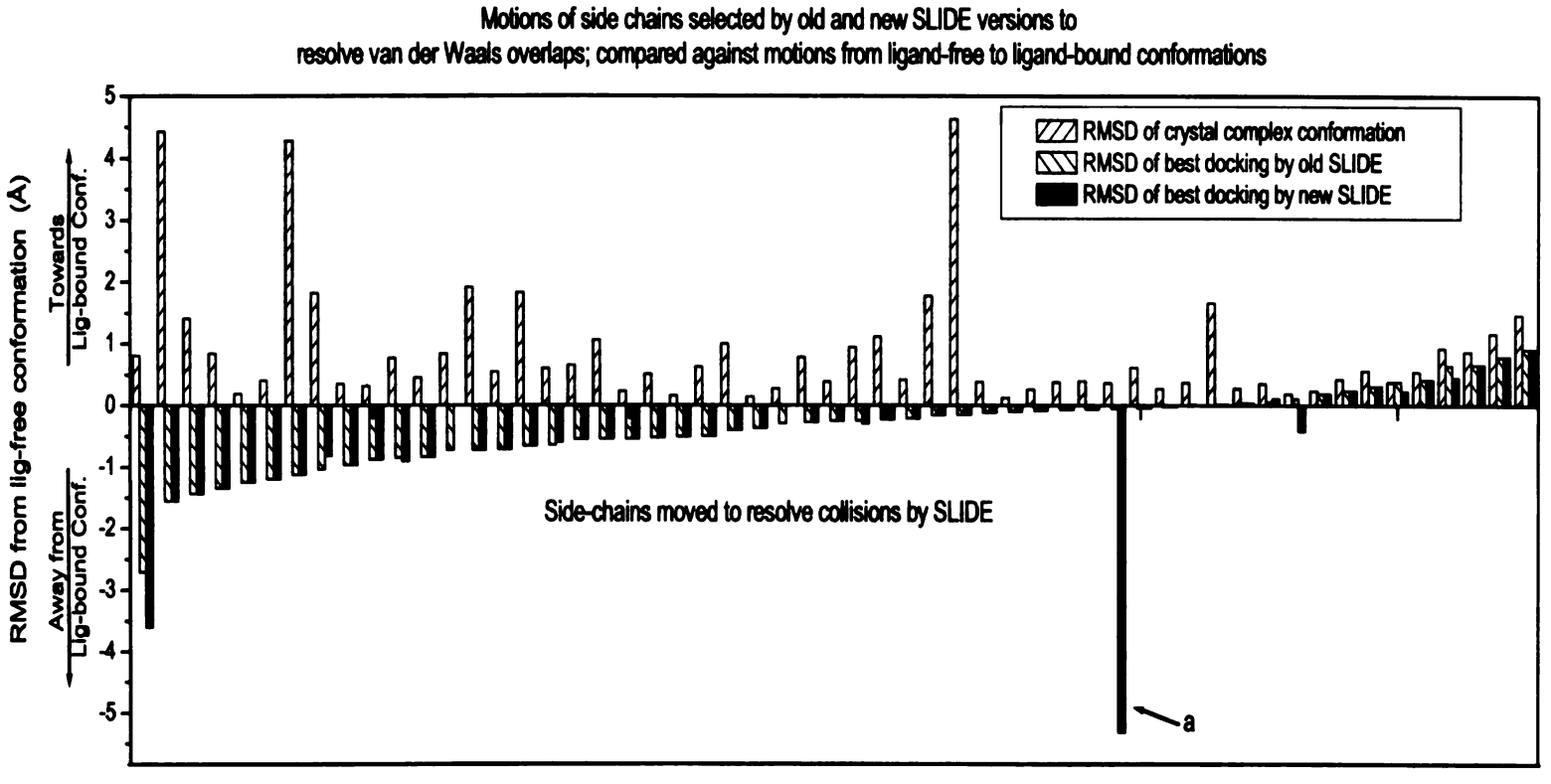


Figure 5.5: Comparison of side-chain motions performed during van der Waals collision resolution and induced-fit development in old and new SLIDE with nature of motions from ligand-free to ligand-bound conformations. The comparison is restricted to best dockings, determined by ligand RMSD relative to the position in the crystal complex, generated by old and new SLIDE. A positive sign of displacement RMSD on Y-axis indicates SLIDE moved the ligand-free side chain closer to the ligand-bound conformation, while a negative sign indicates that SLIDE moved the ligand-free side-chain conformation further from the ligand-bound conformation. The side-chain motion labelled 'a' represents the residue 1gta-Tyr104, which was moved by both old and new SLIDE to resolve a collision. Subsequently, the side chain's hydroxyl oxygen was determined as unsatisfied by the new SLIDE, which then sampled rotamers for the side chain to eventually select a conformation that helped the hydroxyl oxygen make a hydrogen bond. In the crystal complex, the side-chain conformation is slightly adjusted to make an aromatic interaction with the ligand.

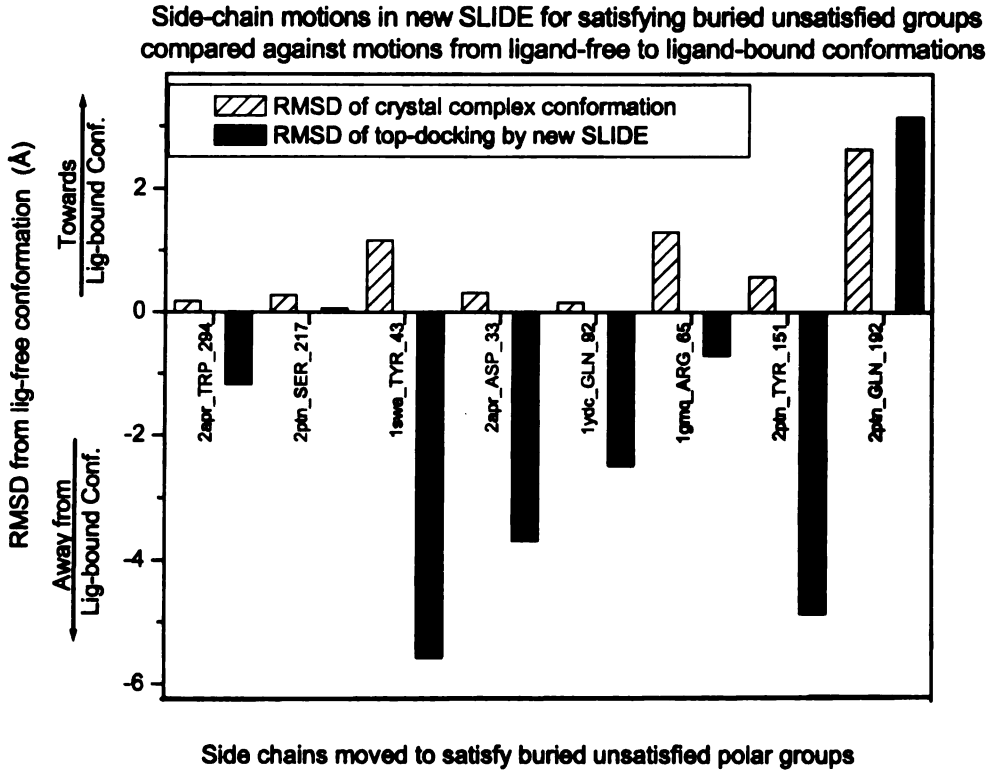


Figure 5.6: Side-chain motions performed in new SLIDE to optimize interactions in the same 24 complexes. For side chains with unsatisfied buried polar groups, rotamers were sampled and steric overlap-free, score-maximizing conformations were selected. As explained earlier, positive displacement indicates that the motion moved the side-chain conformation closer to the ligand-bound conformation from the ligand-free conformation; negative displacement indicates the opposite.

case of 1gta-Tyr104, which was subsequently optimized by new SLIDE for better interactions.

The new SLIDE also moves side chains through rotamer sampling for unsatisfied interfacial groups after the induced-fit docking. In eight cases, where side chains were moved only for optimizing interactions, are presented in Figure 5.6. While many more side chains were detected and sampled for better rotamers (data not shown for brevity), if those rotamers caused unresolvable collisions, they were reverted to the initial state and re-sampled.

Another set of 30 large-motion cases presented in Figure 5.7 remained undetected by both SLIDE versions; new SLIDE did not detect them because they did not involve unsatisfied polar groups. This presents an opportunity for investigating into other causes for large motions. Based on molecular graphics inspections of these complexes, possible reasons include the forming of π-cation or aromatic interactions, large main-chain induced motions or even high B-factor values contributing to imprecise side-chain atom positions.

New SLIDE identified similar side chains as old SLIDE (primarily to develop protein-ligand induced fit), and was also able to identify some of the large-motion cases which old SLIDE did not detect. Some additional, experimentally observed large motions may be detected if the candidate side chain selection criterion were expanded to consider groups besides unsatisfied polar side chains, including side chains having repulsive contacts or with solvent-exposed hydrophobic regions.

2. *Does rotamer sampling in new SLIDE model side-chain motions similar to motions from ligand-free to ligand-bound conformations ?*

   While identifying the correct set of side-chains for rotamer sampling is

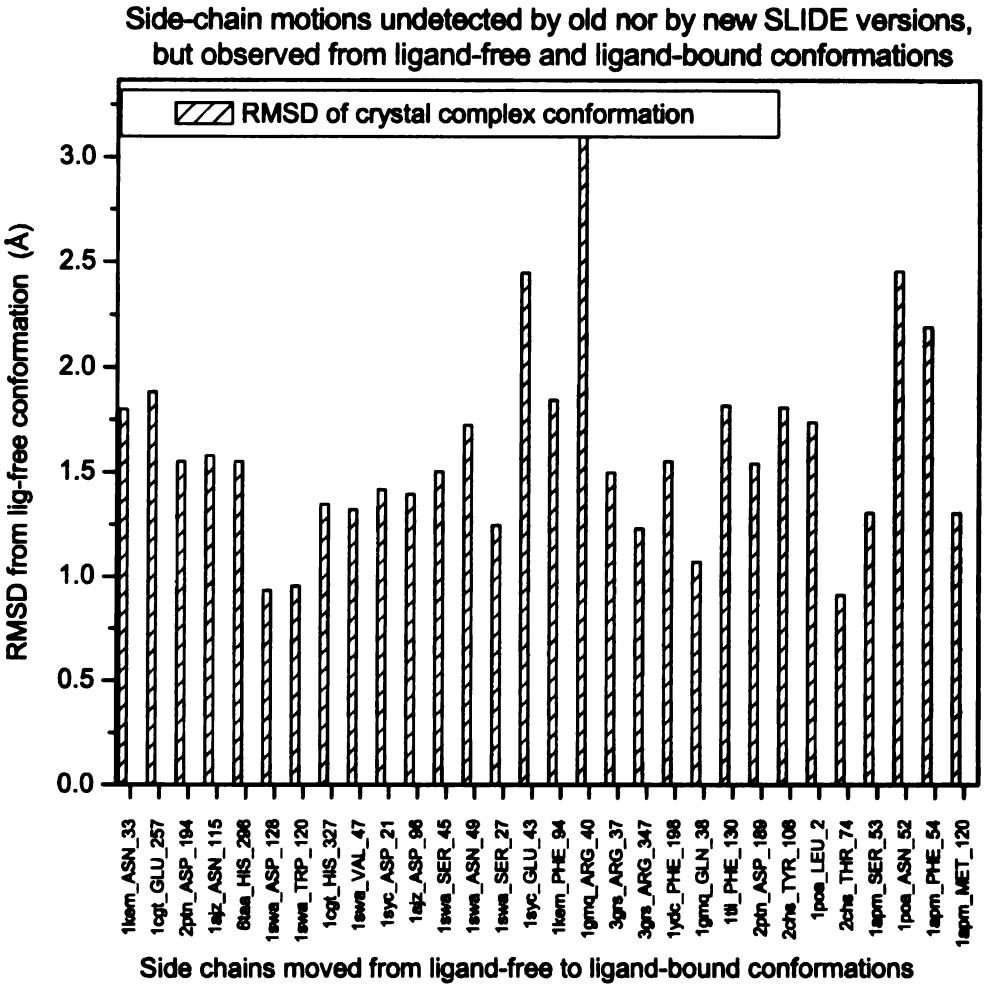


Figure 5.7: Large natural side-chain motions undetected by old or new SLIDE versions. These side-chains appear to have undergone large displacements for reasons other than inter-atomic collisions or to satisfy unsatisfied atoms. Some reasons observed in molecular graphics inspection of the complexes relative to ligand-free conformations, include achieving π-cation or aromatic interactions, large main-chain induced motions, even high B-factor values contributing to imprecise side-chain atom positions.

crucial, it is also important to compare the motions of these side chains in the crystal structures against the motions modeled by SLIDE. Two aspects of side-chain motions were analyzed - displacement and direction of the motion.

- **Displacement Magnitude**

  Does rotamer sampling achieve motions of similar magnitude as observed between ligand-free and ligand-bound crystal structures ?

  The RMSD between ligand-free and ligand-bound or docked, side-chain conformation is taken as the measure of side-chain displacement. For induced-fit side-chain motions presented in Figure 5.5, while displacement magnitudes are similar for old and new SLIDE versions, these magnitudes do not match those from crystal structures consistently; experimentally observed displacements vary more in magnitude. For score-enhancing rotamer sampling for unsatisfied side chains, again the side-chain displacements achieved by SLIDE do not agree with displacements from crystal structures. Investigations of the large displacements achieved through rotamer substitutions in Figure 5.6 show that most of the cases are tyrosine residues which were detected as unsatisfied after induced-fit docking. Rotamers helped satisfy the hydrogen-bond potential of the hydroxyl oxygen. While scores improved due to improved interactions, sampling also moved the tyrosine side chains further away from their true ligand-bound conformations.

- **Direction of Motion**

  Does rotamer sampling model side-chain motion closer to the ligand-bound conformation, as observed from comparing ligand-free and ligand-bound crystal structures ?

More importantly, through Figures 5.5,and 5.6, we learn that side chains often were moved further from their crystal structure positions. The motion was deduced to be more correct if the RMSD between docked and ligand-bound side-chain conformations was less than the RMSD between ligand-free and ligand-bound side-chain conformations; otherwise the motion was considered incorrect. For induced-fit docking, the smallest angle of the available angles is usually taken to resolve a collision, as explained in Figure 3.4. However, the choice of the smallest to minimize conformational effort while resolving the collisions often leads to angular choices that move side chains away from the desired position, as observed in Figure 5.5. In Figure 5.6 too, most of the side chains were positioned further from their ligand-bound positions, though the score improved by satisfying the unsatisfied polar groups.

5.4 Analysis

An interesting observation is that while new SLIDE successfully enhances scores across all dockings, and often for the best dockings too, many of its motions, whether driven by collision-resolution or score-maximization, move side chains through larger than necessary distances. There are, in effect, many more ways of moving side chains incorrectly than correctly, which leads to the observed results in Figures 5.5 and 5.6. The main reason is that the unsatisfied side chains were moved too much.

Reasons why other rotamers were rejected were investigated for each of the sampled rotamer for any unsatisfied side chains in the 24 structures. Each sampled rotamer is denoted as a filled circle in Figure 5.8, with its color conveying whether the rotamer was selected or rejected due to steric overlaps or decrease in score. As

various rotamers were substituted and collision resolution was attempted, rotamers were rejected either due to decrease in score or unresolvable collisions. The base conformation for each rotamer sampling represents the collision free state of the docked complex before the start of rotamer sampling.

Valuable insights can be drawn which help explain score improvement despite the increase in RMSD relative to the ligand-bound conformation:

- **A** is an example, in Figure 5.8, of side-chain rotamer samplings where none of the rotamers, be they closer or further relative to ligand-bound conformation, could achieve an overlap free conformation. Nature may resolve such cases with through more complex motions.

- **B** is an example of a side chain where the base conformation 'b' was far from the ligand-bound conformation, and a rotamer was found that was closer to the ligand-bound conformation. However, the score improvement for the most correctly positioned rotamer was equivalent to the score improvements for other overlap-free rotamers that were further from ligand-bound conformation. The final rotamer choice is based on the maximum probability after score-based mean-field optimization, which depends on both the rotamer probability and the score improvement. The higher-probability rotamer happened to be less native-like.

- **C** is an example of side-chain cases where a score-improving rotamer close to the ligand-bound conformation is sampled, but rejected due to unresolvable steric overlaps. This can happen in SLIDE's minimal rotation paradigm, since a small rotation for any of the single bonds may cause difficult-to-resolve new collisions in a tightly packed interface, while nature may resolve such collisions by a slight backbone motion.

- **D** is an example of side chain case where scoring function evaluated the same

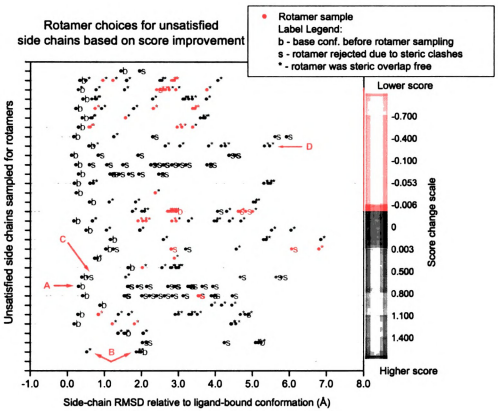


Figure 5.8: *This image is presented in color.* Selection or rejection reasons for each of the sampled rotamers for the unsatisfied side chains in the best dockings of the 24 structures used in experiments. X-axis represents the RMSD between the docked side-chain conformation and the ligand-bound conformation. Y-axis entries are the cases of unsatisfied side chains; their names have been omitted for clarity. For each side chain, the sampled rotamers are represented as filled circles. Symbol 'b' represents side-chain conformation after the induced-fit docking and before the rotamer sampling, 's' indicates the rotamer was rejected due to unresolvable steric overlaps, while '*' means that steric overlaps, if any, were resolved. The color of the filled circle represents the change in score - from brown towards blue indicates the rotameric conformation enhanced the score compared to starting or the base conformation 'b', while brown towards red depicts a decrease in score.

score for each overlap-free rotamer. In such cases, the likelihood of a rotamer being chosen depends on the backbone probability used in the score-based mean-field maximization. A rotamer further from the ligand-bound conformation may be chosen due to having a high probability.

5.5 Conclusion

One of the most important observation involves "B" cases. Representing the side-chain conformation after preliminary induced-fit docking, these conformations are already very close, within 1.0 Å, of the ligand-bound conformations in most of the cases. This is observed to be true for almost all of the unsatisfied polar side chains. This implies that while rotamers help sample a larger 3D space, instead a local 3D sampling of rotations should first be explored to satisfy unsatisfied polar groups, as this approach is more likely to stay close to the true ligand-bound conformation. This also conveys that nature too may resist big rotameric shifts in conformation, since that involves overcoming high strain energy barriers due for the rotamer transitions, and substantial repacking of the interface. Hence though rotamer sampling helped improve scores and hydrogen-bond interactions, across a substantial number of dockings, it was observed that side-chain conformation prediction actually worsened due to ambiguity in choice of the correct rotamer given that several rotamers could result in the same score improvement. Furthermore, even the most positionally correct rotamer was often further from the ligand-bound position than was the initial, collision-resolved side chain. The metrics for choosing a rotamer were to maximize the interactions as measured by the new scoring function, and to choose rotamers that are observed the most frequently in nature. There can be no a priori knowledge within rotamer sampling or optimization to recognize which side-chain conformations are close to ligand-bound conformations. Recognizing that the target side-chain con-

formations are typically within 1 Å RMSD of the initial position, finer and more local conformational sampling may be all that is required.

Chapter 6

Summary and Future Directions

Even though enhancing protein side-chain interactions remains the goal of this work, the ultimate goal of docking tools is to model motions to predict crystal complex configuration using apo protein conformation and unbound ligands. Towards these ends, our key contributions have been:

- Discovering that the intra-protein hydrogen bond network in the binding site largely remains conserved upon ligand binding. Most of the rearrangements in the hydrogen bond network increase the number of hydrogen bonds or serve as a ligand-recognition mechanism by replacing protein-water hydrogen bonds with protein-ligand hydrogen bonds.

- Observing, and corroborating, that SLIDE's paradigm of minimal rotations for induced-fit docking[50] performs well for preserving most of the hydrogen bond network. However, favoring minimal rotations misses 15% of the cases where larger motions are needed to improve docking chemistry, as found from comparing ligand-free and ligand-bound crystal structures.

- Analyzing ligand-free and ligand-bound crystal structures for the reasons behind large-angle rotations in side chains upon ligand binding. For half of the

cases, large rotations aided in satisfying buried polar groups which would have remained unpaired in the protein-ligand interface if they had remained in their ligand-free conformations. The current criteria for selecting side chains for conformational optimization - steric complementarity and satisfying unpaired hydrogen-bond donors and acceptors - could be easily expanded to include other cases of suboptimal chemical complementarity like side chains with repulsive contacts or side chains with solvent exposed hydrophobic groups.

- Implementing in SLIDE an infrastructure for modeling these larger motions through the use of rotamer libraries.

- Determining that rotamer-based sampling typically moves side chains too much, and it can be difficult to determine the correct rotamer when several rotamer choices improve the protein-ligand complementarity score equivalently. In general, the induced-fit docked conformation proved to be closer to the ligand-bound conformation than were any of the choices in the rotamer library. Thus, local rotational sampling (going beyond minimal rotations for removing van der Waals collisions) is likely to be a better approach for repositioning interfacial side chains to satisfy their hydrogen-bonding potential.

6.1 Future Directions

Future work can build upon the insights and algorithmic and geometric framework that has resulted from this work. Specifically:

- *Local, directed rotations.* Akin to directed rotations in induced-fit to resolve collisions, local rotations can be used to explore potential hydrogen bonds and salt bridges with nearby protein and ligand atoms.

- *Ligand flexibility.* While optimizing side-chain placement has been the focus in this work, the ligand's rotatable bonds present equally probable opportunities to enhance binding-site chemical complementarity. Exploring motions in the ligand in balance with those in the protein would also help model ligand flexibility in the docking process.

- *Unbiased protein template.* While we conducted experiments using an active-site template derived from ligands known to bind the protein in nature, SLIDE's typical use is to find new potential ligands. Hence unbiased templates (which include no knowledge of known ligands' interactions) should also be used to validate performance improvements from such local sampling.

- *Residue selection criteria.* As mentioned before, about half of the large side-chain motions in crystal structures were not detected by the current criteria to select residues for flexible modeling. These motions can be investigated further to learn the reasons for their motions and define criteria to identify such residues.

- *Handling water molecules.* Experiments in this work did not consider water molecules during the docking process. Including interfacial conserved waters should help in correctly identifying if side chains have unsatisfied polar groups or not. This should improve results by avoiding search for better geometries of polar side chains already satisfied by water, as well as recognizing water-exposed hydrophobic side chains in the protein-ligand interface.

- *Scoring function discrimination between alternative conformations.* The current scoring function was trained on crystal complexes with known binding affinities. These complexes are unlikely to offer many cases with repulsive contacts or unsatisfied polar atoms left due to rotations or water displacements, even though such scenarios are bound to occur frequently during docking. A scoring function, trained on data that includes favorable as well as unfavorable con-

tacts may better differentiate good conformations from bad ones in the same neighborhood.

With these thoughts on future opportunities, we would like to conclude this thesis. This work is still in progress, and we are excited to pursue the above directions and continue working towards improving state of the art in protein-ligand docking.

BIBLIOGRAPHY

[1] Adrian A. Canutescu, Andrew A. Shelenkov, and Roland L. Dunbrack Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, 12:2001–2014, 2003.

[2] Heather A. Carlson, Kevin M. Musukawa, Kathleen Rubins, Fredric D. Bushman, William L. Jorgensen, Robert D. Lins, James M. Briggs, and J. Andrew McCammon. Developing a Dynamic Pharamacophore Model for HIV-1 Integrase. *Journal of Medicinal Chemistry*, 43:2100–2114, 2000.

[3] H. Claussen, C. Buning, M. Rarey, and T. Lengauer. FlexE: Efficient molecular docking considering protein structure variations. *Journal of Molecular Biology*, 308(2):377–395, 2001.

[4] Bassil I. Dahiyat, D. Benjamin Gordon, and Stephen L. Mayo. Automated design of the surface positions of protein helices. *Protein Science*, 6:1333–1337, 1997.

[5] R. L. DesJarlais, R. P. Sheridan, G. L. Seibel, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *Journal of Medicinal Chemistry*, 31(4):722–729, 1988.

[6] Johan Desmet, Marc De Maeyer, and Ignace Lasters. The Dead-End Elimination Theorem: A New Approach to the Side-Chain Packing Problem. *The Protein Folding Problem and Tertiary Structure Prediction, Birkhauser*, pages 307–338, 1994.

[7] I. D.Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *Journal of Computational Chemistry*, 19:269–288, 1982.

[8] Matthew D. Eldridge, Christopher W. Murray, Timothy R. Auton, Gaia V. Paolini, and Roger P. Mee. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design*, 11:425–445, 1997.

[9] T. Ewing and I. D. Kuntz. Critical evaluation of search algortihms for automated molecular docking and database screening. *Journal of Computational Chemistry*, 18:1175–1189, 1997.

[10] D. Fischer, S. L. Lin, H. L. Wolfson, and R. Nussinov. A geometry based suite of molecular docking processes. *Journal of Molecular Biology*, 248:459–477, 1995.

[11] D. K. Gehlhaar, G. M. Verkhivker, P. A. Rejto, C. J. Sherman, D. B. Fogel, L. J. Fogel, and S. T. Freer. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming. *Chem. Biol.*, 2(5):317–324, 1995b.

[12] Holger Gohlke, Manfred Hendlich, and Gerhard Klebe. Knowledge-based Scoring Function to Predict Protein-Ligand Interactions. *Journal of Molecular Biology*, 295:337–356, 2000.

[13] P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry*, 13(7):849–857, 1985.

[14] Jaap Heringa and Patrick Argos. Strain in Protein Structures as Viewed Through Nonrotameric Side Chains: I. Their Position and Interaction. *Proteins: Structure, Function, and Genetics*, 37:30–43, 1999.

[15] Jaap Heringa and Patrick Argos. Strain in Protein Structures as Viewed Through Nonrotameric Side Chains: II. Upon Ligand Binding. *Proteins: Structure, Function, and Genetics*, 37:44–55, 1999.

[16] Brandon M. Hespenheide, A. J. Rader, M. F. Thorpe, and Leslie A. Kuhn. Identifying protein folding cores from the evolution of flexible regions during unfolding. *Journal of Molecular Graphics and Modelling*, 21:195–207, 2002.

[17] Thomas Huber, Andrew E. Torda, and Wilfred F. van Gunsteren. Optimization Methods for Conformational Sampling Using a Boltzmann-Weighted Mean Field Approach. *Biopolymers*, 39:103–114, 1996.

[18] OpenEye Scientific Software Inc. Quality Atomic Charges, Proton Assignment and Canonicalization . *http://www.eyesopen.com/docs/html/quacpac*, 2004.

[19] Richard M. Jackson, Henry A. Gabb, and Michael J. E. Sternberg. Rapid Refinement of Protein Interfaces Incorporating Solvation: Application to the Docking Problem. *Journal of Molecular Biology*, 276:265–285, 1998.

[20] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe. Protein flexibility predictions using graph theory. *Proteins: Structure, Function and Genetics*, 44:150–165, 2001.

[21] G. Jones, P. Willett, and R. C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology*, 245:43–53, 1995.

[22] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 3:727–748, 1997.

[23] D. E. Koshland Jr. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of National Academy of Science, USA*, 44:98–123, 1958.

[24] Roland L. Dunbrack Jr. Rotamer libraries in the 21^{st} century. *Current Opinion in Structural Biology*, 12:431–440, 2002.

[25] D. Keller, M. Shibata, E. Markus, R. Ornstein, and R. Rein. Finding the global minimum: A fuzzy end elimination implementation. *Protein Engineering*, 8:893–904, 1995.

[26] P. Koehl and M. Delarue. Mean-field minimization methods for biological macromolecules. *Current Opinion in Structural Biology*, 6:222–226, 1996.

[27] L. A. Kuhn, C. A. Swanson, M. E. Pique, J. A. Tainer, , and E. D. Getzoff. Atomic and Residue Hydrophilicity in the Context of Folded Protein Structures. *Proteins: Structure, Function, and Genetics.*, 23:536–547, 1995.

[28] I. Lasters and J. Desmet. The fuzzy-end elimination theorem: correctly implementing the side chain placement algorithm based on the dead-end elimination theorem. *Protein Engineering*, 6:717–722, 1993.

[29] Ming Lei, Maria I. Zavodszky, Leslie A. Kuhn, and Michael F. Thorpe. Sampling Protein Conformations and Pathways. *Journal of Computational Chemistry*, 25(9):1133–48, 2004.

[30] D. M. Lorber and B. K. Shoichet. Flexible ligand docking using conformational ensembles. *Protein Science*, 7:938–950, 1998.

[31] I. K. McDonald and J. M. Thornton. Satisfying hydrogen bonding protein in proteins. *J. Mol. Biol.*, 238:777–793, 1994.

[32] E. C. Meng, B. K. Shoichet, and I. D. Kuntz. A geometric approach to macromolecule-ligand interactions. *Journal of Computational Chemistry*, 13(6):505–524, 1992.

[33] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19:1639–1662, 1998.

[34] Rafael Najmanovich, Joseph Kuttner, Vladimir Sobolev, and Marvin Edelman. Side-Chain Flexibility in Proteins Upon Ligand Binding. *Proteins: Structure, Function, and Genetics*, 39:261–268, 2000.

[35] R. Norel, D. Fischer, H. Wolfson H, and R. Nussinov. Molecular surface recognition by a computer vision based technique. *Protein Engineering*, 7:39–46, 1994.

[36] Protein Structural Analysis and Design Lab., MSU. Quick Guide to SLIDE Version 2.30a. *http://www.bch.msu.edu/~kuhn/projects/slide/user_manual.html*, 2005.

[37] Olivier Roche, Ryuichi Kiyama, and III Charles L. Brooks. Ligand-Protein DataBase: Linking Protein-Ligand Complex Structures to Binding Data. *Journal of Medicinal Chemistry*, 44:3592–3598, 2001.

[38] Adrian Roitberg and Ron Elber. Modeling side chains in peptides and proteins: Application of the locally enhanced sampling and the simulated annealing methods to find the minimum energy conformations. *Journal of Chemical Physics*, 95:9277–9287, 1991.

[39] Paul C. Sanschagrin. Computational Techniques for Modeling Protein-Ligand Interactions and their Application to Serine Proteases and Asparaginyl-tRNA Synthetase. *Ph. D. dissertation*, Michigan State University, 2001.

[40] Lana Schaffer and Gennady M. Verkhivker. Predicting Structural Effects on HIV-1 Protease Mutant Complexes With Flexible Ligand Docking and Protein Side-Chain Optimization. *Proteins: Structure, Function, and Genetics*, 33:295–310, 1998.

[41] Volker Schnecke and Leslie A. Kuhn. Flexibly Screening for Molecules Interacting with Proteins. *Rigidity in Theory and Applications*, Plenum Publishing, New York, 385-400, 1999.

[42] Volker Schnecke and Leslie A. Kuhn. Virtual screening with solvation and ligand-induced complementarity. *Perspectives in Drug Discovery and Design*, 20:171–190, 2000.

[43] Volker Schnecke, Craig A. Swanson, Elizabeth D. Getzoff, John A. Tainer, and Leslie A. Kuhn. Screening a Peptidyl Database for Potential Ligands to Proteins With Side-Chain Flexibility. *Proteins: Structure, Function, and Genetics*, 33:74–87, 1998.

[44] B. K. Shoichet, D. L. Bodian, and I. D. Kuntz. Molecular docking using shape descriptors. *Journal of Computational Chemistry*, 13:380–397, 1992.

[45] D. Stickle, L. Presta, K. Dill, and G. Rose. Hydrogen bonding in globular proteins. *J. Mol. Biol.*, 226:1143–1159, 1992.

[46] P. Tuffrey, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid determination of protein side-chain conformations. *Journal of Biomolecular Structural Dynamics*, 8(6):1267–89, 1991.

[47] G. Vriend. WHAT IF: A molecular modeling and drug design program. *Journal of Mololecular Graphics*, 8:52–56, 1990.

[48] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry*, 47:2977–2980, 2004.

[49] H. J. Wolfson and Y. Lamdan. Geometric hashing: A general and efficient model-based recognition scheme. *IEEE Internation Conference on Computer Vision, Tampa, Fl.*, pages 238–249, 1988.

[50] Maria I. Zavodszky and Leslie A. Kuhn. Side-Chain Flexibility in Protein-Ligand Binding: The Minimal Rotation Hypothesis. *Protein Science*, 14:1104 – 1114, 2005.

[51] Maria I. Zavodszky, Paul C. Sanschagrin, Rajesh S. Korde, and Leslie A. Kuhn. Distilling the Essential Features of a Protein Surface for Improving Protein-Ligand Docking, Scoring, and Virtual Screening. *Journal of Computer-Aided Molecular Design*, 16:883–902, 2002.