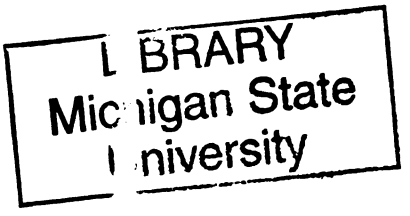


This is to certify that the
dissertation entitled

Modeling Flexibility in Protein-Ligand Recognition

presented by

Maria Ildiko Zavodszky

has been accepted towards fulfillment
of the requirements for the

Ph.D degree in Biochemistry and Molecular Biology

Major Professor's Signature

April 29, 2003

Date

MODELING FLEXIBILITY IN PROTEIN-LIGAND RECOGNITION

By

Mária Ildikó Závodszky

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Biochemistry and Molecular Biology

2003

ABSTRACT

MODELING FLEXIBILITY IN PROTEIN-LIGAND RECOGNITION

By

Mária Ildikó Závodszky

The function of many proteins is to recognize and bind peptides and other small molecules. Understanding the way these proteins work implies understanding the driving forces of protein-ligand interactions. This, in turn, is necessary to find new, specific ligands for proteins that are potential targets for disease therapy, or to help elucidate the function of the increasing number of proteins with known structure, but yet unknown function. Computational methods are well suited for analyzing the large amount of existing experimental data to identify the underlying principles of protein-ligand interactions. These principles, combined with efficient algorithms, are built into docking and screening schemes used to predict binding orientations and interactions of partner molecules providing a working hypothesis for further experiments. At the same time, these computational tools facilitate screening of large databases for reducing the large number of organic molecules to a smaller set of potential ligands to be tested for binding to the target of interest.

SLIDE, the protein structure-based ligand docking and screening tool developed in our laboratory, handles ligands and protein side chains flexibly and has the possibility of taking water-mediated interactions into account. It is capable of screening a database of 100 000 molecules in 1-2 days on a desktop computer. SLIDE was used to propose a

viable model of the ternary complex of R67 dihydrofolate reductase • folate • cofactor, taking into account existing experimental data. The results are in good agreement with the predictions of another widely used docking tool, DOCK, and propose specific interactions that can be tested by mutagenesis.

Improved representation of the binding site, using a knowledge-based approach, coupled with the realistic modeling of protein side-chain and ligand flexibility in SLIDE allowed the identification of new ligands for thrombin. This was achieved by screening the Available Chemical Directory, followed by the experimental measurement of the binding constants of the predicted top scoring ligand candidates using Isothermal Titration Calorimetry. Two of the top scoring ligand candidates were found to have binding affinities in the micromolar range for human thrombin.

As part of this thesis work, a new approach toward modeling main-chain flexibility in docking was proposed: Flexibility analysis of the target protein was performed using the graph-theoretic algorithm FIRST, followed by the generation of alternative conformations for the predicted flexible regions with ROCK, a conformer searching program based on a random walk sampling of the rotatable bonds. A representative set of the conformational ensemble generated this way was used as targets for docking with SLIDE. ROCK is uniquely suited for flexibly handling ring structures and can be used to model the flexibility of large circular ligands, as well, as demonstrated on the case of cyclosporin. The use of this combined method to perform fully flexible docking is illustrated for cyclophilin A – cyclosporin, while addressing the question of how much flexibility of the interacting molecules is tolerated without hindering molecular recognition.

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor, Dr. Leslie A. Kuhn, for her encouragement, support and guidance throughout my graduate carrier. She has been a great example of professionalism, enthusiasm for science and healthy optimism, a pleasure to work with. I have found not only an excellent advisor, but also a good friend in her.

I want to thank Dr. Michael Thorpe for the opportunity to work with him. He encouraged me to rise above the details, look at the big picture. The interesting conversations with him gave me valuable insight, from a physicist's point of view, of how science works.

I extend my thanks to the members of my committee, Dr. Shelagh Ferguson-Miller, Dr. Jack Preiss and Dr. Gregory Zeikus for providing encouragement and valuable feedback on my work.

I am thankful to collaborators Dr. Elisabeth Getzoff and Dr. John Tainer from the Scripps Research Institute, and Dr. Michal Zolkiewski from Kansas State University for giving me the opportunity to use their resources to screen the Available Chemical Directory and perform the ITC experiments, respectively.

Scientific work these days is usually a group effort. My thanks go out to the former members of the Kuhn lab, Dr. Volker Schnecke, Dr. Mike Raymer, Dr. Paul Sanschagrin, Dr. Brandon Hespenheide, Rajesh Korde and the former member of the Thorpe group, Dr. A.J. Rader for their invaluable help and advice. I would also like to

thank the present members of the Kuhn lab, Litian He, Chetan Sukuru, and Sameer Arora for creating an enjoyable working environment. I am especially thankful to Ming Lei from the Thorpe group for the excellent collaboration over the ROCK project. He was always ready to help, debate over ideas and approaches, and answer my suggestions promptly.

Finally, I would like to thank my friends and family, especially my husband Peter for making it all possible and worth while.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Images in this dissertation are presented in color.

LIST OF ABBREVIATIONS

ACD	Available Chemicals Directory
CSD	Cambridge Structural Database
CypA	cyclophilin A
DHF	dihydrofolate
DHFR	dihydrofolate reductase
DMSO	dimethyl sulfoxide
FIRST	Floppy Inclusions in Rigid Substructure Topography
GA	genetic algorithm
GST	glutathione S-transferase
HD exchange	hydrogen-deuterium exchange
ITC	isothermal titration calorimetry
MD	molecular dynamics
$NADP^+$	nicotinamide adenine dinucleotide phosphate, oxidized form
NADPH	nicotinamide adenine dinucleotide phosphate, reduced form
NMN	nicotinamide mononucleotide
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
PDB	Protein Data Bank
RMSD	root mean square deviation
ROCK	Rigidity Optimized Conformational Kinetics

SLIDE	Screening for Ligands with Induced-fit Docking Efficiently
THF	tetrahydrofolate

Chapter 1

Introduction

1.1 Modeling Protein-Ligand Interactions

Life can be viewed as interconnecting series of specific binding steps. The molecular organization of living matter implies the dependence of all biochemical processes on molecular recognition, from protein-ligand interactions through macromolecular associations to the intriguing process of protein folding. This thesis deals with modeling the molecular recognition between proteins and small molecule ligands that is of great practical significance for finding new specific ligands to proteins that are potential targets for disease therapy. Detected patterns in binding preferences can also help elucidate the function of the increasing number of proteins with known structures, yet unknown functions.

Scientists have been puzzled by the specificity and accuracy of molecular recognition since the beginning of modern biochemistry. Advances in technology facilitated rapid advancement in this area over the past few decades. Knowledge about the features contributing to protein-ligand interactions is derived from experimental data,

most importantly from the exponentially increasing number of protein structures solved by X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR) and deposited into the Protein Data Bank. (Berman et al., 2000). Calorimetric methods (Jelesarov and Bosshard, 1999), often combined with mutagenesis, provide useful information about the energetics of association of biomolecules. Reciprocal burial of hydrophobic surface patches as well as shape and chemical complementarity of the interacting partners along the interface are traditionally believed to be the critical contributors to the specificity of the binding process (Halperin et al., 2002; Kuntz et al., 1999; Sobolev et al., 1996). Nevertheless, proteins bind ligands even if their shapes in free form do not seem to be optimized for complementing each other. The differences observed between the ligand-free and ligand-bound structures of a number of proteins (Betts and Sternberg, 1999) point to the importance of accounting for the flexible changes occurring upon binding. According to the original concept of induced fit (Koshland, 1958), the match becomes perfect after these changes are induced in the protein by the ligand itself. The newly emerging view of the pre-selection of a complementary conformer, based on experimental and theoretical considerations (Bosshard, 2001; Ma et al., 2002), is an alternative to the classical induced fit, arguing that all proteins exist in ensembles of substates, presenting to the incoming ligand a range of binding shapes.

1.2 Computational Docking and Screening

A computational biochemist creating a docking program has to find solutions to the following problems: (1) accurate representation of the shape, chemistry and flexibility of the protein binding site and the ligand; (2) efficient search algorithm to explore the

orientational and conformational space of the ligand in the binding site, and (3) prediction of the correct docked ligand orientation. In addition, when the goal is to identify new ligands for a protein of interest, in which case the process is called screening, potential ligands have to be separated from other molecules based on their estimated binding affinity, providing a reasonably short list of ligand candidates for experimental testing. These problems are strongly interrelated: predicting the correct binding orientation depends on the proper representation of the binding site, and a list of plausible ligand candidates could not be provided without a reasonable estimate of the binding affinity, nor could the binding affinity be estimated without a reasonably correct prediction of binding orientation.

Although every screening program includes a docking step, screening is different from docking to some extent: the goal of screening is to identify ligands from a large database of typically diverse molecules. When a large database of small molecules is searched for potential ligands, the computational time spent on docking one molecule must be very short. Current docking algorithms spend a relatively long time (ranging from minutes up to hours) on docking one ligand with high accuracy. If only one minute were spent on docking each ligand when searching through a database of about 100 000 molecules, such as the Cambridge Structural Database (CSD) (Allen, 2002), the screening time would be more than two months. The most efficient way to overcome this problem is to rule out the unfeasible ligand candidates as early as possible from the screening process and spend the most time-consuming final docking step on promising candidates. The computational intensity and relative ranking of different ligands are problems that make screening a more challenging problem than docking.

1.3 Overview of Current Approaches in Docking

There are a number of excellent reviews in the literature summarizing the principles of docking (Abagyan and Totrov, 2001; Halperin et al., 2002; Lengauer and Rarey, 1996) and the different approaches used for screening large databases for lead generation (Abagyan and Totrov, 2001; Klebe, 2000). This section is intended to give only a brief overview of the main steps involved in the docking and screening processes.

1.3.1 Binding Site Representation

All-atom representation of both the binding site and the ligand is used throughout the simulation only when docking is done with molecular dynamics (MD) simulation. The protein is usually held fixed, while the orientational and conformational space of the ligand is sampled with a Monte Carlo (Carlson et al., 2000) or a genetic algorithm (GA) based method (Taylor and Burnett, 2000). The energy of the system is monitored, and a docked ligand orientation is considered to be a possible solution when a local energy minimum is found. These types of methods are the most accurate, but their high computational cost makes them a non-practical alternative for screening.

While maintaining the all-atom representation for the small ligands, fast docking programs use a reduced representation of the binding site in order to minimize the cost of the conformational search that follows. Discrete representations come in different flavors: geometric, with added chemical features, knowledge based or grid based. There is a whole spectrum of binding site representations between the geometric and the knowledge-based method, where a limited number of points are laid down to

trace the protein binding surface while the density of the points in certain areas is determined by knowledge about favored hydrogen-bond geometry, for example. A more detailed description of the currently used methods for binding site representation is provided in Chapter 3.

1.3.2 Orientational and Conformational Search

A rigid-body transformation superimposes the ligand over the binding site in all possible orientations that result in no deep interpenetrations between the molecules' van der Waals surfaces. When the molecules are represented by discrete points indicating the position of atoms capable of participating in potentially key interactions, it is usually sufficient to match three non-collinear points from the two molecules to perform the 3D alignment. A triplet of non-collinear points bears just enough chemical and spatial information about a molecule to indicate a possible nontrivial match to another molecule with a complementary set of three points. In the same time, it is easier to match triplets of points sets of four or more. The result is a large number of possible matches due to the combinatorial complexity of matching every possible atom triplet from the ligand to every possible triplet point describing the binding site. For example, assuming the binding site is described by only 100 points and the ligand by 10 points, there are 100x99x98 = 970 200 possible protein triangles to be matched to 10x9x8 = 720 ligand triplets, which would result in approximately $7x10^8$ computations. Geometric hashing algorithms are used to reduce the complexity of the problem by replacing the exhaustive search of matching every property of an object A to every property of all the objects from set B with a hierarchical search of matching one property at a time and eliminating

mismatches at each step. In the case of SLIDE, for example, all the possible template point triangles describing the protein binding site are organized in index tables based on individual simple properties like chemical labels attached to the points, length of shortest side, length of longest side, and perimeter of the triangle. As a first step, the chemical labels of one possible ligand interaction-point triplet are compared to the chemical labels of all template triangles. Only template triangles with matching chemistry are kept for the next step of matching the length of the shortest side, and during the third step, only triangles with similar shortest sides are kept for comparing the length of the longest side, and so on. The number of matches to be checked is reduced at each level of the index table, resulting in much faster execution times compared to exhaustive matching.

There are two main forms of docking: redocking and predictive docking, with redocking being far simpler. This is done by taking the ligand structure out of a crystallographic complex, and docking it back into its target structure, with both molecules initially possessing their favored conformation for binding to each other. Predictive docking, in which the free structure of the ligand is docked into the unliganded, apo structure of the target protein, is much more complex. In this case, the orientational search of the ligand has to be complemented with the exploration of the internal conformational space available for both the protein and the ligand to find the appropriate conformers that complement each other the best. Most current methods treat the ligand flexibly while keeping the protein rigid. The ligand is either incrementally built up in the binding site, or internal dihedral rotations are used to fit the ligand into the rigid binding site. Although better than completely rigid docking, the shortcoming of this approach is unrealistically placing all the burden of induced conformational change onto the ligand.

Another method to take into account induced fit upon binding is by allowing a certain amount of van der Waals overlap between protein side-chains and the ligand. This is often called soft docking.

Analysis of conformational changes on complex formation for a representative set of 39 pairs of ligand-free and ligand-bound structures (Betts and Sternberg, 1999) showed that about 50% of the proteins undergo substantial main-chain and side-chain conformational changes when binding ligands. This induced fit is often modeled by selecting alternative side chain conformations for the binding-site residues from a rotamer library or by performing directed rotations of rotatable bonds in the protein side chains and flexible ligand portions to resolve collisions after the ligand is transformed into the binding site. Inducing main-chain flexibility changes while performing the docking is too expensive computationally, so efforts are directed toward generating a conformational ensemble of the protein, and using this set as the target for the docking instead of one single structure. This approach is also following the line suggested by a number of theoreticians and experimentalists (Bosshard, 2001; Ma et al., 2002), who argue that the idea of the ligand selecting a complementary conformer from the preexisting native state ensemble of the protein is at least an alternative to the classical induced fit, where ligand binding triggers the conformational changes in the binding partners necessary to create a good steric complementarity. A more detailed analysis of handling protein side-chain and main chain flexibility in docking is presented in Chapters 4 and 6.

1.3.3 Scoring

Docking programs usually return a number of possible docked orientations for each ligand. A scoring function is employed to select the best docking among all. When known ligands are docked to their targets, the scoring function is expected to give the best score to the docking closest to the orientation of that ligand seen in the crystal or NMR structure of the protein-ligand complex. Also, when multiple ligands are docked to a single target, the scoring function should rank them according to their binding affinity. Theoretically, free energy calculations combined with MD simulations were shown provide reliable ranking for some systems, but they are too time-consuming, and as such, not a practical alternative (Miyamoto and Kollman, 1993; Pearlman and Charifson, 2001). Instead of calculating the binding affinity from first principles, docking programs use scoring functions to estimate the tightness of binding from structural parameters. Empirical scoring functions estimate the free energy of binding as a sum of several terms, each of them describing the contribution of one type of interaction to binding, such as van der Waals interactions, hydrogen bonds, ionic interactions, etc. (Bohm, 1994; Rognan et al., 1999; Schapira et al., 1999; Wang et al., 2002). Knowledge-based scoring functions, on the other hand, use statistical preferences derived from pair-wise interatomic distances and frequencies observed in crystal structures of protein-ligand complexes to determine the contribution of individual ligand atoms to the final score (Gohlke et al., 2000; Mitchell et al., 1999; Muegge and Martin, 1999). Scoring is one of the most challenging problems in the field, and there is no existing scoring function that performs consistently well across various systems (Halperin et al., 2002; Tame, 1999). The correct docking, or the one closest to the crystal structure position, is usually near the top of the list, but

buried among the large number of false positives (poor or approximate dockings given very favorable scores). Similarly, true inhibitors are often given smaller scores than inactive compounds when a mixed database of known ligands and decoy molecules is screened against a protein target. Consensus scoring has been suggested by several authors as a feasible way to ease this problem, resulting in improvements of up to 65-70% in hit rates (Bissantz et al., 2000; Charifson et al., 1999).

1.4 SLIDE

The docking and screening software **SLIDE** (**S**creening for **L**igands by **I**nduced-fit **D**ocking, **E**fficiently) developed in our laboratory (Schnecke et al., 1998; Schnecke and Kuhn, 1999, 2000) models flexible protein-ligand interactions based on steric complementarity combined with hydrogen bonding and hydrophobic interactions. SLIDE efficiently reduces the large number of ligand candidates to a manageable number by using geometry indexing and distance geometry filtering on discrete representations of the protein and ligand candidates. Approximately 100,000 small molecules can be screened and docked by SLIDE in one to two days on a typical desktop workstation.

A novel feature of SLIDE amongst geometric (rather than MD) methods is that it can take into account solvation. Consolv, a k-nearest-neighbor classifier developed in our laboratory (Raymer et al., 1997), is applied to predict conservation of binding site water molecules upon ligand binding. Waters predicted to be conserved are included as part of the binding site, although they can be displaced by a ligand atom at a later step if this results in greater molecular complementarity.

SLIDE was the first method to balance protein and ligand flexibility in docking. Due to this feature, it can identify and correctly dock 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 (Schnecke et al., 1998; Schnecke and Kuhn, 1999, 2000). Scoring of the docked protein-ligand complex by SLIDE is based on the number of hydrogen bonds and the hydrophobic complementarity between the ligand and its protein environment. The main steps involved in screening with SLIDE are shown in Figure 1.

The purpose of the work described in this thesis was not only to enhance the performance of SLIDE by improving the representation of the protein binding site and including protein main chain flexibility into the docking process, but to apply it to solving biologically relevant problems.

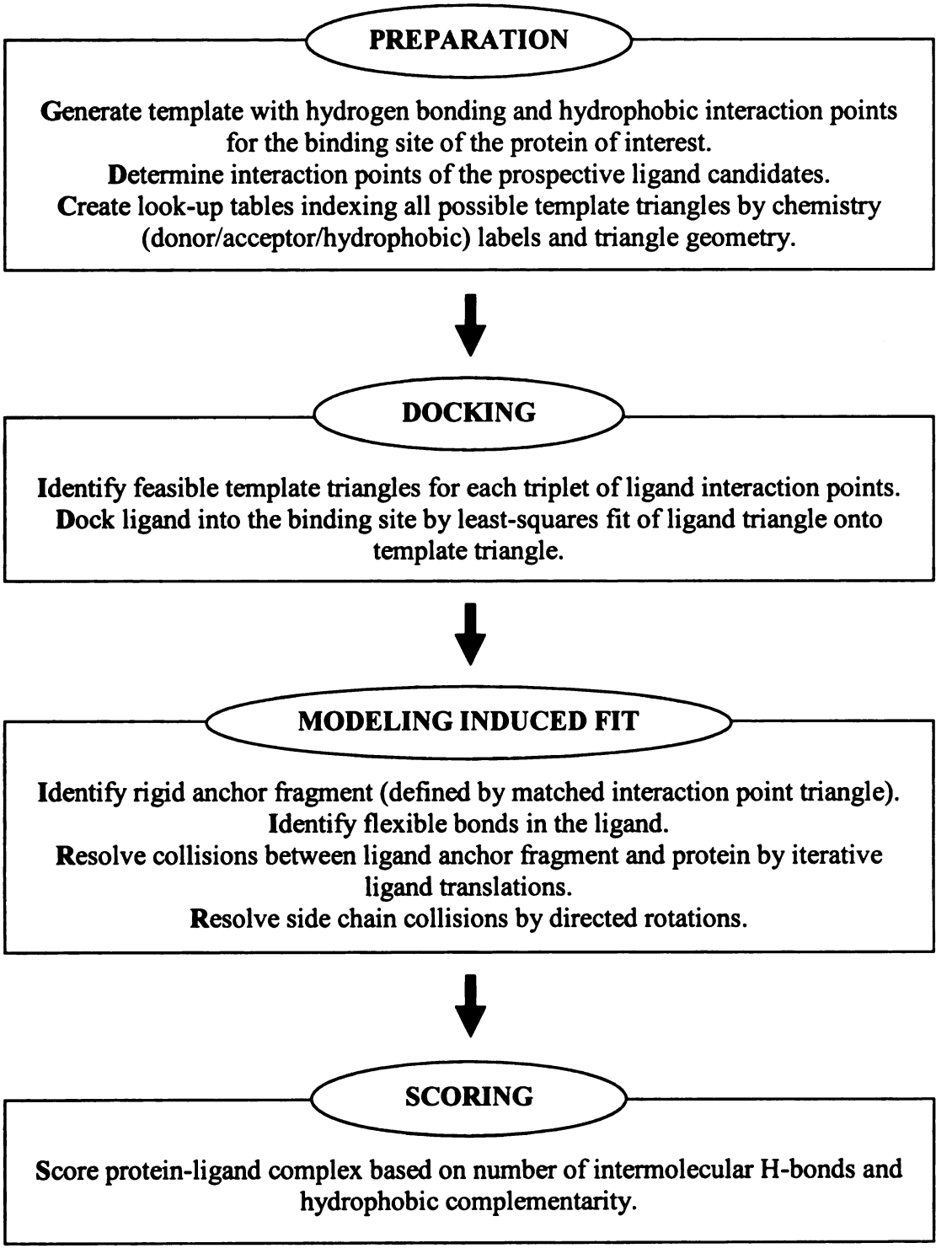


Figure 1.1. An overview of the SLIDE screening and docking algorithm.

Chapter 2

Predicting the Ternary Complex of R67 DHFR • NMN • Folate with SLIDE

The research presented in this chapter has been previously published in:

Howell, E.E., Shukla, U., Hicks, S.N., Smiley, R.D., Kuhn, L.A., Zavodszky, M.I. One site fits both: a model for the ternary complex of folate + NADPH in R67 dihydrofolate reductase, a D2 symmetric enzyme. *J. Comput. Aided Mol. Des.* 15:1035-52, 2001.

2.1 Introduction

R67 dihydrofolate reductase (DHFR) is a novel enzyme that confers resistance to the antibiotic trimethoprim. The crystal structure of R67 DHFR displays a toroidal structure with a central active-site pore. This homotetrameric protein exhibits 222 symmetry, with only a few residues from each chain contributing to the active site, so related sites must be used to bind both substrate (dihydrofolate) and cofactor (NADPH) in the productive R67 DHFR•NADPH•dihydrofolate complex. Whereas the site of folate binding has

been partially resolved crystallographically, an interesting question remains: how can the highly symmetrical active site also bind and orient NADPH for catalysis? Since computational docking tools are optimally suited for predicting such biologically important protein-ligand complexes, I employed our docking program SLIDE to model this ternary complex. Approaching a problem with different methods followed by a comparison of the results has the potential of providing a more confident prediction by supplying the complementary pieces of the whole puzzle. I compared the SLIDE results with the model predicted by Dr. Elisabeth Howell using DOCK, another method for docking flexible ligands into proteins using a quite different algorithm (Howell et al., 2001). One of the strengths of SLIDE is the balanced protein-ligand flexibility modeling, whereas DOCK explores the ligand conformational space more thoroughly. The two programs also employ different scoring functions to rank the dockings.

Dihydrofolate reductase (DHFR) catalyzes the reduction of dihydrofolate (DHF) to tetrahydrofolate (THF) using NADPH as a cofactor. This enzyme is essential in folate metabolism since tetrahydrofolate is required for the synthesis of thymidylate, purine nucleosides, methionine, and other metabolic intermediates; thus, DHFR has been a prime target for anticancer and antibacterial therapy. Whereas chromosomal DHFR has been extensively studied and was one of the first successful targets for structure-based drug design, the plasmid R67 encoded DHFR has only recently been characterized. R67 DHFR is of special interest because it can transfer resistance between bacteria against the antibiotic trimethoprim. This DHFR has an entirely different sequence and fold from chromosomal DHFR (Narayana et al., 1995). R67 DHFR is a homotetramer in which each short chain forms a five-stranded β-barrel also found in SH3 domains (Narayana et

al., 1995) and a variety of other proteins including the Tudor domain of human survival motor neuron protein 1, ferredoxin thioredoxin reductase, nitrile hydratase, two of the 50S ribosomal proteins, and HIV integrase (Holm and Sander, 1996).

Type II DHFR, typified by R67 DHFR, is a dimer of dimers as shown in Figure 2.1. The central pore forms the active site, and the high degree of symmetry means that each of the four subunits contributes the same few residues to the binding surface. R67 DHFR is unlike the chromosomal enzyme in another respect. There are three different ligand binding combinations available to its active site: 2 folate/DHF, or 2 NADPH, or 1 folate/DHF plus 1 NADPH (Bradrick et al., 1996). The latter is the productive ternary complex. Thus, each half of the pore can bind either NADPH/$NADP^+$ or folate/DHF, a very different binding strategy than observed for chromosomal DHFR. Crystallographically defining the positions of bound ligands has proven especially difficult for the plasmid encoded enzyme, as the four-fold symmetry within the pore results in a four-fold dilution of the electron density. For example, if one ligand is bound, there is an equal probability that this binding will be in any one of the four equivalent sites within the pore for each of the individual protein copies in the crystal lattice. This effectively dilutes the observed electron density to an average over these four states. The symmetry and small size of the pore also means that the same residues (possibly from different chains) must contribute to the binding of both folate and NADPH. Thus, R67 DHFR is a fascinating system for studying how evolution can select a limited number of residues to co-optimize the catalytically productive binding of two quite different ligands, folate and NADPH.

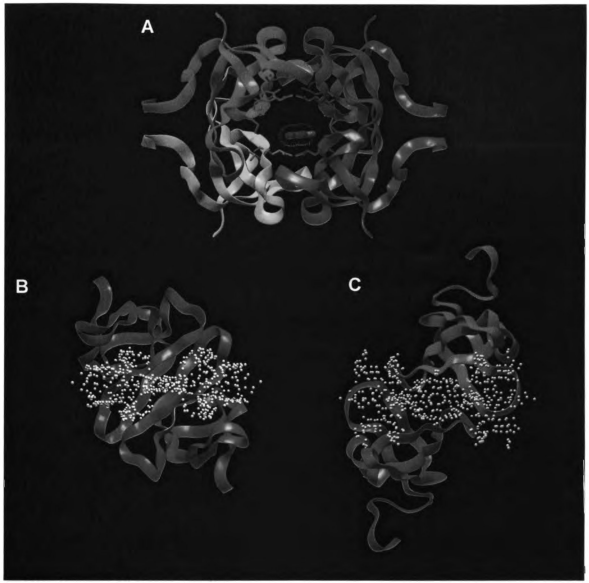


Figure 2.1. Panel A: The structure of R67 DHFR is a homotetramer formed by a dimer of dimers in which all four subunits (shown in red, yellow, blue, and green ribbons) contribute equally to create the symmetry related binding sites for folate and NADPH. The pteridine ring of the bound folate (shown in tubes in the central pore) and the side chains of the residues lining the pore are also shown. The view is down a twofold symmetry axis. Panels B and C describe a reverse image of the active site generated using the SPHGEN subroutine of DOCK on the 1VIE DHFR coordinates from the PDB. Water molecules were removed from the PDB file prior to running SPHGEN. Each sphere point corresponds to a possible atom position for docked ligands. In Panel A, the sphere cluster would fill the active site pore. Two perpendicular orientations of the protein chains and sphere cluster are shown in panels B and C.

A previous model of the ternary complex was given in Narayana et al. (Narayana et al., 1995). However, that model contained three bound ligands (2 folate + 1 NADPH), inconsistent with more recent solution studies indicating only two ligands are bound (Bradrick et al., 1996). Also the model by Narayana et al. positioned the productively bound folate molecule parallel to and above the NADPH molecule. This would predict numerous interligand NOEs, which are not observed in NMR experiments (Li et al., 2001). However, the partial density available for folate in the binary crystallographic complex provides a valuable guide to its favored position within the pore. The possibility that NADPH could interact in R67 DHFR in the same orientation relative to folate as it does in the chromosomal DHFR crystal structures was evaluated. However, due to steric limitations within the small pore of R67 DHFR, this binding mode is not feasible. Because the chromosomal DHFR complexes do not explain how the substrate and cofactor bind in R67 DHFR, and this ternary complex has so far proven crystallographically inaccessible, docking methods were used to predict their interactions in R67 DHFR. The predicted interaction of NADPH with folate in R67 DHFR were then compared with their orientation in chromosomal DHFR, and related to the effects of site-directed mutants on ligand binding.

2.2 Methods

DOCK v4.0 and SLIDE v1.1 were utilized to predict the binding modes of NADPH and folate in the active-site pore of R67 DHFR. DOCK v4.0 uses van der Waals interactions in its scoring and allows ligand flexibility (Ewing et al., 2001). SLIDE v1.1 includes protein side-chain flexibility, ligand flexibility, probabilistic inclusion of active-site

bound water molecules, and a scoring function with hydrophobic interaction and hydrogen bond terms (Schnecke and Kuhn, 2000). The structures of folate and NADPH and their atom labeling conventions are given in Figure 2.2.

SLIDE (Schnecke and Kuhn, 2000) is described in Chapter 1, section 1.4.

CONSOLV, a k-nearest-neighbor-based classifier (Raymer et al., 1997), was used to identify binding-site waters likely to be conserved upon ligand binding based on their mobility and their favorable interactions with the protein. CONSOLV labeled each bound water molecule in the 1VIF R67 DHFR structure according to its probability of being conserved upon ligand binding, and these values were used by SLIDE to appropriately incorporate bound water molecules or to penalize their displacement by non-polar ligand atoms.

DRUGSCORE is a knowledge-based scoring function (Gohlke et al., 2000) that was shown to discriminate efficiently between well-docked ligand-binding modes and computer-generated artifacts. DrugScore was used in addition to the built-in scoring function of SLIDE to score and rank all docked ligand orientations with a suitable distance between the C4 atom of the NADPH nicotinamide ring and the C6 of the folate pteridine ring (<5.0 Å).

LIGPLOT (Wallace et al., 1995) was used to create the figures showing the protein-ligand and ligand-ligand hydrogen bonds and hydrophobic interactions.

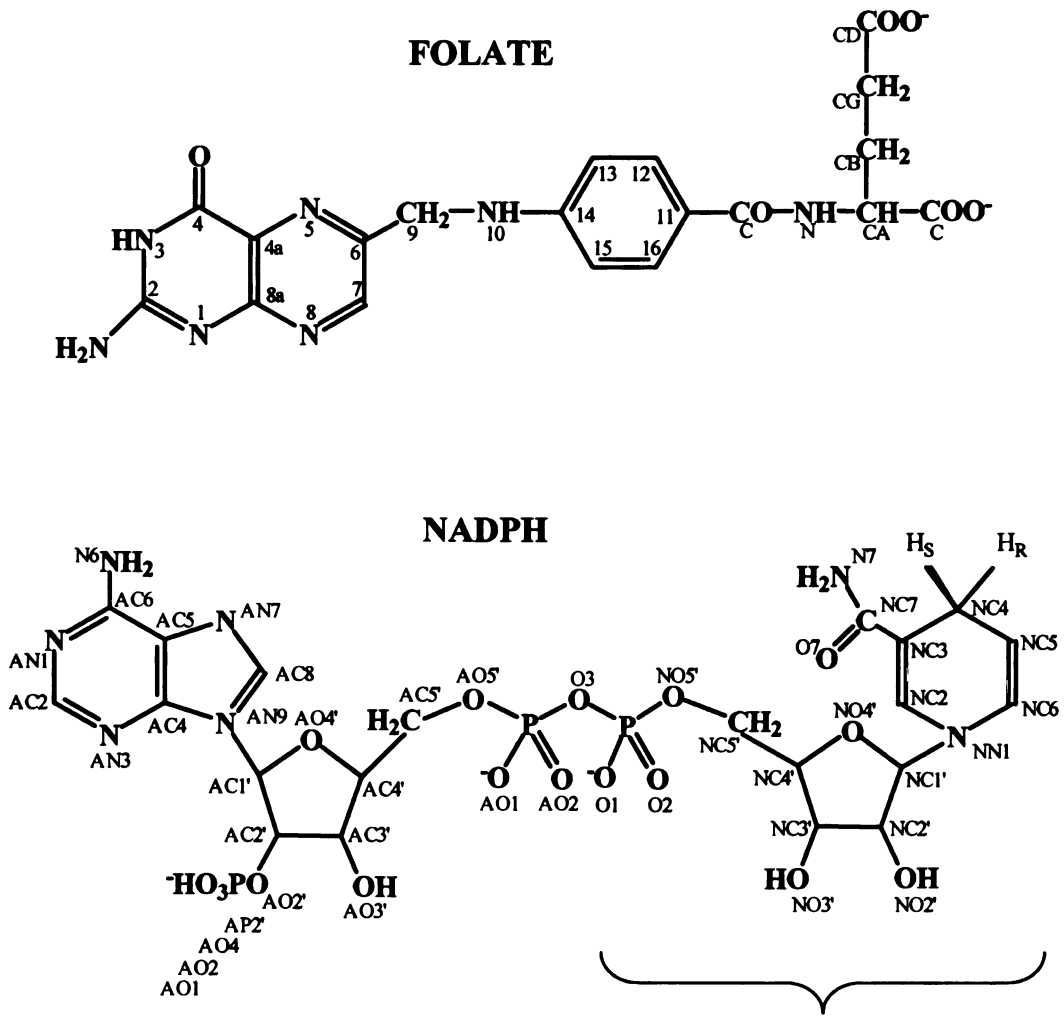


Figure 2.2. The structures of folate and NADPH. Reduction of folate across the C7-N8 bond yields dihydrofolate. During catalysis, the A or *re* hydrogen (H_R) on C4 of the nicotinamide ring faces the *si* face of the folate pteridine ring, which accepts a hydride at C7. The hydride would approach the *si* face of the pteridine ring from beneath the plane of the paper. The NMN moiety of NADPH is indicated by the bracket.

The coordinates of apo R67 DHFR as well as a binary complex with 2 folates bound are available as 1VIE and 1VIF (Narayana et al., 1995) at the PDB. In the present study, the structure 1VIF was used. The coordinates of the NADPH molecule were taken from the TRIPOS database for the DOCK experiment. SLIDE handles ligands as flexible molecules, but it avoids large conformational changes compared to the starting conformation. To include a broad range of energetically favorable starting conformations in docking with SLIDE, 59 NADPH molecules were extracted from crystal structures of various protein-NADPH complexes from the PDB. The nicotinamide ring is *syn* with respect to the ribose ring in 14 of these NADPH conformations and it is *anti* in 45 of them.

2.3 Results

2.3.1 Active Site Symmetry and Docking Constraints

A reverse image of R67 DHFR's active site was generated using the DOCK subroutine, SPHGEN. Two orientations of this image, given in Figure 2.1.B and C, show the symmetry associated with the pore as well as its size. If the ligand were small with respect to the binding site, four symmetry related sites could potentially be occupied. A larger ligand would reduce the number of possible binding sites because of steric hindrance. Binding of the ligand near the center of the pore, as is the case with Fol I from the crystal structure, is expected to have a similar effect by breaking the 222 symmetry, limiting the number of possible bound molecules to at most two, which is consistent with the experimental results (Bradrick et al., 1996).

Several constraints obtained from experimental data were used in preparing the docking experiments and in screening the docked ligand conformations to eliminate unlikely binding modes:

1. Isothermal titration calorimetry (ITC) data show a total of two ligands bind (Bradrick et al., 1996). The combinations are two folates or 2 NADPHs or 1 NADPH + 1 folate. Binding of two NADPH molecules shows negative cooperativity (24), suggesting the first molecule binds at or near the center of symmetry and impedes binding of a second molecule at a symmetry related site. Binding of two folate molecules shows positive cooperativity, indicating there are interactions between the bound folate molecules that enhance affinity.

2. Interligand NOE (ILOE) data from Li et al. (Li et al., 2001) show few ILOE's, suggesting the ligands are bound in extended conformations on opposite sides of the pore and meet somewhere in the middle of the pore.

3. From fitting the electron density, two folate molecules were modeled in asymmetric positions in 1VIF (Narayana et al., 1995). Fol I is bound productively with its *si* face exposed, whereas Fol II has its *si* face against the side of the pore, making it unavailable to receive a hydride. For this reason, Fol I was used to dock NADPH to the binary complex of R67 DHFR-folate.

4. For docking of folate or its analogues, the docked pteridine ring should conform to the observed electron density in the crystal structure (Narayana et al., 1995). This flat density was observed at the center of the pore near the Gln 67 residues, which form the "floor" and "ceiling" of the binding site. Density for the *p*-aminobenzoic acid–Glu (PABA-Glu) tail was not observed in the crystal structure, indicating disorder.

2.3.2 Docking of NMN into R67DHFR•Fol I Using SLIDE

All SLIDE dockings with a distance of 5 Å or less between the C4 of the nicotinamide ring of NMN and the C6 of the folate pteridine ring involved in hydride transfer were analyzed. There are four possible orientations: the nicotinamide ring can be *syn* or *anti* with respect to its ribose ring, and in both cases either the *pro-R* (A-side) or the *pro-S* (B-side) hydrogen can point toward the pteridine ring. These orientations are named *syn R, syn S, anti R and anti S*, respectively. Among the docked orientations, 39 adopted a *syn R* conformation, 4 were in *syn S*, 20 in *anti R*, and 12 in *anti S*. This distribution indicated a preference for the *syn* R orientation of NMN to interact with the R67 DHFR-Fol I complex, especially given that there were about three times as many *anti* conformers as *syn* conformers in the input data set of NADPH molecules. The *syn R* orientation is the one most consistent with the experimental results (Brito et al., 1990; Li et al., 2001)

In addition to the built-in scoring function of SLIDE, DrugScore (Gohlke et al., 2000) was used to evaluate these NMN dockings. DrugScore calculates an empirical intermolecular potential, with the best scores having the largest negative values, whereas the best SLIDE scores have the largest positive values (greater hydrophobic complementarity and number of intermolecular hydrogen bonds). Most of the high-scoring ligand orientations (lower right in Figure 2.3) were in *syn* conformation with the *R*-side hydrogen of the nicotinamide ring directed toward the folate. These orientations had the best scores with both scoring functions, except for one *anti R* orientation, which obtained an unusually high score with DrugScore. The available version of DrugScore does not consider water-mediated interactions, and therefore preferred dockings of

NADPH closest to the wall of the binding site such as this one. The *anti S* orientations, which obtained high scores from SLIDE but not DrugScore, had a larger number of hydrogen bonds formed between the P_i of NMN and various protein residues, but the nicotinamide ring formed at most one hydrogen bond with the protein. However, to have a well-defined stereochemistry between NADPH, folate, and the protein, some specific hydrogen bonding is expected between the head of the NADPH molecule and DHFR. The docked NMN in *syn R* orientation best fulfills this requirement by forming three hydrogen bonds between the O7 and N7 atoms of the nicotinamide head and the backbone hydrogen and nitrogen of Ile 68A, as well as the backbone oxygen of Val 66A, the latter being mediated by a water molecule (W 121A).

For waters bound in the DHFR•Fol I crystallographic complex, CONSOLV was used to predict their probability of conservation upon NADPH binding, based on the favorability of their interactions with the protein. After eliminating those water molecules that were found to be too close (<2.5 Å) to a protein or folate atom, only 11 water molecules were predicted to be more than 50% likely to be conserved inside the pore. Performing the docking experiment with the conserved water molecules included as part of the binding site did not result in significantly different dockings. The preference for the *syn R* orientation of the docked NMN was slightly higher compared to the dockings without waters, accounting for 62% of the docked conformations that have a high SLIDE ranking.

A number of water molecules were found to be important in anchoring the docked NMN to the protein (Figure 2.4.A), similarly to water molecules 121 and 124 (Figure 2.4.B) which have been suggested to form a bridge between the pteridine ring of folate

and the backbone of the R67 DHFR (Narayana et al., 1995). However, these water molecules were predicted to be only moderately conserved by CONSOLV. The explanation of this finding originates in the symmetry of the binding site: the productively bound pteridine ring can occupy any of the four symmetry related positions in the R67 DHFR tetramer structure, and by doing so it displaces different water molecules in different tetramers in the crystal lattice. As a result, many water molecules from the crystal structure of the R67 DHFR-folate complex (PDB entry 1VIF) have high temperature factors. In predicting conserved waters, CONSOLV weighs temperature factors heavily, so most of these waters were predicted to be only 28 - 55% conserved.

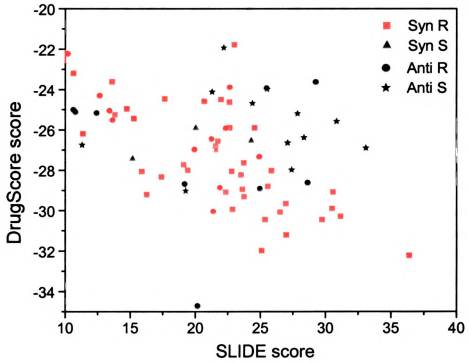


Figure 2.3. Comparison of the scores for well docked NMN molecules (consistent with experimental constraints and a distance of less than 5.0 Å between C4 of NADPH and C6 of folate) obtained with two different scoring functions: those of SLIDE and DrugScore. Because the currently available version of DrugScore does not include water-mediated contacts, these dockings did not include water molecules from the binding site, though similar dockings were found with water molecules included.

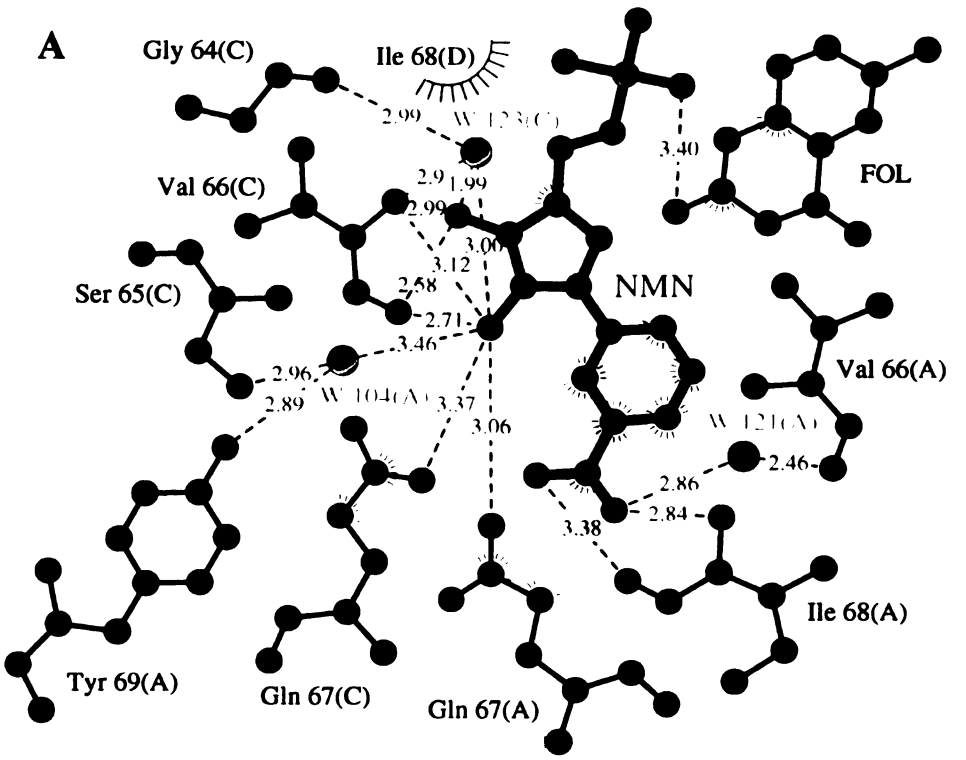
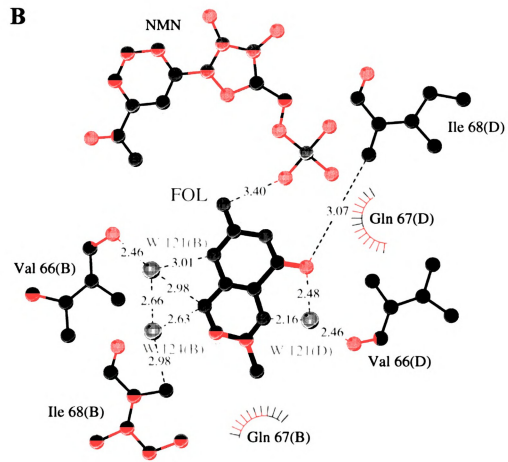


Figure 2.4.A. Potein-ligand and ligand-ligand interactions from the R67 DHFR•Fol I•NMN ternary complex for the NMN docked in the R67 DHFR•Fol I structure (drawn by LIGPLOT). The position of the NMN molecule in this complex corresponds to the top scoring docking obtained with SLIDE. W denotes water molecules.



KEY TO SYMBOLS

FOL folate; NMN nicotinamide mono-nucleotide
(A), (B), (C), (D) subunit identifiers in R67 DHFR

Bonds in ligand
Bonds in other molecules
3.0 Hydrogen bond and its length

His 53 Non-ligand residues involved in hydrophobic contact(s)
Corresponding atoms involved in hydrophobic contact(s)

Figure 2.4.B. Protein-ligand and ligand-ligand interactions from the R67-DHFR•FolI•NMN ternary complex for the pteridine ring. The position of the NMN molecule in this complex corresponds to the top scoring docking obtained with SLIDE. W denotes water molecules.

There is good agreement between the predictions of SLIDE and DOCK: both predict *syn R* to be the most likely orientation of the NMN molecule relative to folate. The position of the nicotinamide ring in the top scoring orientations (using SLIDE's scoring function) *syn R* is very similar to the top orientation produced using DOCK (Figure 2.5). The largest differences are found in the position of the phosphate group of NMN, which is understandable given the large space available and the absence of constraints because of the missing tail of the NADPH. The non-hydrogen atom RMSD between the top NMN orientations obtained with DOCK and SLIDE is 1.5 Å (Figure 2.5). The SLIDE scores and DrugScore scores for these two top dockings are 28.8 and –34,1300 for the DOCK docking and 36.4 and –32,2246 for the SLIDE docking.

The protein-ligand interactions generated by LIGPLOT (Wallace et al., 1995) for the R67 DHFR•Fol I•NMN ternary complex are shown in Figures 2.4.A and 2.4.B for NMN and the pteridine ring, respectively. The position of the NMN molecule corresponds to the top scoring NMN docking obtained with SLIDE, and the Fol I position from the crystal structure is used. A comparison of the contacts for NMN and folate shows that symmetry related residues were involved in binding both ligands. For example, Gln 67 from both the B and D subunits made several contacts with the pteridine ring, while Gln 67 from the A and C subunits made several contacts with the nicotinamide ring. Utilization of symmetry related residues during binding was also observed for Ile 68. Fol I binding involved Ile 68 from the D subunit which interacted with the pteridine ring, while Ile 68 from the A and D subunits interacted with the nicotinamide and ribose groups. Numerous van der Waals contacts and a hydrogen bond were also predicted between the ligands, as shown in Figures 2.4A and 2.4B. Positive

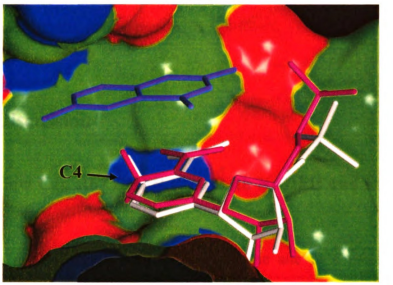


Figure 2.5. The NMN portion of NADPH docked into the binding site of R67 DHFR in *syn* R orientation next to the pteridine ring of folate (purple, at top). The solvent accessible molecular surface of the binding site is colored according to atom type: carbon is green, oxygen is red and nitrogen is blue. The top scoring orientation of NMN obtained with SLIDE (obtained with the water-mediated template and ranked 1st by SLIDE and 3rd by DrugScore) is shown in white and that obtained with DOCK is shown in magenta. Hydrogen atoms are shown only for the C4 of NADPH, which donates the hydride to reduce folate.

cooperativity has been previously observed between R67 DHFR•NADPH and DHF (Bradrick et al., 1996). The proposed interactions between NMN and Fol I may describe how positive cooperativity between NADPH and folate is generated.

One of the significant differences between SLIDE and DOCK is that SLIDE allows protein flexibility upon docking by balancing ligand and protein side chain rotations to resolve van der Waals overlaps, whereas DOCK more thoroughly explores ligand conformational space. In the case of R67 DHFR, there were only slight movements of two Gln 67 residues from subunits A and C, resulting in displacements of

less than 0.5 Å away from the docked NMN molecule, maintaining the original hydrogen-bonding pattern of the protein.

2.4 Discussion

2.4.1 How Can R67 DHFR Bind Both NADPH and Folate?

There are a number of cases in which the same site in a protein is designed to accommodate binding of several different ligands. Binding of diverse peptides to the major histocompatibility complex is achieved by having a number of specific binding pockets available for different side chains as well as by making key interactions to the peptides' backbones (Fremont et al., 1992; Matsumura et al., 1992). Binding of different unfolded protein chains to GroEL is proposed to be accomplished mainly by hydrophobic interactions where more flexibility is allowed (Chen and Sigler, 1999). To bind various sugars, the maltodextrin transport/chemosensory receptor uses aromatic rings to interact with the sugar ring faces (Quiocho et al., 1997). Binding of various peptides to oppA, a peptide transporter, utilizes numerous intermediary water molecules (Sleigh et al., 1999), as does binding of various fatty acids to adipocyte lipid-binding protein (LaLonde et al., 1994), binding of various sugars to arabinose binding protein (Quiocho et al., 1997), and high-affinity binding of a proteinaceous inhibitor, BLIP, to β-lactamases with diverse sequences (Strynadka et al., 1994). These are all mechanisms to facilitate numerous binding modes.

Hot spots for protein-protein interactions have been noted and evaluated by mutagenesis and statistical analysis (Bogan and Thorn, 1998; DeLano et al., 2000; Hu et al., 2000). A general trend proposed is the presence of residues that are amphipathic or can make hydrophobic and hydrogen-bonding interactions. For example, Tyr, Trp and Arg have a large hydrophobic component to their side chains as well as the ability to provide polar interactions. The residues that provide binding contacts in the center of R67 DHFR's active site pore include Ser 65, Val 66, Gln 67, Ile 68 and Tyr 69. The side chains of Ser 65 and Gln 67 are polar, while those of Val 66 and Ile 68 are hydrophobic. However, since Val 66 and Ile 68 present both their hydrophobic side chains as well as their backbone NH- and carbonyl groups for potential interactions, they can mediate both hydrophobic and polar interactions on the active site pore surface. Similarly, the side-chain methylene groups of Gln 67 also comprise part of the binding surface.

From Figures 2.4.A and 2.4.B, it is clear that the same residues are likely to be involved in binding both NADPH/NMN and folate/DHF. Utilization of protein symmetry is the mechanism by which this is achieved. For example, Gln 67 from subunits A and C make contacts with the NMN moiety while Gln 67 from subunits B and D make contacts with folate. This trend is also apparent with Val 66, Ile 68, Tyr 69, and Lys 32 residues. When symmetry operations are performed on the docked folate and NMN conformers, it is clear that while the binding sites are not identical, they overlap to a great extent. Three of the four symmetry related sites (generated by symmetry rotations) are shown in Figure 2.6. Two of the symmetry related sites compare the Fol I and NMN (top scoring conformer from DOCK) binding modes while the third compares NMN and Fol II (the non-productively bound folate in 1VIF). The fourth symmetry related site is

empty, precluding a comparison. Polar atoms that occupy similar positions in panel A are N5 of Fol I and N1 of the nicotinamide ring of NMN. In panel B, the C4 oxygen (Fol I) and the carboxamide oxygen (NMN), the N1 (Fol I) and N1 (NMN) as well as the N3 (Fol I) and carboxamide nitrogen (NMN) atoms occupy similar positions. Finally in panel C, the corresponding pairs of polar atoms that are close in space include: the C4 oxygen (Fol II) and the carboxamide oxygen (NMN) as well as the N1 (Fol II) and the N1 (NMN) atoms. This comparison supports a variation of hot spot binding, in which a few residues are responsible for most of the binding through making both polar and hydrophobic interactions with a small molecule ligand, rather than a protein (DeLano et al., 2000).

The number of similar docking orientations of the NMN fragment of NADPH indicates some alternative possibilities for hydrogen bonding to DHFR. This is also consistent with some mobility of bound NADPH, which in turn may explain the lower catalytic efficiency of R67 DHFR(Dion-Schultz and Howell, 1997; Reece et al., 1991). Because of the high degree of symmetry associated with the binding site of R67 DHFR, the catalytically productive folate•NADPH complex can bind in four equivalent positions, such that both molecules can be positioned at either side of the pore. The position adopted by NADPH independent of folate might well be different from the optimal position when folate is present, for two reasons: because folate creates a new chemical and structural environment that can favor a different placement of NADPH, and because the symmetry of the pore tells us that there may be several favored, overlapping optimal placements for folate and NADPH (Figure 2.6). Therefore, it seems that co-optimization of both ligands' binding is important.

Figure 2.6. Overlap of the NMN binding site with Fol I and Fol II sites. While two molecules do not bind in the same site concurrently, the symmetry of R67 DHFR implies that the same site must be used at different times for both NADPH and folate (in different halves of the pore or in different copies of the protein). Here, the top-scoring orientation of NMN from DOCK (Figure 3A) is compared (by symmetry operations) with the crystallographic orientation of Fol I or Fol II in the same site. Their substantial overlap corresponds to the region in which residues must be co-optimized for NADPH and folate binding. NMN atoms are labeled in yellow while Fol I or Fol II atoms are shown in white. In panel A, the closest protein atoms for interaction with the N1 (NMN) and N5 (Fol I) nitrogens are the carboxamide groups of the Q67 residues (3.69-4.46Å distant). In panel B, the closest protein atoms for interaction with the N1 ligand nitrogens are again the Q67 carboxamide groups (3.68-3.93Å). For interaction with the O4 (Fol I) or O7 (NMN) oxygens, the backbone NH from I68 lies nearby (3.07-3.25Å). The N3 (Fol I) or N7 (NMN) atoms come closest to the backbone oxygen of I68 (3.57-4.75Å). In panel C, the backbone NH of I68 is close (2.90-3.25Å) to the O4 (Fol II) or O7 (NMN) oxygens while the backbone oxygen of I68 could interact with the N5 (Fol II) or the N7 (NMN) atoms (2.68-3.28 Å). The closest protein atoms for interaction with the N1 nitrogens are again the carboxamide groups from the Q67 pairs (3.68-4.37Å). A similar comparison of the overlap between the Fol I and Fol II sites is shown in Figure 4b of Narayana et al. (Narayana et al., 1995).

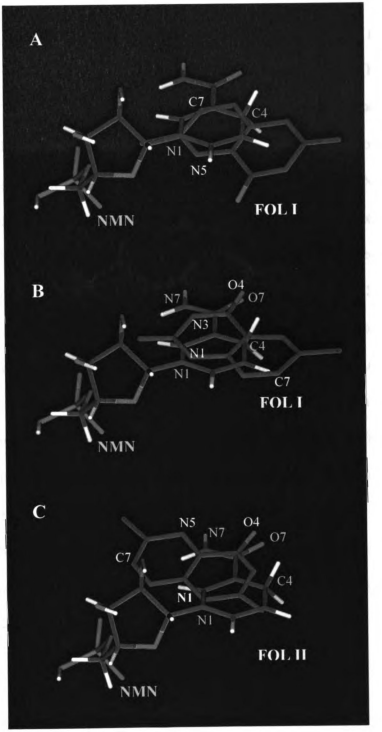

A
C7
C4
N1
N5
NMN
FOL I
B
N7
O4
O7
N3
C4
N1
N1
C7
FOL I
NMN
C
N5
N7
O4
O7
C7
C4
N1
N1
FOL II
NMN

Figure 2.6.

2.4.2 Relationship to Mutagenesis Results

Mutagenesis of R67 DHFR has been performed to evaluate the roles of many of the pore residues in ligand binding: Lys 32, Ser 65, Gln 67, Ile 68 and Tyr 69 (Park et al., 1997; Strader et al., 2001). The effects of mutations are consistent with the docked interactions of NADPH and folate (Figure 2.4). Mutating Ser 65 to Ala does not affect catalytic efficiency, suggesting it does not interact directly with the ligands. NMN docking by SLIDE predicted the Ser 65 side chain hydrogen bonds to a water molecule that participates in NMN binding; however, this water site is also stabilized by interactions with Tyr 69 and could persist in the absence of interactions with Ser 65. Gln 67 hydrogen bonds directly with NMN and makes hydrophobic interactions with folate in the docked ternary complex. Ile 68 makes direct hydrogen bond and hydrophobic interactions with NMN, as well as water-mediated interactions with folate. Tyr 69 participates in water-mediated interactions with NMN. As shown in Table 1, mutations at any of these residues (except S65) alter the K_m values for both ligands. The changes in K_m vary over three orders of magnitude (from 100 fold tighter to 10 fold weaker), however the ability of the mutations to preferentially alter NADPH vs. DHF binding appears marginal (Park et al., 1997; Strader et al., 2001). These data support a dual role for these active-site residues in binding both ligands.

Table 2.1. A comparison of steady state kinetic values for R67 DHFR variants at pH 7.0.

DHFR Species	k_{cat} (s^{-1}) (pH 7)	$K_{m\,(DHF)}$ (μM)	$K_{m\,(NADPH)}$ (μM)
Wt R67 DHFR[a]	1.3 ± 0.07	5.8 ± 0.02	3.0 ± 0.06
S65A R67 DHFR[b]	1.1 ± 0.10	4.0 ± 0.51	2.9 ± 0.57
Q67H R67 DHFR (pH 8)[c]	0.022 ± 0.003	0.16 ± 0.01	0.028 ± 0.001
I68M R67 DHFR[b]	0.17 ± 0.03	25 ± 3.0	21 ± 3.0
Y69F R67 DHFR[b]	2.5 ± 0.04	44 ± 2.1	66 ± 2.6

[a] Values from reference Reece et al., 1991.
[b] Values from reference Strader et al., 2001.
[c] Values from reference Park et al., 1997.

2.4.3 A Model for Hydride Transfer

The picture emerging from the docking studies using SLIDE and DOCK (Howell et al., 2001) is that folate and NADPH approach the catalytic site from the opposite ends of the R67 DHFR binding pore, with the *pro-R* side of the nicotinamide ring of NADPH turned toward the *si* face of the pteridine ring of folate. The hydride transfer distances between C7 of the pteridine ring and C4 of the nicotinamide ring, which participate in the reduction of folate (Figure 2.5) are predicted to be between 3.72 - 3.93 Å. These distances are longer than the 2.6-2.7 Å predicted by *ab initio* calculations (Castillo et al., 1999; Wu and Houk, 1987) and from a model of the transition state in *E. coli* DHFR (Bystroff et al., 1990). No docking method would probably be able to reproduce the

distances predicted by *ab initio* calculations for transition state complexes, but it is possible to reproduce crystal structure orientations with differences in intermolecular distances of approximately 0.2 Å. When testing the capacity of SLIDE to reproduce the crystal structure orientation of $NADP^+$ from a chromosomal DHFR in complex with folate and $NADP^+$ (PDB code 1RA2), the docked orientation of $NADP^+$ closest to the crystal structure position resulted in a C4-C7 distance of 3.45 Å, comparable to the 3.21 Å value found in that same crystal structure. The greater distances observed in the R67 DHFR dockings imply either a low rate of hydride transfer or an interligand chemical attraction that shortens the distance.

Molecular dynamic studies suggest that in general, enthalpic contributions to catalysis predominate over entropic contributions (Bruice and Benkovic, 2000). However in R67 DHFR, a range of similar docking modes is predicted for the ligands, or perhaps an unusual degree of mobility (Howell et al., 2001). Both these options likely result from the use of symmetry related residues. The ability of the PABA-Glu tail of folate and the 2',5'-ADP tail of NADPH to remain flexible but still maintain favorable electrostatic interactions may enhance binding through entropic as well as enthalpic contributions. An additional consequence of alternate binding modes for the ligand tails (or an enhanced mobility) might be to prevent binding of two molecules in one half of the pore, and instead steer binding to one molecule in opposite sides of the pore.

2.5 Conclusions

The evolution of catalytic activity is the focus of many recent research articles. One perspective suggests new enzymes evolve by gene duplication followed by accumulation of mutations. This approach takes advantage of structural and mechanistic similarities in generating different catalytic activities and suggests a certain level of catalytic promiscuity (Babbitt and Gerlt, 1997; O'Brien and Herschlag, 1999). In addition, catalytic antibodies might be expected to provide insight into the process of enzyme evolution. They appear to adopt predominately a lock and key strategy towards binding transition state analogs. Also, a comparison of different catalytic antibodies that catalyze the same reaction suggests they mostly converge to the same binding site motif (Karlstrom et al., 2000; Mader and Bartlett, 1997; Smithrud and Benkovic, 1997). In contrast to these evolutionary strategies, the results of DOCK and SLIDE showing the favored orientation of NADPH relative to folate in R67 DHFR indicate this enzyme has adopted a novel, yet simple approach: the utilization of symmetry related residues to bind both NADPH/$NADP^+$ and folate/DHF using a range of interaction types through a limited number of amphipathic residues. This symmetry is used to generate a hot-spot surface that accommodates numerous, different interactions.

Chapter 3

Distilling the Essential Features of a Protein Surface for Improving Protein-Ligand Docking, Scoring, and Virtual Screening

The research presented in this chapter has been previously published in:

Zavodszky, M.I., Sanschagrin, P.C., Korde, R.S., Kuhn, L.A. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening *J. Comput. Aided Mol. Des. in press.*

3.1 Abstract

For the successful identification and docking of new ligands to a protein target by virtual screening, the essential features of the protein and ligand surfaces must be captured and distilled in an efficient representation. Since the running time for docking increases exponentially with the number of points representing the protein and each ligand candidate, it is important to place these points where the best interactions can be made

between the protein and the ligand. This definition of favorable points of interaction can also guide protein structure-based ligand design, which typically focuses on which chemical groups provide the most energetically favorable contacts. In this chapter, a method of protein template and ligand interaction point design that identifies the most favorable points for making hydrophobic and hydrogen–bond interactions by using a knowledge base is presented. The knowledge-based protein and ligand representations have been incorporated in version 2.0 of SLIDE and resulted in dockings closer to the crystal structure orientations when screening a set of 57 known thrombin and glutathione S–transferase (GST) ligands against the *apo* structures of these proteins. There was also improved scoring enrichment of the dockings, meaning better differentiation between the chemically diverse known ligands and a ~15,000-molecule dataset of randomly-chosen small organic molecules. This approach for identifying the most important points of interaction between proteins and their ligands can equally well be used in other docking and design techniques. While much recent effort has focused on improving scoring functions for protein-ligand docking, our results indicate that improving the representation of the chemistry of proteins and their ligands is another avenue that can lead to significant improvements in the identification, docking, and scoring of ligands. This work is the result of a group effort. My roles were to develop the knowledge-based hydrogen bonding template, design the experiments to test the new method, analyze the results, and write the manuscript.

3.2 Introduction

3.2.1 The Evolution of Protein Surface Representations in SLIDE

Two methods to generate a template for the binding site of interest were initially implemented in SLIDE: small, biased, pharmacophore-like templates, and unbiased, grid-based approaches. The biased template is based on known ligand binding modes and consists of coordinates of ligand atoms making hydrogen bonds or engaging in hydrophobic interactions with the protein of interest, as seen in crystal structures of protein-ligand complexes. This pharmacophore-like representation of binding determinants is biased towards known ligands and is especially appropriate when the aim is to identify other molecules that make similar interactions. When the goal instead is to identify new classes of ligands or help define the ligand specificity for protein structures with unknown functions, an unbiased, thorough representation of the potential ligand-binding site is preferable. Therefore, SLIDE also has an option to automatically generate an unbiased template based on a ligand-free structure of the protein. To generate an unbiased template in version 1 of SLIDE, the binding site was filled with a large number of points, initially located on a fine grid with a spacing of 0.3–0.7 Å (Figure 3.1.A) (Schnecke and Kuhn, 2000). Initial experiments with random placement of the points showed significant under-representation of some areas in the binding site, so the grid-based approach was adopted instead. Only points located 2.5 to 5.0 Å from the nearest protein atom were kept. Each point was then checked to determine if it could serve as a hydrogen bond donor, acceptor, or form a hydrophobic interaction with the protein, and was either labeled as such, or eliminated from the set. All points of the same class were

then clustered using complete linkage clustering to reduce the number of template points to 150 or fewer.

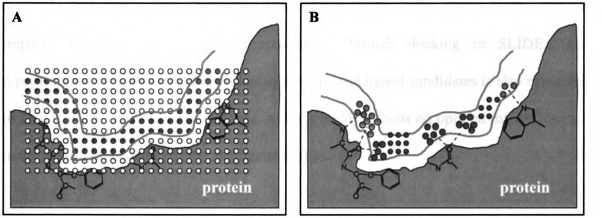


Figure 3.1. Comparison of the grid-based (A) and knowledge-based (B) template generation methods. Template points are generated on a grid in version 1 of SLIDE. The method implemented in SLIDE, version 2, uses a knowledge base to define points where optimal protein–ligand interactions can be made, based on points where the ligand could make optimal hydrogen bonds and hydrophobic interactions with the protein. Template points are colored according to their type: green for hydrophobic, red for acceptor, blue for donor, and purple for donor and/or acceptor points.

Improving the success rate of docking known ligands to a protein structure that does not already have correct side-chain conformations for that ligand (e.g., an "*apo*" structure of the protein, solved in the ligand-free state) was the motivation for the present work, which is aimed at defining protein templates that capture optimal points for interacting with the protein. Knowledge bases of hydrogen-bonding geometry around protein groups (Ippolito et al., 1990; McDonald and Thornton, 1994) allow us to focus now on optimal (rather than just feasible) positions for hydrogen bonding. Significantly hydrophobic positions at the protein surface can also be distinguished from the

background level of solvent-exposed carbon atoms, based on the local enhancement of hydrophobic atoms. Similarly, the interaction points on ligand candidates can be sampled to have similar density and chemistry to the hydrophobic and hydrogen-bonding assignments in the protein template. While this work has been driven by the aim to improve the modeling of protein recognition through docking in SLIDE, this representation of key interacting groups in proteins and ligand candidates is also expected to be useful for other docking methods, and to provide a focus on optimal interactions to make in structure-based protein and ligand design.

3.2.2 Other Approaches for Discrete Representation of Protein Binding Sites

Reduced representations of protein binding sites have been developed by other groups for use in modeling protein recognition. Typically, the protein's binding site is discretized to a set of 100 or fewer interaction points to enable fast comparison between the protein and each ligand. Many of these methods use reduced representations to aid in matching the protein and ligand surfaces. The initial, computationally complex search of the 6 degrees of rotational and translational freedom of the ligand relative to the protein is reduced to a problem of matching a set of N points on the ligand to the best-matching subset of N points from M points on the protein. N and M typically must be small due to the factorial complexity of the number of ways of matching N points to a larger set of M points. In the case of SLIDE, 3-point subsets of N interaction points on the ligand are tested for matching to all 3-point subsets of a set of typically 100-150 template points representing the protein.

In the case of DOCK (Kuntz et al., 1982), the earliest protein-ligand docking technique, the binding site is filled with spheres, whose centers serve as possible ligand atom positions. Chemical properties or other characteristics can be associated with the spheres, and a sphere with a particular characteristic can only be matched with a ligand atom of complementary character (Shoichet and Kuntz, 1993). Jones et al. (Jones et al., 1995) identify solvent–accessible hydrogen–bond donor and acceptor atoms within the active site of the protein and associate virtual points with each hydrogen and lone pair of these atoms, enabling the genetic algorithm employed by GOLD (Jones et al., 1997) to transform the ligand into the binding site by minimizing the least–square distance between protein virtual points and similarly defined ligand virtual points. Ruppert et al. (Ruppert et al., 1997) coat the protein's binding site surface with probes of three types, hydrophobic, acceptor and donor, which could potentially interact with the protein. These probes can serve as potential alignment points for ligand atoms and are scored to represent the probe's affinity for the protein. High affinity probe-clusters identify sticky spots, or regions of strongest potential binding. This method can also be used to find binding pockets on the surface of a protein. FlexX (Kramer et al., 1999) uses a multi-layered representation of the binding site adopted from its predecessor LUDI (Bohm, 1992): interaction types are arranged on three levels depending on their directionality, with H-bonds being the most directional at level three and hydrophobic interactions the least directional at level one. Each group capable of forming an interaction is characterized by an interaction center and a surface, the latter being approximated by a finite number of points. Ligand interaction centers are superimposed over these points and aligned, giving preference to higher-level interaction points over lower-level ones. In

an approach related to that of SLIDE, Fischer et al. (Fischer et al., 1993; Fischer et al., 1995) describe the surfaces of the protein and ligand by a set of critical points and their normals, then apply geometric indexing to dock the ligands into the protein by matching the critical points and vectors.

Grid-based representations are also used to map favorable points of interaction with proteins. In preparation for docking with AutoDock (Morris et al., 1998), the protein binding site is placed in a grid. The protein-ligand pair-wise interaction energies are precalculated at each grid point for each possible ligand atom type and are stored in a look-up table for use during the docking simulation. The Grid technique developed by Boobbyer et al. (Boobbyer et al., 1989) calculates for each grid point an empirical energy designed to represent the interaction energy of a chemical probe group, such as a carbonyl oxygen or an amine nitrogen atom, around the target molecule. This function is used to determine the sites where ligands may bind to the target, such as a protein.

Finally, knowledge bases of the frequency of pair-wise atomic or functional group interactions deduced from the crystallographic protein structures in the PDB (Berman et al., 2000) and small organic molecule structures in the Cambridge Structural Database (CSD) (Allen, 2002) can be used to map favorable sites for ligand interactions with proteins. Relibase (Bergner et al., 2001), a database system of protein-ligand interactions from the PDB, has been used to derive atomic potentials between protein and ligand atom groups for use in DrugScore (Gohlke et al., 2000a). DrugScore can then calculate "hotspots" for interactions with different ligand atom types, which are displayed as contour maps within the binding site (Gohlke et al., 2000b). Similarly, the SuperStar software (Verdonk et al., 2001), based on pair-wise interaction frequencies in the CSD

database, can calculate hotspots for the binding of 16 probe atom types to proteins. A recent paper analyzes how the interaction maps developed from PDB versus CSD data complement each other (Boer et al., 2001). Another knowledge-based approach was taken by Moreno and Leon (Moreno and Leon, 2002) to describe the binding site for DOCK: templates of attached points or contact points are constructed for each amino acid type, representing the geometry of the interactions observed in the different protein-ligand complexes from the PDB.

In this chapter, it is shown how a knowledge-based approach for describing favorable interaction sites on proteins and ligands can improve the performance of SLIDE when a database of known ligands combined with a random selection of CSD compounds is screened against two protein targets, thrombin and glutathione S-transferase (GST).

3.3 Methods

3.3.1 Knowledge-Based Representation of Protein Binding Sites

Because grid placement of hydrophobic and hydrogen-bond points is not always optimal with respect to protein interactions, here we describe the development of a knowledge-based approach to placing points in an unbiased template. Geometrically favored subsites for ligand hydrogen-bonding atoms are assigned based on the distance and angle to protein hydrogen-bonding partners (Figure 3.1.B). After identifying the protein atoms capable of hydrogen bonding, a number of template points are placed at and around the optimal hydrogen bonding position for each of these atoms, using the geometries shown

in Figure 3.2. The template points belonging to one hydrogen-bonding protein atom are separated by ~1 Å and are placed at a distance of 2.9 Å (for Asp, Glu, Lys, Thr and Tyr side chains) or at 3.0 Å (for all the other side chains and backbone oxygen and nitrogen) from the protein donor or acceptor atom. The parameters for optimal hydrogen bonding geometry were taken from the literature (Ippolito et al., 1990; McDonald and Thornton, 1994). The points are labeled as donors, acceptors or donor/acceptors, depending on the role an atom at this position would have in hydrogen bonding to the protein. A donor template point, for example, is located near an acceptor protein atom, such as a backbone carbonyl oxygen, and represents a favorable placement for a ligand atom acting as an H–bond donor. A donor/acceptor point is defined in two cases: when a ligand atom at that point could make favorable hydrogen bonds with separate hydrogen-bond donor and acceptor atoms in the protein, or when it could interact with a group that both donates and accepts hydrogen bonds (e.g., -OH in the side chains of Ser, Thr, or Tyr). Template points that overlap with those belonging to neighboring atoms (template points separated by < 1 Å) are clustered and relabeled, and points closer than 2.5 Å to a protein atom are discarded. The clustering of hydrogen-bonding template points reduces the number of points by about 10–25%. Points generated by the clustering of a donor and an acceptor point are relabeled as donor/acceptors.

Hydrophobic template points are generated using a grid for initial point placement, as before, but the criteria have been updated for which of these points should be included to represent favorable sites for ligand interactions. Hydrophobic points are those grid points with a hydrophobic enhancement score of at least 3. This score is defined as the number of carbon atoms minus the number of hydrophilic atoms, such as oxygen or

Figure 3.2. Panel A: Placement of optimal hydrogen-bonding template points in SLIDE. For each polar side chain, the optimal placement of hydrogen-bond donor (D), acceptor (A) and donor and/or acceptor (N) template points is shown with respect to the donor and acceptor atom positions in the side chain. These template points represent positions where a ligand atom matching the template point can form a hydrogen bond with the protein. A ligand atom matching a donor/acceptor (N) template point must be either a donor or acceptor, or both. These optimal distances and angles are consensus values describing preferred geometries (Ippolito et al., 1990; McDonald and Thornton, 1994) observed in high resolution protein structures from the PDB. The positions of hydrogen atoms in the protein are not assumed in template point placement, since these positions are not available in most crystal structures. Instead, the most favorable positions for hydrogen-bonding partners is measured relative to the geometry of the covalent bonds in the side chains (e.g., trans and gauche positions for Lys), as found from analysis of crystallographic data (Ippolito et al., 1990; McDonald and Thornton, 1994). *Panel B*: Three-dimensional example of template point placement relative to a Lys side chain. The template points defined for minimal, sparse, and dense templates are shown, along with the most-preferred distance and angle for hydrogen bonding, as shown above. The default template specification in SLIDE is dense, and thus there are more possible H-bond template point matches, each of which is shifted by a small amount relative to the optimal position and still allows formation of a near-optimal hydrogen bond between the matched ligand atom and the protein. Sparse and minimal hydrogen-bond templates are alternatives that can be used to decrease the number of hydrogen-bond template points when the complete template for a protein, including hydrophobic points, exceeds the practical limit of about 150 points.

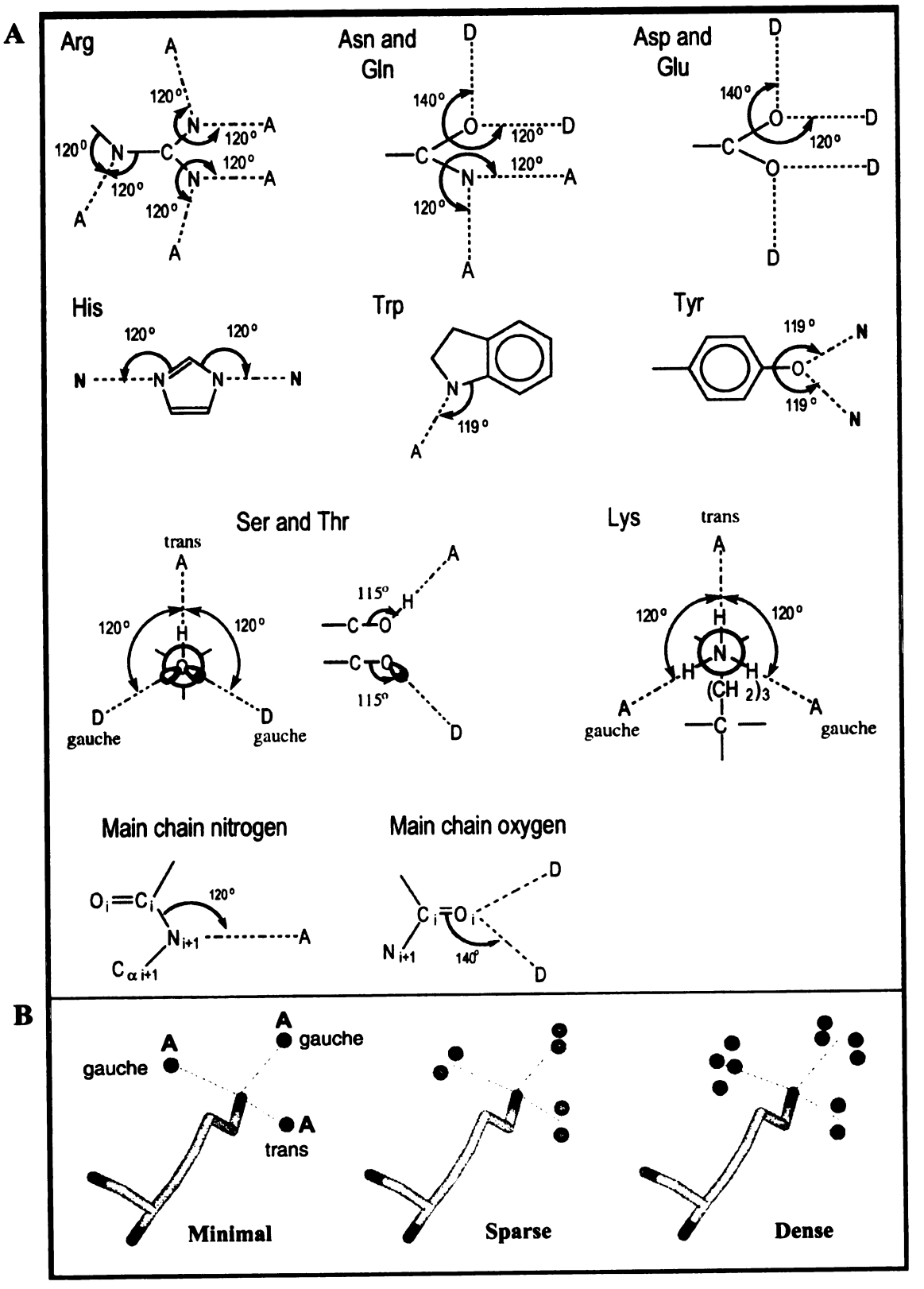

A
Arg
Asn and Gln
Asp and Glu
His
Trp
Tyr
Ser and Thr
Lys
trans
gauche
Main chain nitrogen
Main chain oxygen
B
gauche
trans
Minimal
Sparse
Dense

Figure 3.2

nitrogen, within a spherical shell of radius 2.5–5.0 Å from the template point in question. The cutoff value of 3 was found to define the significantly hydrophobic protein surface patches that complement the hydrophobic groups in ligands for a number of 3D protein-ligand complexes.

After they are generated separately, the H-bonding and hydrophobic template points are merged into one template that can be used for docking with SLIDE. If the total number of template points is much larger than 150 (a practical upper limit given the combinatorics of matching ligand interaction points with template points), then the complete linkage clustering feature can be used to reduce neighboring points of the same class to a single point, the cluster centroid. Complete linkage clustering has the desirable features that the clusters can be defined to not exceed a certain diameter (helping control the separation between centroids), and they are guaranteed to be the most densely occupied set of clusters for that diameter (Sanschagrin and Kuhn, 1998). Typically we use a clustering threshold of 4 Å, resulting in hydrophobic template points separated by about 2 Å. When a clustering threshold of x Å is used with complete linkage clustering (where x is typically chosen between 2 and 4 Å), the average distance between the final template points (the centroids of each cluster) is very close to x/2. For any uniformly distributed set of points clustered by complete linkage, the centroids of the clusters will be separated by half the cluster diameter (called the clustering threshold in this work), on average.

3.3.2 Ligand Interaction Points

Hydrophobic ligand interaction points are assigned using a rule-based approach summarized in Figure 3.3. These rules are designed to place an interaction point at approximately every 1.5 hydrophobic carbon atoms in hydrophobic chains and around the circumference of hydrophobic rings. This density of hydrophobic interaction points is

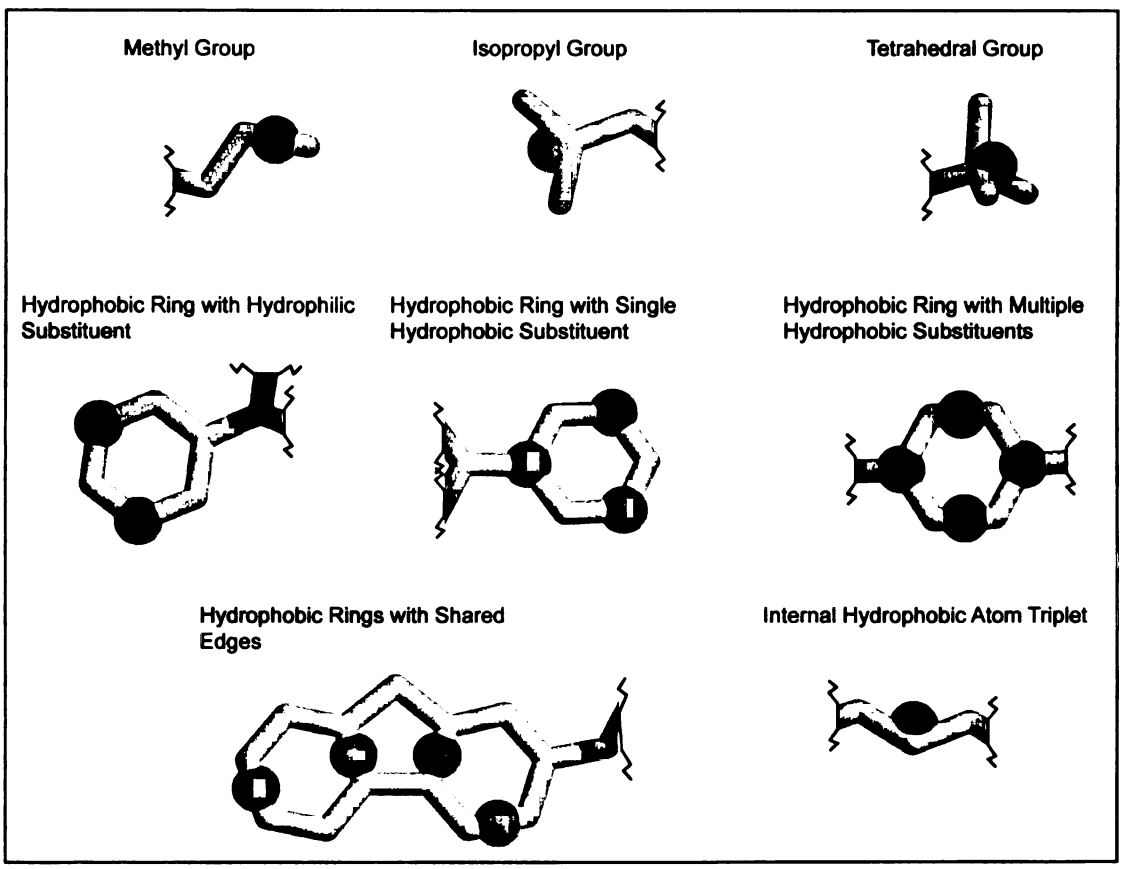


Figure 3.3. Summary of rules for hydrophobic interaction point assignment. The goal is to place a point at approximately every 1.5 carbon atoms, which is commensurate with the default spacing of hydrophobic points in the template. Hydrophobic interaction points are denoted by green spheres, carbon atoms by gray tubes, and nitrogen atoms, representing hydrophilic atoms, by blue tubes.

commensurate with the spacing of hydrophobic points in the protein template, using the default clustering criteria. For this approach, carbon and sulfur atoms bonded only to carbon, sulfur or hydrogen atoms are considered to be hydrophobic. Other atoms are taken as hydrophilic. Hydrogen-bonding interaction points in the ligand are identified as atoms capable of accepting or donating hydrogen bonds, based on the SYBYL atom types in the mol2 file (described at http://www.tripos.com).

3.3.3 Ligand Databases

A combined database of known ligands from the PDB and a subset of 14,691 randomly selected CSD compounds was assembled for alpha-thrombin and π-class human GST. The CSD database was prescreened to exclude molecules with excessive molecular weight as well as those containing unusual atoms. The nonredundant subset of known ligands for thrombin contained 42 molecules taken from thrombin–ligand complexes available from the PDB. To screen for ligands to GST, 15 known ligands with PDB crystal structures in complex with human GST were selected. For both thrombin and GST, ligands from crystal structures with a resolution of 3.0 Å or better were included in the known ligand test set. If a ligand was found in multiple structures, the one with the highest resolution was chosen. To ensure that SLIDE can appropriately model the side-chain conformational changes necessary in nature when proteins bind their ligands, structures of thrombin and π-GST determined crystallographically with ligand-free active sites (*apo* structures) were used as the targets for screening and docking (PDB code 1vr1 for thrombin (Dekker et al., 1999) and PDB code 16gs for GST (Oakley et al., 1998)). This also avoided the docking bias that is implicit in redocking experiments (when the

ligand-bound structure of the protein, already conformationally biased for that ligand, is used as the basis for docking). Because interactions in a mutant protein structure might change the favored orientation of a ligand relative to its orientation in the wild-type protein (and therefore not allow fair comparison of the docking with the crystallographic complex), ligands from complexes containing a mutant version of π-GST were excluded from the analysis. Four of the GST crystal complexes (PDB codes 13gs, 20gs, 21gs and 2gss (Oakley et al., 1997; Oakley et al., 1999)) contained two ligands: glutathione, and a smaller hydrophobic ligand bound to the xenobiotic subsite of the active site. Only the hydrophobic ligands from these structures were included in the screening dataset, and glutathione from the GST–glutathione complex 1aqw (Prade et al., 1997) was used as the single representation of this ligand in the screening set.

3.3.4 Key Template Points

In order to focus the large number of orientations that can result from the screening/docking process on productive binding modes, selected template points can be labeled as key points. Template points from parts of the binding site known to be critical for tight and/or specific binding can be marked as key points, and any docking must then include a match to one (not all) of these points. This ensures that docked molecules will at least partially occupy the targeted site. For thrombin, points in the specificity pocket within 5.0 Å of the carboxyl oxygens of Asp 189 were selected as key points. Assignment of key points in the GST binding site was more challenging, as it is made up of two subsites, one for hydrophobic ligands and the other for glutathione, which is fairly polar. SLIDE was run twice on the known ligands in the case of GST: initially with key

hydrogen bonding points in a 5.0 Å radius sphere around the side chain hydroxyl oxygen of Ser 65 in the deepest pocket of the glutathione binding site, to capture ligands that can bind to this polar site, then with key hydrophobic points in the area between Tyr 108 and Phe 8, the xenobiotic (hydrophobic) binding site. Screening against the CSD ligands was done using the first set of key points in the glutathione-binding pocket, which includes both hydrophobic and hydrogen-bonding interactions.

Using key points is mainly a convenient way to ensure that ligands make interactions in the deep pockets of the binding site, rather than making less favorable, superficial interactions. Placing key points in the deepest pocket of the thrombin active site would be useful, in the absence of any knowledge of thrombin ligand structure or chemistry, to ensure the absence of a significant, destabilizing cavity in the complex. Ensuring that deep pockets are filled is also a widely used approach in structure-based drug design to increase ligand binding affinity and specificity. For GST, the use of key points allows a convenient analysis of ligand binding to the hydrophobic binding site versus binding to the glutathione site, without specifying which ligands favor which site, or how they bind. We can therefore assess the accuracy of ligand specificity as well as docking for GST: hydrophobic ligands should fit and score well in the hydrophobic site, and score poorly if they also dock into the polar site (when key points are included there, instead), and vice versa for the polar ligands. This allows a more sophisticated analysis for GST, making use of both its binding sites. Key points can also hurt docking results, because not all ligands may make one of the chosen interactions and therefore would either not be docked at all, or would be forced to dock by making a non-native interaction. Thus, using key points is only recommended for predicting the docking of ligands if there

is a strong indication as to the location of a key binding pocket within the larger binding site (as is obvious in the case of thrombin, which has a funnel-shaped active site). Another appropriate occasion for including key points is in design applications, when the intent is to control which pocket or binding site is to be probed by a database of ligand candidates or fragments.

3.3.5 Evaluation of New Protein and Ligand Representations in Ligand Screening end Docking

Templates for thrombin and GST were created both with the grid-based and the knowledge-based template generation methods; the knowledge-based templates are shown in Figures 3.4.A and B. Sets of interaction points for the known ligands and the CSD compounds were also identified using both assignment methods. SLIDE was used to screen the known ligands and the CSD compounds against thrombin and GST, first using the grid-based template and the original ligand interaction points, and in a second experiment using the knowledge-based template and the new ligand interaction points. The two methods for representing the protein target and ligand candidates were evaluated in two ways. First, they were evaluated based on how well SLIDE, using these protein and ligand representations, could reproduce the known ligand positions in the structure of the protein-ligand complex. This involved docking the ligands into an *apo* structure of the protein, with side-chain positions not already optimized for the ligands. Secondly, they were evaluated by how well known ligands and nonspecific molecules (in our case, CSD compounds) could be differentiated. The heavy atom root–mean–square–deviation (RMSD) was used to compare the docked ligand orientation with its crystal structure

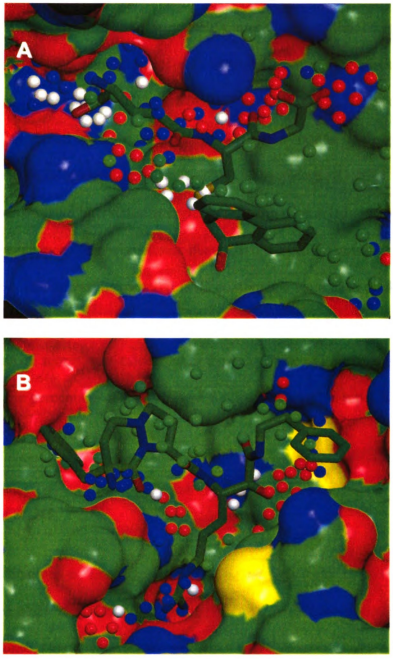


Figure 3.4. Examples of new knowledge-based templates. The Connolly solvent-accessible molecular surfaces (Connolly, 1993) of the GST (A) and thrombin (B) active sites are shown, color–coded according to atom type (green – carbon, blue – nitrogen, red – oxygen, yellow – sulfur). Known ligands from PDB structures 2pgt (A) and 1a5g (B) were docked into the binding site with SLIDE and are shown as tubes, also colored according to atom type. The template points are represented as spheres, with blue representing hydrogen–bond donor points, red for acceptors, whitefor donor/acceptors, and green for hydrophobic interaction points.

position. Because scoring remains a major challenge in the field (Bissantz et al., 2000; Charifson et al., 1999; Stahl and Rarey, 2001), and to ensure that the results were not very dependent on the particulars of one scoring function, the dockings were also evaluated using DrugScore as well as the SLIDE score. While SLIDE scores the protein-ligand complex based on the number of hydrogen bonds and the hydrophobic complementarity (Schnecke and Kuhn, 2000), DrugScore (Gohlke et al., 2000a) calculates protein-ligand interaction energies employing a knowledge-based potential that reflects the frequency of pair-wise atomic distances observed in protein-ligand complexes from the PDB. The known ligands and CSD compounds were each docked, scored, and sorted by score. Then, the enrichment in selecting known ligands from the random database, based on scores, was calculated as the percentage of known ligands captured as a function of the percentage of the database screened, where the top 1% of the database represented the top scoring compounds.

3.4 Results

All four combinations of template and ligand interaction point design were evaluated: grid-based template with original interaction points, grid-based template with new interaction points, knowledge-based template with original interaction points, and knowledge-based template with new interaction points. Both the knowledge-based template design and the new interaction point assignments resulted in improvements individually, but the most improvement was seen upon combining the two. For brevity, only the results obtained with the two most relevant combinations are presented: grid-based protein template with original ligand interaction point assignments (subsequently

referred to as method 1, corresponding to the implementation in SLIDE v.1), and knowledge-based template with new interaction points (method 2, as now implemented in SLIDE v.2).

3.4.1 Thrombin

The 42 known thrombin ligands used in this study are listed in Table 3.1, along with the PDB code of the crystallographic complexes from which they were obtained. SLIDE docked 36 ligands into the binding site of thrombin using both methods. The ligands with no scores listed could not be docked, due to unresolved steric overlaps with the apo-active site thrombin structure (1vr1) except for the case of benzamidine (PDB code 1dwb), which was not docked, due to the unusual proximity of its three interaction points (the two amide N's, and any pair of its three benzene-ring hydrophobic points, were all < 2.5 A apart). This caused benzamidine dockings to not meet a default parameter setting in SLIDE which ensures that the minimum edge of any triangle being matched is > 2.5 Å. This is intended to ensure that ligand dockings are complementary to more than a very local region of the binding site. (If the binding site is small, or the goal is to find small molecules that match very locally, this parameter can be changed easily.) Among the docked ligands, 27 had a heavy atom RMSD smaller than 2.0 Å compared to the crystal structure orientation using method 1, while 33 such dockings were obtained with method 2. As shown in Figure 3.5.A, the dockings were generally closer to the crystal structure position using method 2, as reflected by their lower RMSD values. The mean RMSD for thrombin ligand dockings was 1.83 Å using method 1, and 1.28 Å using method 2. An

example of the typical improvement in the quality of docking for thrombin ligands is shown in Figure 3.5.

Enrichment plots of the percentage of known ligands docked as a function of the percentage of the database screened (CSD plus thrombin ligands) are shown for SLIDE scores (Figure 3.7) and DrugScores (Figure 3.8). Higher enrichment is gained with method 2 compared to method 1, independently of the scoring function used (indicated by a shift to the left of the new curve compared to the original one in panel A in Figures 3.7 and 3.8). This means that more known ligands are returned by SLIDE among the top scoring CSD compounds. Based on the SLIDE score, for example, the percentage of the known ligands that ranked among the top scoring 100 molecules increased from 38% (16 out of 42) to 64% (27 out of 42). The results are very similar when using DrugScores: 67% of the known molecules (28 of the 42) ranked among the top scoring 100 molecules with method 2, compared to 33% (14 of the 42) using method 1.

The score distributions also show that the knowledge–based protein and ligand representations provide a better separation between known ligands and randomly chosen CSD compounds for both the SLIDE scores (Figure 3.7.B and C) and DrugScores (Figure 3.8.B and C). The difference between the mean SLIDE scores of the known ligands and CSD compounds increased from 20.7 score units to 27.1 score units when method 1 was replaced by method 2. DrugScore also mirrors a better discrimination between known ligands and CSD compounds when the knowledge–based method is used.

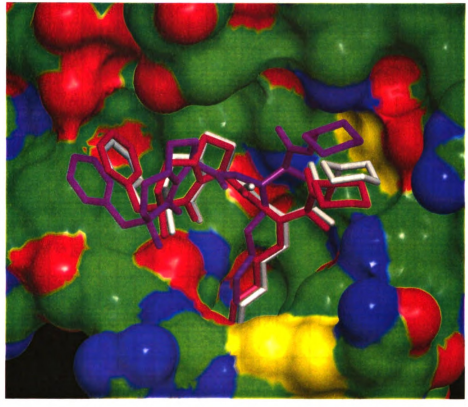


Figure 3.5. Comparing the docked orientations to the crystal structure position of a β–strand mimetic inhibitor (PDB code 1a46) in the binding site of thrombin. The crystal structure position of the ligand is shown in white, and the docked orientation using the knowledge-based method is in magenta (RMSD 1.03 Å), while the docking obtained with the grid-based method is shown in blue (RMSD 2.48 Å). This is representative of the improvement in docking quality observed for the thrombin and GST ligands in general. The view into the thrombin active site is slightly shifted relative to that in the previous panel.

Table 3.1. Comparison of SLIDE scores, DrugScore scores and RMSDs of known thrombin ligands docked into the active site of thrombin by SLIDE using the original template and ligand interaction point generation methods in comparison with the knowledge-based method. The original DrugScore scores are divided by 10^4 to give a comparable order of magnitude to SLIDE scores. For both scores, a larger absolute value means a better score, and "best" corresponds to the docking with the highest score or lowest RMSD (columns 5-8 and 9-10, respectively).

PDB code	Ligand name	DrugScore x10^{-4}	SLIDE score	Best SLIDE score		Best DrugScore x10^{-4}		Best RMSD (Å)	
		crystal structure position		grid-based	knowledge-based	grid-based	knowledge-based	grid-based	knowledge-based
1a2c	Aeruginosin298-A	–41.8	60.4	29.2	23.6	–37.4	–29.8	8.33	8.57
1a3b	Borolog1	–56.5	49.8	50.5	55.0	–48.3	–56.0	1.22	0.30
1a3e	Borolog2	–32.7	32.0	–	–	–	–	–	–
1a46	Beta-strand mimetic inhibitor	–62.0	57.3	43.3	61.2	–37.3	–52.2	2.49	1.03
1a4w	Ans-Arg-2ep-Kth	–48.0	71.5	59.8	61.1	–49.5	–50.5	1.39	0.52
1a5g	Bic-Arg-Eoa	–70.9	60.1	55.8	74.0	–65.5	–62.0	0.56	0.97
1a61	Mol-Arg-Lom	–58.1	58.6	54.1	64.3	–38.1	–54.0	1.90	1.02
1ad8	MDL103752	–72.6	31.8	–	–	–	–	–	–
1ae8	Eoc-D-Phe-Pro-Azalys-Onp	–47.5	36.9	45.5	53.1	–44.6	–46.4	0.32	0.65
1afe	Cbz-Pro-Azalys-Onp	–38.5	21.8	35.4	40.5	–36.2	–33.6	1.26	1.05
1aht	p-Amidino-phenyl-pyruvate	–37.2	30.9	22.7	29.9	–31.6	–34.9	2.40	0.81
1ai8	$PhCH_2OCO$-D-Dpa-Pro-boroMpg	–55.1	44.0	–	–	–	–	–	–
1aix	$PhCH_2OCO$-D-Dpa-Pro-boroVal	–51.4	38.1	–	–	–	–	–	–
1awf	GR133487	–44.8	56.8	44.0	28.0	–43.0	–34.2	1.56	11.29
1awh	GR133686	–44.5	37.0	47.5	–	–46.4	–	0.90	–
1ay6	Hmf-Pro-Arg-Hho	–57.2	72.1	55.1	66.8	–48.5	–54.4	1.01	0.71
1b5g	Bcc-Arg-Thz	–56.7	37.9	32.8	57.8	–30.1	–56.9	8.71	0.40
1ba8	Pms-Ron-Gly-Arg	–51.5	58.5	51.6	57.5	–43.2	–46.4	0.98	0.51
1bb0	Pms-Ron-Gly-3ga	–50.7	54.8	51.2	66.8	–44.7	–50.5	1.14	0.65
1bcu	Proflavin	–30.8	24.0	21.9	25.7	–30.5	–27.0	3.90	2.16

Table 3.1 continued

1bhx	SDZ 229-357	–47.2	49.3	44.2	55.6	–43.4	–51.2	0.75	0.53
1bmm	BMS-186282	–50.3	53.9	47.8	57.5	–41.2	–52.6	0.71	0.38
1bmn	BMS-189090	–55.5	45.7	38.4	43.5	–48.5	–55.9	0.83	0.29
1dwb	Benzamidine	–26.7	15.9	–	–	–	–	–	–
1dwc	MD-805 (Argatroban)	–42.6	52.8	57.3	45.2	–43.3	–43.3	0.56	1.04
1dwd	NAPAP	–60.5	46.9	43.5	52.7	–52.5	–64.3	0.97	0.44
1fpc	Ans-Arg-Epi (DAPA)	–40.4	46.9	57.0	53.2	–40.3	–39.0	0.94	0.90
1hdt	Alg-Phe-Alo-Phe-CH_3	–53.8	53.0	67.1	61.3	–55.2	–53.6	0.43	0.65
1lhc	Ac-D-Phe-Pro-boroArg-OH	–57.3	52.8	37.2	46.2	–45.2	–52.8	1.18	0.67
1lhd	Ac-D-Phe-Pro-boroLys-OH	–51.1	41.3	32.7	48.5	–35.8	–46.9	1.29	0.77
1lhe	Ac-D-Phe-Pro-boro-N-butyl-amidino-Glycine-OH	–59.9	54.2	51.3	55.1	–49.6	–56.3	1.05	0.71
1lhg	Ac-D-Phe-Pro-borohomoornithine-OH	–46.8	42.8	–	37.3	–	–35.2	–	1.32
1nrs	Leu-Asp-Pro-Arg	–51.9	52.8	43.5	57.7	–34.6	–46.6	1.49	0.75
1ppb	PPACK	–50.9	43.8	46.2	44.4	–32.5	–50.7	1.73	0.71
1tbz	Dpn-Pro-Arg-Bot	–62.8	40.2	41.8	66.5	–35.8	–51.4	4.11	1.10
1tmb	Cyclotheonamide A	–68.4	72.3	54.2	57.9	–38.9	–60.4	2.61	1.00
1tmt	Phe-Pro-Arg	–54.7	47.6	48.9	51.4	–36.1	–49.7	2.18	0.54
1tom	Methyl-Phe-Pro-amino-cyclohexylglycine	–49.2	36.1	43.1	45.5	–39.1	–44.9	1.01	0.74
1uma	N,N-dimethylcarbamoyl-alpha-azalysine	–18.9	20.7	12.9	23.0	–15.1	–22.0	2.80	0.94
3hat	Fibrinopeptide A mimic	–51.0	50.4	46.5	39.2	–29.0	–46.7	1.62	0.89
7kme	SEL2711	–60.4	49.1	52.4	60.7	–63.2	–65.1	0.54	0.40
8kme	SEL2770	–59.0	59.2	47.1	58.6	–54.8	–61.5	1.06	0.79

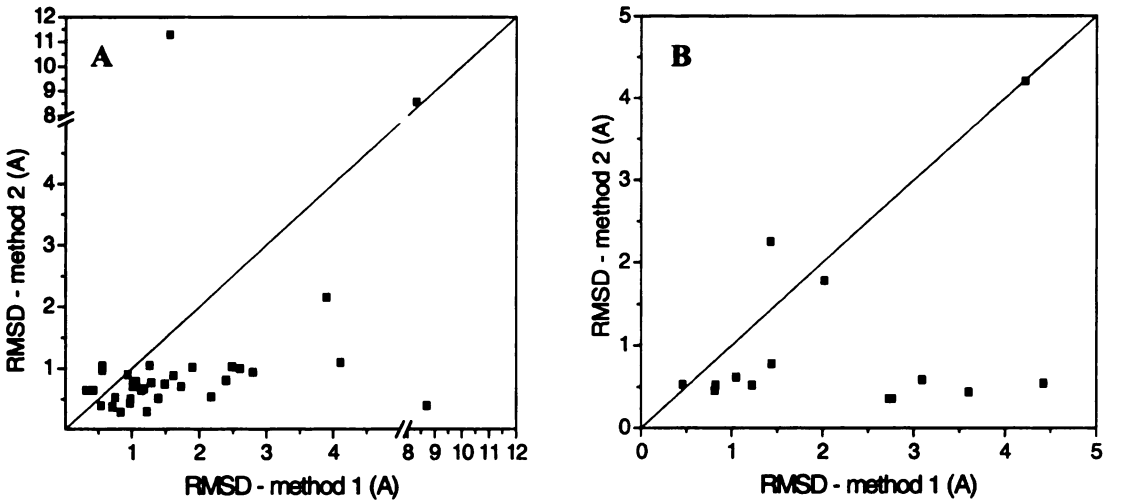


Figure 3.6. Comparing the RMS deviations between the docked orientations of known ligands and their crystal structure positions resulting from the original (1) and the knowledge-based (2) methods of template and ligand interaction point generation in the case of thrombin (A) and GST (B). Ligands docked better (with lower RMSD) with the knowledge-based method are represented by points below the diagonal line. The significant outlier in (A) with RMSD ~11.3 Å is a ligand with a neutral side chain occupying the S1 specificity pocket of thrombin in the x-ray structure of the protein-ligand complex (PDB code 1awf). This is a case in which the inclusion of key points can lead to misdocking. The atypical lack of hydrogen-bonding atoms in the portion of the 1awf ligand that binds to the S1 specificity pocket led to the inability of SLIDE to match this part of the molecule to at least one key point in the S1 pocket. The ligand was thus rotated by SLIDE about 180° compared to its crystal structure position, in order to satisfy the key point matching requirement by placing another, polar side chain into the S1 pocket.

Figure 3.7. Screening and enrichment improvements for thrombin using the knowledge-based template and new ligand interaction point assignments, as reflected by SLIDE scores (A), where a shift to the left of the curve corresponding to the new method indicates slightly improved enrichment. The distributions of SLIDE scores obtained with the grid-based method (B) and the knowledge-based method (C) show that the knowledge-based method gives a better separation between the scores of known thrombin ligands and random CSD compounds, reflected by a greater separation between the means of their score distributions. Curves that do not reach 100% for the "Percent of known ligands retrieved" reflect the fact that some ligands were not docked due to unresolved steric overlaps with the protein or due to the unusual proximity of the ligand interaction points (see text for further details).

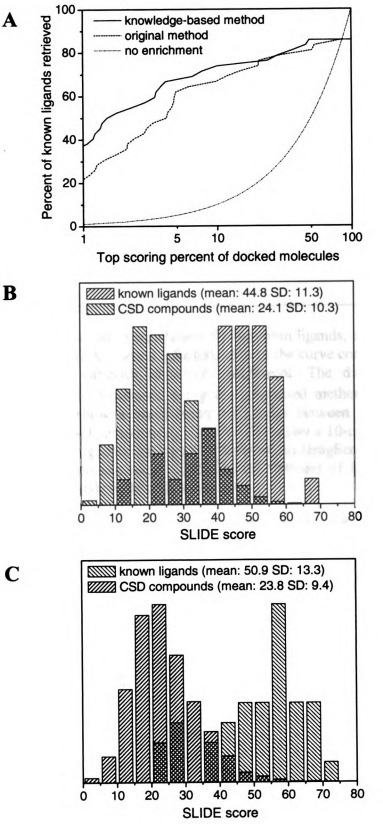

A
knowledge-based method
original method
no enrichment
Percent of known ligands retrieved
Top scoring percent of docked molecules
B
known ligands (mean: 44.8 SD: 11.3)
CSD compounds (mean: 24.1 SD: 10.3)
SLIDE score
C
known ligands (mean: 50.9 SD: 13.3)
CSD compounds (mean: 23.8 SD: 9.4)
SLIDE score

Figure 3.7

Figure 3.8. Significant improvement in enrichment for thrombin ligands, as reflected by the scoring function DrugScore (A), where a leftwards shift of the curve corresponding to the knowledge-based method indicates improved enrichment. The distributions of DrugScore scores (divided by 10^4) obtained using the grid-based method (B) and the knowledge-based method (C) show a much better separation between the scores of known thrombin ligands and CSD compounds. This is reflected by a 10-unit increase in separation between the mean DrugScore for ligands and the mean DrugScore for random CSD compounds. Curves that do not reach 100% for the "Percent of known ligands retrieved" indicate that some ligands were not docked.

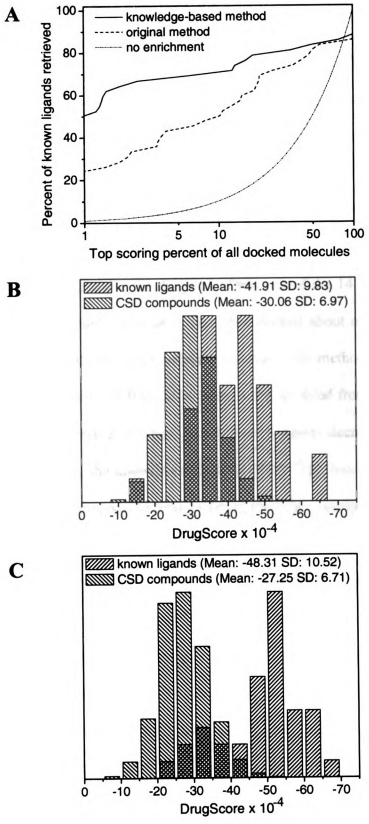

A
knowledge-based method
original method
no enrichment
Percent of known ligands retrieved
Top scoring percent of all docked molecules
B
known ligands (Mean: -41.91 SD: 9.83)
CSD compounds (Mean: -30.06 SD: 6.97)
DrugScore x 10-4
C
known ligands (Mean: -48.31 SD: 10.52)
CSD compounds (Mean: -27.25 SD: 6.71)
DrugScore x 10-4

Figure 3.8

3.4.2 Glutathione S–Transferase

SLIDE was able to find a collision–free orientation for 14 of the 15 known ligands in the active site of GST with method 2, while 13 were docked using method 1 (Table 3.2). The ligand from the crystal complex 19gs could not be docked for the same reason described for benzamidine in the previous section, whereas the reason for failure of chlorambucil (21gs) to dock was the existence of unresolved steric clashes with the protein. Method 2 resulted in better dockings (lower RMSD values), as illustrated in Figure 3.6.B by the majority of points falling under the diagonal. Only one of the 14 docked ligands had a lower RMSD when method 1 was used, two were docked about equally well, while 10 were docked closer to their crystal structure position with method 2. The number of known ligands docked with an RMSD less than 2.0 Å doubled from five to ten, and the mean RMSD between crystal structure and docked positions decreased from 2.15 Å to 1.00 Å upon introducing the knowledge–based method. The four hydrophobic ligands, shown by the crystal complexes to bind to the hydrophobic subsite of GST (13gs, 20gs, 21gs, 2gss), were docked incorrectly (RMSD > 5.0 Å) when polar template points were used as key points. This is not surprising given that these ligands must make interactions in a region different from where the key points were assigned. However, their docking improved substantially when hydrophobic key points were used in the second run with either method of template generation and interaction point assignment. Hydrophobic template points can be used as key points for docking smaller sets of ligands to a protein, but this is not a practical alternative when screening large databases. Since matching three template points is sufficient for docking with SLIDE, using hydrophobic key points when screening a large database can result in docking a very large number of small,

relatively nonspecific, hydrophobic molecules. They could later be eliminated based on their scores, of course, but this would still result in a considerable increase of the running time and output volume.

Only the results of the first run (with hydrogen bonding key points) were used to construct the enrichment plots for GST (Figure 3.9.A). For brevity, only the enrichment plot for DrugScores is shown; the results were substantially similar using SLIDE scores. DrugScores indicate that more of the known ligands were retrieved among the top scoring molecules (Figure 3.9.A), meaning improved enrichment was achieved with method 2 compared to method 1 for GST. When the SLIDE scoring function was used, 73% of the known ligands (11 out of 15) were ranked among the top scoring 100 of all docked molecules when using method 2, compared to 60% (9 out of 15) among the top 100 with method 1. Using DrugScore, the percentage of the known ligands ranking among the top scoring 100 of all the docked molecules increased from 33% (5 out of 15) to 60% (9 out of 15).

The distribution of scores obtained for the docked known ligands and CSD compounds to GST are shown in Figures 3.9.B and C. The difference between the mean scores of the GST ligands and randomly selected GST molecules increased due to the introduction of the knowledge–based method, independently of the scoring function applied: the means were separated by an additional 7.5 score units using SLIDE scores, and by an additional 9.4 X 10^4 units using DrugScore. Although the standard deviations of the DrugScores and SLIDE scores also increased, the increased separation of the means was roughly three times greater than the increase in standard deviations.

Table 3.2. Comparison of SLIDE scores, DrugScore scores and RMSD's of known GST ligands docked into the active site of GST (PDB code 16gs). SLIDE was used with the grid-based template and original ligand interaction point generation methods in comparison with the knowledge-based method. The original DrugScore scores are divided by 10^4 to give a comparable order of magnitude to SLIDE scores. For both scores, a larger absolute value means a better score, and "best" corresponds to the docking with the highest score or lowest RMSD (columns 5-8 and 9-10, respectively).

PDB code	Ligand name	DrugScore x10^{-4}	SLIDE score	Best SLIDE score		Best DrugScore x10^{-4}		Best RMSD (Å)	
		crystal structure position		grid-based	knowledge-based	grid-based	knowledge-based	grid-based	knowledge-based
10gs	Benzylcysteine phenylglycine	–50.2	40.7	42.9	44.6	-23.3	-49.1	2.73	0.36
12gs	S-nonyl-cysteine	-49.7	46.8	40.5	52.2	-24.4	-44.9	2.77	0.36
13gs *	Sulfasalazine	-30.7	34.6	20.6 (38.4)	32.2 (50.9)	-12.4 (-27.3)	-21.7 (-27.7)	8.76 (2.02)	6.42 (1.78)
18gs	1-(S-glutathionyl)-2,4-dinitrobenzene	-41.8	44.5	38.8	47.9	-35.9	-36.2	1.06	0.64
19gs	Phenol-1,2,3,4-tetrabromo- phthalein-3',3"-disulfonic acid ion	-12.6	17.7	–	–	–	–	–	–
1aqv	p-Bromobenzylglutathione	-47.7	43.8	40.4	48.6	-18.2	-43.6	3.61	0.44
1aqw	Glutathione	-36.6	31.8	24.6	37.5	-28.0	-32.7	0.82	0.46
1aqx	S-(2,3,6-trinitrophenyl)cysteine	-46.5	37.4	37.5	42.4	-32.0	-49.6	1.44	0.78
1pgt	S-hexylglutathione	-46.1	49.2	33.1	46.4	-42.9	-38.3	0.46	0.53
20gs *	Cibacron blue	-22.2	21.7	19.5 (50.3)	44.0 (56.2)	-22.1 (-24.6)	-28.2 (-25.9)	5.51 (0.83)	5.48 (0.52)
21gs *	Chlorambucil	-22.0	25.4	– (37.6)	12.1 (35.7)	– (-19.6)	-18.5 (-23.0)	– (4.22)	9.11 (4.21)
2gss *	Ethacrynic acid	-19.7	26.9	10.1 (30.7)	20.3 (36.1)	-14.6 (-24.6)	-18.5 (-19.1)	6.21 (1.43)	5.33 (2.25)
2pgt	(9R,10R)-9-(S-glutathionyl)-10-hydroxy9,10dihydrophenanthrene	-54.8	55.6	35.6	68.5	-28.9	-53.4	4.42	0.54
3gss	Ethacrynic acid-Glutathione conjugate	-52.3	75.0	58.3	71.4	-31.3	-47.7	1.22	0.52
3pgt	Glutathione conjugate of (+)-Anti-BPDE	-51.3	66.1	59.7	64.1	-36.0	-50.6	3.09	0.59

Figure 3.9. Enrichment for glutathione S–transferase ligands, as reflected by the scoring function DrugScore (A), where the significant leftwards shift of the curve corresponding to the knowledge-based method indicates greater enrichment. The distributions of the scores (divided by 10^4) obtained using the original grid-based method (B) and the knowledge-based method (C) again show a better separation between the scores of known GST ligands and CSD compounds, indicated by the large increase of 10 units between the means of these two classes of compounds. Given the smaller sample size (15) of GST ligands, this score distribution is less well defined than those for thrombin (Figures 3.7 and 3.8). However, the same trends in improvement are found for both proteins and both scoring functions. Curves that do not reach 100% for the "Percent of known ligands retrieved" indicate that some ligands were not docked. This percentage decreased with use of the knowledge-based template.

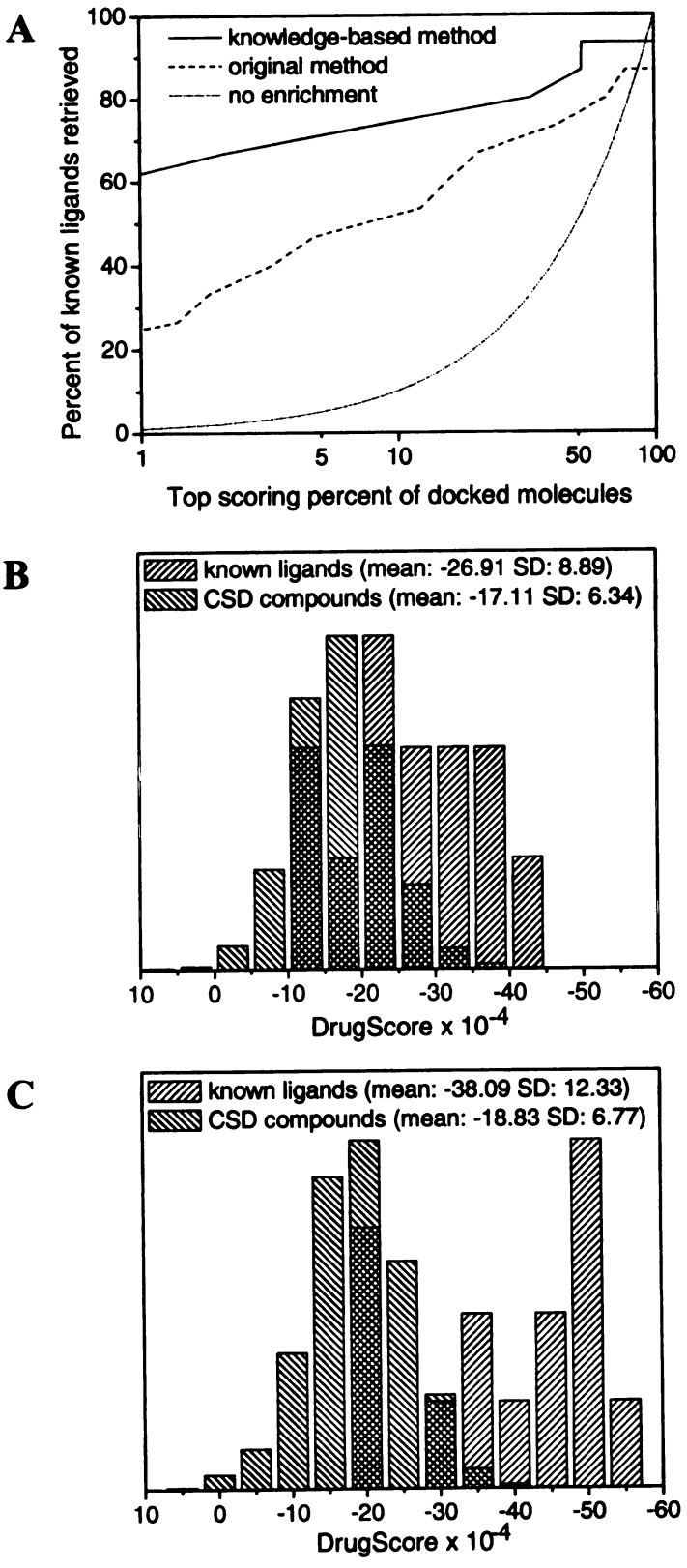

A
100
80
60
40
20
0
Percent of known ligands retrieved
knowledge-based method
original method
no enrichment
1
5
10
50
100
Top scoring percent of docked molecules
B
known ligands (mean: -26.91 SD: 8.89)
CSD compounds (mean: -17.11 SD: 6.34)
10
0
-10
-20
-30
-40
-50
-60
DrugScore x 10-4
C
known ligands (mean: -38.09 SD: 12.33)
CSD compounds (mean: -18.83 SD: 6.77)
10
0
-10
-20
-30
-40
-50
-60
DrugScore x 10-4

Figure 3.9

3.5 Discussion

3.5.1 The Influence of Accurate Binding Site Representation on Docking and Scoring

Because the computation time increases nearly exponentially with the size of the template, a compromise must be reached such that the most important features of the binding site are captured with the smallest possible number of template points. Using a knowledge–based approach for identifying the most favorable hydrogen–bonding subsites in the binding site of the protein proved to be superior over grid–based sampling followed by the selective retention of points where ligand atoms could act as hydrogen–bond donors or acceptors. More known ligands could be docked closer to their known crystal structure positions for both thrombin and GST using the knowledge–based method of template and ligand interaction point generation.

Docking experiments usually return multiple docked orientations per ligand. Ideally, the scoring function will indicate the one closest to the crystal structure by giving it the highest score. Also, when a large database is screened, the scoring function should be able to discriminate between promising ligand candidates and artificial hits. Using the assumption that most CSD compounds are unlikely to be ligands of thrombin and of GST, the ability of SLIDE scores and DrugScores to discriminate between known ligands and CSD compounds was tested. The enrichment plots calculated with both scoring methods showed improvement upon replacing the grid–based template with the knowledge–based one, and the separation of scores between ligands and CSD compounds also increased. The reason for this is the ability of SLIDE to dock ligands better with the

knowledge–based method, with better dockings receiving higher scores, whereas the CSD compounds received roughly the same scores using both methods.

Precise computational prediction of the binding affinities of a series of ligands for an arbitrary protein target cannot be routinely achieved by any method at this time. Particular challenges remain in the handling of interfacial solvation and protein and ligand flexibility, so scoring functions perform best when the details of the protein–ligand complex are well–resolved. Thus, docking presents a particularly hard case for scoring, and consensus scoring by combining several scoring functions has been suggested to enhance hit rates (Bissantz et al., 2000; Charifson et al., 1999; Stahl and Rarey, 2001). To compensate for the shortcomings of using a single scoring function, a second, independent scoring function, DrugScore, was also used to score the ligands docked by

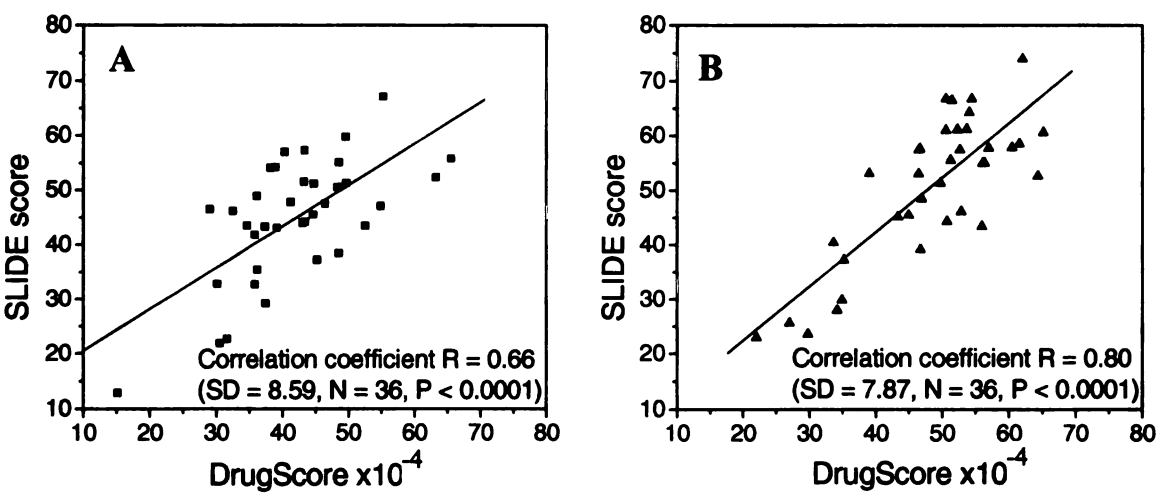


Figure 3.10. Correlation between SLIDE scores and DrugScores of known thrombin ligands with the grid-based (A) and the knowledge-based method (B). The negative DrugScore scores are shown with positive sign for ease of comparison, so that correlation rather than anticorrelation between DrugScores and SLIDE scores is measured.

SLIDE. For thrombin and GST, the two scoring functions showed similar results: increased screening enrichment for known ligands, due to better separation of the ligands from CSD compounds. The correlation between the SLIDE scores and the DrugScores of known ligands also increased (Figure 3.10). This could be due to both scoring functions being trained on correctly positioned ligands from known protein–ligand complexes. They both perform quite well when the ligand is docked correctly, but may show less consistent performance on slightly misdocked molecules. In fact, our analysis on the relationship between RMSD and score (unpublished results) indicates that as a ligand is shifted from its optimal position, the correlation between RMSD and score is quickly lost. Once the ligand is slightly misdocked (say, due to a 1.5 Å shift from its optimal position), its score may be indistinguishable from a that of a poor docking due to misalignment of key hydrogen bonds and hydrophobic interactions. Thus, the score may not suggest that the docking is close to being correct. This problem would be difficult to solve by focusing on improving the scoring function, since even a perfect scoring function would be quite sensitive to a 1.5 Å shift between the interacting protein and ligand groups. However, this problem *can* be addressed by improving the sampling of orientational space and the modeling of flexibility in docking. Better sampling and flexibility modeling result in testing more accurate dockings, increasing the probability that the correct interactions between protein and ligand will be measured and result in high scores. The SLIDE scoring function and flexibility modeling remained the same in versions 1 and 2. Therefore, the improvements in the sampling and representation of protein and ligand chemistry alone account for the significant improvements observed in the scores and docking RMSD values with the new version of SLIDE (see Figures 3.6–3.9).

Both SLIDE score and DrugScore performed significantly better using the knowledge–based protein representation than with the original grid–based template. Regularizing the sampling of hydrophobic interaction points on ligands (another change in version 2 of SLIDE, relative to version 1) also resulted in docking and scoring improvements. One explanation for the observed improvements in scoring could be that neither scoring method was optimized to work with a grid–based template, in which the distances measured between interacting atoms could be non–optimal due to rounding off to the nearest grid point. However, this brings up the important point that the protein template and ligand interaction points in SLIDE are used only for the initial docking of the ligand, whereas scoring by either method is done using the full–atom representation of the ligand docked to the protein, after flexibility modeling (and without reference to the template or interaction points). Thus, *improving the quality of the initial docking,* through improving the representation of the protein and ligand, is what results in the significant improvements in docking accuracy and scoring observed here. These improvements are apparently independent of the scoring function used (DrugScore and SLIDE score were developed using different paradigms, as discussed below) or on the particulars of the protein and its ligands (thrombin and GST are structurally and chemically quite different).

We have no definitive explanation for why SLIDE score and DrugScore results are apparently so correlated for the thrombin ligands (R = 0.80; Figure 3.10.B). DrugScore is derived from the extent to which a given protein–ligand complex shows favored distances between the protein and ligand atoms. Favorability is gauged from pair–wise atomic distance distributions derived from a large set of protein–ligand

complexes from the Protein Data Bank. The SLIDE scoring function is a weighted sum of two terms. The first measures hydrophobic complementarity, calculated as the complementarity in atomic hydrophobicity values of atoms in the ligand with protein atoms that are within a certain radius. This radius was chosen to include the first shell of protein atoms within van der Waals contact of the ligand atom. The atomic hydrophobicity values came from a prior study of the tendency of protein surface atoms to bind water molecules in crystallographic structures (Kuhn et al., 1995). The second term in the SLIDE scoring function, counting intermolecular hydrogen bonds, is based on others' studies of the favored geometries of hydrogen bonds involving protein atoms. Despite counting interactions somewhat differently, SLIDE score and DrugScore are both based on knowledge derived from the geometry of interactions within protein crystallographic structures. This may be the fundamental basis for the observed correlation in their values for the thrombin complexes.

3.5.2 The Role of Flexibility in Docking to Thrombin and GST

Modeling protein flexibility is also very important to accurate docking. Often, validation studies test redocking, in which the ligand is removed from the co–crystal structure, and the separated protein and ligand structures are used to test the docking program's ability to identify the correct ligand binding orientation in the protein. In that case, the protein is guaranteed to be in the correct conformation for the ligand. This simplifies the docking problem, such that only orientational sampling for the ligand is needed. It also assumes that the correct protein conformation is known for that ligand, which is not true when predicting a protein–ligand complex or designing a new ligand. Only 9 of the 42

thrombin ligands could be docked into the *apo* structure without conformational change in the protein or ligand (data not shown), whereas with SLIDE flexibility modeling of the protein and ligand, 36 of 42 (86%) of the ligands could be docked. For GST, 93% of ligands could be docked with flexibility modeling, but only 60% without. Thus, SLIDE models flexibility appropriately, allowing correct docking of the majority (~90%) of thrombin and GST ligands, as well as discriminating well between ligands and non–ligands in screening. Without protein flexibility modeling, for most ligands docking requires using the pre–conformed protein structure for that ligand, or forcing unnatural, additional flexibility within the ligand.

3.5.3 Previous Docking and Screening Validation Studies on Thrombin and GST

A number of groups have done docking and screening method validations on thrombin (Baxter et al., 2000; Fox and Haaksma, 2000; Fradera et al., 2000; Jones et al., 1997; Knegtel et al., 1999; Kramer et al., 1999; Murray et al., 1999; Sotriffer et al., 2002; Stahl and Rarey, 2001), with a focus on how the docking and scoring methods affect the results. In particular, Stahl and Rarey (Stahl and Rarey, 2001) present a detailed analysis of four different scoring functions in combination with the docking tool FlexX, using thrombin as one of their targets. Depending on the scoring function used, 20–70% of the 67 known thrombin ligands are among the top ranking 10% of their screening database of about 10000 compounds. This percentage improves to 80% when using a combined scoring function. Baxter et al. (Baxter et al., 2000) test the docking accuracy of PRO_LEAD on 70 protein–ligand complexes including 6 thrombin structures, resulting

in 79% of the ligands being docked within 2.0 Å RMSD. This program also provides a reasonable separation between the docked scores of the 43 known thrombin ligands and 10000 random molecules from the screening database, with 84% of the known ligands ranking among the top scoring 10% of docked molecules. Knegtel et al. (Knegtel et al., 1999) compare the performance of DOCK 4.0 and FlexX 1.5 by docking 32 known ligands to thrombin. For ~40% of the ligands, fully flexible docking yields orientations within 2 Å of the known binding modes. This increased ligand conformational sampling in DOCK is found to be comparable to rigid docking of about 800 conformers per ligand and increases the docking accuracy somewhat, at the expense of an additional 20 minutes' run time per compound. In another study, Knegtel et al. (Knegtel and Wagener, 1999) use DOCK 4.0 to identify thrombin inhibitors from a database of 32 known inhibitors, ten chemically similar but inactive compounds, and 1000 corporate database compounds. The performance is again scoring–function dependent, with 78–94% of actives being ranked among the 10% best scoring molecules, but neither scoring function gave a good differentiation between actives and inactives among the top scoring compounds. In the results presented here, SLIDE screening on the ~15,000 molecules of the combined thrombin ligand and random CSD compound database identified 64–67% of thrombin ligands (depending on whether SLIDE score or DrugScore was used as the metric) within the top 0.7% of screened compounds. The runtime was about 17 hours for this screening. Although the runtime is determined primarily by the template size, other factors like ligand size and number of rotatable single bonds are also influential. While it is risky to compare methods using different ligand database sizes and degrees of

molecular diversity (as described above), these results give some idea of the state of the art for molecular screening and docking of ligands for thrombin and GST.

Other groups have also investigated the influence of protein or ligand representation on docking results. Fradera et al. (Fradera et al., 2000) test two ligand similarity–driven flexible docking approaches by modifying DOCK 4.0 to include the molecular–field matching program MIMIC (Mestres et al., 1997). The modified methods outperform DOCK by improving the quality of the 31 thrombin ligand dockings by 1 Å RMSD on average and by identifying 1.5–2 times more active molecules among the top–ranked 10% of molecules, for each of the three screening databases used. Their results with MIMIC/DOCK tend to be better than results of DOCK alone and take far less time, but prove to be rather dependent on the choice of the reference ligand. Fox and Haaksma (Fox and Haaksma, 2000) test their approach of combining GRID (Boobbyer et al., 1989) to map the binding site of thrombin and UNITY (TRIPOS, Inc.) to do a flexible 3D database search for benzamidine–based thrombin inhibitors, using a database of in–house thrombin inhibitors and a subset of ACD compounds. The method provides accurate docking orientations for 90% of the x–ray conformations of the known inhibitors, although the docking accuracy drops considerably in the case of CORINA–generated conformers (Sadowski and Gasteiger, 1993).

Glutathione S–transferase has been less widely studied as a docking and screening target, although it has been included in some larger docking validations (Chen and Ung, 2001; Jones et al., 1997; Kramer et al., 1999). There are at least 11 different GST isozymes with different substrate specificities, which complicates the comparisons. Koehler et al. (Koehler et al., 1997) use an interesting approach to decipher the key

determinants of GST isozyme selectivity. Based on finding that glutathione (GSH) binds to all isozymes in a single bioactive conformation, they superimpose the available GST x–ray structures from the PDB using the bound ligands rather than the protein backbones to compare their binding sites. Their conclusion that the shape and surface hydrophobicity of the binding site are the key determinants of differences in ligand specificity between GST isozymes can be exploited in finding new, more isozyme–specific inhibitors by virtual screening. Such isozyme–specific differences would appear directly in SLIDE's knowledge–based protein templates for different GST isozymes, providing a convenient way to screen for ligands that bind well to one template/isozyme but not another.

3.6 Conclusions

Our results show that improving the representation of hydrogen–bonding and hydrophobic interaction points on the ligand and protein by a knowledge–based approach, as implemented in SLIDE, can significantly improve both the quality of docking and the docking scores of known ligands relative to randomly–selected molecules. The resulting unbiased protein template can also provide significant insights into the binding and specificity determinants of the protein, and thus provide a structure–based design template for optimizing ligand functional groups.

Chapter 4

Side-Chain Flexibility in Docking with SLIDE: Testing the *Minimal Rotation Hypothesis*

4.1 Introduction

It is widely accepted that flexibility is indispensable for protein function. The questions are: how much flexibility is needed, in general, for protein-ligand interactions, and how does this flexibility partition between the protein and its ligand? The high computational cost of handling both the ligand and the protein as totally flexible entities requires compromises in modeling protein-ligand recognition, namely including only a certain degree of flexibility in the docking process to maintain a reasonable computational time. The first docking tools, the most widely known of them being DOCK (Kuntz et al., 1982), were designed based on the key-and-lock mechanism of protein-ligand recognition, handling both the ligand and the protein as rigid bodies. Superior to the rigid body docking are the methods holding the protein rigid while allowing ligand flexibility

(Burkhard et al., 1998; Ewing et al., 2001; Goodsell et al., 1996; Kramer et al., 1999; Taylor and Burnett, 2000). DOCK has evolved to become more realistic, too, by handling ligands totally flexibly in its latest 4.0 version (Ewing et al., 2001). The rationale behind this treatment is that ligands are usually much smaller than the protein, so it is computationally less expensive to handle them flexibly. On the other hand, studies of conformational changes accompanying protein-protein (Betts and Sternberg, 1999) and protein-ligand (Najmanovich et al., 2000) associations show that even in the case of proteins with conserved main-chain conformations across crystallographic complexes with various ligands, there are significant side-chain conformational changes in at least 60% of the cases upon ligand binding. These studies provide a conservative estimate of side chain flexibility involved in the recognition process, since side chain conformations were considered to be different only if they were in different low energy states also called rotamers. Nevertheless, they point toward the necessity of also modeling protein flexibility in docking.

Docking and screening tools reach various levels of sophistication trying to achieve this goal. Soft docking (Jiang and Kim, 1991) handles protein flexibility implicitly by allowing a certain degree of interpenetration between the protein and the docked ligand, making the reasonable assumption that the exactly correct conformers of the protein and ligand are not sampled. The docking tool GOLD (Jones et al., 1997) allows rotation of terminal hydrogen atoms on the proteins to optimize fit and hydrogen bonding. The next level of sophistication is reached by using rotamer libraries (Dunbrack, Jr. and Karplus, 1993; Lovell et al., 2000; Tuffery et al., 1991) to sample the low energy conformations available to each side chain while optimizing the shape complementarity between the

protein and the docked ligand (Kallblad and Dean, 2003; Leach, 1994; Leach and Lemon, 1998). Schaffer and Verkhiver (Schaffer and Verkhivker, 1998) improve the rotameric side-chain conformations with an optimization procedure using the dead-end elimination algorithm following the docking which could be an affordable method for fine docking but not for screening. Assuming that crystal structures show the protein side-chains in favorable conformations, an alternative approach to sample the available side-chain conformational space is the use of side-chain conformers from multiple x-ray structures (Claussen et al., 2001; Knegtel et al., 1997). A similar approach was taken by the group of Goodsell to account for protein side-chain motions by combining multiple target structure within a single grid-based look-up table of interaction energies for docking with AutoDock (Osterberg et al., 2002).

SLIDE models flexibility by allowing protein side-chain rotations and full ligand flexibility, assuming that both the protein and the ligand change their unbound conformation as little as necessary to result in an overlap-free docked orientation of the ligand in the protein binding site (Schnecke and Kuhn, 1999; Schnecke and Kuhn, 2000). This hypothesis was tested on a number of different proteins which do not undergo major main-chain conformational change upon ligand binding but show alternative side-chain positions in crystal complexes with different ligands.

4.2 Methods

To examine whether or not side-chain flexibility as modeled by SLIDE is necessary for successful docking, a set of known ligands were docked into the unliganded, apo structure of thrombin (PDB code 1vr1) both by rigid and flexible docking. The two

approaches were evaluated by comparing the number of successful dockings retrieved with and without flexibility, where a successful docking was defined as one with a root mean square deviation (RMSD) of 2.5 Å or less from the crystal structure orientation.

To evaluate the realism of induced fit modeling by SLIDE, the side-chain rotations produced by SLIDE upon docking known ligands into the apo structures of their target proteins were compared to the dihedral-angle differences calculated between corresponding ligand-free and ligand-bound x-ray structures of the proteins. A set of 35 human thrombin (Table 4.1) and 14 human glutathione S-transferase (Table 4.2) crystallographic complexes with known ligands were used in addition to the ligand-free (in the active site) structure of thrombin (PDB code 1vr1) and of GST (PDB code 16gs). In order to avoid a possible bias that could arise from studying only one or two cases, and to ensure the validity of the conclusions across a wide range of proteins, a dataset of 18 ligand-free protein structures with corresponding ligand-bound complexes was also assembled (Table 4.3). Only structures with resolutions of 2.5 Å or better were used. Since this study focused on modeling side-chain flexibility in systems with no significant backbone changes following ligand binding, only ligand-bound and ligand-free protein pairs with backbone superposition RMSD values of $\leq$ 0.5 Å, and pair-wise backbone atom positional deviations of $\leq$ 1 Å were used. To exclude possible errors in determining side chain positions, only protein-ligand crystal complexes with resolution of 2.5 Å or better were included in the analysis.

Table 4.1. Thrombin crystallographic complexes used in testing the minimal rotation hypothesis. The names of the ligands are listed in Table 3.1. Some of the best RMSD values for the dockings differ from those listed in Table 3.1 because a different SLIDE run with slightly different parameter values was performed for this study. The parameter set used in this study was found to allow more known ligands to be correctly docked without increasing the computational time considerably. All the SLIDE parameter values used in the flexible and rigid docking were the identical, except for the number of allowed side-chain rotations, which was set to zero in the rigid docking run. The "-" sign indicates that the ligand could not be docked with RMSD $\leq$ 2.5 Å using these parameters; note that many more ligands could not be docked with rigid docking than with flexible docking.

Table 4.1.

#	PDB code	Resolution (Å)	Best RMSD (Å)	
			Flexible docking	Rigid docking
1	1a2c	2.1	-	-
2	1a3b	1.8	0.30	0.76
3	1a3e	1.9	-	-
4	1a46	2.1	0.35	0.88
5	1a4w	1.8	0.52	-
6	1a5g	2.1	0.97	0.97
7	1a61	2.2	0.96	-
8	1ad8	2.0	0.78	-
9	1ae8	2.0	0.65	0.31
10	1afe	2.0	1.05	-
11	1aht	1.6	0.81	1.11
12	1ai8	1.9	-	-
13	1aix	2.1	1.40	-
14	1awf	2.2	2.11	-
15	1ay6	1.8	0.71	-
16	1b5g	2.1	0.40	-
17	1ba8	1.8	0.51	-
18	1bb0	2.1	0.65	0.65
19	1bcu	2.0	2.16	2.16
20	1bhx	2.3	0.53	0.78
21	1fpc	2.3	0.90	-
22	1lhc	2.0	0.67	0.75
23	1lhd	2.3	0.74	-
24	1lhe	2.2	0.71	0.74
25	1lhg	2.2	1.30	-
26	1nrs	2.4	0.75	-
27	1ppb	1.9	0.71	-
28	1tbz	2.3	1.10	-
29	1tmb	2.3	1.00	-
30	1tmt	2.2	0.54	0.77
31	1tom	1.8	0.74	-
32	1uma	2.0	0.94	0.94
33	3hat	2.5	0.89	-
34	7kme	2.1	0.38	-
35	8kme	2.1	0.79	1.09

Table 4.2. GST crystallographic complexes used in testing the minimal rotation hypothesis. The names of the ligands are listed in Table 3.2.

#	PDB code	Resolution (Å)	Best RMSD (Å)
1	10gs	2.2	0.36
2	12gs	2.1	0.36
3	13gs *	1.9	1.78
4	18gs	1.9	0.64
5	1aqv	1.9	0.44
6	1aqw	1.8	0.46
7	1aqx	2.0	0.78
8	1pgt	1.8	0.53
9	20gs *	2.5	0.52
10	21gs *	1.9	4.21
11	2gss *	1.9	2.25
12	2pgt	1.9	0.54
13	3gss	1.9	0.52
14	3pgt	2.1	0.59

* Ligands that are mainly hydrophobic in character and bind to the hydrophobic subsite of GST. These ligands were docked in a second run, when hydrophobic template points from their respective binding subsite were selected as key points. In docking the other ligands for GST, which bind in the glutathione site, hydrogen bonding template points were selected as key points. Using key points is a convenient way to reduce the number of docked orientations by keeping only those that bind in the correct region within the active site.

Table 4.3. Ligand-free structures and their corresponding ligand-bound complexes used in testing the minimal rotation hypothesis.

PDB code		Protein/Ligand Complex	Resolution (Å)	Best RMSD (Å)	Template Size
Free	Bound				
1ahc	1ahb	alpha-momorcharin/formycin 5'-monophosphate	2.0/2.2	0.94	88
1ajz	1aj2	dihydropteroate synthase/dihydropterine-diphosphate	2.0/2.0	0.75	79
3cox	1coy	cholesterol oxydase/3-beta-hydroxy-5-androsten-17-one	1.8/1.8	1.61	74
1gmq	1gmr	RNase SA/guanosine-2'-monophosphate	1.8/1.8	1.28	87
3grs	1gra	glutathione reductase/glutathione disulfide	1.5/2.0	0.69	139
1kem	1kel	catalytic antibody 28B4 FAB fragment /AAH*	2.2/1.9	0.46	74
2hvm	1llo	hevamine(endochitinase)/N-acetyl-D-allosamine	1.8/1.9	0.67	150
1nsb	1nsc	neuraminidase/N-acetyl neuraminic acid(sialic acid)	2.2/1.7	0.40	74
1swa	1swd	streptavidin/biotin	2.0/1.9	0.62	37
2ptn	1tps	trypsin/inhibitor A90720A	1.5/1.9	0.93	143
1xib	1xid	D-xylose isomerase/L-ascorbic acid	1.6/1.7	2.28	45
1ydc	1ydb	carbonic anhydrase II/acetazolamide	2.0/1.9	1.42	50
2chs	2cht	chorismate mutase/endo-oxabicyclic inhibitor	1.9/2.2	1.02	39
2apr	3apr	acid proteinase/reduced peptide inhibitor	1.8/1.8	0.54	153
1tli	3tmn	thermolysin/Val-Trp	2.0/1.7	0.99	75
2ctv	5cna	concanavalin A/alpha-methyl-D-mannopyranoside	2.0/2.0	1.99	57
2sga	5sga	proteinase A/tetrapeptide Ace-Pro-Ala-Pro-Tyr	1.5/1.8	0.59	126
6taa	7taa	fam. 13 alpha amylase/modified acarbose hexasaccharide	2.1/2.0	0.82	133

* AAH = 1-[N-4'-nitrobenzyl-N-4'-carboxybutylaminomethylphosphonic acid

4.3 Results

Thrombin

The template describing the binding site of ligand-free thrombin consisted of 139 points, with 24 of these points assigned as key points. Key points were selected as template points at a distance of 6.5 Å or less from the CG side-chain carbon atom of Asp189 from the specificity pocket of thrombin (Figure 4.1). From the set of 35 known thrombin ligands, only 13 could be successfully docked (RMSD $\leq$ 2.5 Å) with rigid docking, while 32 could be docked when the protein side chains and ligand were considered flexible (Table 4.1). Most of the side chain rotations performed by SLIDE upon docking these 32 known ligands to thrombin are small (Figure 4.2). As many as 58% of these rotations are 15° or less, and 90% of them are 45° or less (Figure 4.3.A). The dihedral angle differences between protein side chains from the ligand-free and ligand-bound crystal structures of these ligands have a very similar distribution (Figure 4.3.B), with 66% of all dihedral angle differences being 15° or less, and 84% of the differences being 45° or less. Thus, SLIDE is making appropriate magnitudes of rotations for active-site side chains upon ligand binding.

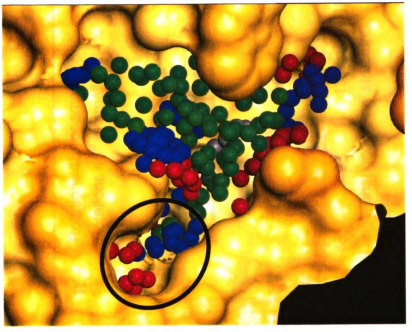


Figure 4.1. The active site of thrombin filled with template points colored according to type: blue for donor, red for acceptor, white for donor/acceptor, green for hydrophobic. The template points from the bottom of the S1 specificity pocket (circled in figure) were marked as key points, meaning that each docked ligand must match at least one of these points.

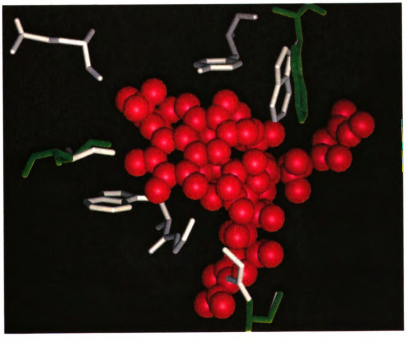


Figure 4.2. Side chains rotated by SLIDE (shown in green) in the active site of thrombin upon docking a known ligand (red spheres). The original positions of the side chains in the ligand-free crystal structure (1vr1) are shown in white.

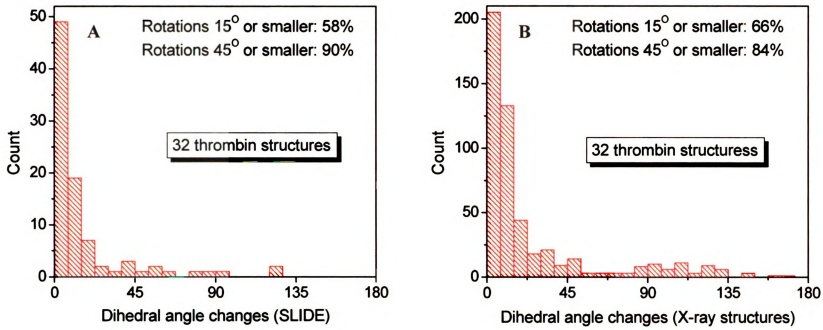


Figure 4.3. Side-chain rotations performed by SLIDE (A) upon docking 32 known ligands into the ligand-free active site of thrombin (PDB code 1vr1), compared to (B) the dihedral angle differences observed between ligand-free and ligand-bound crystal structures.

GST

The template representing the binding site of GST consisted of 120 template points, with 25 key points (16 hydrogen bonding key points used for ligands binding in the glutathione site and 9 hydrophobic key points used for ligands binding in the hydrophobic site), as described in the Methods of Chapter 3. Of the 14 known GST ligands (Table 4.2), 13 were docked successfully into the binding site of the ligand-free crystal structure (PDB code 16gs). The side chain rotations performed by SLIDE for the 13 successful dockings are shown in Figure 4.4.A. Only 4% of the angles rotated by SLIDE were larger than 45°, with 89% of them being smaller than 15°. This result was very similar to the crystal structure dihedral angle differences of the side chains from the binding site of the ligand-free protein and corresponding ligand-bound complexes (Figure 4.4.B), where 85% of the angle differences were 45° or smaller and 96% of them were 15° or smaller.

Eighteen Pairs of Ligand-free and Ligand-bound Proteins

The templates used to represent the binding sites of this diverse set of proteins varied in size from 37 to 153 points. No key points were assigned for these cases. Similarly to thrombin and GST, most side chains (92%) from the binding sites of the apo structures were rotated by SLIDE with 45° or less, with 69% or the rotations being smaller than 15° (Figure 4.5.A). The distribution of the SLIDE-performed side-chain rotations was found to be very similar to the distribution of dihedral-angle differences observed between the apo and ligand-bound crystal structures (Figure 4.5.B), out of which 94% were 45° or smaller and 83% were 15° or smaller.

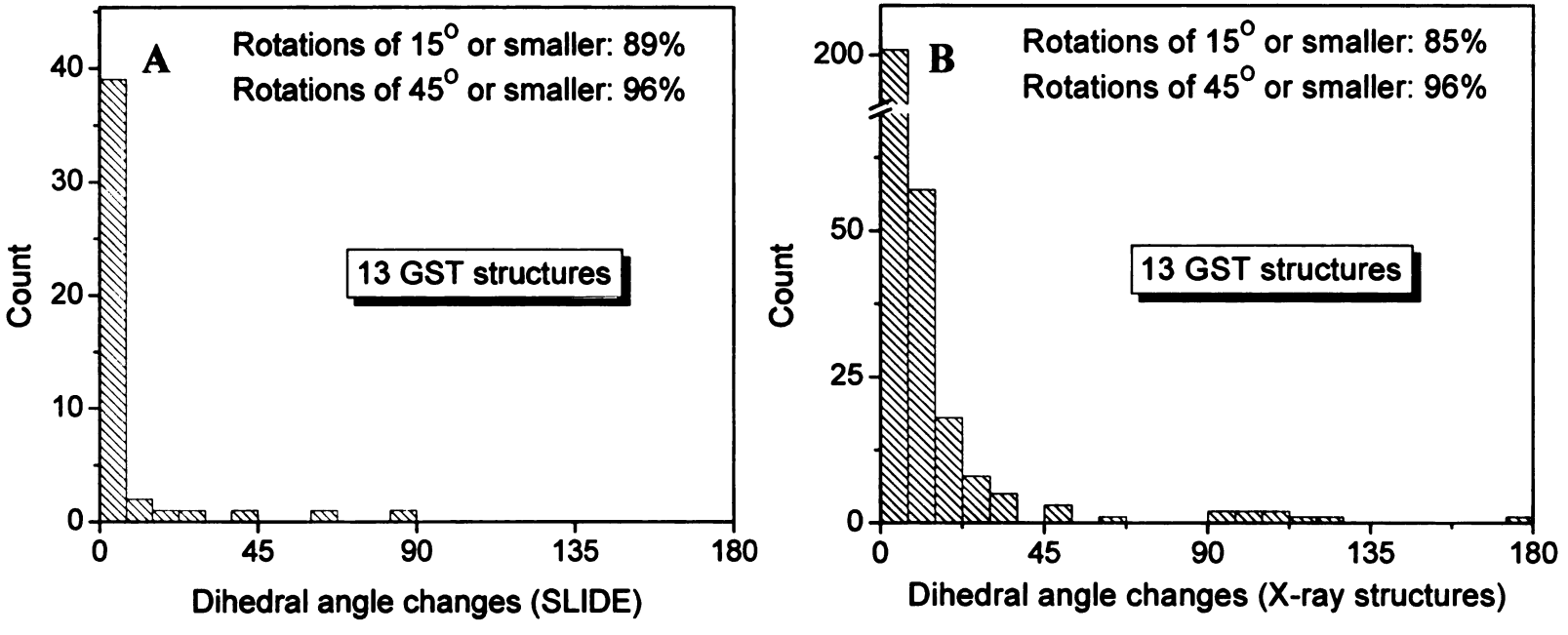


Figure 4.4. Side-chain rotations performed by SLIDE (A) upon docking 13 known ligands into the ligand-free active site of GST (PDB code 16gs), compared to (B) the dihedral angle differences observed between ligand-free crystal structure and the corresponding side chains from the ligand-bound structures.

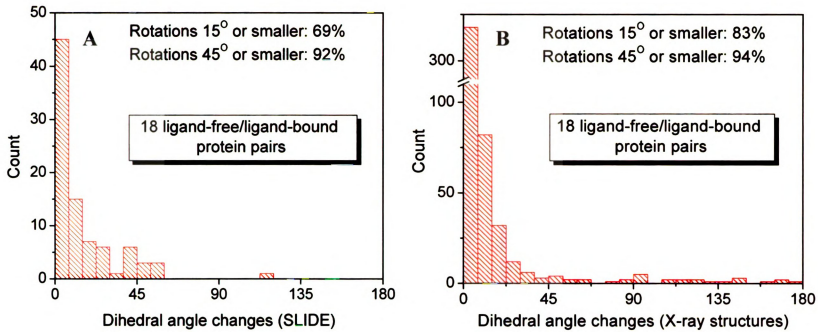


Figure 4.5. Side-chain rotations performed by SLIDE (A) upon docking 18 known ligands into the corresponding ligand-free target structures, compared to (B) the dihedral angle differences between corresponding side chains from the binding sites of ligand-free and ligand-bound structures.

4.4 Discussion

When measuring side-chain dihedral angle differences of ligand-bound and ligand-free proteins, only protein side-chains in direct contact with the ligand in the ligand-bound were taken into account. This was done to ensure that only ligand-induced changes were considered. There is a very good qualitative agreement between the pattern of side-chain rotations that occur upon ligand binding provided by SLIDE and the picture that emerges from comparing ligand-free and ligand-bound protein structures. On average, about 85-90% of side chain rotations are smaller than 45°. Studies of ligand-induced changes in side-chain conformations in protein binding sites usually count only differences larger than 45°, or even 60 or 75° (Betts and Sternberg, 1999; Najmanovich et al., 2000), that would correspond to changes in rotameric states of the side chains. Heringa and Argos on the other hand, observed that ligand binding induces non-rotamericity in the preferred side-chain conformations (Heringa and Argos, 1999). The model of protein side-chain flexibility implemented in SLIDE provided results that are in good agreement with these latter observations. Not only were most of the rotations too small to allow changes in rotameric states, but they were necessary to correctly dock the ligand in about 60% of the thrombin ligands. Even when docking could be achieved with a rigid protein structure, the resulting docked orientation was farther from the correct position in many cases, compared to the orientation resulting from flexible docking.

Comparing the A and B panels of the plots showing the distributions of side chain rotations (Figures 4.3, 4.4, and 4.5) it is noticeable that SLIDE is somewhat more parsimonious than nature by producing a smaller number of rotations in the protein upon binding the ligand. One reason for this is that the larger the number of side chains

allowed to be rotated by SLIDE, the larger the possibility of creating new intramolecular overlaps in the protein. This would ultimately lead to an increase in computational time, limiting the usefulness of the program in screening large databases. Another reason for the above mentioned quantitative discrepancy could be that SLIDE does move side chains away to resolve collisions but does not move them toward the ligand to make new interactions. This is a future improvement to be implemented in SLIDE.

4.5 Conclusions

The assumption that both protein side chains and ligands move as little as necessary in order to achieve a collision-free complex proved to be both reasonable and sufficient to dock most of the known ligands into the binding sites of their target proteins for the systems tested in this study. The results of the ligand-free and ligand-bound crystal structure comparisons underscore that side chain conformational changes are typically not rotameric, but instead involve modest (<15°) changes in side-chain angles.

Chapter 5

Using SLIDE to Find New Ligands for Thrombin

5.1 Introduction

Increase in efficiency and reliability of computational tools has enabled virtual screening to become a valuable method in the pharmaceutical drug discovery process, complementing high-throughput screening (Good, 2001; Schneider and Bohm, 2002; Shoichet et al., 2002; Waszkowycz, 2002). Novel inhibitors have been identified for thrombin (Fox and Haaksma, 2000; Massova et al., 1998), protein tyrosine phosphatase-1B (Doman et al., 2002), various nuclear hormone receptors (Schapira et al., 2000, 2001), human carbonic anhydrase (Gruneberg et al., 2002), and thymidylate synthase (Shoichet et al., 1993) by *in silico* screening of compound databases.

As described in the previous chapters, SLIDE is a computational tool which can efficiently screen databases of hundreds of thousands of molecules to identify feasible ligand candidates for a target protein with known three dimensional structure (Schnecke

and Kuhn, 1999; Schnecke and Kuhn, 2000). The realistic modeling of protein side-chain and ligand flexibility, combined with the improved representation of the binding site by knowledge-based template design has allowed a better discrimination between true ligands and non-specific compounds (Zavodszky et al., 2003). Since experimental testing is a useful complement to modeling, a screening experiment was designed to test the predictive power of SLIDE. After screening the Available Chemicals Directory (ACD; MDL Information Systems, Inc.) to identify new ligands for thrombin, binding affinities were measured for the top scoring candidates using isothermal titration calorimetry (ITC).

The Target: Thrombin

Thrombin is a key player in the blood coagulation cascade: it catalyzes the proteolytic cleavage of the soluble plasma protein fibrinogen to produce fibrin. The linear fibrin monomers are then cross-linked by factor XIII, producing insoluble blood clots. Factor XIII is a transglutaminase, the last enzyme of the coagulation cascade, which is itself activated by thrombin. Thrombin is also a potent platelet activator. Activated platelets adhere to the site of vascular injury, aggregate, and form a plug to reduce blood loss. The coagulant activity of thrombin is kept under control by thrombomodulin, a thrombin binding protein on the surface of endothelial cells. When too much thrombin is generated, thrombomodulin binds to thrombin, dramatically altering its specificity. The complex rapidly cleaves the protein C zymogen to form the anticoagulant, activated protein C. Complex formation between thrombin and thrombomodulin also prevents thrombin from cleaving fibrinogen. Numerous efforts to control the blood clotting process are directed

toward thrombin because of its pivotal role in maintaining the intricate balance between hemostasis and thrombolysis (Davie et al., 1991; Esmon, 1995).

Thrombin is a trypsin-like serine protease with the characteristic Ser-His-Asp catalytic triad at the active site. Its specificity is also similar to that of trypsin, preferentially binding substrates with Lys or Arg residues in its specificity pocket. Two additional binding sites (the fibrinogen binding exosite and the heparin binding site), and the ability to use different combinations of these elements allow thrombin to play a key role in a variety of blood coagulation related processes (Tulinsky, 1996). Biochemical modeling studies are greatly aided by the extent of structural data on thrombin that has become available during the last few years (Stubbs and Bode, 1993; Stubbs and Bode, 1995).

5.2 Methods

5.2.1 Screening the ACD with SLIDE

SLIDE (described in more detail in Chapter 1, section 1.4 and Chapter 3, section 3.3) was used to screen the Available Chemicals Directory (ACD) to identify new ligands for thrombin. After eliminating compounds with fewer than 6 or greater than 200 non-hydrogen atoms, the ligand database contained 214,713 small organic molecules.

DrugScore (Gohlke et al., 2000) was used to rescore the top dockings returned by SLIDE. A short description of DrugScore is provided in Chapter 2, section 2.2.

5.2.2 Isothermal Titration Calorimetry

Isothermal Titration Calorimetry (ITC) is a particularly suitable technique to follow the energetics of an association reaction between macromolecules (Jelesarov and Bosshard, 1999), allowing the measurement of the enthalpy as well as the entropy changes of such interactions. The experiment is performed at a constant temperature by titrating the ligand into the protein solution in the sample cell of the calorimeter. After each step of adding a small aliquot of the ligand, the heat exchange in the sample cell is determined by measuring the electrical power necessary to keep the small temperature difference between the sample cell and the reference cell constant. The integrated heat changes plotted against the molar ratios of the binding reaction show the characteristic sigmoidal curve of the binding reaction. For a single set of identical binding sites, the total heat of the reaction Q can be calculated from the following equation:

$$Q = \frac{1}{2} N[P] \Delta HV \left[1 + \frac{X}{N} + \frac{1}{NK[P]} - \sqrt{\left(1 + \frac{X}{N} + \frac{1}{NK[P]} \right)^2 - \frac{4X}{N}} \right]$$

where N is the number of binding sites, [P] the total protein concentration, ΔH the enthalpy of the binding, V the volume of the calorimetric cell, X the ligand/protein molar ratio, and K the binding constant. Least square fitting of the equation describing the binding process to the experimental data allows determination of the enthalpy of the binding (ΔH), the association constant (K), and the stoichiometry, which reflects the number of ligands binding to one protein molecule (N). The other thermodynamic parameters, the Gibbs free energy (ΔG) and entropy change (ΔS), for the interaction can be calculated from the relationship:

$$\Delta G = -RT\ln K = \Delta H - T\Delta S$$

where R is the universal gas constant, and T is the absolute temperature.

Measurements were carried out using an MCS_ITC instrument from MicroCal (Northampton, Massachusetts). Human α-thrombin (Enzyme Research Laboratories, South Bend, Indiana) was dialyzed overnight at 4°C against TRIS buffer (50mM TRIS, 100 mM NaCl, 0.1% PEG800, pH 7.8) with the buffer changed twice to remove salts and impurities. The third dialyzate was saved and used for making the protein and ligand dilutions. Protein concentrations in the sample cell were in the 0.39 - 0.85 mg/ml range. The ligands were added to the protein solution using a 100μl syringe, with concentrations ranging from 0.4 to 1mM. All the experiments were carried out at 30°C, with both the protein and ligand solutions degassed before measurements. The reference cells contained deionized and degassed water. As a negative control, a buffer-run was performed with each ligand candidate, when the ligand was titrated into the buffer without thrombin. A known thrombin ligand, 4-aminobenzamidine, was used as a positive control. Data analysis was performed with the Origin software supplied with the instrument.

5.3 Results

The ACD screening run to identify potential new ligands for thrombin was completed in approximately two days on a double processor desktop workstation. SLIDE returned 15,474 docked molecules, with an average of two orientations per compound that fit the active site. The top 3000 orientations were rescored with the knowledge–based scoring

function, DrugScore, and these compounds were then ranked according to their consensus score, calculated as the normalized sum of the SLIDE score and DrugScore score:

$$Consensus\_score = \frac{SLIDE\_score}{Max\_SLIDE\_score} + \frac{DrugScore\_score}{Min\_DrugScore\_score}$$

The largest SLIDE score and the smallest DrugScore were used to calculate the normalized scores because SLIDE scores are positive numbers with larger being better, while DrugScore calculates energy-type scores with negative values, where the smaller the value is the better the score. Seven compounds were selected based on their consensus scores (Figure 5.1), excluding closely related molecules and compounds that were difficult to obtain. Molecules that were obviously dyes were also excluded, because they tend to bind to a wide variety of biological macromolecules non-specifically. Eight other compounds were selected based on molecular graphics inspection of shape and chemical complementarity of the ligand with the protein (Figure 5.2). These 15 compounds were chosen to be assayed for binding affinity by ITC. Of the 11 out of 15 compounds that proved to be soluble, two, morelloflavone and new fuchsin, showed micromolar binding affinity to human thrombin and are novel ligands for this protein (Figure 5.3).

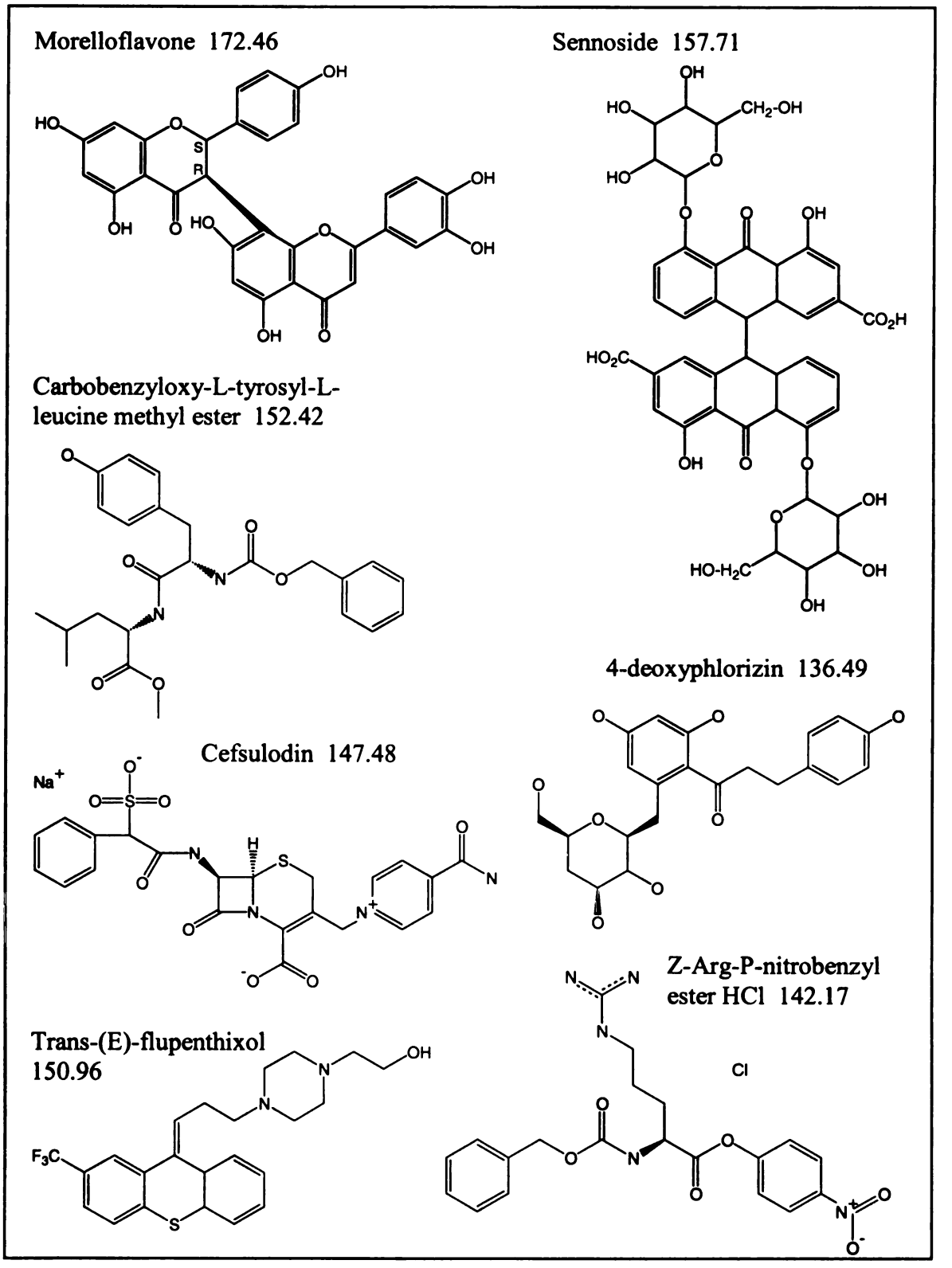


Figure 5.1. ACD compounds selected for testing based on their scores. The numbers next to the ligand names are consensus scores, a normalized sum of SLIDE score and DrugScore, where higher is more favorable.

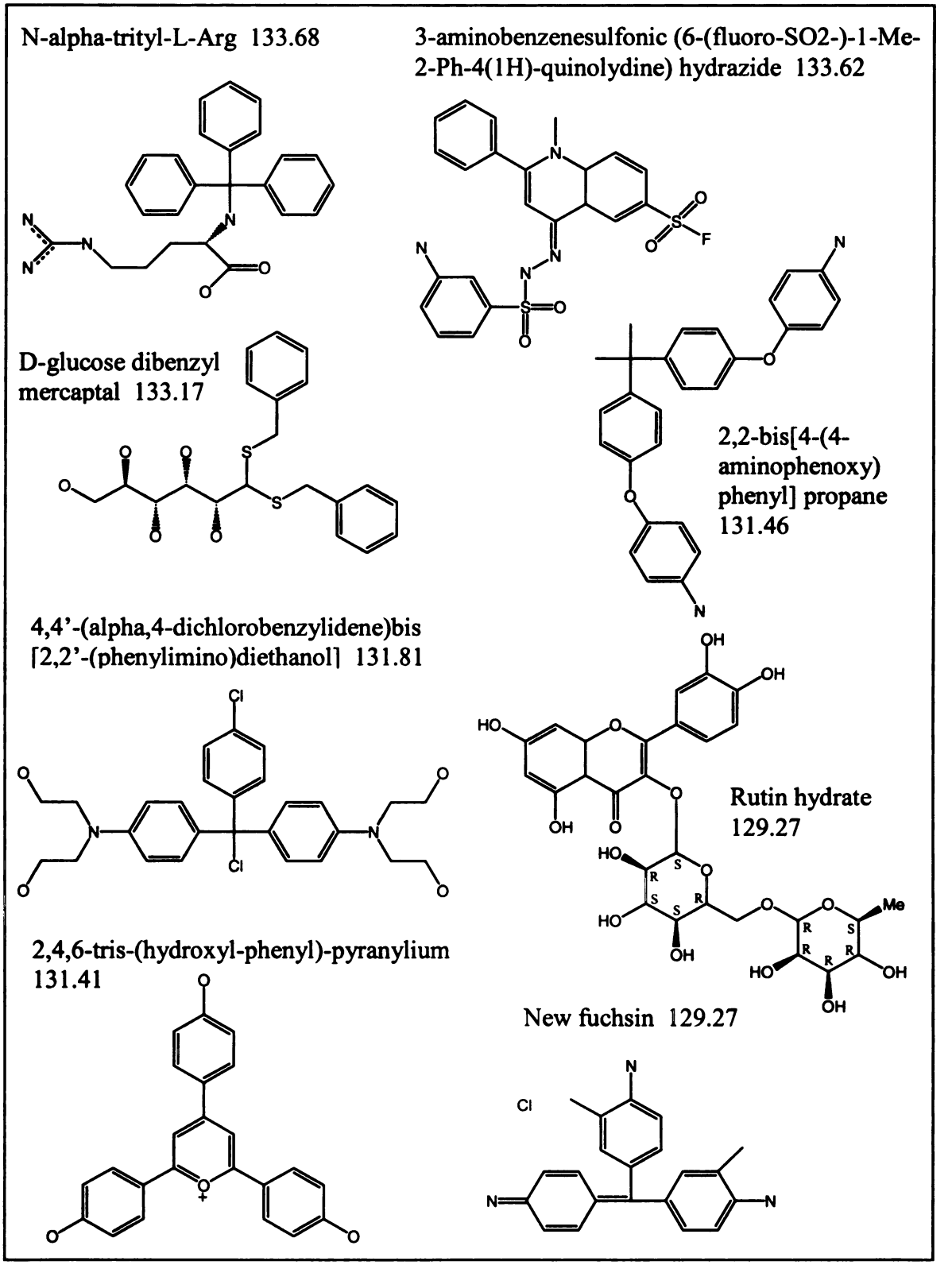


Figure 5.2. ACD compounds selected for testing based on molecular graphics inspection of their docked complexes with thrombin. The numbers next to the ligand names are consensus scores, where higher is more favorable.

Figure 5.3. The recorded heat changes upon successive injections of the ligand into the buffer (negative control, top panel) and the protein solution (middle panel) are shown for each ligand. Morelloflavone and new fuchsin produced larger heat changes (heat of dilution) when injected into the buffer compared to 4–aminobenzamidine, due to the small percentage of DMSO that had to be used to solubilize the first two compounds, while 4–aminobenzamidine was directly soluble in the working buffer. For each ligand, the heat of dilution was subtracted from the corresponding heat of the binding reaction. The integrated heat values plotted against the molar ratios are shown in the lower panels. The red lines represent the least square fitting of the one–binding–site per protein model (N=1) to the experimental data. The parameters calculated from this fitting are the association constant (K, in M^{-1}), the enthalpy (ΔH in cal/mol) and the entropy change (ΔS in cal/mol·K) of the binding reaction. The dissociation constant (K_d in M) is the inverse of K. The shape of the fitted curve depends on the protein concentration, binding constant, and the stoichiometry of the binding reaction.

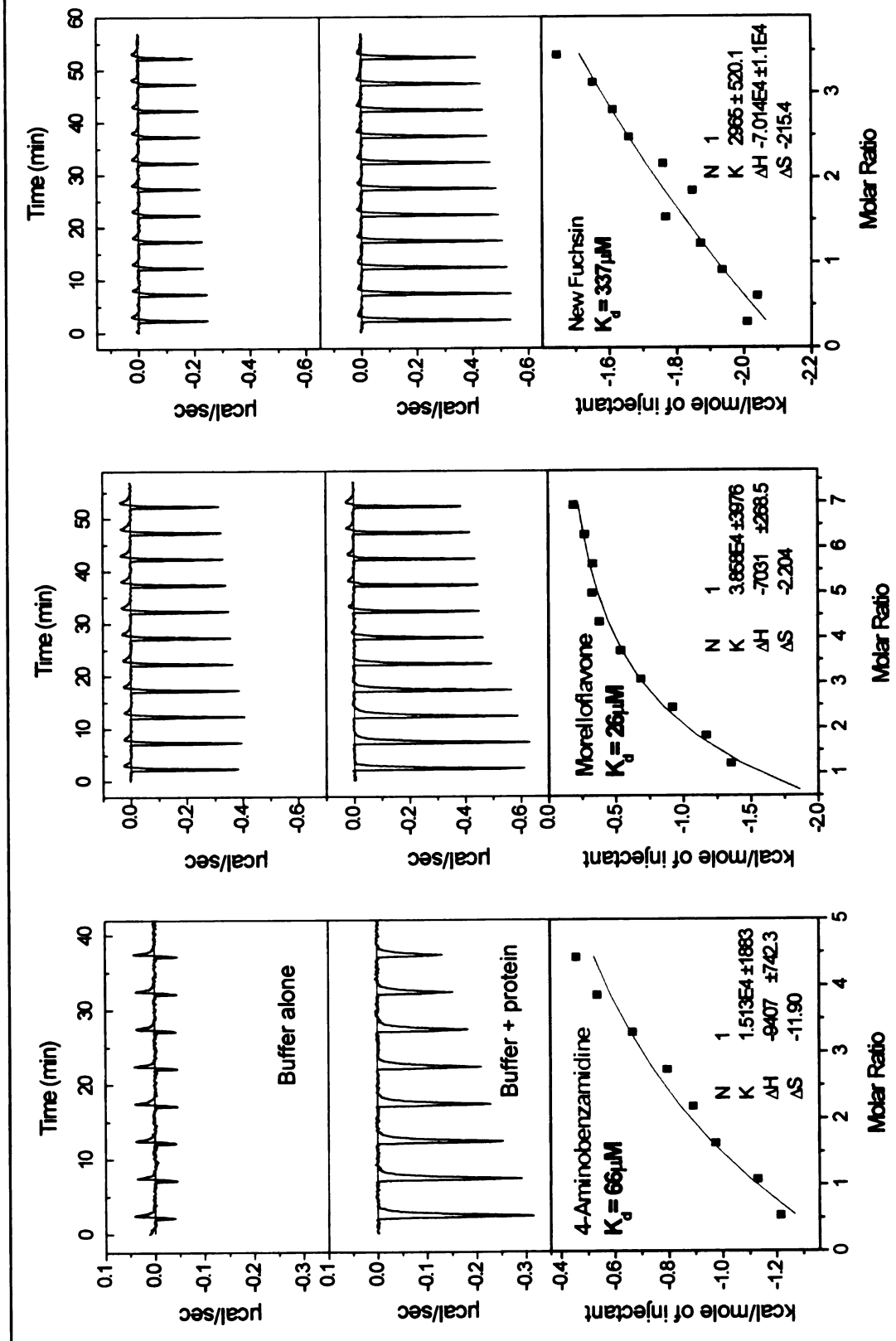

Time (min)
Buffer alone
Buffer + protein
μcal/sec
kcal/mole of injectant
Molar Ratio
4-Aminobenzamidine
K_d = 66μM
N 1
K 1.51354 ±1883
ΔH -9407 ±742.3
ΔS -11.90
Morelloflavone
K_d = 26μM
N 1
K 3.85954 ±3976
ΔH -7031 ±268.5
ΔS -2.204
New Fuchsin
K_d = 337μM
N 1
K 2965 ± 520.1
ΔH -7.014E4 ±1.1E4
ΔS -215.4

Figure 5.3

The predicted binding orientations of these newly identified thrombin ligands mimic the binding modes of known thrombin ligands but have a different molecular scaffold (Figure 5.4).

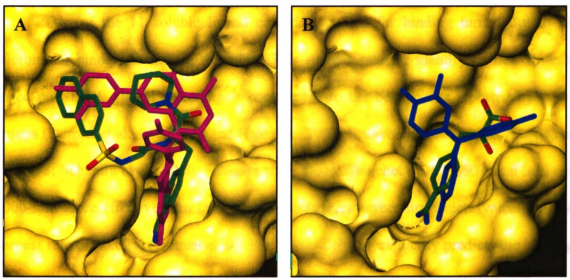


Figure 5.4. Docked orientation of morelloflavone in magenta (A) and new fuchsin in blue (B) in the binding site of thrombin (PDB code 1vr1). The molecules colored by atom types are known thrombin inhibitors from X–ray complexes (PDB codes 1dwd and 1aht, respectively), showing that known thrombin ligands sample similar regions of the binding pocket.

5.4 Discussion

Given the time constraints imposed by the large number of compounds in screening libraries, virtual screening tools can only afford to perform a relatively rough, approximate docking and employ a simple and quick scoring function instead of highly detailed quantum mechanics/molecular mechanics calculations to score the hits. Under

these conditions, finding low affinity binders with novel scaffolds is a realistic expectation from screening random compounds. Micromolar affinity is typical of lead compounds identified by high throughput combinatorial library screening for drug discovery. These leads can then be further modified, with functional groups added or deleted to develop tight and specific inhibitors for the given target protein.

Two of the 11 soluble ligand candidates tested for binding turned out to be micromolar binders to human thrombin. This success rate for identifying new ligands based on SLIDE virtual screening is comparable to the best results reported by other groups (Doman et al., 2002; Fox and Haaksma, 2000; Gruneberg et al., 2002; Massova et al., 1998; Schapira et al., 2000; Schapira et al., 2001; Shoichet et al., 1993), and is about 1000-fold more effective than *in vitro* high-throughput screening, which typically has a success rate of ~0.02% (Doman et al., 2002). SLIDE also explicitly predicts the binding mode between the protein and ligand (Figure 5.4), which will aid in optimizing the new ligands for higher affinity and protein selectivity (e.g., binding to thrombin over other coagulation and digestive serine proteases).

Chapter 6

Modeling Protein Main-Chain Flexibility in Docking

6.1 Introduction

Analysis of conformational changes on complex formation for a representative set of 39 pairs of ligand-free and ligand-bound structures (Betts and Sternberg, 1999) showed that about 50% of the proteins undergo substantial main-chain and side-chain conformational changes when binding the ligand. In another study, focused mainly on evaluating the average number and the type of protein side-chains that undergo major rearrangements upon ligand binding, aside from ubiquitous side-chain movements, Najmanovich and co-workers found backbone displacements larger than 1 Å in 25% of the cases (Najmanovich et al., 2000). This means that in many instances the protein-ligand recognition process cannot be correctly described unless protein main-chain flexibility is taken into account. Excellent reviews have been published recently (Carlson, 2002; Halperin et al., 2002) summarizing the state of the art in flexible docking. Except for

limited cases – simple hinge motions (Sandak et al., 1998), crystallographically determined alternative conformations (Claussen et al., 2001), or small-scale motions typical of molecular dynamics simulations (Carlson et al., 1999; Lin et al., 2002) – main-chain flexibility has not been considered in docking. The various approaches to model side-chain flexibility published in the literature are summarized in Chapter 4, followed by the analysis of how induced fit is modeled for side-chains in SLIDE. This chapter introduces a new and generally applicable method including main-chain flexibility in modeling protein-ligand recognition.

Inducing changes in the protein main chain while performing docking is too expensive computationally, so efforts are directed toward generating a representative conformational ensemble of the protein and using this set as targets for the docking instead of a single structure. This approach is also following the line suggested by a number of theoreticians and experimentalists who argue that the idea of selection of a naturally occurring, fitting conformer is closer to reality than the classical induced fit model (Bosshard, 2001; Carlson and McCammon, 2000; Ma et al., 2002). According to this paradigm, the protein exists in a number of conformations in solution. Ligands of various shapes and sizes can bind to any conformation of the unbound protein, not only to the one with the lowest free energy. A ligand that binds to a less populated conformational state of the receptor with very high affinity can be a stronger binder than one that binds to the lowest energy conformation of the target with lower affinity. Nevertheless, this ligand would be missed if only the lowest energy conformer of the receptor or the average of several low energy structures was used as docking targets.

The set of multiple protein conformers usually come from NMR studies, x-ray structures of the same protein with various ligands, or MD simulations. In their groundbreaking work, Kuntz and co-workers (Knegtel et al., 1997) use ensembles of NMR and x-ray protein structures as targets for docking with DOCK. The binding site is placed on a grid, and intermolecular force field values are calculated at the grid points. Variations among different observable conformations are taken into account by calculating the average of the force field values at each grid point. Two types of averaging are used: energy weighted, and geometry weighted. The first method involves calculating the contribution of each atom from each structure to the potential energy, then calculating a weighted potential by averaging over all structures. Geometry weighted averaging means that the averaging is performed at the structural level by calculating a mean position for every atom of the protein. Although this approach does not include receptor flexibility in a dynamic sense, the composite grid representing the interaction energies of the docked ligands with the different protein conformers is shown to outperform many of the grids derived from individual structures in identifying known inhibitors for the cases studied. Claussen and co-workers use FlexE, an extention of FlexX (Kramer et al., 1999), to dock ligands into a united protein description generated from the superimposed structures of the x-ray srtucture ensemble of the target protein (Claussen et al., 2001). While averaging the similar backbone and side-chain positions, the regions with larger variations are retained in form of conformational libraries. New conformations of the receptor are created by combining compatible conformations of the various flexible regions of the binding site. The method can handle several side-chain conformations and smaller loop (up to three or four amino acid) movements but not

motions of larger backbone segments. Nevertheless, docking into the united protein description with FLEXE did not provide considerably better dockings than docking into the individual crystal structures with FLEXX for the proteins studied.

The use of multiple experimental structures limits the conformational sampling to already observed and existing conformations. Some proteins do not have multiple x-ray structures or are too large or flexible for NMR structure determination. MD simulations can provide novel protein conformers to be used as targets for docking, however, they generate smaller scale movements than may be observed in nature due to their high computational cost. The development of a dynamic pharmacophore model for HIV-1 integrase is described by Carlson et al. by using snapshots of MD simulations and the multi-copy minimization method MUSIC to determine binding regions for probe molecules in the dynamic binding site (Carlson et al., 2000). The drawback of MD simulations is the long time (from weeks to months) required to achieve a good sampling. In fact, it is almost impossible to get beyond microsecond timescale motions.

In this chapter, a new and relatively efficient approach to modeling main-chain flexibility in docking and screening is described. Flexibility analysis from a single conformation of the target protein was performed using the graph-theoretic algorithm FIRST (Jacobs et al., 2001), followed by the generation of alternative conformations for the predicted flexible regions with ROCK (Thorpe et al., 2001), a fast and efficient conformational sampling algorithm. A representative and diverse set of the conformational ensemble generated this way was used as a series of targets for docking with SLIDE. ROCK is uniquely suited for flexibly handling ring structures and can be used to model the flexibility of macrocyclic ligands as well, as it is demonstrated for

cyclosporin. The use of this combined method to perform flexible docking is illustrated on the cyclophilin A – cyclosporin system, while addressing the question of how much flexibility of the interacting molecules is tolerated without hindering recognition.

6.2 Methods

WHAT IF is a program suite for protein structure analysis (Vriend, 1990) used in this study to add the polar hydrogens to the crystallographic structure of the protein. This program was selected because it proved to reliable reproduce hydrogen positions observed in proteins whose structure were determined with neutron scattering.

FIRST (Floppy Inclusion and Rigid Substructure Topography) is a graph theoretical approach to identify rigid and flexible regions based on the bond network in proteins (Jacobs et al., 2001). The bond network consists of covalent bonds, hydrogen bonds and hydrophobic tethers. The algorithm counts the number of internal bond-rotational degrees of freedom in the system and identifies rigid regions as those having no bond-rotational degrees of freedom, in other words having enough constraints to become rigid. Flexible regions are those with remaining degrees of freedom or not enough constraints to become rigid. The number of extra constraints or the number of remaining degrees of freedom is used to calculate the relative rigidity or flexibility index of the region. This computational approach is very fast and is able to reliably predict the conformational flexibility of a protein from a single, static three-dimensional structure (Jacobs et al., 1999; Jacobs et al., 2001).

ROCK (Rigidity Optimized Conformational Kinetics) uses a "random walk" approach to search the conformational space available to proteins represented as bond networks (Thorpe et al., 2001; Thorpe and Lei, 2003). The program keeps bond lengths and coordination angles constant, randomly performing small rotations for the rotatable bonds. It also ensures that all the original bond constraints are obeyed and van der Waals overlaps between atoms are avoided. By using the results of the FIRST flexibility analysis, ROCK generates a set of semi-continuous conformations by sampling only those bonds in the proteins that are predicted to be rotatable. A non-linear constrained optimization algorithm is used for repositioning the side chains not involved in rings, consistent with the new main chain conformation. Only those main-chain conformers obeying the favored Φ,Ψ distribution (Ramachandran and Sasisekharan, 1968) used by the program PROCHEK (Laskowski et al., 1993; Morris et al., 1992) are sampled.

SLIDE (Schnecke and Kuhn, 1999; Schnecke and Kuhn, 2000) was used to perform the docking experiments using the protein conformations generated by ROCK. Descriptions of the algorithm can be found in Chapter 1, section 1.4 and Chapter 3, section 3.3.

The recognition of cyclosporin by cyclophilin A

Cyclophilin A (CypA) is a ubiquitous cytosolic protein composed of 165 amino acids catalyzing cis-trans isomerization in peptides and proteins (Hacker and Fischer, 1993). CypA is also the target for the immunosuppressive drug cyclosporin A, a cyclic undecapeptide, in which 7 of the 11 amide nitrogens are methylated (Figure 6.1). The

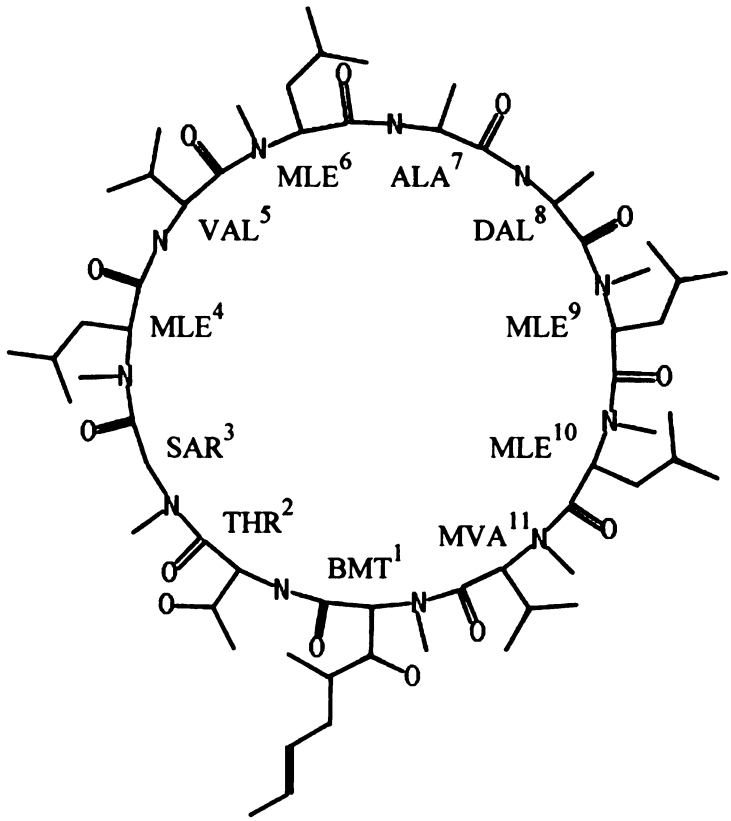


Figure 6.1. The cyclic undecapeptide Thr2-cyclosporin. The only difference between Thr2-cyclosporin and the immunosuppressive drug cyclosporin A is that the second residue of the latter does not have the OH group on the side chain of the second residue. The two cyclosporin molecules have comparable biological activity. The complex of human CypA–Thr2-cyclosporin was selected for this study because of its higher resolution crystal structure (PDB code 1bck, resolution 1.8 Å) compared to the human CypA–cyclosporin A complex (PDB code 1cwa, resolution 2.1 Å).

CypA-cyclosporin complex binds and inhibits the Ser-Thr phosphatase calcineurin, as well as blocks the activation of JNK and p38 signaling pathways, inhibiting T lymphocyte activation (Matsuda and Koyasu, 2000). The role of CypA in immuno-suppression seems to be unrelated to its cis-trans isomerase function (Ke et al., 1994). Another interesting aspect of CypA is that it binds to the HIV-1 Gag protein and is

incorporated into the HIV-1 virion as a necessary element for HIV infection (Saphire et al., 2000). The human CypA – cyclosporin complex serves as an excellent model system for this study for the following reasons: (1) as a peptidyl cis–trans isomerase, CypA has to be flexible enough to accommodate both the cis and the trans conformations of its substrate, (2) since the ligand, cyclosporin, is a cyclic peptide with many bond rotational degrees of freedom, the same method used to model main–chain flexibility of the target protein can be applied to model the flexibility of this peptidyl ring, (3) CypA has been studied extensively, with a number of high resolution x-ray and NMR structures available for both the unliganded CypA as well as complexes of CypA with various peptidyl ligands. This allows comparison of our flexibility predictions with the available experimental data.

6.3 Results

6.3.1 Flexibility Analysis

The CypA structure used as input to FIRST was the 1.8 Å resolution x-ray structure, PDB entry 1bck (Kallen et al., 1998), with the ligand Thr2-cyclosporin removed. Polar hydrogens were added using the program WhatIf (Vriend, 1990). FIRST identified 213 hydrogen bonds satisfying the required geometric criteria, calculated their energy using a modified Mayo potential, and ranked them in order of their decreasing energy, with a maximum (least favorable) H-bond energy of E = –0.1 kcal/mol. A detailed description of modeling hydrogen bonds in FIRST is provided in Jacobs et al., 2001. FIRST also identified 121 hydrophobic tethers, representing hydrophobic interactions between carbon and/or sulfur atoms for which:

$$D \leq R_A + R_B + R$$

where D is the distance between the two interacting hydrophobic atoms, R_A and R_B are their van der Waals radii, and R was empirically defined to be 0.5 Å. The modeling of hydrophobic interactions is a modified version of that described previously (Rader et al., 2002). In this study, only those atoms covalently bond to only carbons and/or hydrogens were considered to participate in hydrophobic interactions, whereas in Rader et al., 2002, any carbon or sulfur atoms with van der Waals surfaces within 0.25 Å were considered engaged in hydrophobic interactions. This change was implemented to provide a more realistic representation of hydrophobic interactions between significantly hydrophobic areas of the protein.

One way to visualize the results of the flexibility analysis is by generating a hydrogen-bond dilution plot (Figure 6.2). The numbering along the top, from left to right, represents the amino acid sequence of the protein (residues 1-165). Each line below is colored according to the rigid cluster decomposition. Thin black lines are the flexible parts, while thick bars represent rigid regions, with identical colors for mutually rigid regions that belong to the same rigid cluster (which may or may not be regions contiguous in sequence). The consecutive lines illustrate the changes in flexibility of the protein as hydrogen bonds are removed in order of their increasing energy, a process analogous to thermal denaturation. These energy values are based on the modified Mayo potential (Hespenheide et al., 2002) and should be considered a reasonable ranking of relative energies, rather than an absolute scale. When the crystal structure (PDB entry 1bck) itself is analyzed (energy of 0, top of the plot) all 213 hydrogen bonds are present and the structure forms one rigid cluster (represented by the red bar), with only a few residues at the N terminus being flexible (represented by the thin black line segment). Moving down the plot, as the energy increases and hydrogen bonds incrementally break, certain regions become flexible. Hydrophobic interactions are maintained intact, because the hydrophobic interaction actually becomes stronger over moderate increases in temperature. Eventually, the system breaks up into two or more independent rigid regions represented by segments of different colors connected by flexible regions. The first and second columns in Figure 6.2 list the index number (rank, from strongest to weakest, where the strongest H-bond is given index 1, etc.) and energy of those hydrogen bonds whose breakage induces a rigid to flexible change in the structure. The last two columns specify the residue numbers of the hydrogen donor (blue) and acceptor (red) of

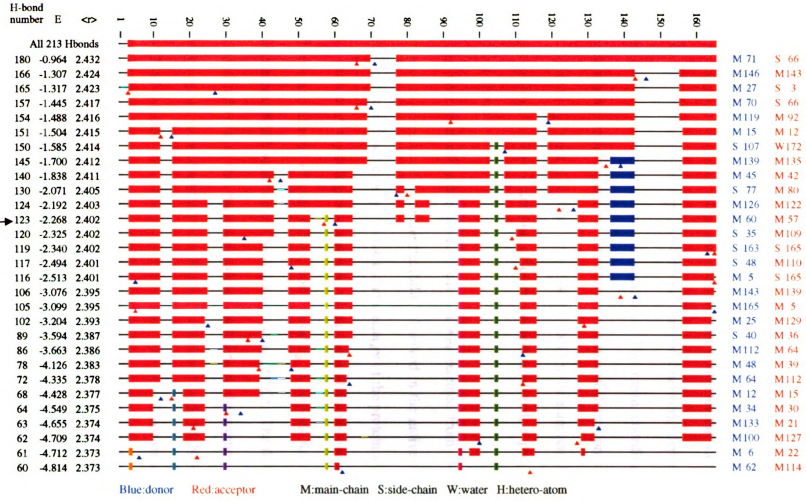


Figure 6.2. Hydrogen-bond dilution plot of the unliganded human CypA.

the respective hydrogen bond, and the donor and acceptor positions are also shown by carets beneath the rigid cluster decomposition plot. The third column lists the mean coordination number at the current H-bond energy. The mean coordination number $\langle r \rangle$ is the average number of covalent and non-covalent bonds for the atoms in the protein, and provides an overall description of the protein bond network, depending strongly on the number of bonds present in the structure. The mean coordination is a useful parameter when comparing rigid to flexible transitions in different proteins (Rader et al., 2002). Moreover, Rader et al. find $\langle r \rangle = 2.405 \pm 0.015$ to be a universal value describing the rigid-to-flexible phase transition of every protein analyzed, a property shared with amorphous glasses.

An energy cutoff value of -2.3 kcal/mol was selected from the hydrogen bond dilution plot as corresponding to the flexibility observed in the native state of the CypA protein. This energy corresponds to the thermal energy of the protein and was selected to reflect a state near physiological conditions where the protein has one rigid core but the outer loops are flexible (Figure 6.3.A), corresponding to regions found flexible in the well-determined NMR structure of CypA (Ottiger et al., 1997). The flexibility properties of the bond network at this particular energy cutoff were determined by FIRST. The following regions were identified to be flexible: residues 12-15, 24-29, 43-47, 54-60, 65-76, 79-82, 87-94, 101-107, 116-127, 133-135, and 143-155. Three strands of the β-sheet forming the bottom of the binding site (Figure 6.3.A) are rigid, while the loops surrounding the incoming ligand are flexible. To study the effect of the ligand binding on the flexibility of the target protein, FIRST analysis was also performed on the CypA-

cyclosporin complex (Figure 6.3.B). Cyclosporin rigidifies part of the CypA binding site, especially the 87-94 strand and loops 24-29, 87-94, and 116-127.

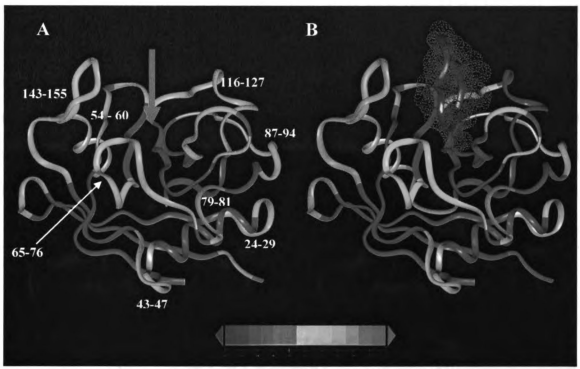


Figure 6.3. Ribbon diagram of the ligand-free (A) and the ligand-bound (B) CypA structures colored by flexibility index. Grey regions are isostatic or just rigid, blue regions are overconstrained, having more than enough bonds to make them rigid, while yellow to red regions are flexible. The red arrow in panel A indicates the location of the binding site, which is occupied by the cyclosporin ligand (colored green) in panel B.

6.3.2 Conformer Generation

The results of the FIRST analysis for CypA (list of hydrogen bonds stronger than -2.3 kcal/mol, list of hydrophobic tethers, flexibility index of each bond) were used as the input for ROCK to generate alternative conformations for the flexible regions. Two ROCK runs were performed: one with dihedral angle rotation steps of maximum 5

degrees, the second one with steps up to 10 degrees. Given the random walk nature of the conformational sampling with ROCK, runs with different angle step have the potential of sampling different areas of the conformational space. Each run generated 600 conformers. The 20 most distinct conformations from each of the two runs were combined and the 12 most distinct conformers (Figure 6.4.A) of these 40 structures were identified and used as targets for docking with SLIDE. The good stereochemistry of these conformers was confirmed with PROCHECK, and the Ramachandran plots for the x-ray structure and the most distinct conformer from the x-ray structure are shown in Figure 6.5, with the other conformers being of similar quality.

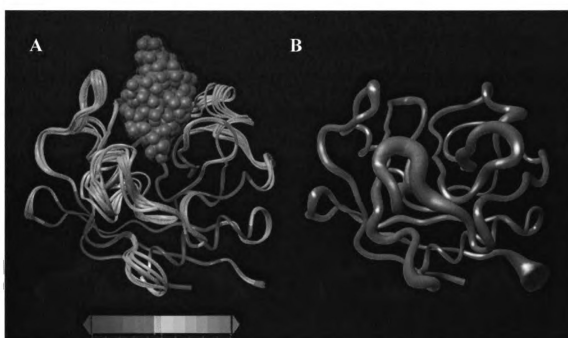


Figure 6.4. The 12 most distinct conformers of CypA generated by ROCK (A). The ligand, in magenta, indicates the location of the binding site. The ribbon diagram of the lowest energy NMR structure of free CypA (PDB code 1oca) is shown in panel B. The thickness of the tube is proportional to the maximum deviation of the backbone Cα atoms from the average Cα position of the 20 energy-minimized NMR structures from PDB entry 1oca.

The set of 12 most distinct conformers was identified based on the backbone RMS deviations of residues 42-46, 67-75, 79-81, 120-124, and 148-149 relative to the crystal structure (PDB entry 1bck), ranging from 0.94 to 1.34 Å overall. The movements of these regions were monitored because these are the main regions predicted to be flexible by FIRST surrounding the binding site and they were also identified as the ones with the most significant backbone differences in NMR structures (Ottiger et al., 1997). The conformers generated by ROCK (Figure 6.4A) sample approximately the same conformational space as the 20 lowest energy NMR structures of CypA (Figure 6.4B). The regions with the largest movements modeled by ROCK are the regions with the most variations among the individual NMR structures, with the ROCK conformers showing somewhat larger backbone deviations. The backbone RMS deviations of the ROCK conformers for the whole sequence of CypA were in the range of 0.50-0.76 Å. To illustrate the range of motions captured by ROCK, the maximum Cα deviations from the x-ray structure observed in the 12 most distinct conformers were plotted for each residue (Figure 6.6.A). As a comparison, the deuterium exchange rates of the backbone amide protons of free CypA are shown in panel B of Figure 6.6. The HD exchange data was generously provided by Marcel Ottiger and Kurt Wüthrich (Ottiger et al., 1997).

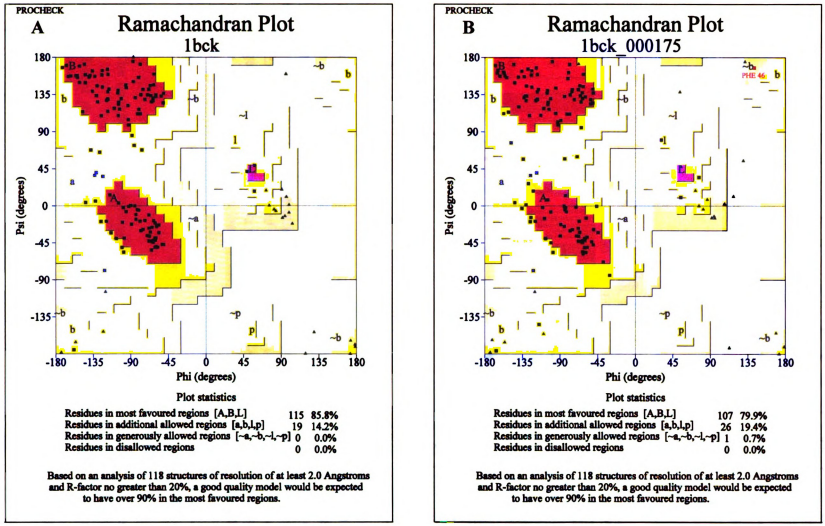


Figure 6.5. Ramachandran plots of the crystal structure of CypA (A) and of the most distinct conformer generated with ROCK (B). Red areas correspond to most favored core regions, dark yellow to allowed regions, and pale yellow to generously allowed regions.

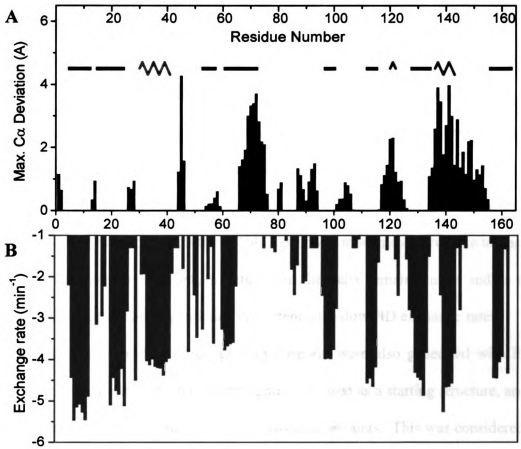


Figure 6.6. Maximum deviations of backbone Cα atom positions compared to x-ray structure positions (PDB code 1bck) seen among the most distinct CypA conformers generated with ROCK (A). As a comparison, the hydrogen-deuterium (HD) exchange rates of the backbone amide protons of free CypA are shown in panel B (Ottiger et al., 1997). Exchange rates of –1 indicate that the exchange was too fast to be observed. Exchange rates of -5 indicate very slow exchange. The locations of the regular secondary structure elements are given in the top panel, where blue lines indicate β-sheets and the red zig-zags corresponds to α-helices.

The backbone RMS deviations of the ROCK conformers provide a measure for how different these conformers are, in average. Even if RMSD values are relatively small (0.94 to 1.34 Å overall for the flexible regions), individual backbone atom deviations can be much larger. The largest Cα deviations of the ROCK conformers of up

to 4.28 Å from the crystallographic structure were observed in the loop regions surrounding the binding site (residues 43-47, 65-76, 116-127, 143-155). These are also the regions with very fast HD exchange rates (Figure 6.6) observed with NMR. The only apparent discrepancy between the FIRST/ROCK conformational predictions and the NMR data is found for helix 136-143, which has low HD exchange rates, suggestive of maintaining rigidity, but large Cα deviations, indicating larger movements. This helix is predicted to be an independent rigid cluster by FIRST, with the ability to move as a rigid unit relative to the rest of the protein. This rigid body movement gives rise to backbone deviations compared to the x-ray structure, but the helix remains intact and so do the intrahelical hydrogen bonds, prediction consistent with slow HD exchange rates.

Conformers for the cyclic ligand cyclosporin were also generated with ROCK. The protein-bound conformation of cyclosporin was used as a starting structure, and only the covalent bond lengths and angles were used as constraints. This was considered to be a reasonable approach given there is only one hydrogen bond in this peptide and no intramolecular hydrophobic tethers were identified by FIRST. Generating conformers starting from the unbound form of cyclosporin was also considered, but later dismissed because the unbound conformations of cyclosporin and its derivatives have a *cis* amide bond between residues 9 and 10, while the bound conformations are all-*trans* (Figure 6.7.A). The peptide bonds are locked in ROCK, so the correct protein-bound conformation would have never been sampled; allowing flips between *cis* and *trans* peptide bonds can be incorporated in a future version of ROCK.

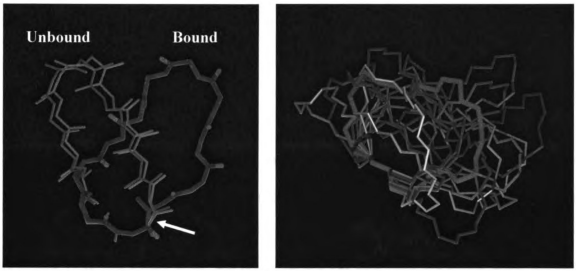


Figure 6.7. (A) Comparing the crystal structures of free (CSD codes ZAJDUJ and KEPNAU) and protein-bound conformations (PDB entry 1bck) of cyclosporin. The arrow indicates the location of the peptide bond between residues 9 and 10, which is *cis* in the free structures but *trans* in the protein-bound one. (B) A subset of cyclosporin conformers generated with ROCK from the protein-bound conformation (taken from PDB entry 1bck). The x-ray structure of cyclosporin is shown in red.

A total of 3000 cyclosporin conformers were generated in three separate runs using small angular steps. The runs were different in the step sizes of the dihedral angle rotations (2.0 and 5.0 degrees) and the maximum percentage of bonds that could be rotated at each step (10 and 20%). Since cyclosporin is a relatively large and flexible ligand, and docking thousand of conformers to 13 CypA targets (12 conformers plus the original x-ray structure) would be very time consuming, the most distinct 395 conformers out of the 3000 cyclosporin structures were selected for docking. The backbone RMS deviations of these conformers compared to the x-ray structure of the protein-bound cyclosporin (PDB code 1bck) were in the range of 0.45-8.60 Å, with Cα deviations up to

14.17 Å (Figure 6.7.B) indicating a diverse set of conformers and good sampling of the available conformational space.

6.3.3 Docking

CypA and cyclosporin conformers were used for docking with SLIDE to probe the range of flexibility consistent with molecular recognition. The templates created for the x-ray structure of CypA (PDB entry 1bck) and the 12 ROCK-generated conformers included 77 to 108 points each. All the hydrogen bonding template points were assigned as key points to assure that only those dockings with at least one hydrogen bond between protein and the ligand were generated. The results of the SLIDE dockings are summarized in Table 6.1. Docked ligands maintaining interactions with the rigid base of the binding site were considered to be correct dockings (Table 6.2). A docking was classified as correct if the contacts listed in Table 6.2 were maintained with a maximum distance of 5 Å (5.5 Å for the hydrophobic contact). This approach of using known recognition determinants (Kallen et al., 1998) to identify good dockings was employed since scoring functions, trained on correct dockings, do not perform well at distinguishing slight misdockings from gross misdockings (Zavodszky et al., 2003).

As can be seen in Table 6.1, every protein conformation could recognize and accommodate at least one ligand conformer (CypA_389 and CypA_548) although not necessarily the x-ray conformation. Main chain deviations at Cα positions of up to 4.28 Å were tolerated in the protein binding site, which was able to accommodate a wide range of cyclosporin ligand conformations with backbone RMSD values up to 4.25 Å compared to the x-ray conformation. Upon docking, key interactions with the more rigid

portions of the binding site were maintained by residues 1, 2 and 11 of cyclosporin, while the effector loop protruding from the binding site could flex considerably, reaching up to about 10 Å in Cα deviations (Figure 6.8).

Table 6.1. Results of docking cyclosporin conformers into CypA conformers.

CypA			Well docked cyclosporin conformers				
Conformer ID	RMSD (Å)	Max. Cα dev. (Å)	# well docked conformers	RMSD (Å)		Maximum Cα deviation (Å)	
				Before docking	After docking	Before docking	After docking
x-ray str.	0.00	0.00	77	0.00 - 2.97	0.41 – 3.28	5.84	5.89
073	1.24	2.82	63	0.00 – 3.35	0.70 – 5.96	6.18	10.41
088	1.17	3.32	45	0.00 – 3.08	0.56 – 3.05	5.21	5.86
118	0.94	3.97	75	0.00 – 4.25	0.61 – 4.63	6.52	7.05
164	1.11	3.09	10	0.74 – 3.54	1.17 – 3.33	5.46	5.70
175	1.73	3.71	66	0.00 – 3.17	0.64 – 5.62	6.04	10.26
269	1.04	2.89	12	1.76 – 3.04	1.98 – 3.32	5.09	6.68
389	1.32	2.71	1	2.93	2.46 – 2.51	5.09	4.40
431	1.08	2.67	13	0.00 – 3.54	1.25 – 3.56	5.46	5.95
457	1.05	2.04	62	0.00 – 3.48	0.73 – 4.33	5.84	6.86
493	1.22	4.28	15	0.84 – 3.09	1.53 – 3.76	4.96	6.08
548	1.12	1.98	1	2.07	3.93	3.92	6.34
592	1.34	2.60	9	1.10 – 2.88	1.76 – 4.67	4.92	8.35

Table 6.2. Interactions monitored to identify correct cyclosporin dockings.

		CypA	Cyclosporin	X-ray distance
1	Hydrogen bond	Gln63: NE2	BMT1: O	3.18 Å
2	Hydrogen bond	Asn102: O	THR2: N	2.97 Å
3	Hydrogen bond	Asn102: N	MVA11: O	3.54 Å
4	Hydrophobic interaction	Phe113: CD1	MVA11: CG1	3.44 Å

Figure 6.8. Cyclosporin conformers (green tubes) docked into the binding sites of CypA conformers (ribbons colored by flexibility index where yellow to red is increasingly flexible and grey to dark blue is increasingly rigid). The red tube is the x-ray conformation of cyclosporin docked into the x-ray conformation of CypA.

6.4 Discussion

In addition to studying the flexibility of CypA starting from the PDB structure 1bck, the FIRST analysis was also performed on the ligand-free crystal structure 2cpl, with a resolution of 1.63 Å (Ke, 1992). The results obtained were in excellent agreement with those for 1bck, indicating that the FIRST results are not very dependent on the particularities of the individual x-ray structures, if they have good stereochemistry.

The good agreement between the flexibility predictions of FIRST and the NMR data indicating which regions are the most flexible in CypA (Figures 6.4 and 6.5), as well as previous results on other proteins (Jacobs et al., 2001; Rader et al., 2002; Thorpe et al., 2000), suggest that FIRST is a reliable method to predict the flexible regions of a protein from a single x-ray structure. Further studies are needed, however, to identify a consistent way of identifying energy cutoff values corresponding to the native state of various proteins.

ROCK is a unique tool in its ability to sample flexibility in multiple interlocked ring systems. The protein main-chain conformers generated for CypA are of good stereochemical quality and span a conformational space very similar to that of the NMR solution structures. A very long MD simulation would be the only other computational method that could provide a similar degree of conformational sampling, but it would require several orders of magnitude longer time than ROCK. Comparisons between MD and ROCK sampling are ongoing in the research groups of our collaborators Michael Thorpe and David Case. The shortcomings of this method are the lack of a timescale associated with the modeled motions, as well as the lack of an energy function that allows assessment of the relative likelihood of the generated conformers. Also, an efficient way

of repositioning the side chains on the modified main chains should be implemented to find not only a feasable side chain conformation for each residue, but to identify the most favorable one. These aspects will be the focus of our future work on ROCK.

FIRST provided flexibility analysis combined with ROCK, for probing the range of possible motions, can have other applications besides providing a set of conformers for modeling main chain flexibility in docking. This approach also has great potential for studying entropy changes upon macromolecular association, as well as for studying allostery.

The suggestion of using multiple protein conformations is only the first step toward realistic modeling of main chain flexibility in docking. When the conformational space to be sampled is large, a large number of individual conformations should be used to ensure uniform sampling, which is not a practical solution. The idea of the combinatorial joining of the discrete conformations of different segments seems to be a sound one (Kramer et al., 1999), and it is worth pursuing.

6.5 Conclusions

Comparing our results to NMR data on CypA indicates that employing FIRST flexibility analysis of the target protein, followed by generating conformers for the predicted flexible regions with ROCK, is a realistic approach to model the flexibility of the protein main chain. Protein conformers with good stereochemistry are generated given only one starting structure. These conformers can be successfully used for docking experiments to model protein main chain flexibility. This method also provides the unique opportunity to study the effects of ligand binding on the entropy of the system. The docking

experiments of cyclosporin conformers to CypA conformers using SLIDE confirmed our hypothesis that there is a considerable amount of flexibility tolerated in the protein-ligand recognition process, reflected by the fact that multiple target structures accommodate a wide range of ligand conformers while maintaining key interactions.

Chapter 7

Summary and Future Directions

7.1 Summary of Advances Made

Our early work on of predicting the binding mode of NMN in the active site of R67-DHFR suggested that the binding site representation in SLIDE needed improvement. This was done, and we envisioned that the next major advance in protein-ligand docking was to incorporate full flexibility of the protein, including main chain motion.

The docking experiments of NMN into the highly symmetric active site pore of R67-DHFR described in Chapter 2 led to the conclusion that a limited number of symmetry-related amphipathic residues allowed binding of either of the two ligands, NADPH or folate, using the same binding site residues. Using internal symmetry in the protein to generate hot-spots that accommodate a number of different interactions was proposed as a novel evolutionary strategy used by this enzyme to confer antibiotic resistance to bacteria.

To improve the quality of docking and scoring in SLIDE, a new, knowledge-based approach to representing the protein binding site for docking was evaluated in Chapter 3. Instead of randomly or uniformly sampling the binding site, template points were placed at positions where the strongest hydrogen bonds with optimal geometry could be made between the protein and its ligand. A better description of the binding site led not only to better dockings but also to improvements is scoring, as reflected by the ability of SLIDE to better differentiate between known ligands and nonspecific molecules in the case of thrombin and GST. The reason behind scoring improvements was that scoring functions trained on correct x-ray binding orientations performed much better when the dockings were close to the correct position, while performing poorly on misdocked ligands.

An evaluation of modeling side-chain flexibility by SLIDE was presented in Chapter 4. This model of induced fit was built upon the hypothesis that both proteins and ligands change their conformations as little as necessary to resolve the interatomic collisions arising upon their association. This assumption was shown to be both necessary and sufficient to dock most known ligands to their target proteins correctly, at least when main-chain conformational change is not significant. The side chain rotations performed by SLIDE were shown to mimic the differences observed in side chain conformations in the binding sites of ligand-free and ligand-bound proteins. The systems studied included thrombin with 35 known ligands, human glutathione S-transferase with 14 known ligands, and a set of 18 diverse protein-ligand complexes. Our results reinforced earlier findings that ligand binding induces nonrotamericity in the target protein (Heringa and Argos, 1999), meaning that smaller rotations are found that do not

necessarily match distribution of the favored side-chain conformations in ligand-free proteins.

Every model, no matter how intuitive and elegant it is, should also relate usefully to experimental work. To assess the predictive power of SLIDE, it was used to screen the Available Chemicals Directory to predict new ligands for thrombin. These predictions were tested by measuring the binding affinity of the ligand candidates with isothermal titration calorimetry. As shown in Chapter 5, two of the molecules tested, morelloflavone and new fuchsin, turned out to be new ligands having micromolar affinities for thrombin. The main role of virtual screening is to identify leads that can be further developed into drugs, and micromolar affinity is typical for lead compounds identified by *in vitro* screening, too. Thus, SLIDE can discover promising ligand candidates from screening a large database.

A novel approach was suggested to model main-chain flexibility in docking in Chapter 6. The most distinct feature of this method is that it did not require lengthy MD simulations, nor was it restricted to available experimental structures to provide alternative main-chain protein conformations. This approach, combining the conformational sampling tool ROCK with SLIDE, was applied to explore the amount of flexibility allowed during the recognition process of cyclosporin by human cyclophilin A. As a first step, the flexibility of CypA was assessed based on the network of covalent bonds, hydrogen bonds and hydrophobic interactions identified in the crystallographic structure using the graph theoretic algorithm FIRST. As a next step, the program ROCK was employed to explore the conformational space available for the regions predicted by FIRST to be flexible, generating feasible main-chain CypA conformers with good

stereochemistry. Comparing the flexibility predictions of FIRST and the range of motions sampled by ROCK to existing NMR data on CypA showed very good agreement. This approach was shown to be not only a possible but also a realistic way to mimic the protein main-chain motions observed in NMR experiments. ROCK was also used to generate a set of conformers for the cyclic ligand cyclosporin. Main chain flexibility in protein-ligand recognition was modeled by docking a wide range of cyclosporin conformers into the CypA conformers with SLIDE. The docking results confirmed the hypothesis that a considerable amount of flexibility can be tolerated without hindering the protein-ligand recognition process.

7.2 Interesting Problems Remaining to be Solved

In an attempt to model a natural phenomenon or find an acceptable explanation to a puzzling problem, usually more questions than answers arise. In this last part, I enumerate some of the new problems that surfaced during my work and which need solutions in the near future.

During docking with SLIDE, both the protein and the ligand are handled as flexible molecules. However, the initial matching of ligand interaction point triangles to template point triplets rigidifies part of the ligand, which can bias the docking toward the starting ligand conformation. This is not a problem if the ligand is in a conformation not too far from the binding conformation. Many times that is not the case. According to the calculations of Bostrom et al. (Bostrom et al., 1998), the difference between the conformational energy of the free ligand in solution and that of the bioactive conformation is $\leq$ 3 kcal/mol for most ligands studied, indicating that the energy

required to distort the ligand too much for binding would decrease the binding affinity, and as such is not favorable. It is relatively safe to assume, that a ligand can exist in multiple low energy conformations in solution, so a possible approach would be to sample the conformational space of the ligand, generate a number of low energy solution conformers, and use them as input to SLIDE.

Water molecules are known to mediate protein-ligand interactions in many cases. Often, correct docking cannot be achieved without taking interfacial water into account. SLIDE has the possibility of taking water molecules into account at the protein-ligand interface, keeping or replacing them during the docking process based on Consolv (Raymer et al., 1997) or other predictions. SLIDE currently seems to penalize too much for displacing a water molecule predicted to be conserved upon ligand binding. Further studies are needed to find an appropriate way of handling this issue. In addition, we need to handle the more general case of new bound water molecules being recruited into the binding interface. An additional functionality is needed that would detect the existence of a cavity between the protein and the ligand appropriate for accommodating a water molecule and place water there such to form a bridge between the interacting partners, when energetically favorable.

One of the greatest challenges of the docking/screening field is predicting the binding affinity of the docked ligand to its target protein. This is done by either free energy calculations or by employing simple physical or empirical scoring functions. None of the methods reported so far performs satisfactorily across various systems. An adequate scoring scheme would facilitate identification of the best docked orientation among the tens or even hundreds of possible orientations produced by a docking

experiment. It would also allow ranking of ligand candidates according to their binding affinity, which would make virtual screening a more reliable tool in searching for ligands. The usefulness of a good scoring method would not be restricted to protein-ligand docking. The same principles of interaction apply to protein-protein docking and protein folding, for example. It is widely known that the bigger challenge in the protein folding field is not generating possible structures but ranking them and identifying the most native-like ones.

Docking a ligand into a rigid protein is of limited use. While it is relatively easy to sample side-chain conformations on the fly, main-chain conformational sampling is more computationally intensive, and it is more feasible to generate a set of possible main chain conformations in advance to use them as a library of alternative structures during docking. Docking into individual conformers might still be very time consuming, especially if a thorough sampling has to be done. Combining the available main-chain conformations in a combinatorial way seems to be a good way to speed up the docking and explore new conformations. In addition, associating energy values with the main chain conformers would help concentrating on the energetically most favorable ones instead of taking into account all the stericly feasible conformers. A major improvement regarding flexibility modeling in SLIDE also could be made by invoking dynamics not only to move protein and/or ligand parts out of the way to resolve inter-atomic collisions, but to move them toward making better interactions as well.

In conclusion, this thesis shows that computational methods can be effective tools for modeling interactions between flexible biological molecules, and they can be used successfully to predict binding orientations of ligands relative to their target proteins, as

well as to find new ligands for proteins. Computational modeling of biomolecular interactions is one of the fastest evolving fields of science today. Further improvements in algorithms, combined with the rapidly growing amount of new information accumulated about the structure, function, and interaction of bio-molecules assures an important role of computational modeling of protein-ligand recognition in the future of biological sciences.

BIBLIOGRAPHY

Abagyan, R. and Totrov, M. 2001. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol. 5*:375-382.

Allen, F.H. 2002. The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallogr. B 58*:380-388.

Babbitt, P.C. and Gerlt, J.A. 1997. Understanding enzyme superfamilies: Chemistry as the fundamental determinant in the evolution of new catalytic activities. *J. Biol. Chem. 272*:30591-30594.

Baxter, C.A., Murray, C.W., Waszkowycz, B., Li, J., Sykes, R.A., Bone, R.G., Perkins, T.D., Wylie, W. 2000. New approach to molecular docking and its application to virtual screening of chemical databases. *J. Chem. Inf. Comput. Sci. 40*:254-262.

Bergner, A., Gunther, J., Hendlich, M., Klebe, G., Verdonk, M. 2001. Use of Relibase for retrieving complex three-dimensional interaction patterns including crystallographic packing effects. *Biopolymers 61*:99-110.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res. 28*:235-242.

Betts, M.J. and Sternberg, M.J. 1999. An analysis of conformational changes on protein-protein association: Implications for predictive docking. *Protein Eng 12*:271-283.

Bissantz, C., Folkers, G., Rognan, D. 2000. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem. 43*:4759-4767.

Boer, D.R., Kroon, J., Cole, J.C., Smith, B., Verdonk, M.L. 2001. SuperStar: Comparison of CSD and PDB-based interaction fields as a basis for the prediction of protein-ligand interactions. *J. Mol. Biol. 312*:275-287.

Bogan, A.A. and Thorn, K.S. 1998. Anatomy of hot spots in protein interfaces. *J. Mol. Biol. 280*:1-9.

Bohm, H.J. 1992. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Des 6*:61-78.

Bohm, H.J. 1994. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des 8*:243-256.

Boobbyer, D.N., Goodford, P.J., McWhinnie, P.M., Wade, R.C. 1989. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J. Med. Chem. 32*:1083-1094.

Bosshard, H.R. 2001. Molecular recognition by induced fit: How fit is the concept? *News Physiol Sci. 16*:171-173.

Bostrom, J., Norrby, P.O., Liljefors, T. 1998. Conformational energy penalties of protein-bound ligands. *J. Comput. Aided Mol. Des. 12*:383-396.

Bradrick, T.D., Beechem, J.M., Howell, E.E. 1996. Unusual binding stoichiometries and cooperativity are observed during binary and ternary complex formation in the single active pore of R67 dihydrofolate reductase, a D2 symmetric protein. *Biochemistry 35*:11414-11424.

Brito, R.M., Reddick, R., Bennett, G.N., Rudolph, F.B., Rosevear, P.R. 1990. Characterization and stereochemistry of cofactor oxidation by a type II dihydrofolate reductase. *Biochemistry 29*:9825-9831.

Bruice, T.C. and Benkovic, S.J. 2000. Chemical basis for enzyme catalysis. *Biochemistry 39*:6267-6274.

Burkhard, P., Taylor, P., Walkinshaw, M.D. 1998. An example of a protein ligand found by database mining: Description of the docking method and its verification by a 2.3 Angstrom X-ray structure of a thrombin-ligand complex. *J. Mol. Biol. 277*:449-466.

Bystroff, C., Oatley, S.J., Kraut, J. 1990. Crystal structures of Escherichia coli dihydrofolate reductase: The NADP+ holoenzyme and the folate.NADP+ ternary complex. Substrate binding and a model for the transition state. *Biochemistry 29*:3263-3277.

Carlson, H.A. 2002. Protein flexibility and drug design: How to hit a moving target. *Curr. Opin. Chem. Biol. 6*:447-452.

Carlson, H.A., Masukawa, K.M., McCammon, J.A. 1999. Method for Including the Dynamic Fluctuations of a Protein in Computer-Aided Drug Design. *J. Phys. Chem. A 103*:10213-10219.

Carlson, H.A., Masukawa, K.M., Rubins, K., Bushman, F.D., Jorgensen, W.L., Lins, R.D., Briggs, J.M., McCammon, J.A. 2000. Developing a dynamic pharmacophore model for HIV-1 integrase. *J. Med. Chem. 43*:2100-2114.

Carlson, H.A. and McCammon, J.A. 2000. Accommodating protein flexibility in computational drug design. *Mol. Pharmacol. 57*:213-218.

Castillo, R., Andrés, J., Moliner, V. 1999. Catalytic mechanism of dihydrofolate reductase enzyme: A combined quantum-mechanical/molecular-mechanical

characterization of transition state structure for the hydride transfer step. *J. Am. Chem. Soc. 121*:12140-12147.

Charifson, P.S., Corkery, J.J., Murcko, M.A., Walters, W.P. 1999. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem. 42*:5100-5109.

Chen, L. and Sigler, P.B. 1999. The crystal structure of a GroEL/peptide complex: Plasticity as a basis for substrate diversity. *Cell 99*:757-768.

Chen, Y.Z. and Ung, C.Y. 2001. Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. *J. Mol. Graph. Model. 20*:199-218.

Claussen, H., Buning, C., Rarey, M., Lengauer, T. 2001. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol. 308*:377-395.

Connolly, M.L. 1993. The molecular surface package. *J. Mol. Graph. 11*:139-141.

Davie, E.W., Fujikawa, K., Kisiel, W. 1991. The coagulation cascade: Initiation, maintenance, and regulation. *Biochemistry 30*:10363-10370.

Dekker, R.J., Eichinger, A., Stoop, A.A., Bode, W., Pannekoek, H., Horrevoets, A.J. 1999. The variable region-1 from tissue-type plasminogen activator confers specificity for plasminogen activator inhibitor-1 to thrombin by facilitating catalysis: release of a kinetic block by a heterologous protein surface loop. *J. Mol. Biol. 293*:613-627.

DeLano, W.L., Ultsch, M.H., de Vos, A.M., Wells, J.A. 2000. Convergent solutions to binding at a protein-protein interface. *Science 287*:1279-1283.

Dion-Schultz, A. and Howell, E.E. 1997. Effects of insertions and deletions in a beta-bulge region of Escherichia coli dihydrofolate reductase. *Protein Eng 10*:263-272.

Doman, T.N., McGovern, S.L., Witherbee, B.J., Kasten, T.P., Kurumbail, R., Stallings, W.C., Connolly, D.T., Shoichet, B.K. 2002. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem. 45*:2213-2221.

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

Esmon, C.T. 1995. Thrombomodulin as a model of molecular mechanisms that modulate protease specificity and function at the vessel surface. *FASEB J. 9*:946-955.

Ewing, T.J., Makino, S., Skillman, A.G., Kuntz, I.D. 2001. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des 15*:411-428.

Fischer, D., Lin, S.L., Wolfson, H.L., Nussinov, R. 1995. A geometry-based suite of molecular docking processes. *J. Mol. Biol. 248*:459-477.

Fischer, D., Norel, R., Wolfson, H., Nussinov, R. 1993. Surface motifs by a computer vision technique: Searches, detection, and implications for protein-ligand recognition. *Proteins 16*:278-292.

Fox, T. and Haaksma, E.E. 2000. Computer based screening of compound databases: 1. Preselection of benzamidine-based thrombin inhibitors. *J. Comput. Aided Mol. Des 14*:411-425.

Fradera, X., Knegtel, R.M., Mestres, J. 2000. Similarity-driven flexible ligand docking. *Proteins 40*:623-636.

Fremont, D.H., Matsumura, M., Stura, E.A., Peterson, P.A., Wilson, I.A. 1992. Crystal structures of two viral peptides in complex with murine MHC class I H-2Kb. *Science 257*:919-927.

Gohlke, H., Hendlich, M., Klebe, G. 2000a. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol. 295*:337-356.

Gohlke, H., Hendlich, M., Klebe, G. 2000b. Predicting binding modes, binding affinities and `hot spots' for protein-ligand complexes using a knowledge-based scoring function. *Perspectives in Drug Discovery and Design 20*:115-144.

Good, A. 2001. Structure-based virtual screening protocols. *Curr. Opin. Drug Discov. Devel. 4*:301-307.

Goodsell, D.S., Morris, G.M., Olson, A.J. 1996. Automated docking of flexible ligands: Applications of AutoDock. *J. Mol. Recognit. 9*:1-5.

Gruneberg, S., Stubbs, M.T., Klebe, G. 2002. Successful virtual screening for novel inhibitors of human carbonic anhydrase: Strategy and experimental confirmation. *J. Med. Chem. 45*:3588-3602.

Hacker, J. and Fischer, G. 1993. Immunophilins: Structure-function relationship and possible role in microbial pathogenicity. *Mol. Microbiol. 10*:445-456.

Halperin, I., Ma, B., Wolfson, H., Nussinov, R. 2002. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins 47*:409-443.

Heringa, J. and Argos, P. 1999. Strain in protein structures as viewed through nonrotameric side chains: II. Effects upon ligand binding. *Proteins 37*:44-55.

Hespenheide, B.M., Rader, A.J., Thorpe, M.F., Kuhn, L.A. 2002. Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graph. Model. 21*:195-207.

Holm, L. and Sander, C. 1996. Mapping the protein universe. *Science 273*:595-603.

Howell, E.E., Shukla, U., Hicks, S.N., Smiley, R.D., Kuhn, L.A., Zavodszky, M.I. 2001. One site fits both: A model for the ternary complex of folate + NADPH in R67 dihydrofolate reductase, a D2 symmetric enzyme. *J. Comput. Aided Mol. Des 15*:1035-1052.

Hu, Z., Ma, B., Wolfson, H., Nussinov, R. 2000. Conservation of polar residues as hot spots at protein interfaces. *Proteins 39*:331-342.

Ippolito, J.A., Alexander, R.S., Christianson, D.W. 1990. Hydrogen bond stereochemistry in protein structure and function. *J. Mol. Biol. 215*:457-471.

Jacobs, D.J., Kuhn, L.A., Thorpe, M.F. 1999. Flexible and rigid regions in proteins.

Jacobs, D.J., Rader, A.J., Kuhn, L.A., Thorpe, M.F. 2001. Protein flexibility predictions using graph theory. *Proteins 44*:150-165.

Jelesarov, I. and Bosshard, H.R. 1999. Isothermal titration calorimetry and differential scanning calorimetry as complementary tools to investigate the energetics of biomolecular recognition. *J. Mol. Recognit. 12*:3-18.

Jiang, F. and Kim, S.H. 1991. "Soft docking": Matching of molecular surface cubes. *J. Mol. Biol. 219*:79-102.

Jones, G., Willett, P., Glen, R.C. 1995. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol. 245*:43-53.

Jones, G., Willett, P., Glen, R.C., Leach, A.R., Taylor, R. 1997. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol. 267*:727-748.

Kallblad, P. and Dean, P.M. 2003. Efficient conformational sampling of local side-chain flexibility. *J. Mol. Biol. 326*:1651-1665.

Kallen, J., Mikol, V., Taylor, P., Walkinshaw, M.D. 1998. X-ray structures and analysis of 11 cyclosporin derivatives complexed with cyclophilin A. *J. Mol. Biol. 283*:435-449.

Karlstrom, A., Zhong, G., Rader, C., Larsen, N.A., Heine, A., Fuller, R., List, B., Tanaka, F., Wilson, I.A., Barbas, C.F., III, Lerner, R.A. 2000. Using antibody catalysis to study the outcome of multiple evolutionary trials of a chemical task. *Proc. Natl. Acad. Sci. U. S. A 97*:3878-3883.

Ke, H. 1992. Similarities and differences between human cyclophilin A and other beta-barrel structures. Structural refinement at 1.63 Å resolution. *J. Mol. Biol. 228*:539-550.

Ke, H., Mayrose, D., Belshaw, P.J., Alberg, D.G., Schreiber, S.L., Chang, Z.Y., Etzkorn, F.A., Ho, S., Walsh, C.T. 1994. Crystal structures of cyclophilin A complexed with cyclosporin A and N-methyl-4-[(E)-2-butenyl]-4,4-dimethylthreonine cyclosporin A. *Structure 2*:33-44.

Klebe, G. 2000. Recent developments in structure-based drug-design. *J. Mol. Med. 78*:269-281.

Knegtel, R.M., Bayada, D.M., Engh, R.A., von der, S.W., van Geerestein, V.J., Grootenhuis, P.D. 1999. Comparison of two implementations of the incremental construction algorithm in flexible docking of thrombin inhibitors. *J. Comput. Aided Mol. Des 13*:167-183.

Knegtel, R.M., Kuntz, I.D., Oshiro, C.M. 1997. Molecular docking to ensembles of protein structures. *J. Mol. Biol. 266*:424-440.

Knegtel, R.M. and Wagener, M. 1999. Efficacy and selectivity in flexible database docking. *Proteins 37*:334-345.

Koehler, R.T., Villar, H.O., Bauer, K.E., Higgins, D.L. 1997. Ligand-based protein alignment and isozyme specificity of glutathione S-transferase inhibitors. *Proteins 28*:202-216.

Koshland, D.E.J. 1958. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA 44*:98-104.

Kramer, B., Rarey, M., Lengauer, T. 1999. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins 37*:228-241.

Kuhn, L.A., Swanson, C.A., Pique, M.E., Tainer, J.A., Getzoff, E.D. 1995. Atomic and residue hydrophilicity in the context of folded protein structures. *Proteins 23*:536-547.

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E. 1982. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol. 161*:269-288.

Kuntz, I.D., Chen, K., Sharp, K.A., Kollman, P.A. 1999. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U. S. A 96*:9997-10002.

LaLonde, J.M., Bernlohr, D.A., Banaszak, L.J. 1994. X-ray crystallographic structures of adipocyte lipid-binding protein complexed with palmitate and hexadecanesulfonic acid. Properties of cavity binding sites. *Biochemistry 33*:4885-4895.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M. 1993. Procheck - A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography 26*:283-291.

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

Leach, A.R. and Lemon, A.P. 1998. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins 33*:227-239.

Lengauer, T. and Rarey, M. 1996. Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol. 6*:402-406.

Li, D., Levy, L.A., Gabel, S.A., Lebetkin, M.S., DeRose, E.F., Wall, M.J., Howell, E.E., London, R.E. 2001. Interligand Overhauser effects in type II dihydrofolate reductase. *Biochemistry 40*:4242-4252.

Lin, J.H., Perryman, A.L., Schames, J.R., McCammon, J.A. 2002. Computational drug design accommodating receptor flexibility: The relaxed complex scheme. *J. Am. Chem. Soc. 124*:5632-5633.

Lovell, S.C., Word, J.M., Richardson, J.S., Richardson, D.C. 2000. The penultimate rotamer library. *Proteins 40*:389-408.

Ma, B., Shatsky, M., Wolfson, H.J., Nussinov, R. 2002. Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations. *Protein Sci. 11*:184-197.

Mader, M.M. and Bartlett, P.A. 1997. Binding energy and catalysis: The implications for transition-state analogs and catalytic antibodies. *Chem. Rev. 97*:1281-1302.

Massova, I., Martin, P., Bulychev, A., Kocz, R., Doyle, M., Edwards, B.F., Mobashery, S. 1998. Templates for design of inhibitors for serine proteases: Application of the program DOCK to the discovery of novel inhibitors for thrombin. *Bioorg. Med. Chem. Lett. 8*:2463-2466.

Matsuda, S. and Koyasu, S. 2000. Mechanisms of action of cyclosporin. *Immunopharmacology 47*:119-125.

Matsumura, M., Fremont, D.H., Peterson, P.A., Wilson, I.A. 1992. Emerging principles for the recognition of peptide antigens by MHC class I molecules. *Science 257*:927-934.

McDonald, I.K. and Thornton, J.M. 1994. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol. 238*:777-793.

Mestres, J., Rohrer, D.C., Maggiora, G.M. 1997. MIMIC: A molecular-field matching program. Exploiting applicability of molecular similarity approaches. *J. Comp. Chem. 18*:934-954.

Mitchell, J.B.O., Laskowski, R.A., Alex, A., Thornton, J.M. 1999. BLEEP - potential of mean force describing protein-ligand interactions: I. Generating potential. *J. Computat. Chem. 20*:1177-1185.

Miyamoto, S. and Kollman, P.A. 1993. Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches. *Proteins 16*:226-245.

Moreno, E. and Leon, K. 2002. Geometric and chemical patterns of interaction in protein--ligand complexes and their application in docking. *Proteins 47*:1-13.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G., Thornton, J.M. 1992. Stereochemical quality of protein structure coordinates. *Proteins 12*:345-364.

Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., Olson, A.J. 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem. 19*:1639-1662.

Muegge, I. and Martin, Y.C. 1999. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem. 42*:791-804.

Murray, C.W., Baxter, C.A., Frenkel, A.D. 1999. The sensitivity of the results of molecular docking to induced fit effects: Application to thrombin, thermolysin and neuraminidase. *J. Comput. Aided Mol. Des 13*:547-562.

Najmanovich, R., Kuttner, J., Sobolev, V., Edelman, M. 2000. Side-chain flexibility in proteins upon ligand binding. *Proteins 39*:261-268.

Narayana, N., Matthews, D.A., Howell, E.E., Xuong, N.H. 1995. A plasmid-encoded dihydrofolate reductase from trimethoprim-resistant bacteria has a novel D2-symmetric active site. *Nat. Struct. Biol. 2*:1018-1025.

O'Brien, P.J. and Herschlag, D. 1999. Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol. 6*:R91-R105.

Oakley, A.J., Lo, B.M., Nuccetelli, M., Mazzetti, A.P., Parker, M.W. 1999. The ligandin (non-substrate) binding site of human Pi class glutathione transferase is located in the electrophile binding site (H-site). *J. Mol. Biol. 291*:913-926.

Oakley, A.J., Lo, B.M., Ricci, G., Federici, G., Parker, M.W. 1998. Evidence for an induced-fit mechanism operating in pi class glutathione transferases. *Biochemistry 37*:9912-9917.

Oakley, A.J., Rossjohn, J., Lo, B.M., Caccuri, A.M., Federici, G., Parker, M.W. 1997. The three-dimensional structure of the human Pi class glutathione transferase P1-1 in complex with the inhibitor ethacrynic acid and its glutathione conjugate. *Biochemistry 36*:576-585.

Osterberg, F., Morris, G.M., Sanner, M.F., Olson, A.J., Goodsell, D.S. 2002. Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins 46*:34-40.

Ottiger, M., Zerbe, O., Guntert, P., Wuthrich, K. 1997. The NMR solution conformation of unligated human cyclophilin A. *J. Mol. Biol. 272*:64-81.

Park, H., Bradrick, T.D., Howell, E.E. 1997. A glutamine 67 → histidine mutation in homotetrameric R67 dihydrofolate reductase results in four mutations per single active site pore and causes substantial substrate and cofactor inhibition. *Protein Eng 10*:1415-1424.

Pearlman, D.A. and Charifson, P.S. 2001. Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system. *J. Med. Chem. 44*:3417-3423.

Prade, L., Huber, R., Manoharan, T.H., Fahl, W.E., Reuter, W. 1997. Structures of class pi glutathione S-transferase from human placenta in complex with substrate, transition-state analogue and inhibitor. *Structure. 5*:1287-1295.

Quiocho, F.A., Spurlino, J.C., Rodseth, L.E. 1997. Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor. *Structure. 5*:997-1015.

Rader, A.J., Hespenheide, B.M., Kuhn, L.A., Thorpe, M.F. 2002. Protein unfolding: Rigidity lost. *Proc. Natl. Acad. Sci. U. S. A 99*:3540-3545.

Ramachandran, G.N. and Sasisekharan, V. 1968. Conformation of polypeptides and proteins. *Adv. Protein Chem. 23*:283-438.

Raymer, M.L., Sanschagrin, P.C., Punch, W.F., Venkataraman, S., Goodman, E.D., Kuhn, L.A. 1997. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *J. Mol. Biol. 265*:445-464.

Reece, L.J., Nichols, R., Ogden, R.C., Howell, E.E. 1991. Construction of a synthetic gene for an R-plasmid-encoded dihydrofolate reductase and studies on the role of the N-terminus in the protein. *Biochemistry 30*:10895-10904.

Rognan, D., Lauemoller, S.L., Holm, A., Buus, S., Tschinke, V. 1999. Predicting binding affinities of protein ligands from three-dimensional models: Application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem. 42*:4650-4658.

Ruppert, J., Welch, W., Jain, A.N. 1997. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci. 6*:524-533.

Sadowski, J. and Gasteiger, J. 1993. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev. 93*:2567-2581.

Sandak, B., Wolfson, H.J., Nussinov, R. 1998. Flexible docking allowing induced fit in proteins: Insights from an open to closed conformational isomers. *Proteins 32*:159-174.

Sanschagrin, P.C. and Kuhn, L.A. 1998. Cluster analysis of consensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity. *Protein Sci. 7*:2054-2064.

Saphire, A.C., Bobardt, M.D., Gallay, P.A. 2000. Human immunodeficiency virus type 1 hijacks host cyclophilin A for its attachment to target cells. *Immunol. Res. 21*:211-217.

Schaffer, L. and Verkhivker, G.M. 1998. Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization. *Proteins 33*:295-310.

Schapira, M., Raaka, B.M., Samuels, H.H., Abagyan, R. 2000. Rational discovery of novel nuclear hormone receptor antagonists. *Proc. Natl. Acad. Sci. U. S. A 97*:1008-1013.

Schapira, M., Raaka, B.M., Samuels, H.H., Abagyan, R. 2001. In silico discovery of novel retinoic acid receptor agonist structures. *BMC. Struct. Biol. 1*:1-7.

Schapira, M., Totrov, M., Abagyan, R. 1999. Prediction of the binding energy for small molecules, peptides and proteins. *J. Mol. Recognit. 12*:177-190.

Schnecke, V. and Kuhn, L.A. 1999. Database screening for HIV protease ligands: The influence of binding-site conformation and representation on ligand selectivity. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 242-251.

Schnecke, V. and Kuhn, L.A. 2000. Virtual screening with solvation and ligand-induced complementarity. *Perspectives in Drug Discovery and Design 20*:171-190.

Schnecke, V., Swanson, C.A., Getzoff, E.D., Tainer, J.A., Kuhn, L.A. 1998. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins 33*:74-87.

Schneider, G. and Bohm, H.J. 2002. Virtual screening and fast automated docking methods. *Drug Discov. Today 7*:64-70.

Shoichet, B.K. and Kuntz, I.D. 1993. Matching chemistry and shape in molecular docking. *Protein Eng 6*:723-732.

Shoichet, B.K., McGovern, S.L., Wei, B., Irwin, J.J. 2002. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol. 6*:439-446.

Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D., Perry, K.M. 1993. Structure-based discovery of inhibitors of thymidylate synthase. *Science 259*:1445-1450.

Sleigh, S.H., Seavers, P.R., Wilkinson, A.J., Ladbury, J.E., Tame, J.R. 1999. Crystallographic and calorimetric analysis of peptide binding to OppA protein. *J. Mol. Biol. 291*:393-415.

Smithrud, D.B. and Benkovic, S.J. 1997. The state of antibody catalysis. *Curr. Opin. Biotechnol. 8*:459-466.

Sobolev, V., Wade, R.C., Vriend, G., Edelman, M. 1996. Molecular docking using surface complementarity. *Proteins 25*:120-129.

Sotriffer, C.A., Gohlke, H., Klebe, G. 2002. Docking into knowledge-based potential fields: A comparative evaluation of DrugScore. *J. Med. Chem. 45*:1967-1970.

Stahl, M. and Rarey, M. 2001. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem. 44*:1035-1042.

Strader, M.B., Smiley, R.D., Stinnett, L.G., VerBerkmoes, N.C., Howell, E.E. 2001. Role of S65, Q67, I68, and Y69 residues in homotetrameric R67 dihydrofolate reductase. *Biochemistry 40*:11344-11352.

Strynadka, N.C., Jensen, S.E., Johns, K., Blanchard, H., Page, M., Matagne, A., Frere, J.M., James, M.N. 1994. Structural and kinetic characterization of a beta-lactamase-inhibitor protein. *Nature 368*:657-660.

Stubbs, M.T. and Bode, W. 1993. A player of many parts: The spotlight falls on thrombin's structure. *Thromb. Res. 69*:1-58.

Stubbs, M.T. and Bode, W. 1995. The clot thickens: Clues provided by thrombin structure. *Trends Biochem. Sci. 20*:23-28.

Tame, J.R. 1999. Scoring functions: A view from the bench. *J. Comput. Aided Mol. Des 13*:99-108.

Taylor, J.S. and Burnett, R.M. 2000. DARWIN: A program for docking flexible molecules. *Proteins 41*:173-191.

Thorpe, M.F., Hespenheide, B.M., Yang, Y., Kuhn, L.A. 2000. Flexibility and critical hydrogen bonds in cytochrome c. *Pac. Symp. Biocomput.* 191-202.

Thorpe, M.F. and Lei, M. 2003. Macromolecular flexibility. *Philosophical Magazine Special Edition*:1-8.

Thorpe, M.F., Lei, M., Rader, A.J., Jacobs, D.J., Kuhn, L.A. 2001. Protein flexibility and dynamics using constraint theory. *J. Mol. Graph. Model. 19*:60-69.

Tuffery, P., Etchebest, C., Hazout, S., Lavery, R. 1991. A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn. 8*:1267-1289.

Tulinsky, A. 1996. Molecular interactions of thrombin. *Semin. Thromb. Hemost. 22*:117-124.

Verdonk, M.L., Cole, J.C., Watson, P., Gillet, V., Willett, P. 2001. SuperStar: Improved knowledge-based interaction fields for protein binding sites. *J. Mol. Biol. 307*:841-859.

Vriend, G. 1990. What If - A Molecular Modeling and Drug Design Program. *Journal of Molecular Graphics 8*:52-56.

Wallace, A.C., Laskowski, R.A., Thornton, J.M. 1995. LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. *Protein Eng 8*:127-134.

Wang, R., Lai, L., Wang, S. 2002. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des 16*:11-26.

Waszkowycz, B. 2002. Structure-based approaches to drug design and virtual screening. *Curr. Opin. Drug Discov. Devel. 5*:407-413.

Wu, Y.D. and Houk, K.N. 1987. Theoretical transition structures for hydride transfer to methyleneiminium ion from methylamine and dihydropyridine. On the nonlinearity of hydride transfers. *J. Am. Chem. Soc. 109*:2226-2227.

Zavodszky, M.I., Sanschagrin, P.C., Korde, R.S., Kuhn, L.A. 2003. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. *J. Comput. Aided Mol. Des. in press*.