This is to certify that the

dissertation entitled
COMPUTATIONAL TECHNIQUES FOR MODELING PROTEIN-LIGAND INTERACTIONS AND THEIR APPLICATION TO SERINE PROTEASES AND ASPARAGINYL-tRNA SYNTHETASE

presented by
Paul C. Sanschagrin

has been accepted towards fulfillment
of the requirements for

DOCTOR OF PHILOSOPHY degree in BIOCHEMISTRY

Leslie A. Kuhn
Major professor

Date 12/13/01

COMPUTATIONAL TECHNIQUES FOR MODELING PROTEIN-LIGAND INTERACTIONS AND THEIR APPLICATION TO SERINE PROTEASES AND ASPARAGINYL-tRNA SYNTHETASE

By

Paul C. Sanschagrin

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

2001

ABSTRACT

COMPUTATIONAL TECHNIQUES FOR MODELING PROTEIN-LIGAND INTERACTIONS AND THEIR APPLICATION TO SERINE PROTEASES AND ASPARAGINYL-tRNA SYNTHETASE

By

Paul C. Sanschagrin

This thesis describes several techniques for modeling protein-ligand interactions, including interactions between water molecules and proteins, as well as application of these techniques to thrombin, trypsin, and asparaginyl-tRNA synthetase. A complete-linkage hierarchical cluster analysis technique for determining the degree of conservation of water sites in a series of related protein structures is presented. This technique was applied to the study of conserved water binding sites in the serine proteases thrombin and trypsin, with analysis of the implications for conserved water sites in the active site and nearby sodium ion site in thrombin. Cluster analysis was also used to identify conserved water binding sites in bovine pancreatic trypsin inhibitor (BPTI) and in the trypsin-BPTI complex. The conserved sites in the trypsin-BPTI complex were compared to those in trypsin and BPTI, showing that only about half of the interfacial water sites in the complex exist in the free form of either protein, while the remaining half are recruited or shuffled upon complex formation. The results of cluster analysis also allow inclusion of highly conserved water sites in protein and drug design.

In addition to examination of water molecules as protein ligands, modeling sites of favorable potential interactions with proteins in the SLIDE technique for computational ligand screening was improved. SLIDE is a multi-step algorithm which eliminates potential ligand molecules from a screening database via increasingly more stringent and computationally expensive steps to yield a ranked list of potential ligands for the protein target. Improvements made to the modeling of sites of favorable hydrophobic interaction for both the protein template and the potential ligand molecules are described and evaluated. Additional improvements made by my colleague Maria Zavodszky to the description of hydrogen bonding points in the protein template are also briefly described and evaluated. These improvements were tested using 42 thrombin and 16 glutathione S-transferase complexes. Both the protein template and database molecule representation improvements yield ligand dockings that are closer to those seen in the crystallographic complex. An enrichment of known ligands selected from a set of molecules from the Cambridge Structural Database (CSD) is also shown.

Application of SLIDE to asparaginyl-tRNA synthetase from *Brugia malayi*, a human pathogenic nematode, was performed. Screening against the CSD identified three potential ligands of particular interest: variolin B, with possible antitumor and antiviral properties; cercosporamide, which has known phytotoxic and fungitoxic properties; and phlorizin, a sodium/glucose transport inhibitor. Suggested binding modes for two of 16 *in vitro* high-throughput screening hits are also described.

For my wife, Suzanne. Without your support, encouragement, understanding, and love over the years I could not have dreamed to make it this far.

ACKNOWLEDGMENTS

I would like to first thank my advisor, Dr. Leslie Kuhn for her support, encouragement, and guidance throughout my graduate career. I consider myself lucky to have met such an intelligent and energetic person at the right time in my education. She kindled my interest in research and persuaded me to continue into graduate school. The opportunity to work with such a dynamic person at this stage is one I will always treasure. I would also like to thank the other members of my committee, Dr. Doug Gage, Dr. Michael Garavito, Dr. Bill Punch, and Dr. John Wilson for their guidance during my graduate tenure and for keeping my disparate research interests on a single track. Additional appreciation goes to Dr. Michael Kron, who piqued my interested in working with asparaginyl-tRNA synthetase and provided valuable input on the synthetase portion of my thesis work, and to Dr. Stephen Cusack, who provided the structure of the *Brugia* AsnRS. Also, thanks go to Dr. Laurie Kaguni and Dr. Li Fan for their great assistance in the phage library project I undertook, which went infinitely further with their help than it would have otherwise.

I would also like to thank the American Heart Association and their Michigan affiliate for funding under grants to Dr. Kuhn.

I also am thankful for current and past members of the MSU Protein Structural Analysis and Design group, including Carrie Barkham, Sridhar Venkataraman, Vishal Thakkar, Trevor Barkham, Rajesh Korde, and Brandon Hespenheide for providing an exciting and stimulating research environment and for their many ideas. I am especially grateful to Volker Schnecke for helping me to start on the main portion of my thesis research, to Maria Zavodszky for working with SLIDE alongside me and ensuring that I had a good grasp of the theory and methods behind the work, and to Michael Raymer for his immeasurable help in getting started in computational research, his innumerable ideas and careful criticisms, and his important friendship.

Finally, I would like to thank my family, who made me realize both the importance and the joy of learning so many years ago. I would not have come this far if it wasn't for you. Of special importance is the support in so many ways my wife, Suzanne, has given me over the last 9 years. Her constant encouragement and understanding have helped me get past the times of slow progress and difficult work and have meant more to me than I could ever express.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

ACD	Available Chemicals Database
ADN	atomic density
AHP	atomic hydrophilicity
AsnRS	asparaginyl t-RNA synthetase
AvgPrBVAL	average protein B-value, average B-value of the protein environment surrounding a water molecule
BPTI	bovine pancreatic trypsin inhibitor
BVAL	B-value, Debye-Waller factor
CASP2	Second Meeting on the Critical Assessment of Techniques for Protein Structure Prediction
CMC	Comprehensive Medicinal Chemistry (database)
COX-2	cyclooxygenase-2
CSD	Cambridge Structural Database
DES	diethylstilbestrol
DFT	discrete Fourier transform
DHFR	dihydrofolate reductase

DME	distance matrix error
ER	estrogen receptor
GA	genetic algorithm
GST	glutathione S-transferase
HTS	high-throughput screening
IleRS	isoleucyl-tRNA synthetase
LGA	Lamarckian genetic algorithm
MHC I	major histocompatibility complex class I
MMFF	Merck molecular forcefield
MMP3	stromelysin-1 (matrix metalloproteinase 3)
MOB	mobility, normalized protein mobility measure
PDB	Protein Data Bank
PLP	piecewise linear potential
PMF	potential of mean force
PPACK	phenyl-prolyl-arginiyl-chloroketone (Phe-Pro-Arg-chloroketone)
PrBHBD	protein hydrogen bonds, number of hydrogen bonds formed between a water molecule and a protein
RMSD	root mean square deviation
SAS	solvent accessible surface
S-AMP-Asn	S-adenosyl-asparagine
SG-DOCK	similarity-guided docking

SLIDE	Screening for Ligands with Induced-fit Docking Efficiently
SP-DOCK	similarity-penalized docking
TPrBVAL	total protein B-value, sum of the B-values for protein atoms in a water molecule's environment
WatHBD	water hydrogen bonds, number of hydrogen bonds formed between a water molecule and other water molecules
WDI	World Drug Index

Chapter 1

Computational Docking and Screening Methods – A Review of Algorithms, Scoring Functions, and Applications

1.1 Protein-Ligand Binding Sites

Proteins perform many of the processes that are responsible for giving life to the cell. While proteins play important structural roles, much of the interest in proteins involves the study of their roles in catalytic and signaling events, both of which involve binding to other molecules. The molecular partners of proteins range in size from single water molecules and metal ions to large, multimeric protein complexes. Understanding the mechanisms of protein binding is a key to understanding the function of proteins and the particular niche each protein inhabits in the large, extraordinarily complex system of a living organism.

The classic view of protein-ligand binding is the "lock-and-key" concept (Fischer,

1894), in which the ligand "key" acts as the complement to the binding site "lock". Part of the interaction between protein and ligand is simply the steric fit of the two pieces, i.e., similar to a key fitting a lock. Of course, chemistry also plays an important in establishing complementarity between the proteins and their ligands. Examination of protein surfaces showed that, on average, 57% of the surface involves non-polar residues (Miller et al., 1987). Interaction between proteins in regions of non-polar character are predominantly steric in nature. Given that the remaining 43% of the surface of a protein consists of polar and charged residues, it is not surprising that interactions between these types of residues also play a significant role in ligand binding. Since the non-polar interactions are generally non-specific in nature, the polar and charged residue interactions provide the specificity of interaction. A study of 15 protease-inhibitor complexes and 4 antigen-antibody complexes determined these complexes generally form a large number of protein-ligand hydrogen bonds (Janin and Chothia, 1990).

Since various residue types can contribute similar chemical properties to a ligand binding site, it is interesting to examine if there is a preference for any specific residue types in ligand-binding sites. In a study of 46 monomeric proteins, Miller and colleagues (1987) showed that some residue types are significantly overrepresented (lysine, serine, and glycine) compared to the overall amino acid composition of proteins, while others, (methionine) are significantly underrepresented. A similar analysis focused on protein-ligand binding sites in 50 crystallographic complexes was performed (Villar and Kauvar, 1994). They showed an overabundance of tryptophan, histidine, and tyrosine located close to the ligand, indicating that these may play an important role in ligand binding. Proline

is found to be significantly underrepresented in the binding site relative to its abundance in proteins overall, perhaps due to its general structural role, which may be important internal to the protein and away from the binding site. Lysine is also significantly underrepresented in ligand-binding sites, while it constitutes a higher than expected proportion of the protein surface. In addition to examination of the relative abundance of each amino acid type in ligand-binding sites relative to the over-Al protein surface, they studied the percentage of the binding site, defined as those amino acids within 4.0 Å of the ligand, made up of each of the amino acids. This examination showed that glycine, serine, arginine, and tyrosine predominate in the binding site. The distributions for the binding sites are significantly different from those for the protein surface and for the protein overall.

The key may not be quite the perfect match to the lock as a simple lock-and-key model would suggest, even with the inclusion of chemistry as part of the lock and the key. Many proteins exist which bind a variety of ligands. The major histocompatibility complex class I (MHC I) binds a large variety of peptides with high affinities (Wilson and Fremont, 1993). One way to extend the "lock-and-key" concept is that of a flexible lock and/or a flexible key, i.e., the protein and/or ligand changes conformation during ligand binding. A recent study of 39 complexes showed that proteins generally undergo some conformational change upon ligand binding (Betts and Sternberg, 1999), including moderate backbone conformational changes in addition to side-chain movements. Inclusion of protein and ligand flexibility greatly increases the complexity of computational models for protein-ligand binding. Several docking and screening methods, including their approaches to handling flexibility, are discussed below.

1.2 Computational Docking Methods

Computational docking can be described as the process of modeling the binding orientation of a specific ligand to a specific protein of interest, i.e., a "receptor". Docking and screening methods (screening is described below in Section 1.3) provide for a further understanding of the mechanisms involved in protein-ligand binding in general, as well as helping to understand the details of the interactions in a specific protein or protein-ligand complex of interest. In addition to gathering such knowledge for understanding the processes, this knowledge can be used to improve the design of ligands with high specificity and high affinity toward a specific target.

1.2.1 Ligand Manipulation Docking Methods

The classical algorithm implemented for computational docking is that of DOCK (Kuntz et al., 1982; DesJarlais et al., 1988; Shoichet et al., 1992; Ewing and Kuntz, 1997). DOCK operates by generating a set of spheres to describe the volume, or negative image, of the binding site and uses the centers of these spheres as sites for matching to ligand atoms. Sets of receptor spheres are matched to sets of ligand atoms to generate a ligand orientation, which can then be scored according to their complementarity with the protein. The early internal DOCK scoring function, GRID (Meng et al., 1992), is a grid-based scoring function in the method of Goodford and colleagues (1985). Later implementations have used more robust scoring functions which are also grid-based. It is possible to use the ligand docking method of DOCK with an externally supplied scoring function. The initial

implementation of DOCK used only steric fit and electrostatics as a determinant for ligand docking, but later versions implemented chemical type matching to better model chemical complementarity between ligand and receptor groups, including hydrogen bonding interactions. It should be noted that DOCK uses only rigid-body translations and rotations, including no flexibility within the docking algorithm.

Several improvements have been made to the DOCK algorithm since its inception, including extension to include information about known ligand docking orientations through the development of similarity-penalized docking (SP-DOCK) and similarity-guided docking (SG-DOCK; Fradera et al. 2000). During docking with the SG-DOCK algorithm, the docking score is weighted according to the similarity, defined by MIMIC (Mestres et al., 1997) to a known ligand structure and/or pharmacophore structure, the spatial arrangement of key ligand functional groups identified by analysis of a set of known ligands, for each scoring during the docking. This causes the dockings to be biased towards that of the known ligand/pharmacophore. When using SP-DOCK, the final scores of dockings performed without a pharmacophore bias are weighted by the similarity, which results in a resorting of the final docking orientations. The similarity measure plays a much more important role in the SG-DOCK procedure as it can drive the orientational and conformational search.

Another adaptation of DOCK involves the observation that protein-protein interfaces generally contain more hydrophobic contacts than other surface regions. Vakser and Aflalo (1994) implemented an algorithm which reduces the model of the ligand protein surface to include only points attributed to hydrophobic atoms. This algorithm resulted in only a

small improvement in docking predictions for three of the four cases tested by the authors compared to standard DOCK, but this method uses a reduced surface representation, an important consideration for computationally intensive docking and screening methods, and is more tolerant to conformational changes.

Another rigid body docking procedure, developed for protein-protein docking, is the PUZZLE algorithm (Helmer-Citterich and Tramontano, 1994), which maps protein surfaces into two-dimensional matrices, consisting of distances between adjacent points on the surface along the edge of a surface slice of fixed height, and then identifies matching submatrices. The PUZZLE algorithm has been modified to include a more comprehensive scoring function and improved mapping of small protein surfaces, dubbed ESCHER (Ausiello et al., 1997). This algorithm operates by slicing each of the proteins, describing a polygon for each slice, and then finding complementary shape matchings for these polygons by way of translations and rotations of one relative to the other. Alternate sets of polygons are generated by three-dimensional translation and rotation in fixed increments. Matching polygons are then scored based on steric and electrostatic parameters. A shortcoming of this procedure is the relatively coarse set of three-dimensional rotations employed, only taken at 10° increments. The ESCHER program is generally able to achieve correct dockings for well-buried ligands, but fails to do so for ligands which are bound to a shallow binding site.

Additional rigid-body docking algorithms have been developed. One is that developed by Fischer and colleagues (1995). This algorithm is a geometric-based approach, in which the protein and the ligand are represented by "critical" points and sets of points

are matched using geometric hashing, whereby the geometry for the spatial arrangement of points of potential interaction are precomputed and entered in a lookup table for later reference. Once possible matchings are identified, they are scored based on the contact area between molecules and their electrostatic interactions. This procedure was able to dock test ligands to within 1.5 Å root-mean-square deviation (RMSD) relative to the position of the ligand in the crystallographic complex in 18 of 19 test cases, though often the best docking in terms of RMSD was not ranked near the top based on scoring. Scoring of computational dockings, equivalent to predicting binding affinity, is a key component of identifying the "best" docking and remains a significant challenge. A discussion of scoring functions is presented below in Section 1.4. Another rigid-body docking algorithm is the LIGIN program developed by Sobolev et al. (1996). In LIGIN, a complementarity function including terms for favorable contacts, unfavorable contacts, and the contact surface is defined. Minimal receptor flexibility is modeled by allowing the user to define one or more residues whose side chains, from C_β to the side-chain termini, are ignored in calculation of the complementarity function. LIGIN was tested on a set of 14 complexes and was able to dock all ligands with reasonable RMSDs relative to the crystal structure.

More complex approaches to handle conformational flexibility in ligand docking have been developed. In the latest version of DOCK, limited ligand flexibility is modeled by the use of rigid-body dockings of ligand conformers, with later versions of DOCK including an internal conformation generator using a genetic algorithm (GA; Oshiro et al. 1995). An alternative to stochastic sampling of the rotational degrees of freedom is to use previously observed low energy states. The algorithm of Leach (1994) implements such a technique

to explore both protein side chain flexibility, via the use of rotamer libraries, and ligand flexibility, via the use of conformational analysis. Instead of relying on rigid-body dockings of ligand conformations, flexibility is addressed more directly in the latest version of DOCK using an incremental construction algorithm based on that developed by Leach and Kuntz (1992). Incremental construction docking begins by placing a fragment of the ligand into the binding site and then adding functional groups to build up the ligand. Incremental discussion methods are further discussed below in Section 1.2.3.

A further derivative of DOCK is FLOG (Miller et al., 1994). FLOG uses the same matching approach of DOCK, but expands the types of atoms assigned based on their chemistries. It also includes some ligand flexibility by including generated ligand conformations in the database, but the docking procedure remains a rigid-body one. The authors of FLOG developed an enhanced grid-based scoring function which includes electrostatic, hydrogen bonding, hydrophobic, and van der Waals potentials. FLOG is able to select known inhibitors from a large database of drug-like molecules.

In all docking algorithms, there is a tradeoff between how extensively and accurately the orientation and conformational space is explored and the computational requirements to perform the exploration. Several methods use stochastic approaches to the docking problem to achieve a higher degree of accuracy at the expense of exploration. A popular algorithm is AutoDock (Morris et al., 1996), which employs a Monte-Carlo simulated annealing method to sample binding orientations and ligand conformations, by randomized rotation of rotatable torsions in the ligand. AutoDock is inexpensive, easy to use, achieves reasonable dockings, and has been applied to several cases (Goodsell et al., 1996). An extension to

AutoDock replaces the Monte-Carlo search with a Lamarckian genetic algorithm (LGA), a hybrid GA and local search method (Morris et al., 1998). The LGA differs from a traditional GA by its performing a finer local search on the orientation of the ligand relative to the protein and of the ligand's conformation.

A second stochastic approach based on the use of a GA was developed by Jones and colleagues (1995). Their approach uses a simple GA operating on rotational angles in the protein, rotational angles in the ligand, and on hydrogen bonds from ligand to protein and protein to ligand. The fitness or scoring function encompasses terms for the hydrogen bond energy between protein and ligand, for the van der Waals energy between protein and ligand, and for the internal van der Waals energy of contacts within the ligand. Improvements to this algorithm resulted in the development of the GOLD algorithm (Jones et al., 1997), with changes in the representation of angles and hydrogen bonds and inclusion of a more robust scoring function. The GOLD algorithm achieved "acceptable" dockings for 71 of 100 test cases. Similar to GA approaches is the evolutionary programming approach AG-DOCK (Gehlhaar et al., 1995b; Verkhivker et al., 1999), which is able to generally correctly dock, defined here as docking within 1.5 Å of the complex crystal structure, methotrexate into dihydrofolate reductase (DHFR) and a proprietary ligand, AG-1343, into HIV protease.

Another approach is the ICM algorithm (Abagyan et al., 1994; Totrov and Abagyan, 1997), which uses a complex procedure to finely dock a ligand. In this approach the molecules are represented by a set of internal variables (rotatable angles, e.g., rotatable side-chain single bonds) for their relative positions and conformations, which are randomly

changed, energy minimized, and selected via the Metropolis algorithm for each step. This approach successfully docked lysozyme to its antibody (Totrov and Abagyan, 1994) and β-lactamase and its inhibitor (Strynadka et al., 1996), but the study presented here docked only three of eight ligands used in the Second Meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP2; Moult et al. 1997) with an RMSD relative to the crystal structure of less than 5.0 Å. Generally, dockings with RMSDs lower than 2.0 to 2.5 Å are considered as correct.

The SLIDE algorithm, which constitutes a major focus of this thesis work, belongs to this first, ligand manipulation, class of docking algorithm types and is described in Chapter 3.

1.2.2 Recombination Docking Methods

An alternative to the approaches described above, in which the docking is performed via manipulation of a ligand molecule, is the approach of placing each piece of the ligand independently and then linking the docked fragments to form a docking of the complete ligand. One implementation of this approach is the empirically based GEMINI algorithm (Singh et al., 1991), which docks peptidyl ligands. In this method, a database of side-chain packing arrangements generated from 52 protein structures (Singh and Thornton, 1990) is used to map potential orientations of each peptide ligand side chain relative to the protein binding-site side chains. The three-dimensional distribution for each ligand residue/protein residue type pair is orientated based on the protein side chain's positions. These distributions can then be superimposed to identify potential orientations of the ligand

side chains which simultaneously occur in regions of high frequency in the distributions. Some simple constraints on covalent bond formation to other ligand side-chains, i.e., along the peptidyl ligand chain, can be included to yield a small set of possible ligand binding orientations and conformations. The technique has the limitation that it is applicable to only peptidyl and, perhaps, a few other limited types of ligands as an empirical database of side-chain or functional group orientations must be known. This requires that a statistically significant number of example interactions exist in the structural database, which is the case for peptidyl packings as one can use the proteins in the Protein Data Bank (PDB; Berman et al. 2000; Bernstein et al. 1977). However, for many ligands of interest, there may be only a few or no structures with similar groups bound.

The algorithm developed by Sandak and colleagues (1998b) is another recombination-based approach. In this method, the ligand is represented by a set of predefined hinges and set of interaction points, representing sites of favorable interaction. Each triplet within a part of the ligand, i.e., portions of the ligand on one side of the hinge, are matched by triplets of interaction points in the protein. The best dockings of each piece and hinge orientation are tabulated, recombined, and then scored using an interatomic contact scoring function. This algorithm can be applied to hinges in either the ligand or the receptor, but not both, during a docking run. Acceptable dockings were achieved for synthetic peptides binding to calmodulin and to HIV protease (Sandak et al., 1998a), which undergo a clamping motion upon ligand binding.

1.2.3 Incremental Construction Docking Methods

A third docking approach is that of incremental construction, which initially docks only a portion of the ligand and then builds up the remaining ligand groups based on this initial placement. The classical algorithms for this docking approach are GROW (Moon and Howe, 1991), for peptidyl ligands, and FLEXX (Rarey et al., 1996a), for generalized ligands. In FLEXX, the flexibility of the ligand is described by torsion angle rotations at discrete steps as in the MIMUMBA algorithm (Klebe and Mietzner, 1994). In the first stage of docking, the base, or anchor, fragment is chosen, which is then placed into the binding site. This placement in done based on local interaction surfaces as described in Rarey et al. (1996b), involving a pose, or orientation, clustering step. Once the anchor fragment is placed, in one or more orientations, a tree representing the possible conformations for adding the remaining pieces of the ligand is constructed. Each node of the tree represents a set of currently docked fragments and their orientations and conformations, with the base node(s) representing the base fragment docking(s). Branches from a node represent the next ligand fragment to add, with each branch including the identity of the fragment and the rotation about the newly formed bond, set to one value in a set of fixed increments of the rotation angle. Ligand groups connected by a rigid, i.e., nonrotatable, bond are considered a single fragment, so the bond formed by all additions is rotatable. For example, branches from a current node may include a methyl group and a rotation angle of 10°, a methyl group and a rotation angle of 20°, etc. to 360° (36 branches), and a second set of branches from the same node for a carboxyl group and rotation angles of 10–360° (36 branches), for a total of 72 branches. Due to the computational complexity of exploring the complete tree,

only the top k, with k set to 500 in the work presented by the authors, scoring branches from each node are explored. The solution set, consisting of all dockings, is then clustered to reduce dockings that are very similar which arose from independent traversals of the docking tree. Scoring is done with a energy function derived from that of Böhm (1992a,b). This algorithm was tested on a set of 19 protein-ligand complexes which had between 1 and 8×10^{10} theoretical ligand conformations, i.e., combinations of incremental torsional angle rotations in the ligand, including ligands which may contain internal steric collisions. In 15 of 19 cases, a docking with RMSD from the crystallographic ligand of less than 1.0 Å was achieved, but in some cases the lowest energy docking was significantly different from the crystallographic complex, with RMSDs up to 4.5 Å. Also, the energy scoring function used in this work generally did not correlate well with the experimental binding energy.

A later study on the performance of FLEXX using a test set of 200 protein-ligand complexes was done (Kramer et al., 1999). For this test set, the top ranked docking was within 2.0 Å RMSD of the ligand in the crystallographic structure in 47% of the cases, and a docking within 2.0 Å of the crystallographic ligand position was found in 70% of the cases if all dockings are examined, not just those with the best score. In general, the FLEXX algorithm achieves better dockings with simpler ligands. More complex ligands, those with more than 15 components, yielded a correct docking in only 25% of the cases. This study also showed that, in general, the algorithm is able to cross dock most ligands, i.e., dock a ligand from one crystallographic structure into the conformation of the protein extracted from another crystallographic structure that contains the protein bound to a different ligand.

To address side-chain and main-chain flexibility of the receptor structure in docking, an

extension to FLEXX, FLEXE (Claussen et al., 2001), was created. FLEXE docks ligands into an ensemble of structures of the receptor instead of a single receptor binding site. The binding site ensembles can be from different crystallographic structures, as in the study described here, from a homology model with uncertain side-chain positions, from a series of molecular dynamics time steps, or from another source. The key component of this algorithm is that multiple conformations of the protein can be used as a docking target simultaneously. In this algorithm, the receptor structures are merged, with regions of similar conformations reduced to a single structure and regions with dissimilar conformation constituting alternate positions. While this algorithm may handle some backbone movement in addition to side-chain rotations, the authors claim it is not able to work with large domain movements and limit their test set to protein ensembles with similar backbone traces. FLEXE was able to generate a docking ranked in the top ten potential dockings which was within 2.0 Å of the crystallographic structure in 67% of the 105 structure test set, compared to 63% with the FLEXX algorithm. In addition, FLEXE was able to effectively cross dock two potent inhibitors which FLEXX was not.

1.2.4 Water in Computational Docking

Water molecules are known to play a key role in many protein-ligand complexes, reviewed by Ladbury (1996), but are often ignored in computational docking approaches as they are difficult to model. Many of the scoring functions do consider the energy of desolvation, generally as a function of buried and exposed hydrophobic surface area of the ligand and/or protein, but there are two problems with using this as the only water modeling technique:

(1) it assumes that the best protein-ligand interface will be completely desolvated, while in many instances water molecules are retained in the interface and contribute key hydrogen bonding interactions, and (2) the shape of the binding pocket surface to which the ligand binds in reality, which includes the water molecule(s), is different from the surface used in the docking, which does not include the water molecule(s). The second point is especially important as steric fit between the ligand and the protein is often a key component of the docking and scoring procedures.

Since there are no covalent constraints on their positions relative to the ligand or the protein, treatment of water molecules in docking procedures must be different from the procedures used to model either the protein receptor or the ligand molecule. One approach is to retain bound water molecules from the free protein structure as a fixed part of the protein target in the docking process. However, this can lead to overrepresentation of interfacial water as roughly two-thirds of the water molecules in the binding site of the protein are lost upon ligand binding (Raymer et al., 1997). A method to model protein-bound water sites conserved upon ligand binding is to predict which bound water molecules will be conserved and which will be displaced. This approach is used with the *k*-nearest-neighbor genetic algorithm application *Consolv* (Raymer et al., 1997), allowing those water sites predicted to be conserved to be included in docking, while removing those predicted to be displaced. Because this prediction is not 100% correct and will depend, to some extent, on the ligand shape and chemistry, water molecules wrongly predicted as conserved may incorrectly bias the docking. A better method may be to allow the docking algorithm to displace any water molecule, but penalize the displacement of water molecules according

to their predicted likelihood of being conserved, as is done with SLIDE (Section 3.2.4). FLEXX uses a different idea to address the problem; it places water molecules at favorable sites during the docking process by placing "particle phantoms" at favorable positions in the binding site prior to docking. These phantoms can be turned on during docking when they can make favorable interactions to the protein and to the growing ligand and turned off if they are involved in steric collisions or fail to meet angular hydrogen bond constraints (Rarey et al., 1999). Another method is that developed by Jackson et al. (1998), which uses a finely spaced grid of potential hydration sites around the ligand as it is being docked. The energy of the docked ligand can be computed including hydration at subsets of the grid points, yielding favorable hydration sites for the docking.

1.3 Computational Screening

Computational screening is the process of identifying molecules that bind to a protein of interest from a database using computational methods, as reviewed in Walters et al. (1998). It can be considered as a computational equivalent to traditional high-throughput screening. In theory, any of the docking methods described above can be used in a screening mode by attempting to dock all molecules in a database, such as the Cambridge Structural Database (CSD; Allen and Kennard 1993). In practice, however, this method is not practical due to the time needed to dock each molecule. Most of the docking algorithms described above are reported to take several minutes to hours to dock a single ligand. Most of the molecular databases of interest contain 100,000-plus molecules. Even if the fastest method is used, at one minute per ligand, a screening of 100,000 molecules would take over two months

of computational time. When factoring in the desirability of docking 10–20 different conformers of each ligand to better represent their flexibility, the computational time expands to years for a single database screening. One approach is to develop a docking algorithm designed for screening, such as the SLIDE algorithm originally presented by Dr. Volker Schnecke and extended as described in this thesis (Chapter 3).

An alternative screening approach to database docking is closely related to the incremental construction docking algorithms. In such an approach, the ligand backbone is initially docked, followed by the addition of various functional groups, akin to combinatorial chemistry methods. This approach is used in LUDI (Böhm, 1992a,b), Grow (Moon and Howe, 1991), SPROUT (Gillet et al., 1993, 1994), GroupBuild (Rotstein and Murcko, 1993b), BUILDER (Roe and Kuntz, 1995), HOOK (Eisen et al., 1994), and SMOG (DeWitte and Schaknovich, 1996). A technique in which the ligand is built up atom by atom instead of fragment by fragment is employed in the Genstar (Rotstein and Murcko, 1993a), Legend (Nishibata and Itai, 1993), MCDNLG (Gehlhaar et al., 1995a), and CONCEPTS (Pearlman and Murcko, 1995) programs.

Closely related to these incremental construction methods for screening is the FLEXS algorithm (Lemmen et al., 1998). This method requires knowledge of the binding orientation for at least one ligand, which is used as the rigid component in flexible alignment. The structures in the screening database are broken into fragments, as in the previous incremental construction screening methods; however, the base fragments, and later the flexible groups, are placed based on matching chemical properties, such as hydrophobicity, hydrogen bonding character, and partial charge, with the known ligand instead of the protein

binding site. FLEXS was able to extract and highly rank ligands with known function against the fibrinogen receptor from a database of 984 drug-like compounds. In terms of docking ability, FLEXS was able to reasonably reproduce dockings resulting from superposition of crystal structures for test set consisting of 14 protein targets and a total of 284 superimposed ligands.

1.4 Docking and Screening Scoring Functions

One of the difficulties in computational docking and scoring methods is that of accurate scoring, measuring the affinity between protein and ligand. The scoring function is key to being able to identify the most realistic ligand dockings, in docking experiments, or the most promising potential ligands for further study, in screening experiments (Greer et al., 1994). Scoring functions can be classified into two general categories: molecular mechanics-based and empirical.

Molecular mechanics-based scoring functions are those that describe the energy of ligand binding in terms of a physical chemistry function summing the component binding energies, such as van der Waals contacts, electrostatic interactions, and covalent bond stretching, bending, and torsional energies. Such scoring functions includes the forcefield-based measures such as AMBER (Weiner et al., 1984, 1986), which is not commonly used in docking and screening methods as it has been tuned to reproduce protein and nucleic acid energies, rather than arbitrary small organic ligands. A commonly used molecular mechanics-based scoring function for docking and screening is the Merck Molecular

Forcefield (MMFF; Halgren 1996), which was developed as a combined "organic/protein" forcefield for molecular dynamics of such systems. One problem with this scoring function is that it is highly detailed. While this feature enables it to be relatively accurate, it also makes the function computationally expensive, which causes significant problems when scoring a large number of docking orientations for many potential ligands. One method to reduce the computational cost is to precompute the atomic potential for the protein on a grid, so that only the ligand potential changes with each new cycle. This is done with GRID (Meng et al., 1992), implemented in DOCK, and FLOG (Miller et al., 1994) scoring functions. Other grid-based scoring algorithms developed by Blom and Sygusch (1997) and Mandell et al. (2001) in DOT uses a discrete Fourier transform (DFT) correlation approach to solve a moderately complex molecular mechanics-based scoring function with reasonable computational cost. Docking experiments done with this scoring function on four complexes were able to dock the ligand molecule close to the observed orientations in the crystal structure. The docking time of the DFT correlation approach is generally low, but much too slow, on the order of 2 to 30 hours per ligand depending on grid resolution and ligand size, to be of use in screening experiments.

Given their extensive use in docking and screening algorithms, a brief note about the general methodology behind grid-based scoring procedures is warranted. In general, a three-dimensional grid is established in the binding site to calculate energies. The initial step is to precompute the energy of interaction between the target and each type of atom residing at each grid point. After this step, the scoring function constitutes a lookup table of the energy for each type of atom at each grid point. When a ligand is placed into the binding

site in an arbitrary orientation during the docking procedure, each atom of the ligand is assigned to the nearest grid point. The energy for the docking orientation is then calculated as the sum of the energies for those grid points to which ligand atoms have been assigned, using the assigned ligand atom type energy calculated for the grid point. Since grid-based scoring functions are precomputed prior to docking and screening and simply looked up during the run, they are very fast to compute during the actual docking or screening run. The drawback of grid-based scoring methods is that they use only an approximate ligand atom position. Problems resulting from this approximation can be limited through the use of a finely-spaced grid; however, this can greatly increase the precomputation time, especially with complex scoring functions, and somewhat increase the run time, due to the ligand atom to grid point assignment step.

Most of the commonly used docking and screening tools implemented use empirically tuned scoring functions. One of the earliest such scoring functions is SCORE1, developed by Böhm for the LUDI program (Böhm, 1992a, 1994). SCORE1 seeks to calculate the binding affinity, $\Delta G_{binding}$, as the sum of the energy from polar interactions, including hydrogen bonding and ionic contact; nonpolar interactions, constructed from the surface area buried in the complex; and flexibility, defined as an "energy" for loss of the ability to rotate a rotatable bond in the ligand. This scoring function was generally able to predict binding energies close to those experimentally observed. Examination of complexes with large deviations between the predicted and observed binding energies led to the development of SCORE2 (Böhm, 1998), which includes additional parameters for penalizing cavities in the binding site, an improved electrostatics model, an term for aromatic group

interactions, and a term for desolvation effects. This improved scoring function improved the correlation between predicted and observed K_i values and reduced the standard deviation in predicted K_i from 1.7 orders of magnitude to 1.3 orders of magnitude.

Other similar energy-based scoring functions have been developed. One example is the VALIDATE scoring function developed by Head et al. (1996), which combines molecular mechanics approaches with empirical descriptors. The enthalpy of binding in the protein-ligand complex is computed from a molecular mechanics forcefield, while additional properties, including fixing of rotatable side chains, buried surface area complementarity, and steric compatibility, are used to estimate the entropy of binding. Overall the VALIDATE scoring functions consists of 12 terms, and while accuracy is quite good, the estimated cross-validation error for a set of 51 complexes in predicted K_i was 1.1 order of magnitude, the time to compute the score is over 1000 times slower than the original Böhm SCORE1 function. Another similar scoring function is that developed by Jain (1996), which includes terms for hydrophobic complementarity, for hydrophilic complementarity (hydrogen bonding and salt bridges), for electrostatic repulsion, for desolvation, and an entropic term. This function was tested on a set of 34 protein-ligand complexes and resulted in a mean predicted K_d error of 0.7 orders of magnitude. Yet a third is the piecewise linear potential (PLP; Gehlhaar et al. 1995b,Verkhivker et al. 2000), which simplifies the energy function to only four terms to yield a piece-wise linear approximation of the hydrogen-bond and lipophilic interaction wells. A fourth approach is the ChemScore algorithm developed by Murray and colleagues (Eldridge et al., 1997; Murray et al., 1998), which includes terms for hydrogen bonding interactions, metal interactions, hydrophobic interactions, and a term

for loss of entropy, i.e., rotational freedom. This function predicts the observed affinities with a cross-validated error of 1.3 orders of magnitude.

Several additional methods have been developed which take advantage of the structural information available in the PDB. In all of these methods, the key is the creation of a statistical distribution for various contacts between atom types in the ligand and in the protein. One such example is that developed by Mügge and Martin (1999), which uses the statistical information to derive a set of free energies of protein-ligand atom pair interactions, or potentials of mean force (PMFs), which are summed to provide the final score. The method of deriving the PMFs implicitly includes entropic and desolvation effects. Tests performed on a database of eight protein-ligand complex sets, which each contained between 11 and 77 complex structures, showed reasonable correlation between the score and observed binding affinities. The authors note that one set gave poor correlation, likely due to the large and variate size of the inhibitors.

Other methods which use statistical information derived from structures reduce the representation further away from one of energetics to include only the distribution information. An early implementation of such a method is that developed by Klebe (1994), which extracts information on interaction angles from structures in the Cambridge Structural Database. This method was able to reasonably predict the key binding sites for methotrexate in dihydrofolate reductase (DHFR), for a peptidic inhibitor in endothiapepsin, and for tyrosinyladenylate in tyrosyl-tRNA synthetase. This method has been extended by Nissink et al. (2000) through the use of Isostar (Bruno et al., 1997), which tabulates the orientations of specific chemical groups with respect to another specific chemical group of interest

from PDB and CSD structural data. In this method, the orientation data are used to derive propensity plots, or smoothed spatial distribution functions, which can be used as a basis for scoring the spatial relationships in computationally docked ligands. A second extension of Klebe's work is the development of DrugScore (Gohlke et al., 2000a,b). In DrugScore, the model of spatial relationships is further reduced to a set of one-dimensional radial distribution functions for each set of atom-type interactions. DrugScore was shown to be an improvement when examining the RMSD between the top-ranked ligands and the ligands in the crystallographic structures during docking experiments with FLEXX (Kramer et al., 1999). The top-ranked docking was within 2.0 Å RMSD from the crystallographic structure in 73% of 91 test protein-ligand complexes using DrugScore, versus only 54% for the FLEXX scoring function. DrugScore resulted in standard deviations in predicted pK_i values of 0.7 to 2.2 on a test group of 9 test sets, consisting of between 16 and 71 complexes, though linear correlations with observed binding affinities were poor for some of the test cases.

Stahl and Rarey (2001) compared several scoring functions in terms of the docking algorithm of FLEXX for a computational screening procedure. The test database consisted of seven proteins, with 36 to 128 known inhibitory compounds per protein. The FLEXX, PLP, DrugScore, and PMF scoring functions were tested. In general, the FLEXX scoring function performs best with compounds for which the binding is dominated by hydrogen bonds, such as for thrombin and neuraminidase, but poorly with compounds whose binding is predominated by hydrophobic interactions, such as for cyclooxygenase-2 (COX-2). In contrast, the DrugScore algorithm performed well with the COX-2 screening, but poorly

for neuraminidase. The PLP algorithm tended to perform well with shallow sites that still had significant hydrogen bonding interactions, while the PMF algorithm performed poorly with very narrow and/or restricted binding sites.

Given the preponderance of scoring functions available, one may consider a method of combining a set of functions to overcome shortcomings in any one particular scoring function. This approach was implemented by Charifson et al. (1999) and by Stahl and Rarey (2001), who both showed improvements in screening efficiency, in terms of ranking active compounds highly compared to inactive compounds. Both tested against a set of diverse targets, with consensus scoring providing an overall increase in performance across the set. In some cases, the consensus method performed worse for a specific target, but without *a priori* knowledge of which scoring function may be best suited for the target of interest, it is not possible to improve on the overall performance. When examining combinations of only two scoring functions, as is done in the thesis work presented here (Section 4.3.3), there is a very limited number of ways to combine the scoring function results. If more scoring functions are included, the possible ways to combine them increases. Wang and Wang (2001) present a computational experiment to explore the effects of different combination methods of the resulting overall selection of ligands from a computational screening run. They constructed a set of virtual ligands by assigning a "true energy", i.e., equivalent to the experimentally observed energy, based on a Gaussian distribution. This experimental dataset was then used to create ten predicted datasets by adding an random error to each experimental score, representing the error introduced by each of ten independent scoring functions. They showed that use of increasing numbers of scoring functions

results in an increase in the ability to include an "active" ligand in the top 100 ligands from the screening database using two combination methods, but the increase in performance becomes significantly less when including more than four to five scoring functions. The use of additional scoring functions is undesirable for docking and screening applications since each additional scoring function evaluated will require additional computational time, which is at a premium during docking and screening runs. Another concern arises out of statistical pattern recognition and that is the fact that, with a finite amount of data, overall scoring accuracy is likely reach a peak and then decline as additional scoring functions, i.e., dimensions, are added to the system. This phenomenon is often termed the *curse of dimensionality* (Jain and Chandrasekaran, 1982). A classic construction to illustrate such a case was presented by Trunk (1979).

One concern with all empirically tuned scoring functions is that of the measurement of the binding affinities. Generally, the observed affinity values are taken from literature sources; however, it is very uncommon to have affinity values measured at standardized conditions, even for a set of ligands to a single protein. By tuning the scoring function to these observed values, one makes the assumption that differences in the observed values due only to differences in affinity for the ligands. One must question the validity of tuning the scoring functions to data which may not be self-consistent, and one can ask how much of the deviation seen between predicted values and observed values is likely due to inaccuracies in the scoring function and how much is due to differences in experimental methods and conditions used to measure the binding affinity.

One interesting idea related to scoring is the use of a post-processing filtering step

(Stahl and Böhm, 1998). A limitation of most scoring functions is the absence of strong penalties for unfavorable interactions, especially in terms of leaving cavities in between the buried surfaces of the protein, which are not generally seen in protein-ligand complexes. The filtering algorithm presented in this study includes four terms: the size of cavities in the internal protein-ligand interface, the portion of solvent accessible surface (SAS) of the ligand which is hydrophobic, the fraction of ligand volume buried in the binding cavity, and the presence of pairs of polar atoms in close contact which do not participate in hydrogen bonds. Testing against a set of 32 complexes docked with FLEXX showed a general decrease in the RMSD of the best ranked ligand and a dramatic increase in the number of complexes for which the the docking closest to the crystal structure was ranked within the top 20 ligand dockings. The authors also note than the docking with the best RMSD relative to the crystal structure was not lost after filtering for any of the 32 complexes. For ligand manipulation docking and screening methods, e.g., SLIDE, a filtering step could be directly incorporated into the scoring function.

1.5 Evaluation of Docking and Screening Methods

As there are several docking and scoring methods available, it is interesting to examine their performance on identical problems. Performance of various scoring functions implemented in FLEXX is discussed above. One of the key studies examining each of the docking algorithms independently is the docking section of CASP2, the results of which were summarized by Dixon (1997). In this study, seven small molecule-protein complexes and one protein-protein complex were used as targets. Target protein structures were pro-

vided to several docking algorithm research groups along with the two-dimensional ligand structures, but no ligand conformational, i.e., three-dimensional, information was given. Of the approaches discussed above, the ICM method of Abagyan et al. (1994), FLEXX (Rarey et al., 1996b), the DFT method of Blom and Sygusch (1997), and LIGIN (Sobolev et al., 1996) were examined. The mean RMSD-based score between the docked and crystallographic ligand orientation for all docking methods ranged from 2.5 to 7.2, indicating that while some targets are easier to dock overall, none regularly dock correctly for every method. The DFT docking method performed poorly, having RMSDs between 15 and 28 Å for the three targets for which predictions were performed. Of the other three approaches, the overall performance were roughly equivalent. Some of the targets were clearly easier to dock, with most of the presented methods docking ligands within 3 Å, while other targets had no dockings within 4 Å. For the protein-protein trial, none of the algorithms achieved a close docking when examining detailed geometry. However, some methods were able to correctly predict some of the interactions which occur in the binding site.

Bissantz et al. (2000) also examined a series of docking algorithms and scoring functions for screening against thymidine kinase and the estrogen receptor. They found that most algorithms were able to extract roughly 70% of known ligands from a database of 990 random molecules and that consensus scoring generally enhanced hit rates. However, they saw no relationship between the accuracy of producing the correct docking orientation and correctly ranking the orientations. It was also not possible to accurately predict binding energies. Of note is the authors' suggestion to use a limited size database and several screening methods to determine the best tool for the protein in question and then to screen

a large database.

Other studies have examined the performance of docking and screening algorithms in more limited ways. Knegtel and Wagener (1999) used the DOCK algorithm to explore the efficacy of screening methods for thrombin inhibitors and for selectivity for progesterone receptor ligands versus estrogen receptor ligands. This study applied two scoring functions: an energy scoring function consisting of the AMBER forcefield (Weiner et al., 1984, 1986) and a chemical scoring function consisting of the AMBER forcefield with the attractive van der Waals interaction energy applied only to interaction between complementary atom types. For thrombin, a set of 32 active inhibitors and a set of ten chemically similar, but inactive, compounds was compared. Rigid-body docking yielded a slight bias towards active compounds in the ligands ranked in the top 100 ligands, but allowing flexibility effectively eliminated this bias, allowed more highly ranked inactive compounds. Neither scoring function had a strong ability to differentiate between active inhibitors and inactive compounds, though the chemical scoring methods had a slight advantage over the AMBER scoring function. Docking against the progesterone receptor was performed on a set of 28 known agonists and 20 chemically similar estrogen receptor ligands. A similar analysis to the thrombin case showed that the both scoring functions had a somewhat more pronounced discriminant ability, but that the energy scoring function performed somewhat better. This is likely due to the highly hydrophobic nature of the progesterone receptor site whose interactions are predominantly based on van der Waals forces. The AMBER forcefield is likely to more accurately model such forces.

A second such study was performed on stromelysin-1 (matrix metalloproteinase 3;

MMP3) by Ha et al. (2000). The authors of this study co-crystallized MMP3 with 6 biphenyl-based inhibitors, all of which dock in highly conserved orientations and are used to construct "correct" dockings for a set of 61 biphenyl ligands with IC_{50} values in the low micromolar to low nanomolar range. DOCK with the PMF scoring function performed best, with a mean RMSD of 1.8 Å between the docked and crystallographic binding modes and oriented nearly all ligands with the biphenyl moiety in the correct binding pocket. FLEXX and DOCK with the AMBER forcefield performed worse and generally produced ligand orientations with the biphenyl group in other, unoccupied pockets around the binding site. However, orientations calculated by FLEXX which did place the biphenyl into the correct pocket generally had a lower RMSD relative to the crystal structure, indicating that FLEXX may be better able to fine tune the ligand orientation.

1.6 Creation of Targeted Computational Screening Databases

Computational screening's requirement for very fast handling of individual ligands puts severe limits on the detail one can use in the algorithm. A possible method to reduce the effective time per molecule and add complexity is to limit the screening to ligands that are of particular interest, for example by removing molecules from the database that do not resemble drug molecules. Bemis and Murcko (1996) showed that 50% of orally deliverable human drug molecules in the Comprehensive Medicinal Chemistry (CMC) database are characterized by only 31 of 1179 graph frameworks, i.e, a connectivity graph ignoring atom

types and bond orders, and 24% are characterized by only 41 of 2506 atom frameworks, i.e., graph frameworks which consider atom types and bond orders. In fact, 8.5% of these drug molecules have a benzene framework. Databases could also be pruned to include only those molecules which chemically resemble known drugs. The well-known Lipinski rule of fives (Lipinski et al., 2001), based on examination of a 2245 orally active subset of the World Drug Index (WDI; Derwent Information, London, UK), gives the following guidelines for molecules that are unlikely to be adequately soluble and permeable to function as oral pharmaceuticals:

- contains more than five hydrogen bond donors,
- contains more than ten hydrogen bond acceptors,
- has a molecular weight greater than 500 Da, or
- has a calculated Log P (CLogP; the calculated octanol/water partition coefficient) greater than 5 (or has a Morigucchi Log P, MLog P >4.15).

Limiting the database to only those molecules which contain a common framework and which meet the Lipinski rule would greatly decrease the number of molecules in most chemical databases, allowing a greater exploration of each of the molecules. Other restrictions on the screening database contents, such as limiting it to molecules which contain a specific functional group, could also be implemented. One disadvantage to these approaches is that any new ligands identified will resemble drug-like molecules and any unusual novel compounds will be overlooked.

1.7 Molecular Clustering

Another method to reduce the effective size of the database is to eliminate molecules which bear close resemblance to other molecules in the database, e.g., by clustering. However, when clustering is used, one must ensure the molecules contained in the clustered database represent a sufficiently broad scope of chemical space. The breadth of this scope is not necessarily in relation to the whole of chemical space, but may be restricted to regions of space with particular interest. Matter (1997) examined the diversity in a set of 1283 biologically active compounds using various similarity measures. Compounds were classified with simple 2D fingerprints or with 2D fingerprints combined with more complex descriptors. Clustering was then performed, taking the structure closest to the center of the cluster as the representative. Chemical space coverage was measured by the percentage of biological classes which were included in the final set of database molecules. The best coverage was achieved using only 2D fingerprints. While using this approach to prune the database achieves a reduction in the number of molecules which must be screened, there is a potential that the representative molecule in a cluster may not be selected during a screening run or may be inactive upon experimental investigation while another molecule contained in the cluster would be a very good ligand. Other clustering methods include the Jarvis-Patrick nonhierarchical and Ward hierarchical methods, their application to chemical structures discussed in Brown and Martin (1996), the clustering of Markush structures using a *k*-means clustering algorithm (Barnard et al., 2000), and clustering using molecular field matching algorithms (Mestres et al., 1997).

A related approach is to restrict the number of molecules which must be analyzed after

a screening run is performed as effective analysis can often be difficult given the large numbers of selected potential ligands. This is an especially serious problem when the screening database contains many compounds which closely resemble known ligands for the targets as any novel potential ligands will be difficult to extract from these known ligand hits. Su et al. (2001) propose a method by which the database is initially grouped into families based on common frameworks. Each family consists of a base fragment along with an ensemble of attached functional groups, each transformed into the same reference frame based on the base fragment. The base fragment is then rigidly docked into the binding site, the ensemble functional groups transformed into the binding site based on this original docking, and then each family molecule is scored independently. The transformation of the ensemble as an entity means that only a single transformation needs to be done instead of a transformation for each database molecule, greatly reducing the amount of necessary calculation. In the final score list, only the best scoring member of each family is identified, meaning that potential ligands previously ranked below a large molecular family are pulled to a higher ranking. The authors show that families which contain known ligands are pulled into ranks which could be considered to be reasonably examined from ranks below what would normally be analyzed. One concern with this approach are that if the best scoring member of a family shows no experimental inhibitory activity, any inhibitors in this family could be missed. These concerns can be reduced, though not eliminated, by analyzing the top n potential ligands in the family. Another concern is that molecules which contain a common base fragment but which are otherwise unrelated are placed in the same family. The use of other modes of clustering could be explored to alleviate this concern.

1.8 Database Comparisons

Aside from the notion of the best method to group the molecular structures in a single database is the question of which database to use. Several studies have compared some of the available databases (Shemetulskis et al., 1995; Cummins et al., 1996; Bernard et al., 1998). The most recent and exhaustive comparison is that done by Voigt et al. (2001). They compared the open NCI database (Milne and Miller, 1986), the publicly available portion of the National Cancer Institute anti-cancer and anti-AIDS screening database; the Available Chemicals Database (ACD; MDL Information Systems, Inc., San Leandro, CA), a database of commercially available compounds; the ChemACX database (CamSoft, Cambridge, MA), a second database of commercially available chemicals; the Maybridge Catalog (Maybridge, Plc, Cornwall, England), a third database of commercially available chemicals; the Ansinex database (Asinex, Ltd., Moscow, Russia), a database of commercially available chemicals with emphasis on compounds from combinatorial chemistry; the Sigma-Aldrich Catalog (Sigma-Aldrich, St. Louis, MO); the World Drug Index (WDI); and the Cambridge Structural Database (CSD). These databases contain between 55,000 and 249,000 available three-dimensional structures. All of the databases have some duplication of entries, ranging from 0.02% for the Asinex database to 13% for the ChemACX database. Diversity analysis showed that the CSD is significantly more diverse than the remainder of the databases, which is not surprising given its origin as a repository for information about all types of organic compounds and not only those commercially available or likely or known pharmaceutical compounds. Combination of all eight databases yields a database of 681,000 unique structures. Given the large number of compounds in these

publicly available databases combined with what may be proprietary databases of several hundred thousand molecules makes it clear the efficiency needed to computational screening techniques.

1.9 Conformer Generation

The final issue to arise in computational docking and screening is that of molecular conformers. Conformer generation is of in importance for two reasons. For many molecular databases, many of the structures do not have three-dimensional coordinates attached to them and would therefore be inappropriate for most docking and screening algorithms. Secondly, most of the structures in the molecular databases are taken from crystal structures of the free ligand and/or from crystal structures of the ligand with other proteins. It is quite likely that the conformation of the ligand as bound to the target of interest will differ from the conformations in the database (Betts and Sternberg, 1999), especially in the case for databases of small molecule crystal structures as the crystal packing forces are large compared to the size of the molecule. Many of the docking and screening algorithms allow for minor conformational changes, but the bound conformation could be significantly different and beyond the range for which the program can compensate for. Common conformation generators generally follow an empirical method, where rotations are based on observed structures, such as protein side-chain rotamer libraries (Maeyer et al., 1997; Dunbrack and Cohen, 1997; Lovell et al., 2000) and the MIMUMBA program (Klebe and Mietzner, 1994); a systematic method, where each rotatable bond is altered by a fixed angle, generating a tree of conformers, as implemented by the systematic conformational search

function in MOE (Chemical Computing Group, Montreal, Quebec); or follow a stochastic method, as implemented in the stochastic conformational search function in MOE. In the stochastic method, each rotatable bond in the molecule is rotated by a random amount, the structure is then energy minimized, and then it is compared to previously generated conformers. If it is close in energy and/or conformation, it is not saved. This process is repeated until a set number of conformers is generated or a set number of failures, i.e., generation of a conformer which is similar to a previously generated conformer, occurs. Instead of random torsion rotations, some algorithms also employ random atom displacements and subsequent energy minimization. Boström et al. (1998) compared energy minimized uncomplexed ligand structures with those bound to the protein and found that, for most, protein-ligand complexes, the energy difference between the bound and free structure is small, $\leq$ 3.0 kcal/mol. This gives a guideline for effective conformer generation for screening and docking methods. On the receptor side, rotations between free and liganded structures are generally small and lie within the range of the SLIDE screening algorithm (Maria Zavodszky, unpublished results).

1.10 Successful Application of Docking and Screening Methods

Structure-based drug design has become a common addition to drug discovery projects, reviewed by Klebe (2000). The first successful application dates back to 1973 (Beddell et al., 1976; Goodford, 1984) when a hemoglobin effector mimic of diphosphoglycerate was de-

veloped. Later, researchers at Dupont-Merck used a pharmacophore model to screen the CSD and identified the DMP-323 cyclic urea inhibitor (Lam et al., 1994), which reached phase I clinical trials. Other successes include identification of an inhibitor with a low micromolar IC_{50} using LUDI (Klebe, 2000), identification of four inhibitors of farnesyl-transferase with moderate to high micromolar IC_{50}'s (Perola et al., 2000), identification of isoform specific inhibitors of adenylyl cyclase through pharmacophore screening (Onda et al., 2001), identification of retinoic acid receptor antagonists using ICM docking methods to screen the ACD (Schapira et al., 2000), and identification of ligands which bind specifically to the RNA hairpin HIV-1 TAR RNA by screening a subset of the ACD using the ICM docking method (Filikov et al., 2000). Many other successes of computational docking and screening are likely to reside within pharmaceutical companies. Given the array of available techniques for computational docking and screening and the dramatic growth in computational power and speed, it is likely that the use and importance of computational docking and screening methods will continue to grow.

1.11 Motivation for this Thesis Work

The thesis work presented in this dissertation seeks to improve techniques for modeling protein-water and protein-small molecule interactions and to apply these techniques to gain knowledge about such interactions in systems of interest. Previous examinations of water molecule binding have been limited to experimental studies, which are often arduous and time consuming, or have relied on using a single crystallographic structure as a reference. Chapter 2 describes a technique applying hierarchical clustering to computationally analyze

water molecule conservation in a series of crystallographic structures without the necessity for assigning a structure as a reference. Application of this technique to the serine proteases thrombin and trypsin shows that it can be effective at assisting in explaining specificity differences between related enzymes and in examining the water content of protein-protein interfaces.

Extension from examination of water molecules as ligands has led to the development of a computational screening technique. Previous ligand docking and screening tools limited the modeling of the conformational changes which occur upon a ligand binding to protein to the ligand molecule or contained very limited protein receptor flexibility, through rotamer libraries. Chapter 3 describes a screening algorithm, SLIDE, which allows for both ligand flexibility and protein side-chain flexibility, without resorting to rotamer libraries. This thesis work focuses on improvements made to the description of the hydrophobic character of the interaction, including results from testing on thrombin and glutathione S-transferase (GST). Chapter 4 describes the application of SLIDE to analyze potential docking orientations of molecules selected by *in vitro* high-throughout screening and to identify a limited set of potential new ligands for asparaginyl-tRNA synthetase.

Chapter 2

Identification of Conserved Water Binding Sites in Proteins

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

2.1 Introduction

2.1.1 The Role of Water Molecules in Proteins

Water molecules play an important role in protein structure and function. In addition to providing the driving force behind protein folding via the hydrophobic effect (Kuntz and Kauzmann, 1974; Eisenberg and McLachlan, 1986), they play a significant role in medi-

ating protein-ligand interactions. In some protein complexes, water molecules play key roles in establishing the specificity of ligand binding, such as the interface water molecules in the Trp repressor which allow the protein to make base specific interactions with the repressor element DNA (Otwinowski et al., 1988; Joachimiak et al., 1994). In thymidylate synthase, a water molecule conserved in all crystal structures has been shown to allow the protein to distinguish between substrate and product nucleotides (Fauman et al., 1994). HIV-1 protease shows a similar highly conserved water molecule (Wlodawer et al., 1989). Inclusion of a carbonyl oxygen group to displace the water molecule and satisfy the water molecule's position in the protein's hydrogen bonding network enabled the construction of a high-affinity inhibitor (Lam et al., 1994).

In contrast to the above function, water molecules have also been found to contribute to the plasticity of ligand binding in some protein complexes, such as in the class I major histocompatibility complex (MHC I) where bound water molecules rearrange to allow the protein to bind to several peptidyl ligands (Wilson and Fremont, 1993). The different peptides have varying side chains, with water molecules bridging the gaps which would otherwise occur between the bound peptide and the protein. Water molecules can also be directly involved in the protein's catalytic function, as in the case of the hydrolytic mechanism for peptide bond breakage by serine proteases (Blow et al., 1969; Perona et al., 1993; Singer et al., 1993). In fact, proteins which are stripped of their primary hydration level are observed to lose catalytic function (Rupley and Careri, 1991).

In addition to playing a such a direct role, water molecules have been shown to stabilize protein structures via formation of extensive hydrogen bond networks (Baker and

Hubbard, 1984) and by filling grooves on the protein surface (Kuhn et al., 1992a). It has been shown crystallographically that bound water molecules remain an integral part of the protein structures, even after repeated rinsing of protein crystals with anhydrous organic solvent (Fitzpatrick et al., 1993; Travis, 1993).

In general, several techniques exist for the identification of bound water molecule sites in a single protein, reviewed by Levitt and Park (1993) and Karplus and Faerman (1994), which measure somewhat different aspects of water binding. Protein structures solved by X-ray and neutron crystallography often assign water molecules bound to very favorable binding sites as the water molecules bound to minimally favored sites and water molecules in the bulk solvent are too mobile to appear as electron or neutron density peaks. A concern with using crystallographic structures for identifying favored water sites is the influence of crystal packing contacts, which can either exclude water, leading to undiscovered sites, or trap water, leading to sites that are not biologically relevant. A second method of identifying favored water binding sites is through NMR, which can measure the time during which a water molecule occupies a given site, i.e., the residence time. Aside from limitation of NMR to small to moderate-sized proteins, water site identification remains a significant challenge. Water sites that are too far from a proton group, water sites that are close to rapidly exchanging protons, and water sites that exchange extremely rapidly cannot be identified. Several NMR studies of protein structure have identified long-lived buried water sites that coincide with crystal structure sites (Otting and Wüthrich, 1989; Clore et al., 1990; Forman-Kay et al., 1991; Xu et al., 1993). Otting et al. (1991) were able to observe some rapidly exchanging surface water molecules with NMR and found that

water molecules that corresponded to surface water sites in the crystal structure had similar residence times to each other.

2.2 Conservation of Water Molecules among Several Crystal Structures of a Protein

Given the multiple and important roles water molecules play in protein structure and function, the ability to quantitatively define conserved water sites from crystallographic protein structures has a number of practical applications, including drug design, allowing the design of ligands that displace conserved bound water molecules (Ladbury, 1996; Wang and Ben-Naim, 1996), and analysis of protein ligand interfaces to identify such sites (Raymer et al., 1997; Sanschagrin and Kuhn, 1998). A typical method for analysis of crystallographic bound water molecules is to use molecular graphics to visualize the water bound in a single protein structure, or small number of closely related structures which have been superimposed, and their proximity to catalytic or ligand-binding residues. As the number of superimposed structures increases, the ability to effectively analyze the conservation of water molecules decreases dramatically. In general, the water molecules located at the same binding site will be somewhat shifted in position due to minor changes in neighboring protein atom positions and due to minor variations in both the actual location of the water molecule in the crystal and variations in its placement by the crystallographer. Visualization of more than a few structures simultaneously will cause the waters to become nearly a continuous shell of hydration, losing all definition of preferred sites.

A second, quantitative approach is to use a single, chosen structure as a reference to judge conservation in a series of related protein structures. This approach has been used to study the solvation of FKBP12 complexes with the immunosuppressant FK506 (Faerman and Karplus, 1995) and the solvation of T4 lysozyme (Zhang and Matthews, 1994). Water site conservation is defined based on sites which occur in the reference also occurring in the remainder of analyzed structures, i.e., are the water sites located in the reference structure also observed in the other structures? This causes the results to be dependent on which structure of a homologous set is chosen as the reference. The Aquarius2 algorithm (Pitt et al., 1993) uses a knowledge base of protein-water molecule interactions, with each interaction tabulated and referenced to a set of common functional groups and side chains, in a series of unrelated protein structures to derive a three-dimensional probability map for locating bound water sites in protein structures in general. Analyses of water molecule binding sites remain subject to limitations in crystallographic fitting and refinement (Levitt and Park, 1993; Karplus and Faerman, 1994), but limitations due to assignment in any given single structure can be minimized through the use of multiple, independently solved structures as a knowledge base for analysis and design.

This section presents work employing the statistical method of complete linkage hierarchical clustering to define consensus water sites in thrombin, trypsin, and bovine pancreatic trypsin inhibitor (BPTI), with the goal of determining the extent to which water sites are conserved for each protein and between the two serine proteases and their relationship to ligand binding. This technique circumvents the problem of using the bound water sites from a single structure as a reference set, because all sites from each of the different pro-

tein structures are equally weighted in cluster analysis. Thrombin was chosen as a protein of focus for several reasons: there are a number of structures solved at good resolution with different ligands bound, thrombin is an important pharmaceutical target for regulating blood coagulation, and highly conserved water molecules are known to surround the binding site of its allosteric regulator, Na^+ (Di Cera et al., 1995; Zhang and Tulinsky, 1997). Thrombin is a serine protease at the junction between blood coagulation and anticoagulation pathways and can initiate both processes, reviewed by Furie and Furie (1988) and Esmon (1992). In addition to binding its receptor, proteolytic substrates, and several physiological inhibitors, thrombin also binds exogenous inhibitors such as hirudin (produced as an anticoagulant agent by leeches) and D-Phe-Pro-Arg chloromethylketone (PPACK), a substrate transition-state analog. Thrombin contains two major ligand binding sites: the active site, which binds fibrinogen at the cleavage site, and an exosite, which provides additional substrate binding surface, enhancing the affinity for fibrinogen and hirudin and its analogs (Vijayalakshmi et al., 1994). The variety of crystallographic protein:ligand complexes available for thrombin provides the ability to study water sites that are conserved regardless of ligand, as well as those water sites that are ligand specific.

Another goal was to identify water sites that are shared by thrombin and trypsin, a serine protease not involved in blood coagulation, in order to identify water sites that are essential in serine proteases and also point to water molecules that are specific to thrombin or trypsin ligand-binding sites. Trypsin is a serine protease which proteolytically activates other digestive proteases. The loop which binds Na^+ in thrombin cannot bind Na^+ in trypsin due to a change in conformation and chemistry associated with the Tyr 255 to Pro sequence

change (Dang and Di Cera, 1996). This results in the absence of allosteric regulation by Na^+ in trypsin. Several high-resolution trypsin structures are available in the Protein Data Bank (PDB), and its water structure has been studied via several techniques including room-temperature and low-temperature X-ray crystallography (Earnest et al., 1991), neutron-diffraction (Finer-Moore et al., 1992), and D_2O-H_2O difference neutron diffraction (Kossiakoff et al., 1992). BPTI is a natural inhibitor of trypsin and its water interactions have been studied using NMR and molecular dynamics (van Gunsteren et al., 1983; Brunne et al., 1993; Denisov et al., 1996) and simultaneous NMR and X-ray diffraction refinement (Schiffer et al., 1994). Several high-resolution structures of BPTI are available in the PDB, along with X-ray diffraction structures of the trypsin:BPTI complex, providing the ability to examine the fate of water molecules bound to free trypsin and free BPTI upon formation of the trypsin:BPTI complex. Given the wealth of structural information available for these serine proteases, they provide an ideal system for testing the technique presented here for determination and analysis of conserved water binding sites.

2.3 Water Site Clustering Methods

2.3.1 Structure Selection

Thrombin, trypsin, and BPTI structures were selected from the Protein Data Bank based upon the absence of unusual crystallization conditions (e.g., low pH), sequence insertions, deletions, or point mutations, and a resolution of $\leq$2.0 Å for trypsin and BPTI and $\leq$2.4 Å for thrombin. Ligand-free structures were selected; no ligand-free structure thrombin struc-

tures were available, but 6 of the 10 structures analyzed here have no ligand in one of the two sites, the active site or the exosite. The availability of 10 thrombin structures for analysis helped compensate for their somewhat lower resolution as compared to the trypsin structures. Visual inspection of the superimposed molecules for each protein eliminated those with regions of large structural deviation likely to affect water site conservation. Only structures with refined water molecule positions were included. The quality of water refinement was assessed using a mobility measure designed to normalize and combine the crystallographic temperature factor (Debye-Waller factor; B-value) and the occupancy as defined below (Craig et al., 1998):

$$\text{Mobility}_{\text{water molecule}} = \frac{\text{B-value}_{\text{water molecule}}/\text{Average B-value}_{\text{all waters in structure}}}{\text{Occupancy}_{\text{water molecule}}/\text{Average Occupancy}_{\text{all waters in structure}}} \quad (2.1)$$

This facilitates comparison of atomic mobility between protein structures refined with different protocols, in particular, those structures in which occupancy as well as B-value were allowed to vary during refinement of the water molecules.

Using this normalization, a water molecule (or other atom) with a high degree of rigidity has a mobility value near 0, a water molecule with average mobility relative to other atoms in the protein has a mobility value of 1, and a highly mobile water molecule has a mobility value greater than 1. In general, if a water molecule's mobility is x, then it is x times as mobile as the average water molecule. In practice, the mobility of a water molecule is determined by its oxygen atom, since hydrogen atom positions are not assigned in the majority of structures. Histograms of the water mobility values for each structure showed

whether there were a number of water sites with high mobility (>2); a preponderance of such sites was found, by analysis of inter-water molecule distances, to indicate water molecules placed too close to each other (<2.6 Å). Such structures were excluded from this analysis. As an example, Figure 2.1 compares the mobility distributions of water sites in two BPTI structures. Structures selected using all the above criteria and selected for this work are presented in Table 2.1.

2.3.2 Hierarchical Clustering

The following steps were performed independently for the thrombin, trypsin, and BPTI structural sets (Table 2.1). The chosen structures were superimposed onto a reference structure using main-chain least-squares superposition in *InsightII* (Accelrys, San Diego, CA) to transform the protein structure and water molecules into the same reference frame. The x, y, z coordinates for these transformed water molecules were then extracted and used for clustering. Clustering is in an iterative process. The first step is to generate a matrix of all inter-element distances. Here, the simple Euclidean distance between points is used, though, in general, any distance metric can be used. The first cluster is then formed from the two closest elements and the distance between this initial cluster and the remaining elements is calculated. Once again, the two closest elements, one of which could be the previously formed cluster, are joined into a new cluster. The process repeats until all elements are joined in a single cluster or, as is the case here, until a distance threshold, representing the maximum distance between any of the elements assigned to a single cluster, is reached. There are several methods of calculating the distance between a cluster of

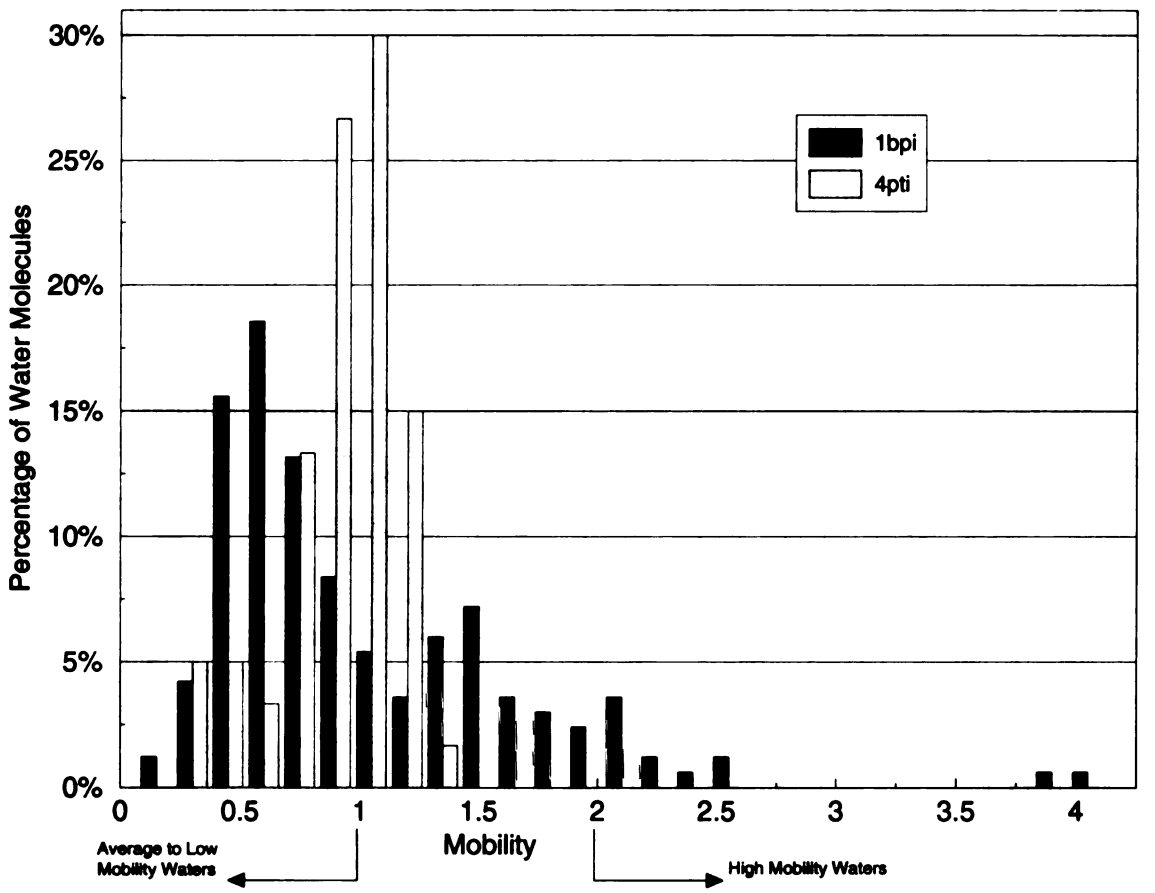


Figure 2.1: Mobility distributions of two BPTI structures, used as a quantitative tool to screen for structures with uncertain water positions. The 4pti distribution is narrow and shows that most water molecules have nearly average mobility and none are highly mobile. The 1bpi distribution is broad and has an extended right tail, indicating the presence of a number of water molecules with high mobility (≥ 2; at least twice as mobile as average). Further analysis showed one-half of these high-mobility sites could be explained by the occupancies of overlapping sites summing to ≤ 1, suggesting that they represent alternate locations of a single water molecule. However, including multiple copies of single water molecules corresponding to their different, partially occupied sites would introduce a statistical bias into the cluster analysis and has been avoided in this work.

Table 2.1: Database of thrombin, trypsin, BPTI, and trypsin:BPTI structures for analysis of conserved water sites

PDB Code	Ligand Binding Site: Active Site	Ligand Binding Site: Fibrinogen Binding Site (Exosite)	Resolution (Å)	Main-Chain RMSD (Å)	Number of Crystallographic Bound Waters
Thrombin Structures[1]					
1hai	PPACK		2.4	0.000	194
1abj	PPACK		2.4	0.694	196
1ppb	PPACK		1.9	0.802	409
1tmb	Cyclotheonamide A	Hirugen	2.3	0.560	239
1hah		Hirugen	2.3	0.345	205
1tmt	——— CGP50,856 ———		2.2	0.458	111
1abi	——— Hirulog-3 ———		2.3	0.409	246
1thr		Hirullin	2.3	0.350	190
1ths		MDL-28050	2.2	0.439	140
1ihs	——— Hirutonin-2 ———		2.0	0.481	146
Trypsin Structures[2]					
1tpo			1.7	1.395	84
2ptn			1.6	0.103	82
3ptn			1.7	0.266	82
BPTI structures[3]					
4pti			1.5	0.000	60
5pti[4]			1.0/1.8	0.403	63
6pti			1.7	0.436	73
9pti			1.2	0.418	67
Trypsin/BPTI complex structures[5]					
2ptc			1.9	0.343/0.479	157
1tpa			1.9	0.336/0.638	159

[1] Superpositions and RMSDs are relative to 1hai.

[2] Superpositions and RMSDs are relative to 1tpo, except for 1tpo which is relative to 1hai.

[3] Superpositions and RMSDs are relative residues 1–46 of 4pti.

[4] Resolution is for X-ray diffraction/neutron diffraction data.

[5] RMSDs are reported for the trypsin chain of the complex superimposed onto 1tpo and for the BPTI chain of the complex superimposed onto 4pti.

multiple elements and other clusters or remaining unclustered elements: (1) the shortest distance between any pair of elements in each of the clusters (single linkage), (2) the distance between the cluster centroids, which are the mean x, y, z coordinates of the elements of each cluster (average linkage), and (3) the maximum distance between any pair of elements in each of the clusters (complete linkage). Complete linkage clustering was chosen for this work as it yields compact, globular clusters and allows the specification of a maximum diameter for any cluster by defining the maximum distance between cluster elements. This ability is useful when defining water sites as it can ensure the water molecules from different structures which contribute to a cluster can form the same approximate hydrogen bonds and are within hydrogen bond forming distance (2.4 Å).

An example cluster analysis for a subset of the water molecules from the BPTI structure set is shown in Figure 2.2. Complete linkage clustering begins by placing the two closest elements together in a cluster; 6pti 108 and 4pti 108 are less than the maximal distance of 2.4 Å apart, the basis for this threshold is given below, and are grouped into a cluster (arbitrarily numbered 109). Next, the distance between this cluster and each of the remaining data elements is computed; for complete linkage clustering, this is defined as the maximum distance between that element and all the elements in the clusters. In Figure 2.2, the distance between 4pti 139, 9pti 103, and 6pti 238 (which are not yet clustered) and cluster 109 is calculated as the distance to 4pti 108, since it is the furthest element of cluster 109. This process is repeated until no further elements can be clustered without exceeding the selected maximum distance, 2.4 Å in this work. Any elements not included in clusters at this point are considered to define single-element clusters; for example, cluster 134 in

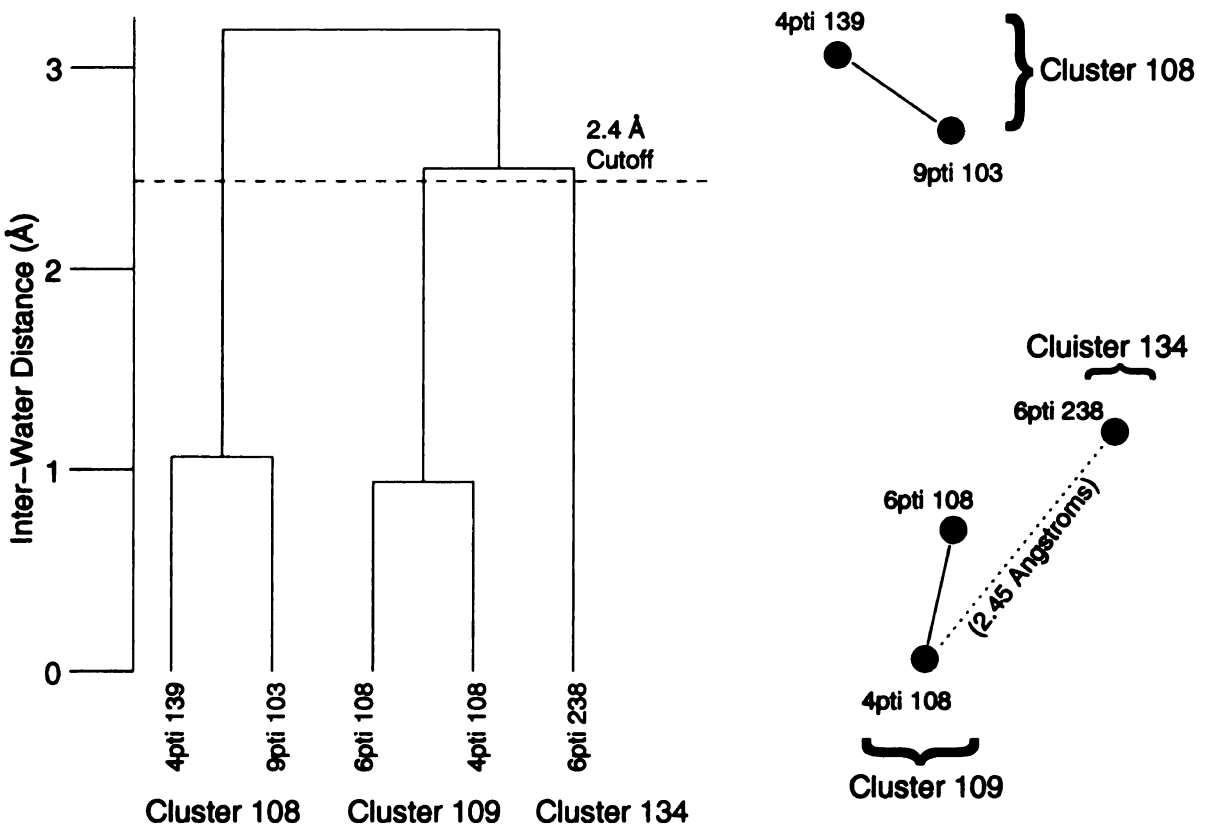


Figure 2.2: Example of complete linkage clustering applied to water sites from several BPTI structures. A portion of the BPTI clustering tree is shown at left, based on the inter-water distances from the superimposed BPTI structures, shown at right. Note that 6pti water 108 and 6pti water 238 are not clustered together even though they are closer than the cutoff distance of 2.4 Å, since 6pti 108 belongs to a cluster in which one water (4pti 108) is too far from 6pti 238 to meet the 2.4 Å threshold. This feature of complete linkage clustering guarantees that no cluster contains water sites separated by more than 2.4 Å. At this distance, all water molecules in a microcluster are overlapping, and it is unlikely that more than one water molecule will be included from any given protein structure (they would be too close). Cluster numbers are arbitrary sequential indices, whereas individual water molecules are labeled by the residue number from the corresponding PDB file.

Figure 2.2 consists only of water molecule 6pti 238.

A maximum diameter of 2.4 Å was chosen, resulting in clusters with a maximum inter-water distance of 2.4 Å, as measured from oxygen center to oxygen center. This value was chosen because water molecules have an approximate effective radius of 1.6 Å, which includes the radius of the oxygen and a correction for the contribution of the hydrogen atoms, whose positions are typically unknown. Thus, if two water molecules are placed with their oxygen atoms at a center-to-center distance of 2.4 Å, their radii will overlap by 50%. This almost always prevents water sites from the same structure from being included in the same cluster, since at <2.4 Å apart, they would be positioned too closely. Complete linkage clustering results in the set of maximally dense clusters (in terms of average number of water molecules per cluster). These will be referred to as "microclusters" to emphasize that all water molecules within a single cluster physically overlap. The WatCH (Waters Clustered Hierarchically) software package developed in this work is implemented in C and has been made available via the internet at `http://www.bch.msu.edu/labs/kuhn/web/software/WatCH/doc.html`.

2.3.3 Crystal Contact Calculation

To observe the possible effects of crystal contacts on water site conservation, crystal contacts in the seven thrombin structures in space group C_2 were calculated using Chain (Sack, 1988), where interactions were included for protein crystal lattice symmetry mate atoms within 4.0 Å. Crystal contact residue and atom lists were generated for each of the crystallographic structures, with water sites represented by the microclusters observed in that

protein. The number of times each microcluster appeared in a crystal contact was calculated for the seven thrombin structures, and software was developed to convert Chain's crystal contact lists and the microcluster lists into InsightII subsets for visualization of the spatial relationship between crystal contacts and microcluster conservation.

2.3.4 Evaluation of Bound Water Environments

The degree of conservation of the water microclusters, each representing a favored site for water binding, was calculated as the number of individual water molecules contained in the microcluster divided by the number of structures used for clustering. To assess the influence of the shape and chemistry of the water binding site on its conservation in different structures, measures of eight environmental features were calculated:

- atomic density (ADN), measured as the number of protein atoms within van der Waals packing distance, 3.6 Å, of the water molecule, which correlates with whether the site is in a groove (high density of protein neighbors) or a protrusion (low density of neighbors) (Kuhn et al., 1992b);
- local atomic hydrophilicity (AHP), measured by the sum of the atomic hydrophilicity of all protein and water atoms within 3.6 Å of the water site (Kuhn et al., 1995);
- crystallographic temperature factor (Debye-Waller factor; B-value; BVAL), a measure of the atom's thermal mobility and spread in the in the electron density, read from the protein's PDB file;
- the number of hydrogen bonds to neighboring protein atoms (PrHBD);

- the number of hydrogen bonds to neighboring water molecules (WatHBD), using a distance of $\leq$3.5 Å between donor and acceptor atoms;
- the water site mobility (MOB), a normalized measure of mobility (see Section 2.3.1);
- the summed B-values for all protein atoms within 3.6 Å of the water site (TPrBVAL); and
- the average B-value for these neighboring protein atoms (AvgPrBVAL).

Several of these features are related, and the goal here was to see determine which features best correlate with degree of water site conservation.

For each microcluster, the value for each of the eight features was averaged over the individual environments of its water molecules. To assess the correlation between conservation of the microclusters and their environments, feature values were also averaged over all microclusters with a given degree of conservation (e.g., those containing waters from 6 of 10 structures).

2.3.5 Calculation of Overlapping Microclusters between Thrombin and Trypsin

Distances were calculated between the centroids of microclusters in the superimposed structures of thrombin and trypsin, and overlapping clusters were defined as those with a centroid-to-centroid distance of $\leq$1.8 Å. With a maximum diameter of 2.4 Å for each microcluster, the microclusters' radii overlap by 50% when their centroids are within 1.8 Å.

To analyze the effect of using different overlap criteria and provide a list of microcluster overlaps between thrombin and trypsin using less stringent criteria, overlapping microclusters were also tabulated using thresholds up to 2.4 Å (where two microclusters would just touch). To determine the significance of the observed number of overlapping sites between thrombin and trypsin, in a separate experiment, microcluster centroids were randomly placed in the thrombin structure at the density of microclusters experimentally observed for thrombin, and the same was done for trypsin. Then, the number of overlaps between thrombin and trypsin was calculated using these random distributions. Because many of the water sites in thrombin and trypsin are buried in the proteins, the microcluster density was calculated based on the number of microclusters per $Å^3$ of protein volume, which was calculated for each protein using the PQMS routine of the Molecular Surface Package, version 2.6 (Connolly, 1983). Random placement of microcluster centroids and subsequent counting of overlaps was repeated 100 times to obtain statistical means and standard deviations for the number of overlaps as a function of overlap criterion (1.8 to 2.4 Å). For analyzing conserved water site proximity to functionally important sites (e.g., residues in the catalytic triad), a distance threshold of 3.6 Å from the microcluster to the functionally important atom(s) was used. Interaction with an active-site or exosite ligand was determined by measuring the distance to all ligands bound in the structure.

Images in this dissertation are presented in color.

2.4 Results

2.4.1 Clustering Statistics

To identify shared versus unique conserved water sites for thrombin and trypsin, complete linkage clustering was performed on their respective water sites (Table 2.2). Clustering of 2,075 water sites from the ten thrombin structures yielded 708 microclusters with an average of 2.93 waters each, indicating that the average water site was observed in 29.3% of the structures. Of the 708 microclusters, 18.5% were found in at least half of the 10 structures. Clustering of 248 water sites from the three trypsin structures yielded 106 microclusters, conserved on average in 78.0% of the structures. Of these microclusters, 56.6% were observed in all three structures. This high degree of conservation was surprising, but two of the structures (PDB codes 1tpo and 2ptn) were solved by the same crystallographers and have very similar water sites; however, mobility plots (data not shown) indicated that the water assignments in both structures were reasonable. (Consideration was given to analyzing additional trypsin structures, but there were only three ligand-free, wild-type bovine structures solved under typical crystallization conditions.) Given the similarity in water assignments for 1tpo and 2ptn, trypsin water sites were considered to be highly conserved only if they appeared in all three structures. A similar analysis of BPTI clustered 263 water sites from four structures into 134 microclusters, with an average conservation of 49.0%. Of these microclusters, 54.5% were found in at least half of the BPTI structures.

Table 2.2: Clustering statistics

Thrombin (10 superimposed structure)		
Number of water molecules	2075	
Number of water clusters	708	
Average conservation (waters/cluster)	2.93	(29.3%)
Number of clusters with ≥50% conservation	131	(18.5%)
Number of clusters with 100% conservation	28	(4.0%)
Mean protein volume (Å^3)	38.27	
Cluster density (clusters/Å^3)	0.0185	
Conserved cluster[1] density (clusters/Å^3)	0.0034	
Trypsin (3 superimposed structure)		
Number of water molecules	248	
Number of water clusters	106	
Average conservation (waters/cluster)	2.34	(78.0%)
Number of clusters with ≥50% conservation	82	(77.3%)
Number of clusters with 100% conservation	60	(56.6%)
Mean protein volume (Å^3)	27.21	
Cluster density (clusters/Å^3)	0.0039	
Conserved cluster[1] density (clusters/Å^3)	0.0030	
BPTI (4 superimposed structure)		
Number of water molecules	263	
Number of water clusters	134	
Average conservation (waters/cluster)	1.96	(49.0%)
Number of clusters with ≥50% conservation	73	(54.5%)
Number of clusters with 100% conservation	18	(13.4%)
Mean protein volume (Å^3)	7.31	
Cluster density (clusters/Å^3)	0.0180	
Conserved cluster[1] density (clusters/Å^3)	0.0100	

[1]Conserved clusters are those with conservation ≥50%.

Table 2.3: Linear correlation coefficients between degree of conservation and environmental features

Environmental Feature	Correlation Coefficients			
	Thrombin	Trypsin	BPTI	Combined
ADN	0.412	0.451	0.418	0.408
AHP	0.483	0.467	0.304	0.475
BVAL	-0.377	0.443	-0.355	-0.474
PrHBD	0.457	0.463	0.439	0.485
WatHBD	0.218	0.075	-0.040	0.177
MOB	-0.525	0.450	-0.458	-0.478
TPrBVAL	0.068	0.371	0.146	-0.061
AvgPrBVAL	-0.387	0.028	-0.328	-0.514

2.4.2 Environmental Analysis

Analysis of water site environments provided insights into the determinants of conserved water binding. All protein-bound microclusters, i.e., those containing at least one water molecule making direct contacts ($\leq$3.6 Å) with the protein, were analyzed. There were 521 protein-bound microclusters for thrombin, 98 for trypsin, and 117 for BPTI. Highly conserved water molecules occupied somewhat different environments than less conserved environments (Figure 2.3). Linear correlation coefficients for each feature are given in Table 2.3. Conserved microclusters had more neighboring protein atoms (atomic density; ADN), made more hydrogen bonds to the protein (PrHBD), and were in a more polar environment, indicated by more hydrophilic neighboring atoms (atomic hydrophilicity; AHP). Most mobility measures, the water site's B-value (BVAL), its mobility (MOB), and the average B-value of the neighboring protein atoms (AvgPrBVAL), were negatively correlated with conservation, indicating that water sites with high conservation tend to reside in less

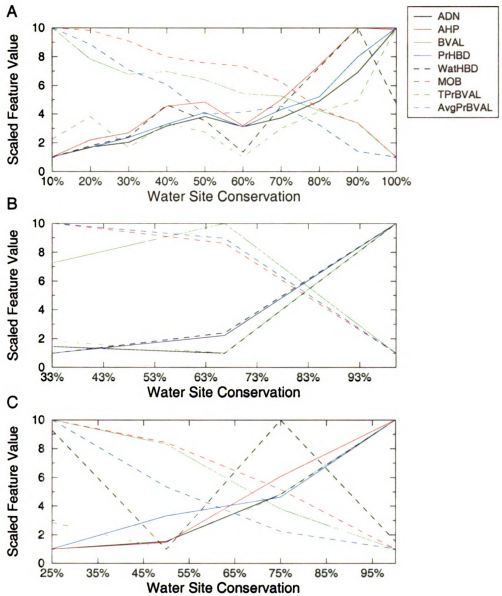


Figure 2.3: Correlation between water site conservation and environmental features for (A) thrombin, (B) trypsin, and (C) BPTI. Shown are the average values of eight environmental features for the water microclusters as a function of their degree of conservation. Features are those described in Section 2.3.4. The feature values have been averaged within microclusters as described in Section 2.3.4, averaged over the microclusters with the given degree of conservation, and normalized to range between 1 and 10 to allow visualization on the same plot. The curve of AHP for trypsin superimposes with that of PrHBD, and is therefore not apparent on the plot. Approximately linear correlation with conservation is seen for many of the features, as described in Results.

mobile portions of the protein. The number of hydrogen bonds to other water molecules (WatHBD) did not correlate strongly with the conservation, suggesting that consensus water sites are not strongly stabilized by hydrogen-bonded water networks.

2.4.3 Effects of Crystal Contacts on Bound Water Conservation

The effects of crystal contacts upon water binding were examined by spatially correlating water site conservation with contacts in the protein lattice. To address whether water sites were preferentially excluded from or trapped in these contacts, the location of conserved water sites along with the crystal contact residues for the seven thrombin structures in the C_2 space group were visualized. Crystal contacts had fewer conserved water sites than surrounding areas, consistent with the observed expulsion of interfacial bound water upon dimerization of chymotrypsin (Blevins and Tulinsky, 1985).

2.4.4 Spatial Analysis of the Conserved Microclusters

To explore how microclusters of different conservation levels are distributed spatially around the protein, molecular graphics visualization was used. For thrombin, a concentration of highly conserved microclusters (in $\geq$50% of the structures; yellow spheres in Figure 2.4) was found near the sodium site but not observed in the active site, perhaps due to water displacement by the presence of active-site ligands in 7 of the 10 structures. Other conserved microclusters were observed in deep grooves or cavities within the protein, as expected from the known correlation between water site conservation and groove

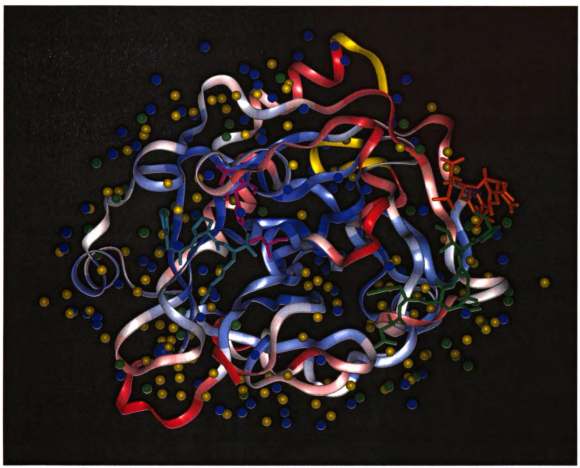


Figure 2.4: Conserved water sites in thrombin. Thrombin microclusters containing water sites from three of the ten structures are colored blue, sites found in four are green, and sites found in at least five are yellow. The backbone ribbon of 1hai is shown colored by B-value (dark blue equals a B-value of 0, white ~30, red ~50, and yellow >50 Å^2). The catalytic triad Asp, Ser, and His side chains are rendered as pink tubes at center. PPACK, an active-site ligand 1hai, is shown in blue tubes, and hirugen, an exosite ligand from 1hah, is shown in green (structurally conserved region) and orange (structurally divergent region) at right. The sodium ion (labeled as water 410 in 1hai) is rendered as a large blue sphere at lower left. A concentration of conserved water sites exist near the sodium site and its channel, at lower left; many other sites are buried.

topography (Kuhn et al., 1992a) and previous studies on the conservation of buried water molecules in serine proteases (Finer-Moore et al., 1992; Rashin et al., 1986; Meyer, 1992; Sreenivasan and Axelsen, 1992). When the exosite ligands were superimposed, a structurally conserved region, comprising the six N-terminal ligand residues (green tubes at the bottom right of Figure 2.4, and a structural variable region, extending from the seventh residue to the C-terminus of the ligand (orange tubes are rightmost edge of Figure 2.4), were found; water sites associated with the structurally conserved region in the exosite ligands were also generally conserved. Similar patterns of buried water site conserved were observed for trypsin.

Given the functional importance of the Na^+ binding site for switching between the coagulant (Na^+ bound) and anticoagulant (water bound) forms of thrombin (Di Cera et al., 1995), this region of the structure was analyzed in detail. The Na^+ sites in structures 1hai and 1hah assigned by Zhang and Tulinsky (1997), which were originally labeled as water molecules in the PDB structures and later confirmed by rubidium replacement to represent a Na^+ site (Di Cera et al., 1995), occur in two overlapping microclusters (centroids 1.2 Å apart) containing the Na^+/water molecules from all 10 structures. The 38 water microclusters in the channel coupling the Na^+ site with the active site are >50% conserved on average, consistent with the recent discovery of this conserved solvent channel (Zhang and Tulinsky, 1997).

2.4.5 Overlapping Water Sites between Thrombin and Trypsin

To define water sites shared between these serine proteases involved in distinct biochemical pathways, overlaps between conserved ($\geq$50%) sites in thrombin and trypsin were evaluated (Table 2.4 and Figure 2.5). The number of overlapping water microclusters in thrombin and trypsin, 37, is statistically significant, since only 7.9 overlaps would be expected if the conserved water sites in thrombin and trypsin were distributed randomly (see 2.3.5). Seven of the conserved microclusters were in the active-site region, four being near at least one of the catalytic triad residues. Three overlapping clusters were near the Na^+ binding site of thrombin, with two additional ones in the surrounding solvent channel. Conservation of solvent in this region (Figure 2.5, lower left), which regulates the coagulant/anticoagulant function of thrombin via Na^+ binding displacement, suggests it may also be important in trypsin. To assess whether water site conservation between thrombin and trypsin is associated with conservation of nearby side chains and their conformations, the 37 shared waters sites were evaluated in the context of PDB structures 1hai (thrombin) and 1tpo (trypsin). Ninety-two percent of the shared water sites had chemically and conformationally similar environments, based on no more than one side-chain substitution and no more than one residue with a significant (1.5–2 Å) shift. Larger shifts were considered structurally dissimilar, yet even substituted side chains tended to be similar through the γ-carbon. Of the 37 shared sites, 38% were structurally very similar, with no side-chain substitutions and no positional shifts exceeding 1.5 Å. Thus, conserved protein structure between thrombin and trypsin largely accounted for their water site conservation, which can be considered a shared feature of their structure and function as serine proteases.

Table 2.4: Overlapping conserved water sites between thrombin and trypsin

Thrombin			Trypsin			Distance between Thrombin and Trypsin Centroids[4] (Å)
Cluster Number	Percent Conserved[1]	Representative Water Residue Number[2]	Cluster Number	Percent Conserved[1]	Representative Water Residue Number[3]	
1021	70	570	25	100	470	0.16
899 LSC[5]	100	417	45	100	415	0.30
996	100	423	22	100	717	0.41
954	100	461	5	100	430	0.43
1197	70	515	54	67	806	0.48
1017 LC	100	407	33 L[5]	100	416	0.52
935 A[5]	100	430	6 A[5]	100	703	0.54
857 A	100	445	20	100	701	0.56
951 A	100	468	19	100	408	0.56
874	100	401	28	100	708	0.56
970	50	551[6]	72	100	752	0.57
885	100	404	18	100	721	0.62
888 LC	100	403	31	100	704	0.65
926	100	414	30	100	429	0.67
1214	80	480	21	67	751	0.72
1075	80	489	59	100	736	0.75
948	50	554[7]	62	67	754	0.75
852	90	441	13	100	473	0.76
1016 A	100	436	10 L	100	410	0.77
963	100	439	17	100	722	0.78
1051	70	455	66	100	728	0.80
878	100	405	9	100	406	0.82
972 SC	90	448	35	100	705	0.86
1032	80	469	55	67	803	0.94
981	100	467	38	100	709	0.96
955	100	412	29	100	716	0.99
1139	70	537	42	100	530	1.01
1150 E[5]	90	507	8	67	738	1.06
832	50	458	65	67	801	1.10

Continued on next page.

Table 2.4 (cont'd)

Thrombin			Trypsin			Distance between Thrombin and Trypsin Centroids[4] (Å)
Cluster Number	Percent Conserved[1]	Representative Water Residue Number[2]	Cluster Number	Percent Conserved[1]	Representative Water Residue Number[3]	
1111	90	546	60	100	744	1.11
1086 SC	70	409[6]	44	100	562	1.14
890	90	406	24	100	516	1.29
870	90	443	4	100	726	1.32
916	80	452	34	100	604	1.40
921	100	413	1	100	746	1.58
1038	90	451	2	100	741	1.61
1108	80	539	16	100	725	1.66
1150	90	507	71	100	733	1.86
1259	70	426	52	67	750	1.93
1053	80	457	56	67	735	2.10
981	100	467	24	100	516	2.13
1170	60	494	66	100	728	2.15
964 SC	90	450	35	100	705	2.15
948	50	554[7]	65	67	801	2.19
1032	80	469	37	100	720	2.25
827	100	446	18	100	721	2.29
995	90	431	27	67	743	2.30
1119	50	505	2	100	741	2.34
857	100	445	10	100	410	2.36

[1]Only clusters with at least 50% conservation are tabulated.

[2]Representative thrombin waters are from 1hai unless there is no member water from 1hai, in which case the source structure is noted.

[3]Representative trypsin waters are from 1tpo.

[4]A line divides the table into highly overlapping water microclusters with centroids $\leq$1.8 Å apart (cluster radii overlap by $\geq$50%), shown in the top section of the table, from somewhat overlapping microclusters with centroids 1.8–2.4 Å apart.

[5]Labels indicate overlapping clusters that interact with (are $\leq$3.6 Å from) active site ligands (L), active-site catalytic triad residues (A), Na^+ channel waters (C), exosite ligands (E), or Na^+ site (S).

[6]Representative thrombin water is from 1hah.

[7]Representative thrombin water is from 1abj.

Several water sites were highly conserved in functionally important regions of thrombin or trypsin, but were not shared between them (Table 2.5). These may contribute to their specificity differences. Four more microclusters were specifically associated with active-site ligands in trypsin than were seen in thrombin, in reflecting the larger inhibitor in trypsin; 13 residues of BPTI interact with trypsin, whereas the thrombin active-site ligands are only three to seven residues long. Five Na^+ binding site and channel clusters were shared between thrombin and trypsin (Table 2.4); however, 15 conserved sites in this region were found only in thrombin (Table 2.5). Combined with the five conserved exosite water positions found uniquely in thrombin and eight active-site water positions found uniquely in trypsin (Table 2.5), it is apparent that bound water can make a significant contribution to ligand specificity.

2.4.6 Contribution of Conserved Water Molecules to the Trypsin:BPTI Complex

Trypsin provides an ideal system to test the applicability of a lock-and-key mechanism for the contributions of protein-bound and ligand-bound water molecules to serine protease complex formation because several high-resolution ligand-free structures are available for trypsin, BPTI, and the trypsin:BPTI complex. Using water microclusters identified for trypsin and BPTI, the conserved water sites from each protein were compared with water sites conserved in the complex structures (2ptc and 1pta). The free trypsin structures were superimposed onto the trypsin chain of the 2ptc complex, and the free BPTI structures were superimposed onto the BPTI chain of 2ptc. Three conserved microclusters from the

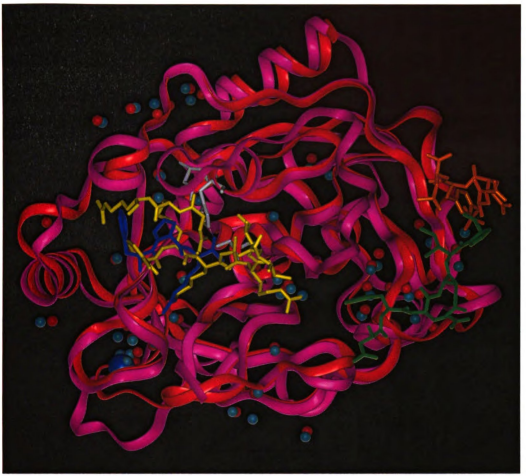


Figure 2.5: Overlapping conserved water sites between thrombin and trypsin. Water sites conserved in at least half of the structures of thrombin (water sites shown as blue spheres) or trypsin (red spheres) are shown. It is important to note that each cluster shown corresponds to one cluster in an overlapping pair; nonoverlapping clusters are omitted from the figure. The backbone of thrombin (represented by 1hai) is shown as a magenta ribbon and the backbone of trypsin (represented by 1tpo) is shown as a red ribbon. Catalytic triad residues are shown as white tubes (center of figure), PPACK (a thrombin active-site inhibitor) is shown in blue, and hirugen (a thrombin exosite inhibitor) is shown in green and orange (structurally conserved and divergent regions, respectively). The region of the trypsin inhibitor BPTI which contacts trypsin is shown in yellow and superimposed from 2ptc; note the conformational similarity between PPACK and BPTI, extending downward from the Pro residue of PPACK. The Na^+ from 1hai is rendered as a large blue sphere at lower left. A concentration of overlaps between conserved thrombin and trypsin water sites is observed near the Na^+ site, despite trypsin having no known functional similarity here; these conserved water molecules form a network which extends towards the active site. There is also a number of overlapping sites located in the exosite, though these are more spatially spread..

Table 2.5: Functionally relevant conserved water sites unique to thrombin or trypsin

	Cluster Number	Percent Conservation[1]	Representative Water Residue Number[2]	
Thrombin				
Active Site Catalytic Triad Residues				
	No Nonoverlapping Conserved Water Sites			
Active Site Ligands				
	1153	100	408	
	1196	90	428	
Exosite Ligands				
	821	60	560	
	949	50	576	
	1179	70	415	
	1241	50	490	(1hah)
	1278	70	496	(1hah)
Na^+ Binding Site				
	1195	80	418	
	976	100	424	
	1121	90	482	
	838	100	514	
Na^+ Channel Waters				
	1001	100	409	
	1195	80	418	
	976	100	424	
	1196	90	428	
	788	70	463	
	914	100	464	
	944	50	474	
	1121	90	482	
	915	90	497	
	838	100	514	
	1229	60	457	(1hah)

Continued on next page.

Table 2.5 (cont'd)

	Cluster Number	Percent Conservation[1]	Representative Water Residue Number[2]
Trypsin			
Active Site Catalytic Triad Residues			
	48	100	747
	80	100	702
Active Site Ligands			
	80	100	702
	61	100	710
	48	100	747
	64	67	807
	23	67	808
	77	100	805

[1]Only waters with $\geq$50% conservation are tabulated
[2]Representative thrombin waters are from 1hai unless noted. Representative trypsin waters are from 1tpo.

free structures overlapped with the conserved water sites in the complex (large spheres in Figure 2.6), two being contributed by trypsin and one by BPTI. Thus, three of the seven trypsin:BPTI interfacial water molecules were donated by the free proteins, while four were newly recruited or shuffled upon complex formation. This contrasts with the contributions of water molecules bound to the free structures of lysozyme and the D1.3 antibody, which contribute 20 of the 25 water molecules observed in the antibody-lysozyme interface (Braden et al., 1995). Thus, the hydration structure of the free protein and ligand and the creation of new environments favorable for water binding upon docking of the protein and ligand should both be considered in inhibitor design. One approach to doing so in the pattern recognition application *Consolv* (Raymer et al., 1997).

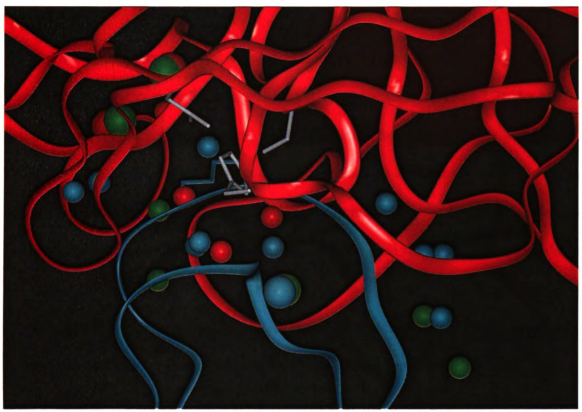


Figure 2.6: Conservation of water sites in the trypsin:BPTI interface. This close-up of the interface between trypsin (red ribbon, from PDB 1tpo) and BPTI (blue ribbon, PDB 4pti), superimposed into the analogous chain of the trypsin:BPTI complex (PDB 2ptc), shows the conservation of water sites between the free structures and their complex. The orientation is approximately a 90° rotation about the horizontal axis relative to Figure 2.5. Conserved ($\geq$50%) water sites from the free trypsin structure are shown as red spheres, those from BPTI shown in blue, and interfacial waters found in both structures of the trypsin:BPTI complex (PDB 2ptc and 1tpo) are shown in green. Water sites overlapping between the complexes and free structures are rendered as large spheres, while non-overlapping sites are rendered as small spheres. The catalytic triad of trypsin is shown in white, and the side chain of the inhibitory Lys 15 from BPTI is shown in cyan. Of the seven trypsin:BPTI interfacial water sites, three are contributed by either trypsin or BPTI.

2.5 Discussion

2.5.1 Conservation of Water Sites in Thrombin and Trypsin

A number of water sites were conserved in at least half of the thrombin and trypsin structures, and several sites were found in all of the structures examined (Table 2.2). An earlier detailed study of the solvent structure of trypsin (Finer-Moore et al., 1992) defined 211 consensus water sites via high-resolution X-ray diffraction data for the waters' oxygen atoms, verified by D_2O-H_2O difference neutron scattering density for the waters' hydrogen atoms. In this work, significantly fewer consensus water sites were identified, 60, perhaps due to comparing three structures. A key goal was to distinguish conserved water sites characteristic of serine proteases in general from those contributing to ligand specificity. Thirty-seven overlapping conserved water sites were found between thrombin and trypsin, four and a half times the number expected for a random distribution of water sites. Finer-Moore et al. (1992) evaluated similarity in solvent structures between pairs of eight trypsin and trypsinogen structures and also found significant similarity between them. Ten of the 37 shared sites observed here were in contact with ligands or associated with the solvent channel proximal to the Na^+ site (Table 2.4). This is consistent with the observation of Krem and Di Cera (1998) that one-third of the conserved internal water sites in serine proteases (Sreenivasan and Axelsen, 1992) are located near the Na^+ site; they proposed that the water structure stabilizes this pocket associated with the substrate specificity (Krem and Di Cera, 1998). Two water sites conserved between thrombin and trypsin in a channel leading from non-catalytic triad Ser 214, which interacts with Asp 102 of the catalytic triad,

were also found. This solvent channel has been proposed as an exit path for the protons produced during catalysis (Meyer, 1992).

2.5.2 Conserved Water Sites and Ligand Specificity

To identify water sites that can contribute to substrate specificity, water sites which were conserved in functionally important regions of thrombin and trypsin but not conserved between the two enzymes were analyzed. The 22 water sites conserved in the active site, Na^+ binding region, and exosite of thrombin but not in trypsin, and the eight active-site water molecules conserved in trypsin but not in thrombin (Table 2.5) are likely to contribute to their different substrate specificities. Design of thrombin inhibitors may be optimized by mimicking these water interaction, as has been achieved for HIV protease (Lam et al., 1994) and cyclophilin-A (Mikol et al., 1995). Results presented in Raymer et al. (1997) on a study of 20 nonhomologous proteins bound to diverse ligands showed that water molecules in ligand-binding sites can be displaced by similarly polar ligand atoms, but also that water-mediated bridges between protein and ligand are ubiquitous, with an average of 19 water-mediated hydrogen-bond interaction between proteins and small ligands. Thus, the positions of conserved interfacial water molecules can used to specify a template of favorable hydrogen bonds for ligands to satisfy, providing another strategy for optimizing ligand design.

2.6

The

defi

pro

wat

of c

nur

cor

ide

Sir

an

2.6 Conclusions

The work presented here demonstrates hierarchical clustering as a useful tool for unbiased definition and analysis of consensus water sites when several independent structures of a protein are available. This approach is particularly useful for resolving the continuum of water site overlaps that occurs when a number of structures are superimposed. Analysis of colocalization between thrombin and trypsin water sites showed a small, but significant number of overlaps, predominantly surrounding the sodium ion site in thrombin and the corresponding region in trypsin. Cluster analysis of water sites and their environments also identified the features associated with highly conserved water sites:

1. a high density of protein atom neighbors, indicating the water site is in a protein groove or cavity instead of being associated with a surface protrusion,
2. several hydrogen bonds being formed to the protein,
3. a hydrophilic environment, and
4. low thermal mobility of the site.

Since cluster analysis is a general statistical method, it is also expected to be useful for analyzing side-chain and ligand atom positions and their chemistries.

Chapter 3

Computational Ligand Screening – An Improved Model of Protein-Ligand Interactions

3.1 Introduction

In addition to examining water molecules as a set of special ligands, it is desirable to expand analysis to small, organic molecules which act as ligands. One approach to this is the development of computational screening techniques which can be used to screen large databases of molecules efficiently using computers versus experimental approaches. One method of computational screening is to approach the problem as an extension to computational docking, which seeks to find the position of a known protein ligand in the protein's binding site. Using such a method for screening would involve docking each of the molecules in the database into a defined binding site and ranking them based on the quality of the docking. While such an approach would work in theory, using the best docking algorithms available, which include full ligand flexibility (Welch et al., 1996; Rarey et al., 1996b,a), would cause

the problem to be computationally intractable as these take at least on the order of a few minutes to dock a single molecule. Such a screen of several hundred thousand molecules would take an infeasible amount of time. Even allowing for just one minute per ligand, screening a database of 100,000 molecules would take 10 weeks. However, one could consider using a simplified docking algorithm which reduces the docking time to a second or less per ligand and enables the screening to completed in a day. Such a technique is presented here.

The classical view of protein-ligand binding, introduced by Fischer (Fischer, 1894), is that of the "lock-and-key", where the ligand fits as a key into the protein lock. However, a more recent study of 39 complexes showed that the "lock" and the "key" are often flexible (Betts and Sternberg, 1999), meaning that a simple steric fit docking is not sufficient to function as a screening method. In an ideal case, both the protein and the ligand would be fully flexible, but this returns to the problem of computational tractability. Instead, a method where sufficient, but not additional, flexibility is included in the model would be optimal. An additional step to further increase the efficiency of the docking is to represent the binding site of the protein by a series of points which reflect the possible interactions which can be made to a potential ligand. In the docking tool DOCK (Shoichet et al., 1992; Shoichet and Kuntz, 1993), this is generally a set of around 100 spheres which constitute a negative image of the binding site. When a docking search is performed, the set of points representing the protein binding site is matched with the set of points representing the possible interactions a potential ligand could make.

3.2 Methods

3.2.1 A General Overview of the SLIDE Method

The overall method of the SLIDE (Screening for Ligands with Induced-fit Docking Efficiently) algorithm involves three stages, all of which will be discussed in detail:

1. assignment of a set of points which represent the types of interactions a molecule could make when bound to a protein (performed once per database of molecules),
2. identification of the sites of favorable interaction in the protein and reduction to a set of favorable template points (performed once per protein of interest), and
3. matching the database molecules, via their interaction points, to the protein, via its template points, i.e., the actual screening process.

Both molecule interaction point and protein template points can be one of four types: (1) a hydrogen bond acceptor, (2) a hydrogen bond donor, (3) a hydrogen bond doneptor (donor and acceptor), or (4) a hydrophobic or non-polar point.

3.2.2 Assignment of Interaction Points to Molecules in the Screening Database

Assignment of interaction points to database molecules is based upon an atom by atom examination of the molecule in question. Hydrogen bonding points are placed at atom

positions which can make a hydrogen bond. A hydrogen bond donor point is assigned to each nitrogen which is bonded to a hydrogen or which is positively charged. A hydrogen bond acceptor point is assigned to each of the following atoms:

- carboxylate oxygen atoms,
- sp^2 oxygen atoms,
- sp^3 oxygen atoms which are not in hydroxyl groups and which are not bound to a nitrogen atom,
- fluorine atoms in a C–F bond, and
- chlorine atoms in a C–Cl bond.

A hydrogen bond doneptor point is assigned to hydroxyl groups as the lone-pair electrons on the oxygen can act as hydrogen bond acceptors while the oxygen can donate the hydrogen to another hydrogen bond acceptor.

Assignment of hydrophobic interaction points is done using a set of rules summarized in Figure 3.1. These rules strive to place a hydrophobic interaction point every 1.5 to 2 carbon atoms along hydrophobic chains and around the edges of hydrophobic rings. The method originally implemented in SLIDE assigned a hydrophobic interaction point to every hydrophobic carbon, i.e., those bonded only to other carbons, hydrogens, or sulfurs, and to the center of hydrophobic rings. This caused a significant overassignment in long aliphatic carbon chains, such as those contained in fatty acid molecules, due to the assignment of points at every carbon position. The previous method also resulted in underassignment

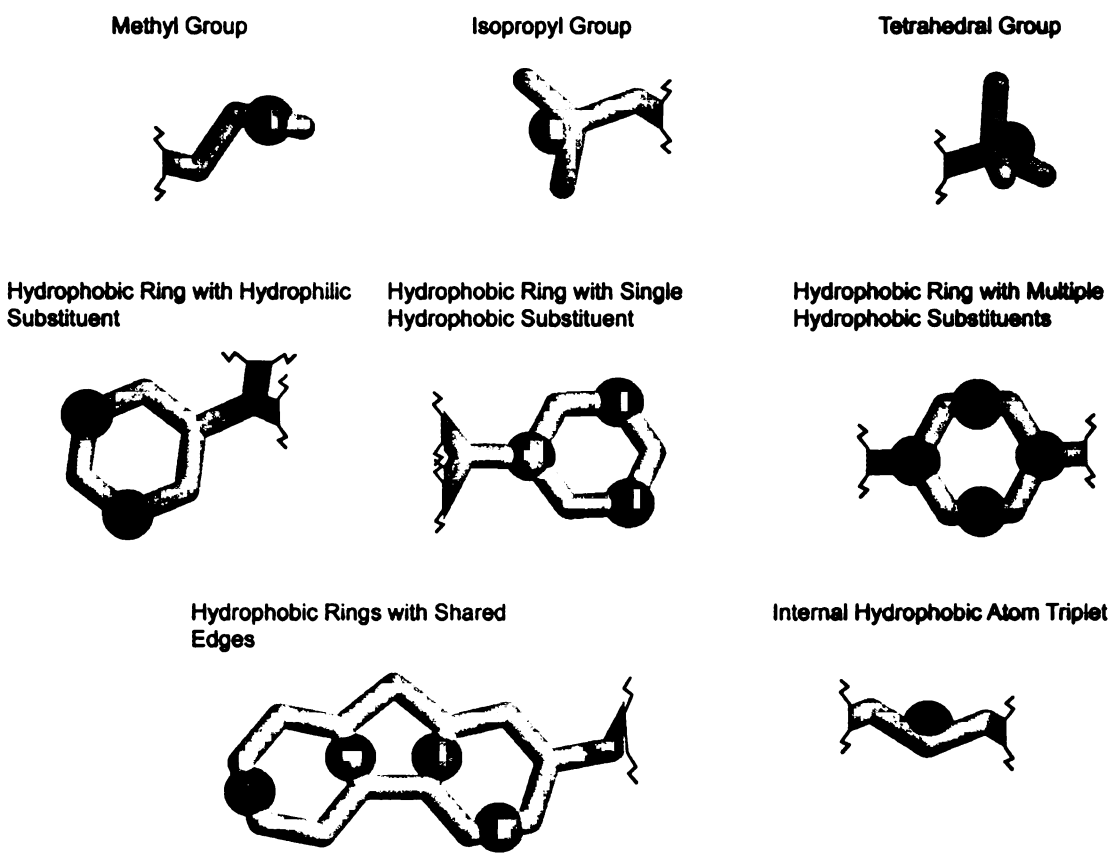


Figure 3.1: Summary of rules used to assign hydrophobic interaction points to molecules in the screening database. The overall goal is to assign a point every 1.5 to 2 carbon atoms.

for hydrophobic rings due to the assignment of a single point in the center of the ring. The new method presented in this work sought to eliminate both the overassignment and underassignment to better represent the hydrophobic character of the molecules.

3.2.3 Identification and Assignment of Protein Template Points

Once the chemistry of molecules in the screening database has been described by the assignment of interaction points, the chemistry of the protein binding site, the template, must be described. This is done by the assignment of a set of template points in the binding site. These points represent favorable interaction positions of potential ligands and represent the negative image, in terms of both shape and chemistry, of the binding site. As for the database molecules' interaction points, the template points are assigned one of four types: hydrogen bond acceptor, hydrogen bond donor, hydrogen bond doneptor (donor and acceptor), or hydrophobic point. There are two methods of placing template points: based on known ligands or in an unbiased approach.

Creation of a Protein Template Based on Known Ligand Binding Modes

If the structure of at least one protein-ligand complex is known, a template can be created which is based on the binding orientation of this known ligand or ligands. This technique is useful when it is desirable to identify potential ligands which represent the chemistry of the known ligand(s) and can be useful to screen a subset of a larger molecular database and/or when a particular set of protein-ligand interactions want to be exploited. The first step of template creation is to assign interaction points to each of the docked ligands, as above in

Section 3.2.2, with each ligand in the reference frame of the protein binding site. These points are then clustered using complete linkage clustering (see Section 2.3.2 for a detailed explanation of complete-linkage clustering) to reduce the set to a representative sample of the ligands' chemistries. In the case where a single ligand is used as a basis for the creation of the template, the resulting template is simply the set of interaction points of the single ligand.

Creation of an Unbiased Template

The unbiased approach to template creation is the preferred method when searching a molecular database for novel potential ligands, i.e., potential ligands which do not resemble ligands in the available protein-ligand complex structures. It is also the only method available when the only available protein structures do not contain a ligand. Hydrogen bond forming points are placed based on geometry of residues residing in the binding site in a technique developed by my colleague Maria Zavodszky.

To identify potential hydrophobic interaction centers, a set of points are initially placed at the vertices of a three-dimensional grid, generally with spacing of 0.5 Å, in the binding site, defined by a box surrounding it. This generally results 10,000-40,000 points, depending on the size of the binding site. An earlier method placed points randomly in this box, but this resulted in uneven sampling which often caused areas of possible interaction to be unrepresented in the resulting set of template points. The set of points is then reduced to include only those within a shell between 3.0 and 5.2 Å from the protein surface. This step generally reduces the set of potential template points to 2,000-10,000. Each of these

remaining points are then checked for hydrophobic character by calculating a hydrophobic enhancement score as follows:

$$\text{Enhancement Score} = \text{Number of Hydrophobic Atoms in Protein Environment} - \text{Number of Hydrophilic Atoms in Protein Environment} \quad (3.1)$$

The protein environment is defined as a sphere of radius 5.2 Å centered on the potential hydrophobic interaction point. This measure encompasses the idea of having a more hydrophobic environment when the point neighborhood contains more hydrophobic atoms. This is in contrast to a measure which involves the average hydrophobic character of the point's protein environment, giving equal weight to an environment with a single hydrophobic atom and one with many hydrophobic atoms. By adjusting the enhancement score used as a cutoff to assign a potential template point as being a hydrophobic template point, the number of hydrophilic atoms allowed within this environment can be adjusted. By examination of the predominantly hydrophobic diethylsilbestrol (DES) ligand of the estrogen receptor (PDB code 3erd; Tanenbaum et al. 1998), a cutoff of 3 was chosen. After determining which points reside in hydrophobic environments, they are clustered using complete linkage clustering, generally with a clustering threshold of 3.0 Å, which provides for an approximate inter-cluster nearest-neighbor distance of 1.5 Å. It is important to note at this point that while the hydrophobic template points are initially placed at grid vertices, the clustering results in them being assigned to arbitrary, non-grid positions. Hydrophobic interaction points which overlap with hydrogen bond interaction points, defined as having

a center-to-center distance of less then 1.5 Å, are eliminated in favor of the geometrically placed hydrogen bond points. The remaining hydrophobic points and the hydrogen bond points constitute the final template.

Earlier Method of Unbiased Template Design

As a major focus of this work is the change in the model of the protein binding site, a brief note on the original method by which the the binding site is modeled is warranted. The original method described the protein binding site as a set of template points which reflect the favored sites of potential interactions with ligands, as the new method does. The original method also assigned points as hydrogen bond acceptor, hydrogen bond donor, hydrogen bond doneptor, or hydrophobic, as does the new method, but these were assigned differently. Instead of being placed at geometrically preferred positions, hydrogen bond points were selected from the shell of all points as those which can form hydrogen bonds to protein atoms. Each type of hydrogen bond point, i.e., acceptor, donor, or doneptor, was then clustered and rechecked for the ability to still participate in a hydrogen bond. The same set of points constituting the shell around the protein was then probed for points which reside in a hydrophobic environment based on the average hydrophobicity of the atoms in the potential point's protein environment. The hydrophobicity for a protein atom in the template point's environment is defined as the average number of instances when a water was bound to the atom per 1000 occurrences of the atom in a study of 53 non-homologous protein structures (Kuhn et al., 1995). As in the new method, points classified as hydrophobic were then clustered and combined with the hydrogen bond points to form

the final template set.

3.2.4 Matching Molecular Interactions to the Template: The Screening Step

Once the protein binding site has been described as a set of template points and each of the database molecules have been described as a set of potential interaction centers, identification of compatible matches can be achieved. As stated in the introduction, the use of a full scale docking approach for each ligand would prove computationally intractable. The approach described here implements several techniques to reduce the overall screening time to enable the screening of databases on the order of a 100,000 molecules in approximately one day.

Use of Hashing Techniques to Rapidly Eliminate Infeasible Dockings

The first step in the screening process is to define a set of four hash tables to describe the triangles present in the set of template triangles. These four hash tables include the following parameters of the template point triangles:

1. the chemical type, i.e., hydrogen bond acceptor, hydrogen bond donor, hydrogen bond doneptor, or hydrophobic, of the three template points which define the triangle (20 hash entries),
2. the perimeter of the template point triangle, generally over a range of 3–25 Å in bins of 0.25 Å (88 hash entries),

3. the length of longest side of the triangle, generally over a range of 1–10 Å in bins of 0.125 Å (72 hash entries), and

4. the length of the shortest side of the triangle, generally over a range of 1–5 Å in bins of 0.125 Å (32 hash entries).

The lengths described above are used in most screening cases, but can be altered for specialized runs. This tabulation, while somewhat computationally costly, is performed only a single time for a screening run.

Identification and Docking of the Anchor Fragment

Each set of three interaction points in a molecule in the database describes an anchor fragment for that ligand (Figure 3.2). An anchor fragment is the substructure of the molecule which is rigid when allowing only torsion angle rotations, i.e., if any of the bonds in the anchor fragment were to be rotated, the triangle defined by the interaction centers would be distorted. The screening process examines all of the triangle mappings in each of the database molecules, leading to an exhaustive approach. Each anchor fragment of a database molecule is used as a potential basis for docking the molecule into the protein binding site. The previously calculated hash tables are used to very quickly eliminate template point triangles which cannot feasibly match the anchor fragment currently being explored, as shown in Figure 3.3. In order to eliminate edge effects that may occur when the measure of a particular geometric property for an anchor fragment triangle lies near the boundary of bins, the template triangles in bins on either side of the matched on are also included.

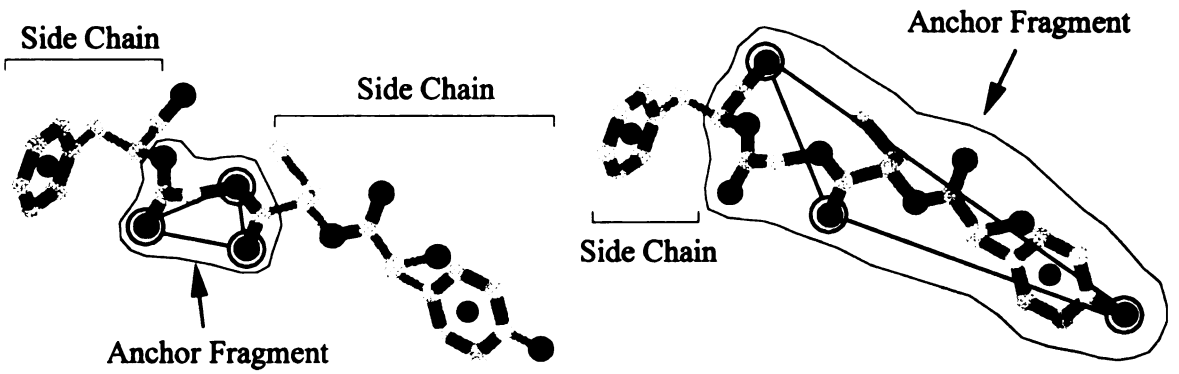


Figure 3.2: Two example anchor fragments for an example molecule. An anchor fragment is defined for each set of interaction point triplets for each molecule in the database. Rotation of any rotatable bond within the anchor fragment would cause a distortion in the anchor fragment triangle, disturbing the initial triangle matching. Portions of the molecule outside of the anchor fragment are ligand side chains and can be rotated to alleviate collisions with the protein.

Once a set of feasible template triangle matches to the anchor fragment is identified, the remaining screening process, summarized in Figure 3.4, is performed. The overall idea is to perform the least computationally expensive steps early on, discarding molecules which fail to meet particular thresholds at each step. In this way, the most costly steps are only performed on molecules most likely to dock. After the set of feasible matching template triangles have been extracted from the hash tables, each template triangle is examined individually. The six possible triangle one-to-one triangle mappings are investigated. Initially, the chemical complementarity of the mappings is checked, e.g., to ensure acceptor database molecule interaction points are mapped onto only template acceptor or doneptor points. Database molecule interaction points are mapped onto the same type of template points since the template represents the negative image of the protein binding site. For all complementary triangle mappings, the distance matrix error (DME) for the side lengths of

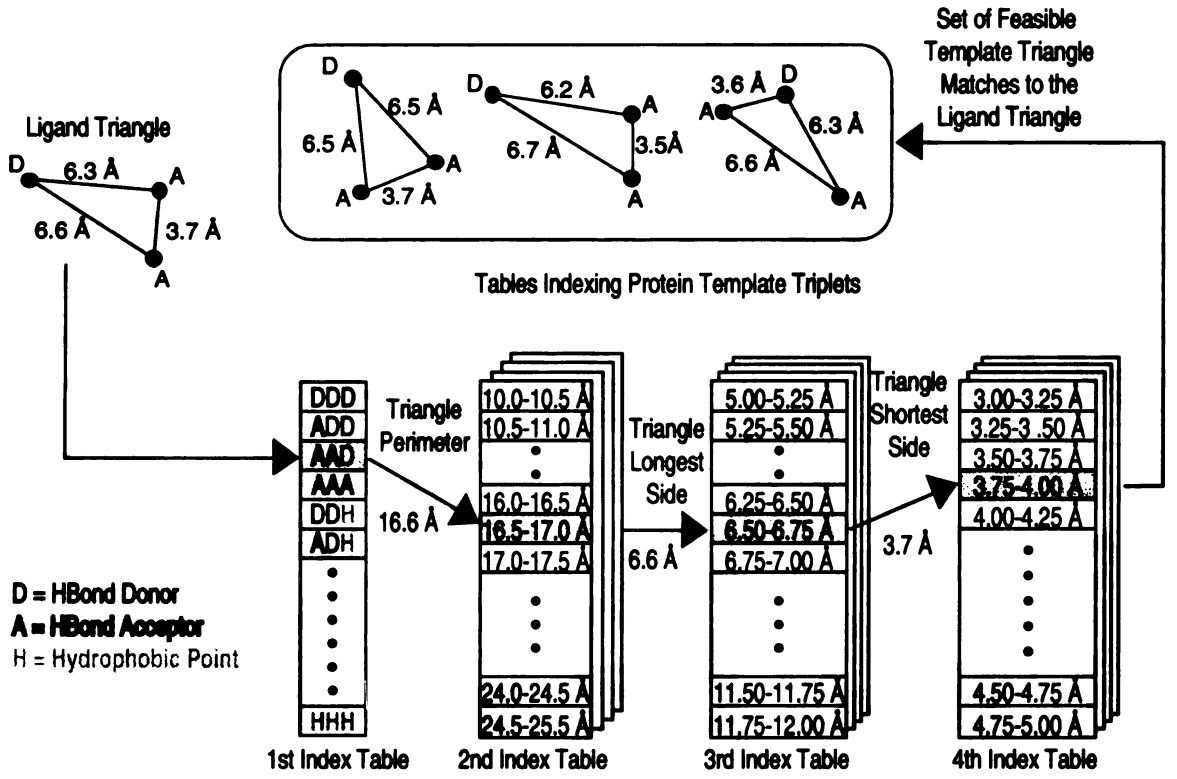


Figure 3.3: Hashing scheme implemented in SLIDE. The template hash tables are calculated a single time at the beginning of the screening run and are used to quickly eliminate template triangles which cannot feasibly match to the ligand anchor fragment triangle under current examination. This initial step reduces the number of template triangles which have to undergo more computationally expensive triangle fitting steps. To eliminate effects which may occur when the measured property is near the boundary between bins, the template triangles in bins adjacent to the matched one are also included as potential further matches.

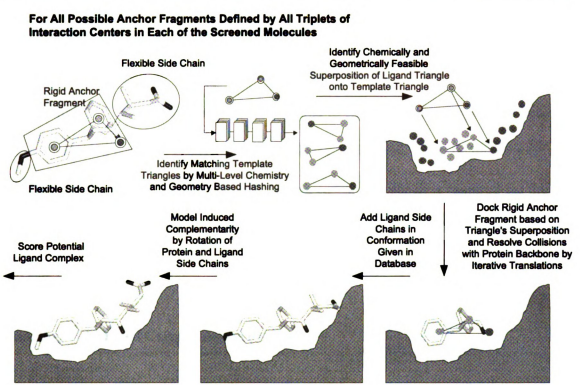


Figure 3.4: Screening algorithm implemented in SLIDE. SLIDE's docking of potential ligands into the binding site is based on mapping triplets of ligand interaction centers (H-bond donors, acceptors doneptors, or hydrophobic) onto triangles of template points located above the protein surface. Feasible template triangles for each possible triplet in a screened molecule are directly accessed via a multi-level hash table, and the corresponding mapping is used to dock the rigid anchor fragment of the potential ligand. Single bonds in the flexible parts of both molecules are rotated to generate a shape-complementary interface, before the complex is scored by the number of intermolecular hydrogen bonds and hydrophobic complementarity of the contact surfaces. In all steps the ligand triplets or dockings that do not meet a particular threshold are discarded.

each molecule side, m_i, and the corresponding template site, t_i is computed as follows:

$$DME = \sqrt{\frac{1}{3}\sum_{i=1}^{3}\left(m_i - t_i\right)^2} \tag{3.2}$$

The DME provides an approximation of the root-mean-square deviation, (RMSD) of the superimposed triangles and is simpler to calculate. Therefore, it can be use as a first approximation to eliminate infeasible matchings and to find the best superposition between the molecule and template triangles. Both the DME and RMSD must be below a defined threshold for the anchor triangle to pass. In both cases, a looser fit is required for the hydrophobic template/molecule interaction point match to allow for the fact that hydrophobic interactions are less specific.

Modeling of Induced Complementarity

Until this point in the algorithm, SLIDE has been working with the database molecule anchor fragment and protein template in a reduced form as a simple triangle of interaction points, but now the algorithm introduces a more realistic model by introducing the atoms included in the database molecule's anchor fragment and the atoms in the protein's main chain and C_β atoms. A check for intermolecular collisions between the anchor fragment atoms and the protein main-chain atoms is performed. If atoms are found to overlap, the anchor fragment is translated away from the protein in the direction which alleviates the collision(s) the minimum distance necessary to remove the overlap(s). This direction and distances can be calculated as the sum of the vectors which lie along the collision axes. This translation is limited, generally to 0.2 Å, to maintain the original triangle matching

and is repeated, up to 100 times, effectively shaking the molecule in the binding site, but in a directed method.

If a docking with no overlaps between the protein main-chain atoms and molecule anchor fragment atoms is found, induced flexibility is modeled by rotation of protein and molecule side chains. In this context, a database molecule's side chains are those fragment connected to the anchor fragment by rotatable bonds (Figure 3.2). If intermolecular overlaps are found between the database molecule and the protein, using a full atom representation, they may be resolved by rotation of a bond that will move either the database molecule atom or the protein side-chain atom involved in the collision. Often times, there are multiple bonds which can be rotated to resolve the collision, each displacing a different set of atoms a different amount.

The approach presented here for modeling induced complementarity and deciding the best bonds to rotate to resolve a set of database molecule/protein collisions is based on mean-field theory (Jackson et al., 1998; Koehl and Delarue, 1994, 1996). This method allows the rotation of the best of any rotatable bond to resolve one or more of the collisions. A key part of this method is the creation of a probability matrix, $P(i, j)$, which describes the probability that a collision i will be resolved by rotation of bond j. Initially, all intermolecular collisions are identified. These form one dimension of the matrix. If more than 20 collisions are identified, the docking is discarded as it is unlikely that this many collisions will be resolvable. All rotatable bonds which can be used to resolve at least one collision and do not cause a new intramolecular collision in the current configuration form the other matrix dimension. It is important to note that there is no differentiation between

database molecule and protein side chains. All rotations which can resolve a particular collision, i.e., all entries $P(i, j)$ such that rotation j resolves collision i, are assigned equal initial probabilities. For each probability entry, $P(i, j)$, a force, $F(i, j)$, is computed that reflects the cost of rotating bond j to resolve collision i. The force assigned in this work is simply the product of the absolute value of the angle of rotation of the bond and the number of atoms which will be displaced by rotating the bond. Such a force penalizes rotating a larger number of atoms a larger number of degrees, as this is more likely to cause additional collisions elsewhere.

After initialization, several iterations of mean-field optimization are performed by updating the probability matrix P to converge to high probabilities for those rotations which provide the lowest cost conformational change for both the database molecule and the protein and which resolve the largest number of collisions. In each iteration, a mean force, $E(i, j)$, is computed for each rotation, as follows:

$$E(i,j) = F(i,j) + \sum_{h \neq i,\, k} dep[(i,j),(h,k)] \cdot P(h,k) \cdot F(h,k) \qquad (3.3)$$

The value of $dep[(i, j), (h, k)]$ is a measure of the dependency between probability entries $P(i, j)$ and $P(h, k)$. It is set to -1.0 if both entries refer to the same bond and both rotations are in the same direction, i.e., $j = k$. If this is the case, two collisions can be resolved at once by a rotation of this single bond. Assignment of a dependency of -1.0 results in a lower mean force, $E(i, j)$, thereby favoring this rotation. If probability entries $P(i, j)$ and $P(h, k)$ refer to the same bond ($j = k$), but the rotations are in opposite directions, the dependency is set to $+1.0$, penalizing this rotation. If bond j lies on the path to bond k,

e.g., bond j is between C_β and C_γ and bond k is between C_γ and C_δ of the same side chain, the dependency is also set to +1.0. This is penalized, since if rotation (i, j) were applied, bond k would be displaced, invalidating the assumptions about the current conformation of the both the database molecule and the protein in the current optimization iteration.

At the end of each iteration, the entries of the probability matrix are updated based on the mean force using the Boltzmann principle:

$$P(i,j) = \frac{e^{-E(i,j)/\mu}}{\sum_k e^{-E(i,k)/\mu}} \tag{3.4}$$

μ is the average value of all computed mean forces. Convergence of the values in the probability matrix is generally seen in fewer than ten iterations, and those rotations with the highest probability are chosen to resolve the collisions. It is necessary to check for negative correlations between bonds again at this time. Although checked for during the mean field optimization, two correlated bonds can receive high probabilities if they are the only bonds which will resolve a particular set of collisions or if alternative rotations are much more expensive. Also, it is not possible to anticipate complex dependencies, e.g., which ligand rotations influence protein bonds related to other collisions, during mean field optimization. Since it is unlikely that all intermolecular collisions can be resolved by a single application of optimization, up to 10 cycles are executed. Database molecule dockings are discarded if they have more than 20 collisions at any time during the optimization or have remaining collisions after 10 cycles of mean field optimization.

Scoring

Once a collision free complex is identified, the final step in determining if the current complex is a valid potential ligand is to calculate the chemical complementarity between the database molecule and the protein. As a first pass, all complexes with poor shape complementarity are eliminated. In the 89 complexes (Eldridge et al., 1997) used to tune SLIDE's scoring function, an average of 88% of the ligand carbon atoms where located within 4.0 Å of a protein atom, and all ligands buried at least 55% of their carbon atoms against the protein surface (Figure 3.5). Based on this observation, all dockings in SLIDE with fewer than 50% of carbon atoms buried, i.e., within 4.0 Å of a protein atom, are discarded.

The complexes are then assessed for chemical complementarity by a scoring function, SCORE(P, M), which consists of a term for the number of intermolecular hydrogen bonds formed between the protein, P, and the database molecule, M, HBOND(P, M), and a term for hydrophobic complementarity between the protein and the database molecule, HPHOB(P, M). For calculation of the number of intermolecular hydrogen bonds, hydrogen bonding is considered for cases where the distance between donor and acceptor is less than 3.5 Å. For proteins and database molecules with no hydrogen atoms provided in the crystallographic structure, the positions of the hydrogens are computed based on known bond angle and length constraints and optimal placement for hydrogen bonding when several positions are possible (Hooft et al., 1996), such as for the hydrogen in a hydroxyl group. For cases when the hydrogen atoms are given in the structure, their positions are taken as given. Rotatable hydrogens are rotated to optimize hydrogen bonding when appli-

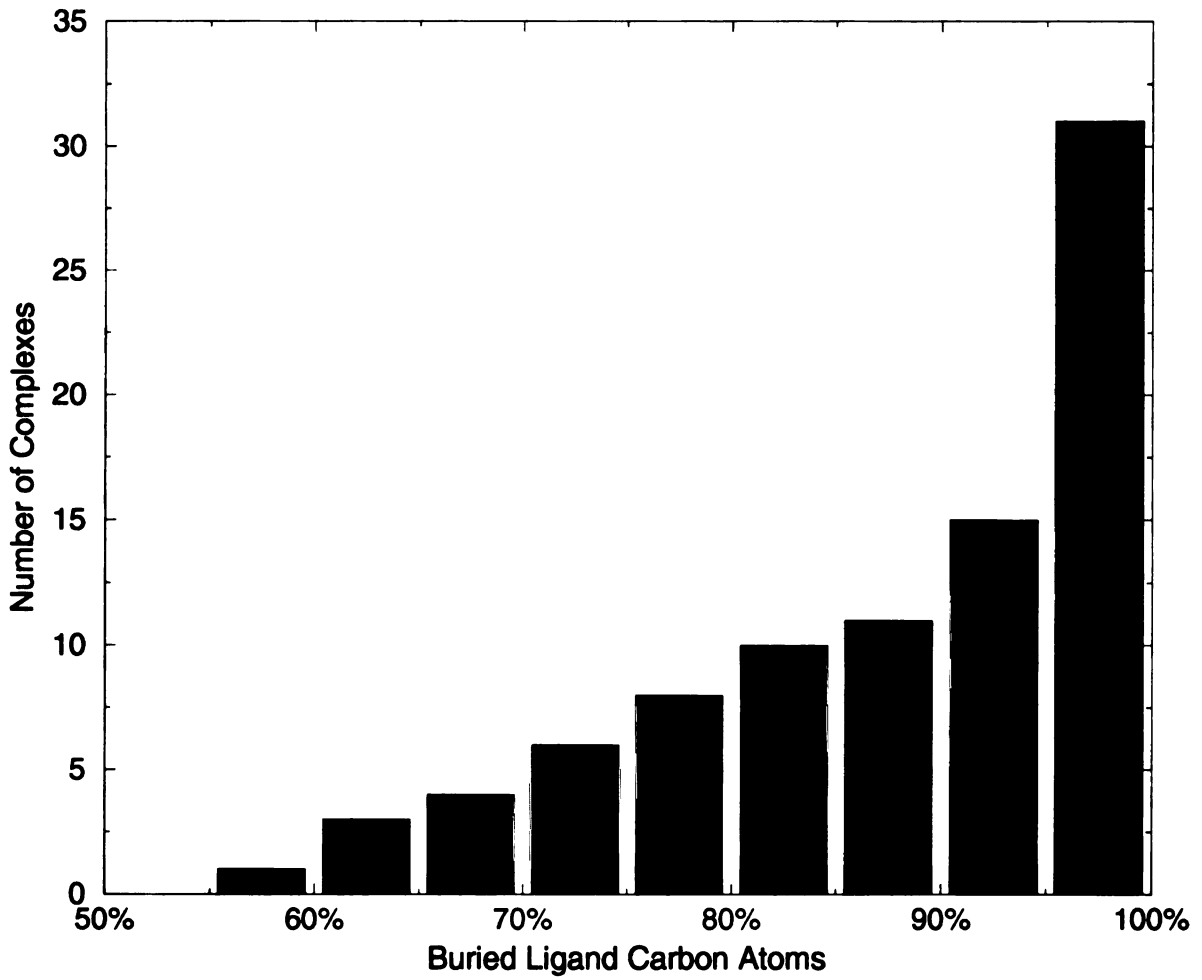


Figure 3.5: Percentage of buried carbon ligand atoms in the 89 complexes used to tune the SLIDE scoring function. Complexes were derived from Eldridge et al. (1997). All complexes had at least 55% of their carbons buried against the protein surface, defined as being within 4.0 Å of any protein atom. Based on this observation, SLIDE rejects any ligand docking in which less than 50% of the carbon atoms are buried as an initial screen before scoring.

cable. Donation to multiple acceptors is allowed if the angular constraints are fulfilled. The following constraints are used to qualify a hydrogen bond: (1) a donor· · ·acceptor distance of 3.5 Å, (2) a donor–hydrogen distance of 1.0 Å, and (3) a donor–hydrogen· · ·acceptor angle of 120° to 180° (Habermann and Murphy, 1996).

Hydrophobic complementarity is calculated based on the residue and atoms type in the protein and the atom type of the database molecule. The values are calculated as the average number of hydrations per 1000 occurrences of the atom and are taken from a statistical study of 56 protein structures (Kuhn et al., 1995). (The values for protein atoms are from Table II and the values for ligand atoms are from Table III). The hydrophilicity values range from 0, maximally hydrophobic, C_α, to 635, maximally hydrophilic, the tyrosyl hydroxyl oxygen atom. The hydrophobic complementarity between database molecule M and protein P is calculated as follows:

$$\mathrm{HPHOB}(P, M) = \sum_{\substack{m_i \in M \\ \#P_i > 0}} \frac{\mathrm{avg}\{h'(M_i), \bar{h}(P_i)\}}{\max\{\mathrm{abs}(h'(M_i) - \bar{h}(P_i)), 32\}} \tag{3.5}$$

where

$$h'(M_i) = \max\{317 - h(M_i), 0\} \tag{3.6}$$

considers only the hydrophobic contribution of the database molecule atoms, M_i, since values larger than 317 refer hydrophilic atoms. If the atom is hydrophilic, i.e., the atom's hydrophilicity value is > 317, $317 - h(M_i)$ is < 0 and the hydrophobic character of the database molecule atom becomes zero. The hydrophobicity, $\bar{h}(P_i)$, of the protein neighborhood P_i for a single database molecule atom, M_i, is the average hydrophobic character

of all protein atoms, p_j, within 4.0 Å of the database molecule atom:

$$\bar{h}(P_i) = \max\left\{\left(317 - \frac{1}{\#P_i} \cdot \sum_{p_j \in P_i} h(p_j)\right), 0\right\} \tag{3.7}$$

The denominator in each element of the hydrophobic score, $\mathrm{HPHOB}(P, M)$ is set to be greater than or equal to 32, which is 10% of the maximum score for a single database molecule atom. This prevents a very few contacts with a large difference from dominating the hydrophobic score. The overall score for a complex is simply a linear combination of the hydrophobic complementarity term and the number of intermolecular hydrogen bonds:

$$\mathrm{SCORE}(P, M) = A \cdot \mathrm{HPHOB}(P, M) + B \cdot \mathrm{HBONDS}(P, M) \tag{3.8}$$

The weights A and B have been chosen to optimize the fit between the scoring function and the affinities of 89 high-resolution complexes (Eldridge et al., 1997). These complexes had an average hydrophobic complementary term of 28.7 and made an average of 7.8 hydrogen bonds. Values of 0.59 and 2.76 for A and B, respectively, give a reasonable approximation to the series of measured affinity values (linear correlation coefficient of 0.615), which yields a relative contribution of 1.3:1.0 of the hydrogen bond term over the hydrophobicity term. The overall goal of SLIDE's scoring function is to provide a relative rank for the potential ligands. At this point, a minimum score cutoff can be used to include only favorable complexes, resulting in a set of 100-500 potential ligands. Optimization of the binding mode and prediction of the binding affinity can be done using a more detailed conformational search and/or docking algorithm on the top-ranked potential ligands.

Another aspect of the scoring function is the modeling of solvation in SLIDE, though not used in the work presented here. In other screening and docking algorithms, water molecules are generally either considered to be a fixed part of the binding site or are ignored. Some approaches seek to identify favorable water molecule positions in the protein binding site prior to docking (Rarey et al., 1999) or seek to solvate the ligand (Shoichet et al., 1999), but the water molecules in question are still fixed in position during the screening. It is well known that while some water molecules are key to providing high affinity binding (Ladbury, 1996), but it has also been shown that generally there are many water molecules which are displaced from the binding site upon binding of the ligand (Raymer et al., 1997). SLIDE uses the approach developed by Raymer et al. (1997) in *Consolv*, which is a ***k***-nearest-neighbor/genetic algorithm application to predict which water molecules are conserved and which are displaced prior to the screening run. *Consolv* uses only information about the protein binding site and uses no information about the bound ligand. The information about *Consolv*'s predictions is used in SLIDE via inclusion of the prediction confidence, the proportion of votes for conservation in the ***k***-nearest-neighbor classifier, for the water molecules predicted to be conserved. SLIDE can displace water molecules which are predicted to be conserved, but a penalty proportional to the confidence of the conserved prediction is assessed in the score. Water molecules which are predicted to be displaced are removed from the binding site before screening.

One post-screening method that can be used is to apply a more costly, but more sophisticated, scoring function to further refine the ranking of potential ligands. The empirically based scoring function DrugScore (Gohlke et al., 2000a,b) is such a scoring function.

DrugScore was created by examination of crystallographic structures for a series of protein-ligand complexes to generate a database of radial distribution functions for each type of possible ligand atom-protein atom pair. The favorability/unfavorability of a SLIDE generated protein-ligand complex can be computed by comparing the distance between each ligand atom-protein atom pair in the docking to the functions derived from the empirical study. Distances which match common distances are considered favorable, while distances which match those only rarely seen are unfavorable.

3.2.5 Testing Databases

To examine the effects of changes to SLIDE, human α-thrombin and glutathione S-transferase (GST) were selected. Thrombin and GST are good cases to test for several reasons: there is a high resolution crystal structure of the ligand-free protein available in the Protein Data Bank (PDB), there are several high resolution protein-ligand complex structures available, and there is a moderate diversity of ligands in the complex structure of each protein. For thrombin, no truly ligand-free structure is available, but this work focuses on active-site ligands so structures with only non-active site ligands can be used as a ligand-free structure. The protein structures, shown in Tables 3.1 and 3.2, were chosen to provide a set of unique ligands so as to prevent biasing the results towards one type of ligand. In cases when more than one PDB entry contained the same ligand, the structure with the best resolution was used.

Images in this dissertation are presented in color.

Table 3.1: 42 Thrombin protein-ligand complexes used for testing SLIDE modifications

PDB Code	Ligand	Resolution (Å)
1a2c	Aeruginosin298-A	2.1
1a3b	Borolog1	1.8
1a3e	Borolog2	1.9
1a46	β-strand mimetic inhibitor	2.1
1a4w	Ans-Arg-2ep-Kth	1.8
1a5g	Bic-Arg-Eoa	2.1
1a61	Mol-Arg-Lom	2.2
1ad8	MDL103752	2.0
1ae8	Eoc-D-Phe-Pro-azaLys-Onp	2.0
1afe	Cbz-pro-azaLys-Onp	2.0
1aht	p-Amidino-phenyl-pyruvate	1.6
1ai8	$PhCH_2OCO$-D-Dpa-Pro-boroMpg	1.9
1aix	$PhCH_2OCO$-D-Dpa-Pro-boroVal	2.1
1awf	GR133487	2.2
1awh	GR133686	3.0
1ay6	Hmf-Pro-Arg-Hho	1.8
1b5g	Bcc-Arg-Thz	2.1
1ba8	Pms-Ron-Gly-Arg	1.8
1bb0	Pms-Ron-Gly-3ga	2.1
1bcu	Proflavin	2.0
1bhx	SDZ 229-357	2.3
1bmm	BMS-186282	2.6
1bmn	BMS-189090	2.8
1dwb	Benzamidine	3.2
1dwc	MD-805 (Argatroban)	3.0
1dwd	NAPAP	3.0
1fpc	Ans-Arg-Epi (DAPA)	2.3
1hdt	Alg-Phe-Alo-Phe-CH_3 (BMS-183507)	2.6
1lhc	Ac-D-Phe-Pro-boroArg-OH	2.0
1lhd	Ac-D-Phe-Pro-boroLys-OH	2.3
1lhe	Ac-D-Phe-Pro-boro-N-butyl-amidino-glycine-OH	2.2
1lhg	Ac-D-Phe-Pro-borohomoornithine-OH	2.2
1nrs	Leu-Asp-Pro-Arg	2.4
1ppb	PPACK	1.9
1tbz	Dpn-Pro-Prg-Bot	2.3

Continued on next page.

Table 3.1 (cont'd)

PDB Code	Ligand	Resolution (Å)
1tmb	Cyclotheonamide A	2.3
1tmt	Phe-Pro-Arg	2.2
1tom	Methyl-Phe-Pro-amino-cyclohexylglycine	1.8
1uma	N,N-dimethylcarbamoyl-α-azalysine	2.0
3hat	Fibrinopeptide A mimic	2.5
7kme	SEL2711.	2.1
8kme	SEL2770.	2.1

3.3 Results

Presented below are the results of a series of tests on the changes made to SLIDE during the course of this research work. Previously published reports have shown that SLIDE can identify known ligands from a large database, can rank known ligands as better potential ligands, and can correctly dock known ligands (Schnecke et al., 1998; Schnecke and Kuhn, 1999, 2000a,b).

3.3.1 Visual Examination of the New Template and Interaction Point Methods

A first step was a visual examination of the template changes and the hydrophobic interaction point assignment changes. Example unbiased templates for the estrogen receptor (ER) in comparison to the diethylstilbestrol (DES) ligand using the original and the modified methods are shown in Figure 3.6 (original method) and Figure 3.7 (new method). The figures show a much better representation in the hydrophobic space occupied by the two

Table 3.2: 16 GST protein-ligand complexes used for testing SLIDE modifications

PDB Code	Ligand	Resolution (Å)
10gs	Benzylcysteine phenylglycine	2.2
12gs	S-nonyl-cysteine	2.1
13gs	Sulfasalazine (SAS)	1.9
18gs	1-(S-glutathionyl)-2,4-dinitrobenzene	1.9
19gs	Phenol-1,2,3,4-tetrabromophthalein-3′,3″-disulfonic acid ion	1.9
1aqv	p-Bromobenzylglutathione	1.9
1aqw	Glutathione	1.8
1aqx	S-(2,3,6-trinitrophenyl)cysteine	2.0
1gss	S-hexylcystine	2.8
1pgt	S-hexylglutathione	1.8
20gs	Cibacron blue	2.5
21gs	Chlorambucil	1.9
2gss	Ethacrynic acid (EAA)	1.9
2pgt	(9R,10R)-9-(S-glutathionyl)-10-hydroxy-9,10 dihydrophenanthrene	1.9
3gss	Ethacrynic acid-Glutathione conjugate	1.9
3pgt	(+)-Anti-BPDE	2.1

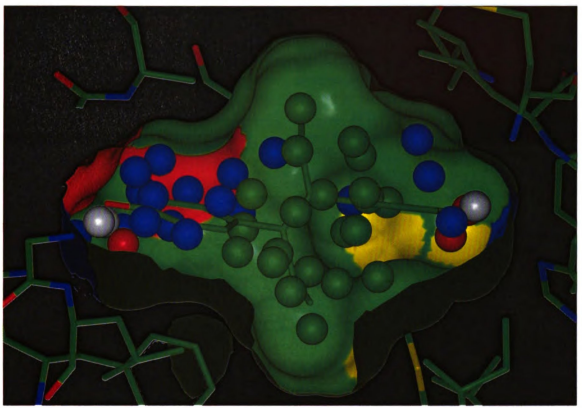


Figure 3.6: Unbiased template for the estrogen receptor generated using the original method. The template points are represented by spheres colored according to type: green, hydrophobic; red, hydrogen bond acceptor; blue, hydrogen bond donor; and white, hydrogen bond doneptor (donor/acceptor). The solvent-accessible ER surface and the diethylstilbestrol (DES) ligand are colored by atom (green, carbon; blue, nitrogen; red, oxygen; yellow, sulfur). The DES ligand is shown for comparison only and is not used in the template generation. Compared to new method in Figure 3.7, one can see the the very poor representation of the hydrophobic area corresponding to the left benzyl ring of DES and the moderately poor representation of the right DES benzyl ring. (planar in the figure).

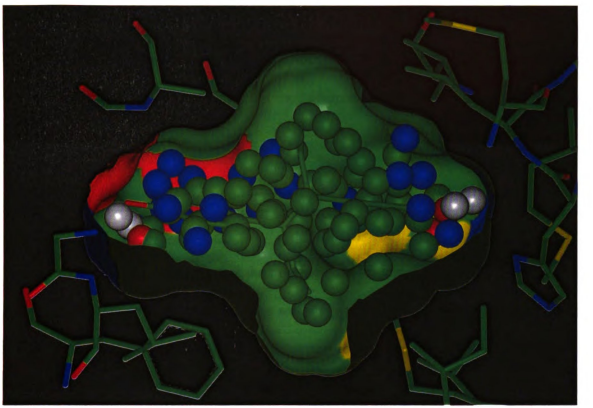


Figure 3.7: Unbiased template for the estrogen receptor generated using the new method. The template points are represented by spheres colored according to type: green, hydrophobic; red, hydrogen bond acceptor; blue, hydrogen bond donor; and white, hydrogen bond doneptor (donor/acceptor). The solvent-accessible ER surface and the diethylstilbestrol (DES) ligand are colored by atom (green, carbon; blue, nitrogen; red, oxygen; yellow, sulfur). The DES ligand is shown for comparison only and is not used in the template generation. Comparison to the original method in Figure 3.6 shows a better representation of the hydrophobic DES rings, especially in the area corresponding to the DES benzyl ring at left.

benzyl rings of the DES ligand.

Additional visual analysis was done for the changed hydrophobic interaction point assignment method. A comparison between assignments using both the original and new methods is shown in Figure 3.8. The new method shows a more balanced hydrophobic representation, especially between the hydrophobic rings and the aliphatic tails. The overrepresentation seen in such tails and the underrepresentation seen in rings has been eliminated.

3.3.2 SLIDE Docking of Known Ligands

While visual examination shows that the new methods for template point assignment and hydrophobic interaction point assignment are likely to be an improvement, further work needed to be done with screening and docking runs to confirm this result. A series of experiments were performed with the original and new template methods and with the original and new hydrophobic interaction point assignment methods. In one set, the template model was held constant while the hydrophobic interaction point method was changed, isolating the effects of using the new interaction point method. In another set, the interaction point model was held constant while the template method was changed from the original one to the new one, isolating the effects of changing the template. In a third set, both the template and interaction point models were changed to judge the combined effects. A graphical summary of these experiments is shown in Figure 3.9.

Initial tests examined dockings of the known ligands with SLIDE compared to the crystallographic structure dockings. Dockings of the known ligands from the complex struc-

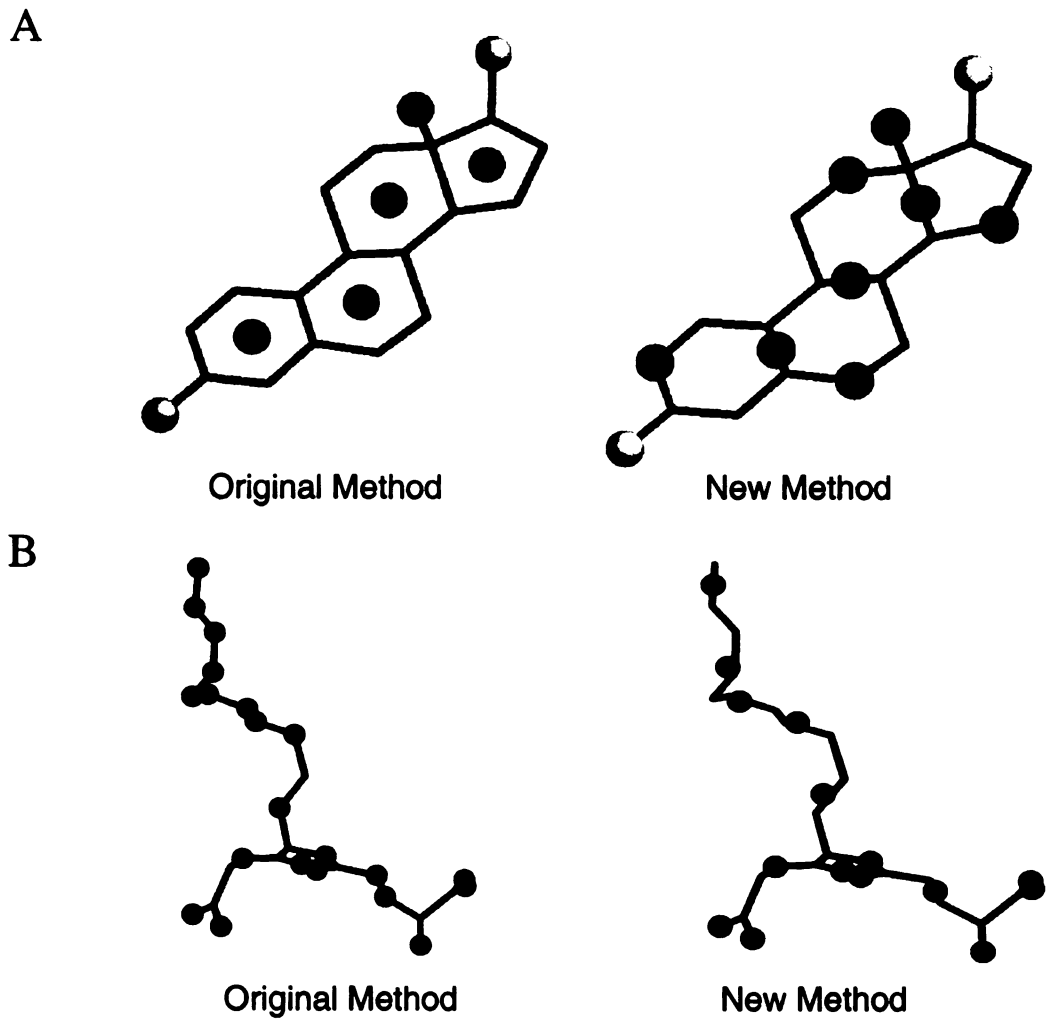


Figure 3.8: A comparison of hydrophobic interaction point assignment methods for (A) estradiol from CSD code BEQJIQ (Parrish and Pinkerton, 1999) and (B) S-nonyl-glutathione from PDB code 12gs (Oakley et al., 1997). Interaction points are represented as spheres, colored by type: hydrogen bond acceptor, red; hydrogen bond donor, blue; hydrogen bond doneptor, white; and hydrophobic, green. The new assignment method provides a more balanced hydrophobic representation of the molecules, especially in the hydrophobic tail of the S-nonyl-glutathione.

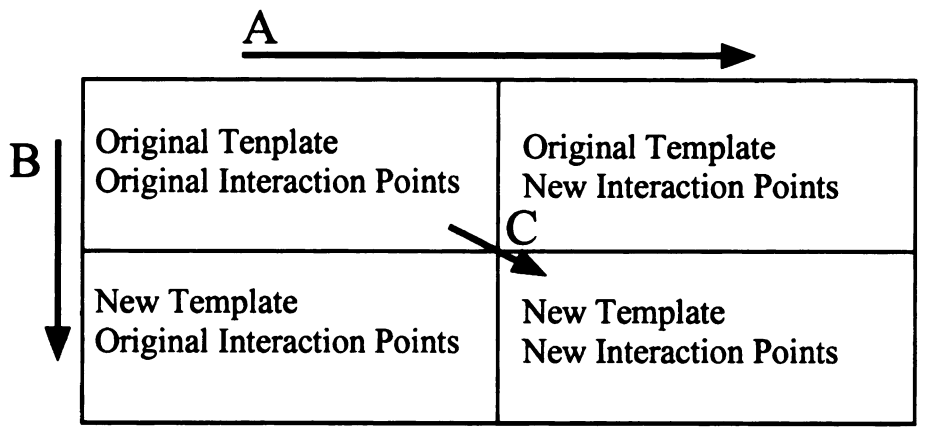


Figure 3.9: Overview of experiments to test interaction modeling modifications made to SLIDE. (A) represents experiments in which the template was maintained as a control and the method for interaction point assignment was altered (results in Tables 3.3 and 3.4). (B) represents experiments in which the interaction points were maintained as a control and the template generation method was altered (results in Tables 3.5 and 3.6). (C) represents tables in which both the template generation method and interaction point assignment method were changed (results in Tables 3.7 and 3.8).

tures were docked into the ligand-free structure via superposition of the protein's active-site residues. Table 3.3 shows the number of known ligands which were docked for the experiments in which the template was held constant but the hydrophobic interaction point assignment method was changed from the original one to the new one. The table is interpreted such that changing the interaction point method to assign interaction points to the known thrombin ligands from the original method to the new method while using the template derived by the original method for the thrombin binding site enables SLIDE to dock two known ligands not docked using the original interaction point method, but one known ligand is no longer docked, yielding a net of +1 (without using an RMSD cutoff). There were 34 known ligands that were docked using both the original and new interaction point methods along with the original template point method. The key lines to note

Table 3.3: Known ligand dockings for constant template point method experiments

Template Used	RMSD Cutoff (Å)*	Dockings Gained	Dockings Maintained	Dockings Lost	Net Docked
Thrombin					
Original	None	2	34	1	+1
	2.5	**2**	**26**	**3**	**-1**
	1.0	6	4	5	+1
New	None	1	33	0	+1
	2.5	**1**	**31**	**0**	**+1**
	1.0	1	19	2	-1
GST					
Original	None	0	15	0	0
	2.5	**2**	**6**	**1**	**+1**
	1.0	1	1	1	0
New	None	0	16	0	0
	2.5	**0**	**12**	**0**	**0**
	1.0	0	12	0	0

*The RMSD cutoff is the maximum RMSD relative to the crystallographic docking that a SLIDE docking must have to be considered successful. None indicates that all dockings were allowed. An RMSD of 2.5 Å means the SLIDE docking must be moderately close to the crystallographic docking, while an RMSD limit of 1.0 Å means only very close dockings are considered successful.

are the RMSD cutoff 2.5 Å lines (shown in boldface). These show data for the dockings which were reasonably close to the crystallographic complex, and that while a docking is sometimes lost, generally additional dockings are successfully achieved using the new interaction point method. One can now ask if there is a quantitative improvement in the dockings of ligands that were docked using both interaction point methods. The data, shown in Table 3.4, clearly indicate that these dockings are better. More dockings had a lower RMSD using the modified interaction point assignment method and the average change was generally near or below zero. Dockings of known ligands using the new interaction point method also generally had improved scores, using either the scoring function implemented in SLIDE or DrugScore, compared to dockings using the original interaction point method.

A set of experiments similar to the above was performed by maintaining the interaction point method constant and altering the template creation method used. Table 3.5 shows the number of known ligands docked in these experiments. This table is interpreted the same as Table 3.3. In these experiments, there are a few cases where changing the template causes a loss of a few ligands, but when examining only the ligands docked at least reasonably well, RMSD cutoff of 2.5 Å, an average of 3.3 additional ligands are docked that were not docked using the original method. It is especially key in the GST experiments where only a few (7 to 8) known ligands are reasonably well docked using the original interaction point method, but 11 to 12 of the 16 (the number maintained + the number gained) are reasonably docked using the new template method. The ability to correctly dock known ligands is an important feature of the SLIDE algorithm and has an

Table 3.4: RMSDs for known ligands docked with both the original and modified interaction point methods for constant template point method experiments

Template Used	RMSD Cutoff (Å)[1]	Dockings with Better RMSD	Dockings with Equal RMSD	Dockings with Worse RMSD	Mean RMSD Change (Å)[2]	Standard Deviation
Thrombin						
Original	None	17	2	15	-0.02	0.05
	2.5	**14**	**0**	**12**	**+0.05**	**0.05**
	1.0	3	0	1	-0.16	0.02
New	None	13	10	10	+0.01	0.05
	2.5	**13**	**10**	**8**	**+0.002**	**0.05**
	1.0	9	5	5	-0.04	0.06
GST						
Original	None	5	0	10	-1.36	0.79
	2.5	**2**	**0**	**4**	**+0.02**	**0.75**
	1.0	0	0	1	0.14	—
New	None	6	6	4	-2.01	0.06
	2.5	**3**	**6**	**3**	**+0.03**	**0.01**
	1.0	3	6	3	+0.03	0.01

For an overview of these experiments, refer to Figure 3.9. Please also note that it is not possible to directly compare values within a column as they relate to varying numbers of docked ligands. The changes reported in each line of this table are equivalent to traversing left to right in Figure 3.9 (A).

[1]The RMSD cutoff is the maximum RMSD relative to the crystallographic docking that a SLIDE docking must have to be considered successful. None indicates that all dockings were allowed. An RMSD of 2.5 Å means the SLIDE docking must be moderately close to the crystallographic docking, while an RMSD limit of 1.0 Å means only very close dockings are considered successful.

[2]RMSD change is scaled such that values < 0.0 are better while values > 0.0 are worse.

Table 3.5: Known ligand dockings for constant interaction point assignment method experiments

Interaction Point Method Used	RMSD Cutoff (Å)	Dockings Gained	Dockings Maintained	Dockings Lost	Net Docked
Thrombin					
Original	None	1	32	3	-2
	2.5	**4**	**27**	**2**	**+2**
	1.0	14	7	2	+12
New	None	1	33	3	-2
	2.5	**6**	**26**	**2**	**+4**
	1.0	12	8	2	+10
GST					
Original	None	1	15	0	+1
	2.5	**4**	**7**	**0**	**+4**
	1.0	10	2	0	+10
New	None	1	15	1	0
	2.5	**4**	**8**	**0**	**+4**
	1.0	10	2	0	+10

Table 3.6: RMSDs for known ligands docked with both the original and modified template method for constant interaction point method experiments

Interaction Point Method Used	RMSD Cutoff (Å)	Dockings with Better RMSD	Dockings with Equal RMSD	Dockings with Worse RMSD	Mean RMSD Change (Å)*	Standard Deviation
Thrombin						
Original	None	27	0	5	-0.67	0.07
	2.5	**23**	**0**	**4**	**-0.38**	**0.02**
	1.0	5	0	2	-0.20	0.19
New	None	25	0	8	-0.77	0.03
	2.5	**20**	**0**	**6**	**-0.37**	**0.12**
	1.0	6	0	1	-0.17	0.001
GST						
Original	None	13	0	2	-1.36	0.31
	2.5	**7**	**0**	**0**	**-0.84**	**0.03**
	1.0	2	0	0	-0.22	0.14
New	None	15	0	0	-2.01	1.04
	2.5	**8**	**0**	**0**	**-0.80**	**0.11**
	1.0	2	0	0	-0.44	0.06

For an overview of these experiments, refer to Figure 3.9. Please also note that it is not possible to directly compare values within a column as they relate to varying numbers of docked ligands. The changes reported in each line of this table are equivalent to traversing top to bottom in Figure 3.9 (B).
*RMSD change is scaled such that values < 0.0 are better while values > 0.0 are worse.

important role in ensuring that ligands are not missed during screening runs. Once again, one can question if these dockings are better, and the answer is yes. Table 3.6 shows that in all cases, there are significantly more ligands are docked with better RMSDs using the new template method compared to the original method. Also, in all of the experiments performed, the mean RMSD decreased, in some cases by a large amount, the most dramatic change being for the GST using the new template method for all docked ligands. Given the relatively small size of these ligands, a drop of mean RMSD of 2.01 Å is very significant.

Table 3.7: Known ligand dockings for combined new template and interaction point methods compared to the original methods

RMSD Cutoff (Å)	Dockings Gained	Dockings Maintained	Dockings Lost	Net Docked
Thrombin				
None	1	33	2	-1
2.5	**4**	**28**	**1**	**+3**
1.0	14	6	3	+11
GST				
None	1	15	0	+1
2.5	**5**	**7**	**0**	**+5**
1.0	10	2	0	+8

Changes in scores also show a similar improvement in known ligand docking.

Now that it has been shown that the new interaction point assignment method and the new template point method are improvements independently, the next experiment compared dockings done with both new methods to dockings done with both original methods. The combination of new methods also shows a significant increase in the ability to dock known ligands, as seen in Table 3.7. Once again, there is generally an increase in the number of known ligands docked, especially when focusing on only those ligands which were reasonably well docked in comparison to the crystallographic docking. This indicates that the combination of new template method and new interaction point method is an improvement over using the original methods for both template generation and interaction point assignment. Also, the combination of new methods generally does better than the introduction of either the new template method or the new interaction point method by itself. Table 3.8 shows the improvements in scores for each of the scoring functions, SLIDE's built-in scoring function, DrugScore, and RMSD, for the combined new methods

versus the combined original methods for the thrombin and GST known ligands. From this, it can be seen that the dockings using the combination of new template and interaction point methods are quantitatively better in all measures with all successful docking RMSD cutoffs. The improvement is more pronounced for the GST ligands, but is still significant for the thrombin ligands. In fact, for GST, 13 of the 16 docked ligands using the combined new methods have RMSDs relative to the crystallographic dockings lower than 1.0 Å, while only two of the dockings using the original methods have RMSDs lower than 1.0 Å. The scores and RMSDs are as good or better than seen using either single change, indicating that using both the new method for template creation and for interaction point identification is better than using either change on its own.

The question may arise as to the nature of the known ligands which cannot dock to the protein. Most of the known thrombin ligand docking failures, 63% of the new method docking attempts and 86% of the original method docking attempts, occur at the side-chain collision resolution stage. This result is consistent with many dockings runs seen during the course of SLIDE development, indicating that this stage is the most critical for docking, which is not surprising given the complexity needed to effectively model induced fit. Other failures commonly occur at the ligand anchor fragment/protein main chain collision resolution stage, which is likely related to the lack of modeling of protein main-chain conformational changes in the available docking and screening tools, including SLIDE.

In addition to quantitative measure of the quality of dockings of known ligands, one can examine the dockings visually. Figures 3.10 and 3.11 show two such comparisons. In both these cases, the docking determined using the new template and interaction point

Table 3.8: Scores for the known ligands docked using the new template and new interaction point methods compared to using the original methods

RMSD Cutoff (Å)	Scoring Function Used	Dockings with Better Score	Equal Score	Worse Score	Mean Score Change[1]	Standard Deviation
Thrombin						
None	SLIDE [2]	27	0	6	+6.8	2.8
None	DrugScore[3]	27	0	6	-59300	6200
None	RMSD	26	0	7	-0.63	0.01
2.5	SLIDE	**22**	**0**	**6**	**+6.1**	**3.2**
2.5	**DrugScore**	**23**	**0**	**5**	**-51800**	**8200**
2.5	**RMSD**	**22**	**0**	**6**	**-0.37**	**0.04**
1.0	SLIDE	5	0	1	+6.0	1.1
1.0	DrugScore	6	0	0	-61400	12400
1.0	RMSD	4	0	2	-0.233	0.18
GST						
None	SLIDE [4]	15	0	0	+13.5	2.45
None	DrugScore[5]	14	0	1	-127700	5100
None	RMSD	14	0	1	-1.59	0.25
2.5	SLIDE	**7**	**0**	**0**	**+9.4**	**1.53**
2.5	**DrugScore**	**6**	**0**	**1**	**-92200**	**30000**
2.5	**RMSD**	**6**	**0**	**1**	**-0.79**	**0.03**
1.0	SLIDE	2	0	0	+13.1	0.17
1.0	DrugScore	1	0	1	-810	47000
1.0	RMSD	1	0	1	-0.15	0.21

For an overview of these experiments, refer to Figure 3.9. Please also note that it is not possible to directly compare values within a column as they relate to varying numbers of docked ligands. The changes reported in each line of this table are equivalent to traversing diagonally upper-left to lower-right in Figure 3.9 (C).

[1] SLIDE scores are scaled such that values > 0.0 are better while values < 0.0 are worse. DrugScore and RMSD changes are scaled such that values < 0.0 are better while values > 0.0 are worse.

[2] SLIDE score is unitless. For comparison, the mean score for known thrombin ligands in the crystallographic dockings is 36.3.

[3] DrugScore score is unitless. The mean score for known thrombin ligands in the crystallographic dockings is -468000.

[4] The mean known GST ligand SLIDE score for the crystallographic dockings is 40.2.

[5] The mean known GST ligand DrugScore score for the crystallographic dockings is -378000.

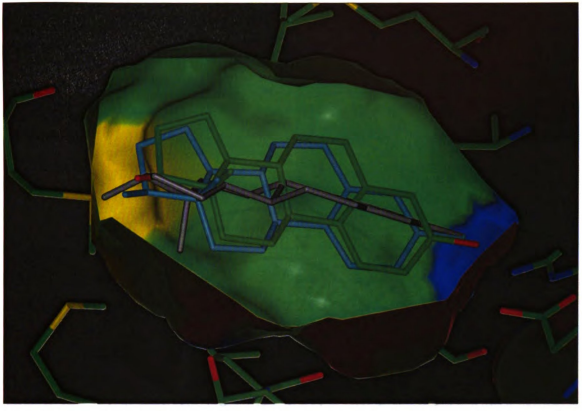


Figure 3.10: Dockings of estradiol to the estrogen receptor. Shown are the dockings of the estradiol ligand computed using the original template and interaction point methods, shown in grey, and computed using the new template and interaction point methods, shown in blue, compared to the position on the crystal structure, colored by atom (taken from PDB 1a52; Tanenbaum et al. 1998). The solvent-accessible surface of the estrogen receptor (ER) binding site is shown colored by atom. The new method docking is significantly improved compared to the original method docking, which is rotated roughly 90° about the long axis of the ligand and can no longer form a hydrogen bond with the protein at the left of the binding site. The new method docking is in a similar position to the known ligand and can make both hydrogen bonds (left and right of the binding site) as seen in the crystal structure complex.

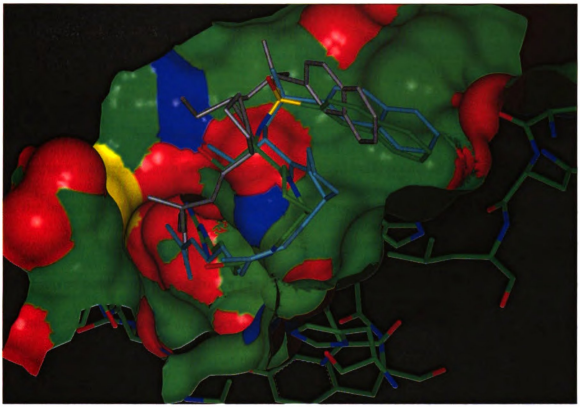


Figure 3.11: Dockings of ligand BMS-182282 to thrombin. Shown are the dockings of a thrombin ligand to the ligand-free thrombin structure using the new template and interaction point methods, shown in blue, and using the original template and interaction point methods, shown in grey, compared to the crystallographically docked ligand, shown colored by atom. The solvent-accessible surface of the thrombin active site is shown colored by atom. The docking computed using the new methods is similar to the crystallographic docking while the docking computed using the old methods is much less similar, especially in the middle region of the ligand where the new method docking and the crystallographic docking track quite closely, but the original method docking follows a significantly different path.

methods is much more similar to the known ligand, from the crystallographic complex for the estrogen receptor or docked into the ligand-free crystal structure via active site superposition for the thrombin, than the docking determined using the original template and interaction point methods.

3.3.3 Improved Enrichment

Now that it has been shown that the new methods to identify points of favorable interaction in the template and in the database molecules improves the docking of known ligands, it remains to show that the new methods result in improved screening results, i.e., better selection of potential ligands from a database of molecules. The way to measure this is to examine if known ligands are generally docked with higher scores than non-ligands, and, therefore, would reside at the top of the screening hit list. To explore this, the first 14691 molecules of the 87326 molecule Cambridge Structural Database (CSD) screening subset were selected. The vast majority of these molecules are unlikely to be true ligands, so comparing the rankings of any hits resulting from screening against these molecules to the rankings of known ligands can yield a measure of how likely one is to select true ligands as the top ranking screening hits. Shown in Figure 3.12 is a plot of enrichment for selection of known ligands over random molecules for thrombin. If no enrichment was seen, the curves would have a slope of 1, which is clearly not the case; therefore significant enrichment was seen for both the original and new template point and interaction point identification methods. The plots do not reach 100% on the vertical axis as they are scaled to set 100% to the full dataset of known thrombin ligands, 42 molecules, but only 35 (original method) and

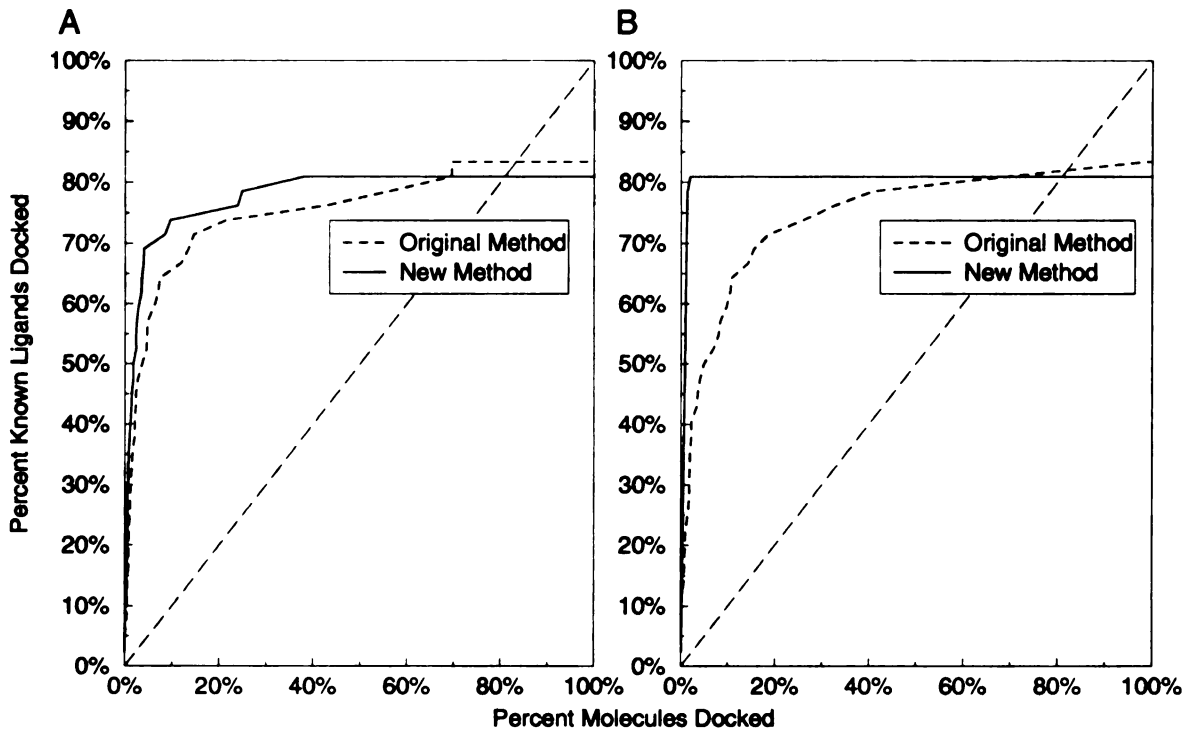


Figure 3.12: Enrichment of known thrombin ligands in a set of random CSD molecules using (A) the SLIDE scoring function and (B) the DrugScore scoring function. The plots are scaled such that if the known ligands were randomly mixed in with CSD molecules in terms of score, i.e., no enrichment was achieved, the plots would have a slope of 1 (demarcated by the thin dashed line). This is clearly not the case for either the original method or the new method for template point and interaction point identification, indicating significant enrichment was achieved. The new method curve is shifted to the left, indicating the new method yields an increase in the enrichment over the original method, especially when the dockings of scored using the DrugScore scoring function (B). The y-axes are scaled such that docking of all 42 thrombin ligands in the database would yield 100%.

34 (new method) ligands are docked. It can also be seen that the enrichment was greater for the new method compared to the original method as the enrichment curve is shifted to the left, i.e., a higher percentage of known thrombin ligands rank above the same number of random molecules for the new method, indicating that the new method is an improvement when measured using the enrichment. It can also be said that the scoring functions are able to distinguish reasonable dockings from questionable ones based on these plots. A similar analysis done for glutathione S-transferase showed a similar increase in enrichment. (Figure 3.13).

A quantitative measure of enrichment, or enrichment factor, can also be calculated. One such measure is derived from that developed by Knegtel and Wagener (1999):

$$F = \frac{N_{\text{act}(p)}/p}{N_{\text{act}}/N} \tag{3.9}$$

where $N_{\text{act}(p)}$ is the number of active compounds/known ligands in the top p ranked ligands and N_{act} is the number of active compounds/known ligands in the complete docking set, i.e., all molecules that were chosen as potential ligands, of size N. This factor is a measure of the proportion of known ligands in the top p selected molecules of the database relative to the proportion of known ligands in the set of all docked molecules. In general, p is selected as a fixed proportion of the database, e.g., the top 1% of the docked ligands. A second metric was developed for this work which calculates the proportion of total ligands docked with higher scores than the top k percent of known ligands:

$$E_k = \frac{k\%}{T\%} \tag{3.10}$$

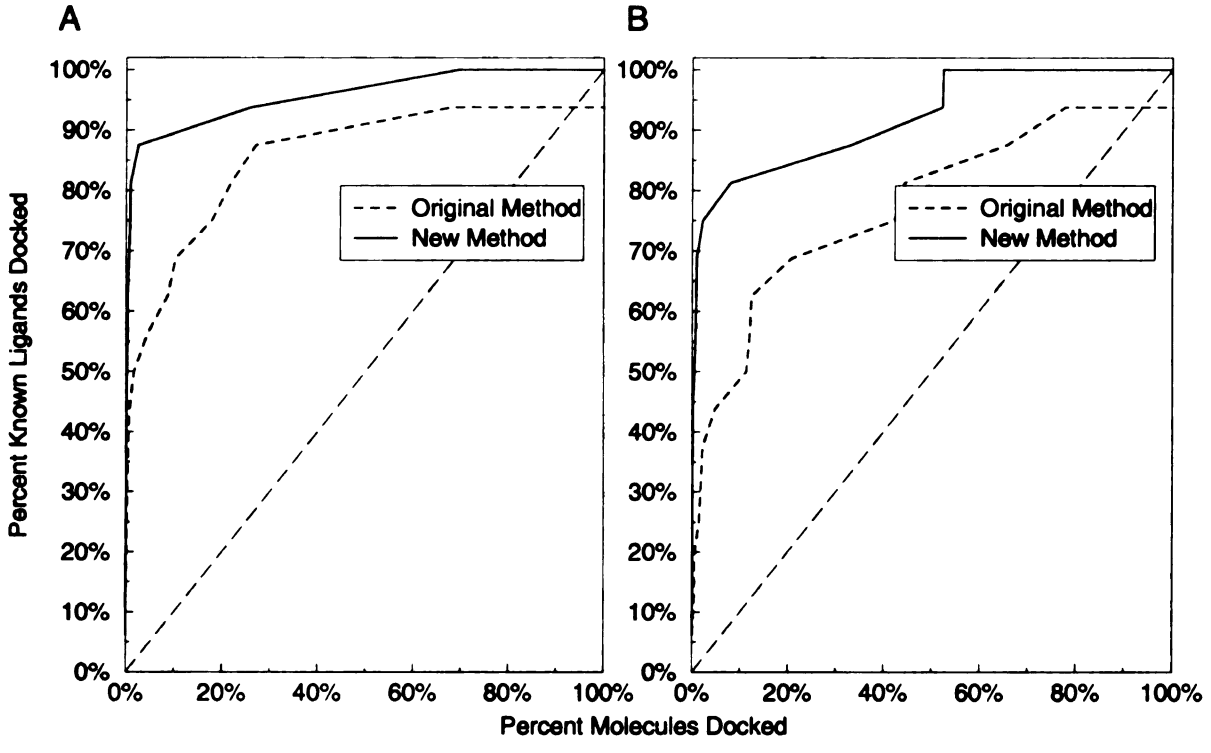


Figure 3.13: Enrichment of known glutathione S-transferase ligands in a set of random CSD molecules using (A) the SLIDE scoring function and (B) the DrugScore scoring function. The plots are scaled such that if the known ligands were randomly mixed in with CSD molecules in terms of score, i.e., no enrichment was achieved, the plots would have a slope of 1 (demarcated by the thin, dashed line). As can be seen, this is clearly not the case for either the original or the new methods for template point and interaction point identification, indicating significant enrichment was achieved. The new method curve (solid line) is shifted to the left, indicating the new method yields an increase in the enrichment over the original method (dotted line).

Table 3.9: Enrichment factors for thrombin and GST test screening runs

Protein	*Scoring Function*	*Original Method*	*New Method*	*Fold Increase*
Knegtel and Wagener Measure				
Thrombin	SLIDE score	28.6	44.1	1.5
	DrugScore	25.7	76.5	3.0
GST	SLIDE score	46.7	62.9	1.3
	DrugScore	20.0	71.9	3.6
New Measure				
Thrombin	SLIDE score	12.1	26.0	2.1
	DrugScore	7.7	74.4	9.6
GST	SLIDE score	6.5	93.6	14.3
	DrugScore	3.0	18.3	6.0

where E_k is the enrichment factor for docking of k percent of the known ligands, generally 70%, and $T\%$ is the percent of total ligand dockings which were ranked higher than the top k percent of docked known ligands. Enrichment factors calculated using this measure are shown in Table 3.9 Both measures of enrichment show quantitatively that the new method yields improved enrichment, sometimes dramatically. It is not informative to compare the enrichment factors calculated here to those determined by Knegtel and Wagener using their method as their database was significantly smaller, only 1000 molecules compared to the roughly 15,000 used here, resulting in their enrichment factors being significantly lower, generally five to ten.

3.3.4 Results Summary

The method used to model the interactions a ligand molecule can make with a protein, i.e., modeling of the ligand's interaction points, has been modified as described in Sec-

tion 3.2.2. This was shown to be an improvement using both visual analysis of example database molecules (Figure 3.8) and quantitative comparison between dockings achieved with SLIDE and dockings achieved via superposition of the binding sites of the ligand-free and complex crystallographic structures. There was generally an increase in the number of known ligands docked (Table 3.3) and the quality of those dockings (Table 3.4). Additional changes were made to the model of the interactions made by the protein with a ligand molecule, i.e., the protein template, as described in Section 3.2.3. These changes were shown to be improvements via both visual analysis of generated templates (Figures 3.6 and 3.7) and by quantitative measures comparing the dockings achieved by SLIDE to those achieved by superposition of the binding sites. As with the new interaction point method, the RMSD between the SLIDE dockings and the crystallographic dockings generally improved using the new method compared to the original method (Tables 3.5 and 3.6). Combination of the new methods also showed an improvement, generally greater than either new method individually, as seen by comparing known ligand dockings (Tables 3.7 and 3.8) and by visual analysis of the dockings (Figures 3.10 and 3.11). Combination of the new methods also improved enrichment of the screening results (Figures 3.12 and 3.13; Table 3.9).

3.4 Discussion

SLIDE is an efficient tool for virtual ligand screening which includes ligand and protein side chain flexibility. Databases on the order of 100,000 molecules can be screened in a few hours to a day, depending on template size and screening parameters. Full ligand docking

is computationally too expensive to be performed on such a large database of molecules, so initial matchings are calculated based on point representations of the protein binding site and of the database molecules' potential interaction centers. This approach enables the elimination of infeasible matchings very quickly, thereby reducing the need to perform the expensive computational operations on these infeasible matchings. The work presented here describes changes made to the model used to identify both the sites of interaction in the protein binding site and on the database molecules.

The method used to identify sites of potential interaction in the protein binding site identifies sites where ligands could make favorable interactions to the protein and mirror the binding site in both shape and chemistry. Each site is represented by a point with an associated chemistry type, hydrogen bond acceptor, hydrogen bond donor, hydrogen bond doneptor, or hydrophobic, and reflects the favorable ligand atom type to place at that site. Altering the definition of a hydrophobic interaction site from one that is in an environment which is hydrophobic on average to one which is in an environment which contains several hydrophobic atoms proved to produce an improved model of the protein template site as evidenced by the improved docking of known thrombin and glutathione S-transferase ligands. While it may seem more logical to base an assignment of hydrophobic character on the average environment of the point in question, there is no sense of the size of the hydrophobic area in question since having an environment consisting of only a single carbon atom is treated as equally hydrophobic as an environment consisting of many carbon atoms. However, it is clear these environments are not equal in terms of their hydrophobic character. Instead, it may be more important to have several hydrophobic atoms in the environment,

allowing the presence some hydrophilic atoms while necessitating an overabundance of hydrophobic atoms. This is partially reflected in the fact that bound water molecules are more often located in depressions in the protein surface than adjacent to protrusions, even when neglecting the hydrophobic character of the water molecules' environments (Kuhn et al., 1992b). One could reasonably expect this observation to be stronger for hydrophobic ligand atoms as there will be a stronger push for isolation from the surrounding solvent. Use of the concept of the number of hydrophobic atoms in the environment could easily be included in docking and screening scoring functions to further differentiate the more favorable potential ligands and/or docking orientations selected. Often, scoring potential ligands can be a key factor in extracting the best of the potential ligands selected, so an improvement such as this in the scoring function could prove to be an important enhancement of the overall screening process.

A change in the model used for sites of potential hydrophobic interaction in the database molecules was also implemented. This change yields a much better balanced representation of the molecule's hydrophobic character than was seen previously. One key feature of the new model is the assignment of hydrophobic points to the edges of ring structures in the molecules. Previously, a single point was placed at the center of each hydrophobic ring. This leads to difficulties with determining the proper orientation in which to dock the ring, i.e., stacked against the protein surface or edge-on relative to the protein surface, and with limited placement of the ring. Since only a single point was assigned to the ring, there was a severe limitation on the way that ring could match a site of favorable hydrophobic interaction in the protein binding site, but it has been shown that while there

are some preferences observed for ring-ring packing interactions, there is also a fair amount of variability (Burley and Petsko, 1985; Mitchell et al., 1997). Given that a 6-membered aromatic ring is approximately 3 Å wide and hydrophobic template points are generally placed at an approximate nearest-neighbor distance of 1.5 Å, there would be only one to two matching sites for a ligand ring structure. By placing points around the edge of the ring, more possibilities are allowed for ring placement as there are more ring points to match to template points. The initial docking of the ring can also shift substantially more by mapping points from different sides of the ring onto the same template point, further increasing the docking space explored. The exploration of this additional docking space is important when searching for potential ligands so as to decrease the probability that a true ligand is missed due to a misalignment of a key feature.

In addition, to change the model for hydrophobic ligand rings, a new model of hydrophobic ligand atoms outside of rings was introduced. This model averages the hydrophobic character of the molecule throughout carbon chains compared to the previous model, which assigned a hydrophobic interaction point to every hydrophobic atom bonded only to other hydrophobic atoms. While one can argue that the reduction in the number of points in this case reduces the docking space sampled, which would be unfavorable, reducing the number of points without significantly reducing the sampled space can be achieved, as happens here. The new model of points, while reduced, still allows for a significant amount of sampling as the points are placed approximately every 1.5 carbon atoms along a chain. This is an approximate spacing to match the approximate 1.5 Å template point spacing, and adequate sampling of the space is still achieved.

The idea of docking space sampling is always a concern when performing computational screening. As the sampling becomes finer, such as with a finer template, the screening time can increase dramatically. In fact, docking tools, such as DOCK (Shoichet et al., 1992; Shoichet and Kuntz, 1993), AutoDock (Morris et al., 1996), and GOLD (Jones et al., 1995), which seek to find the optimal binding orientation of a single ligand, can be thought of as finely sampling the docking space of a particular protein-ligand complex. The key to effective screening methods will be to reduce the time needed to explore an adequate amount of this docking space to provide accurate enough docking orientations to be able to effectively analyze the resulting selected ligands. The ultimate goal of all virtual screening techniques is to provide a set of potential ligands which show binding to the protein of interest and can be used for further ligand optimization or as a set of probes for functional studies, e.g., via differential inhibition. The work presented here comes closer to this goal.

Chapter 4

Computational Screening of Asparaginyl tRNA Synthetase

4.1 Introduction

In addition to development of improvements to the computational screening algorithm SLIDE, work was undertaken to identify potential novel inhibitors of asparaginyl-tRNA synthetase (AsnRS) using SLIDE. Aminoacyl-tRNA synthetases are responsible for catalyzing the addition of an amino acid onto the 3′ ribose of its cognate tRNA via a two-step ATP dependent reaction. Initially, the amino acid is activated by the addition of ATP to form an enzyme-bound aminoacyl adenylate intermediate. The aminoacylation reaction and its specificity are vital to protein synthesis. The 20 aminoacyl-tRNA synthetases can be divided into two general classes, I and II, based on sequence motifs and putative structural domains, as reviewed by Cusack (1995) and Arnez and Moras (1997). In general, class I synthetases contain a Rossmann Fold in their active sites and are active as monomers,

whereas the class II synthetases contain an anti-parallel β-fold and are active as α_2 homodimers or $\alpha_2\beta_2$ heterotetramers. Class II aminoacyl-tRNA synthetases (AARS) are defined by the presence of three class specific motifs of 10 (motif 1), 18-26 (motif 2), and 16 (motif 3) residues in their catalytic domains. It is possible to further divide class II into subclasses IIa, IIb, and IIc based on the presence of specific domains, which generally play a role in anticodon recognition. In addition to these subclass specific domains, a variable insertion of 60 to 280 residues occurs between motifs 2 and 3 in the catalytic domains of class II AARSs. AsnRS is a member of subclass IIb, which also includes aspartyl- and lysyl-tRNA synthetases.

Lymphatic filariasis caused by *Brugia malayi* infection affects an estimated 100 million people worldwide, and more than 1 billion people live in areas where the disease is actively transmitted (Awadzi, 1997). There are currently no effective preventive medicines against filariasis, as none are effective against the larvae, which are transmitted to humans by mosquitos. *Brugia* AsnRS is an excellent target for filarial drug development as it is highly expressed in the worms, has been well characterized biochemically and structurally in several species, and can be recombinantly expressed to facilitate *in vitro* studies. Also, the sequence and structure of the *Brugia* enzyme is different from the human AsnRS, providing for the possibility of identification of inhibitors specific for the *Brugia* synthetase.

Pieces of the protein translation apparatus have long been targets for antibacterial agents, reviewed by Schimmel et al. (1998), including streptomycin and tetracycline, which target the 30 S ribosomal subunit, and erythromycin and chloramphenicol, which target the 50 S ribosomal subunit. Aminoacyl-tRNA synthetases are currently a promising target for

anti-infective agents as an increased number of pathogens become resistant to traditional antibiotics. Targeting new anti-infectives to aminoacyl-tRNA synthetases is promising as the enzymes are specific for the transacylation of tRNAs and are somewhat species specific, reducing the possibility of cross-reactivity with host enzymes and the resulting side-effects. Pseudomonic acid (mupirocin) is a natural product synthesized by *Pseudomonic fluorescens* (Fuller et al., 1971) that has been shown to be an inhibitor of isoleucyl-tRNA (IleRS) synthetase from Gram-positive infectious bacteria, including antibacterial-resistant *S. aureus* (Casewell and Hill, 1985). It has been shown to have an approximately 8000-fold selectivity for pathogen IleRS over mammalian IleRS (Hughes and Mellows, 1980). Several other natural products have been shown to inhibit aminoacyl-tRNA synthetases (Nass et al., 1969; Paetz and Nass, 1973; Tanaka et al., 1969; Ogilvie et al., 1975; Werner et al., 1976; Konrad and Roschenthaler, 1977; Konishi et al., 1989), indicating the potential for use of natural and synthetic products to be effective against aminoacyl-tRNA synthetases.

4.2 Methods

4.2.1 Available Asparaginyl-tRNA Synthetase Structures and Ligands for Virtual Screening Studies

A 1.9 Å structure of *Brugia malayi* asparaginyl-tRNA synthetase complexed with an S-adenosyl-asparagine (S-AMP-Asn) substrate analog was provided by Dr. Stephen Cusack (EMBL, Grenoble, France), shown in Figure 4.1. The active-site pocket of AsnRS, shown in Figure 4.2, is relatively deep and consists of two lobes, one of which binds the adenosyl

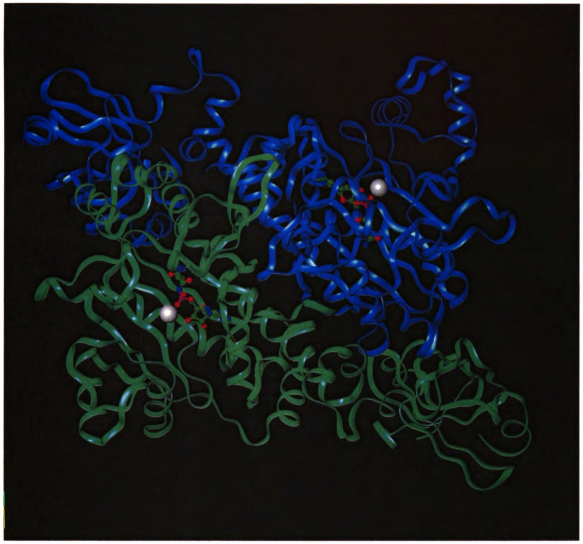


Figure 4.1: Structure of *Brugia malayi* asparaginyl-tRNA synthetase complexed with S-adenosyl-asparagine. Shown is the backbone structure of AsnRS, with one chain of the dimer colored green and the other chain colored blue. The S-adenosyl-asparagine ligand is rendered as ball-and-stick and colored by atom type and the associated Mg ion is rendered as a large white sphere. While there is a substantial amount of contact between the chains of the dimer in the interface, the active site is isolated within each chain and, therefore, only a single chain need be considered during the computational screening procedure.

moiety (right side of Figure 4.2) with the asparaginyl group extending into the other lobe (left side of Figure 4.2). This crystal structure with the ligands removed was used to construct two templates for SLIDE screening. The ligand positions from this structure, along with the positions of the asparagine ligand from *E. coli* AsnRS (PDB code 11as; Nakatsu et al. 1998) and the AMP and asparagine ligands from a second *E. coli* AsnRS crystal structure (PDB code 12as; Nakatsu et al. 1998) were used to generate a ligand-based template. The ligands from the *E. coli* structures were transformed into the same reference frame as the *Brugia* structure via superposition of their active site atoms onto the *Brugia* active site. A second, unbiased SLIDE template (i.e., a template based only on the chemistry of the binding site, incorporating no information about the structures or positions of known ligands bound in the site) was generated using the *Brugia* structure. This template was modified to eliminate points outside of the pocket and to label points in the deep lobes as key points. By labeling points as key points, SLIDE requires any ligand docking to match at least one of these points, ensuring that the ligands are well situated in these deep portions of the pocket. These ligand-based and unbiased templates were used to screen a set of three databases:

1. a database of six known ligands consisting of three asparagine molecules (two taken from PDB files 11as and 12as (*E. coli* AsnRS structures) and a third generated from the asparaginyl portion of the *Brugia* S-adenosyl-asparagine ligand), two AMP molecules (one from 12as and one generated from the AMP portion of the *Brugia* complex ligand), and the S-adenosyl-asparagine ligand from the *Brugia* structure;

2. a database of the 481 conformers generated for a set 16 high-throughput screening

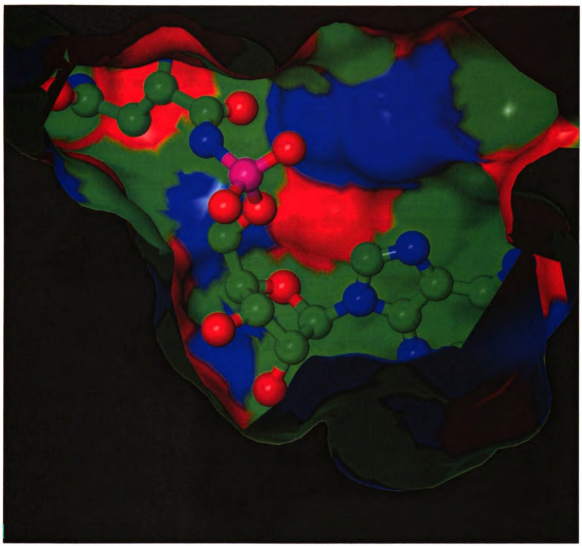


Figure 4.2: Asparaginyl-tRNA synthetase active-site pocket. The pocket's solvent-accessible surface is colored by the type of atom which contributes to it: carbon surface, green; oxygen, red; and nitrogen, blue. The S-adenosyl-asparaginyl ligand is also colored by atom type, with sulfur in magenta. The two lobes of the pocket which hold the adenosyl and asparaginyl moieties are oriented in the upper-left and lower right, respectively. Though difficult to visualize in this two-dimensional figure, these lobes extend deeper than the central portion of the pocket, which forms a ridge between the two deeper parts of the binding site.

ligands (described below); and

3. an 85,000 molecule subset of the CSD of crystal structures of small organic molecules, consisting of those molecules with three-dimensional structures, nonpolymeric nature, and the absence heavy metal atoms, e.g., Fe, Mg, Zn, U.

In order for ligand candidates identified by screening to be guaranteed to have the ability to span the entire binding site, encompassing both lobes of the AsnRS pocket (Figure 4.2), SLIDE was set to require that at least two of the three ligand interaction points matching the template be a minimum of 9.0 Å apart.

4.2.2 High-throughput Screening for Asparaginyl-tRNA Inhibitors and Conformer Generation of Selected Ligands

High throughput screening (HTS) for asparaginyl-tRNA synthetase inhibitors was performed by Discovery Technologies (Allschwil, Switzerland) for our collaborator, Dr. Michael Kron (Department of Medicine, College of Human Medicine, Michigan State University). Screening was performed using a library of 11,700 compounds selected from Bionet, MayBridge, SPECS, Aldrich, Analyticon, and several university compound libraries. Activity assays were performed using unfractionated yeast RNA as a tRNA substrate along with ^{14}C-asparagine under the protocol developed by Dr. Michael Hartlein (EMBL, Grenoble, France). For each potential inhibitor, ("HTS hit") the IC_{50} and relative specificity for the *Brugia* versus human enzyme were determined. Since the HTS ligands were provided as two-dimensional structures, three-dimensional structures must be gener-

ated to perform structural analysis of their mode of interaction with AsnRS. A search of the Cambridge Structural Database (CSD) based on the structure of the HTS identified inhibitors yielded no entries, indicating that the crystal structures are not available for any of these compounds. Initial three-dimensional structures were generated from SMILES codes (a one-dimensional representation of the molecule's chemical structure) and energy minimized using the Molecular Operating Environment (MOE; Chemical Computing Group, Inc, Montreal, Quebec). Each of these minimized structures was used as input for the stochastic conformer generation function in MOE to generate a set of three-dimensional conformers for each of the HTS ligands. This conformer generation was run using a fixed set of parameters that yielded a fine sampling of conformational space.

Images in this section of this dissertation are presented in color.

4.3 Results

4.3.1 High-Throughput Screening

High-throughput screening for asparaginyl-tRNA synthetase inhibitors by Discovery Technologies resulted in a set of 16 potential inhibitors, shown in Figure 4.3. The compounds exhibited a range of structures, though they are generally small, with molecular weights between 178 and 542 Da, and contain some aromatic structure. Inhibitory activity, measured as the concentration of ligand which reduces enzyme activity to 50% of native activity (IC_{50}), ranged from 6.7 to 171 μM. Specificities for the *Brugia* enzyme relative to the human enzyme, measured as the ratio of IC_{50} for human over the *Brugia* IC_50, ranging from

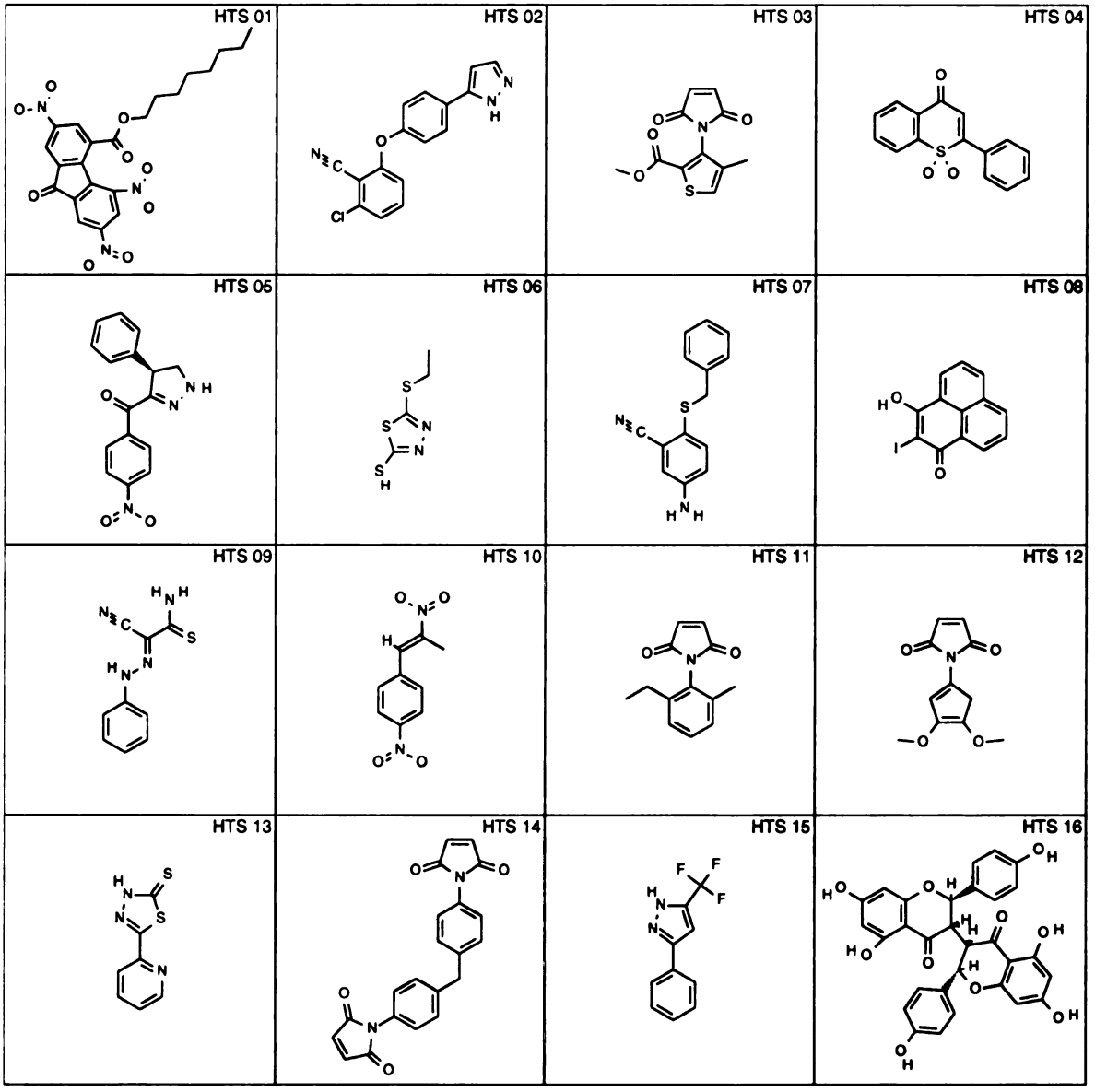


Figure 4.3: Potential asparaginyl-tRNA synthetase inhibitors identified by high-throughput screening. These HTS ligands contain at least some aromatic character while also containing hydrogen bond forming groups. Of particular interest are ligands numbered 2, 7, 13, and 15 which are suggested by Basilea scientists working on AARS inhibitor development as promising compounds (Malcolm Page and Frank Daniels, private communication). Compounds 2, 7, and 15 have been shown to have some toxicity towards either adult worms or larvae.

1.7 to 63. Compounds 2 and 15 were shown to be toxic to adult *Brugia* worms, while compound 7 has been shown to have only moderate toxicity against adult worms and strong toxicity to worm larvae via *in vivo* assays (Michael Kron, unpublished results).

Conformer generation for these 16 HTS ligands resulted in an average of 30 conformers (standard deviation 35.6) per ligand. This large standard deviation reflects the fact that the flexibility of the HTS ligands varies considerably. Ligand 8 is completely rigid with respect to its non-hydrogen atoms and yields only a single conformer, while ligand 7 is quite flexible and yields 100 conformers under the same conformer generation parameters. Generated conformers for two example HTS ligands are presented in Figure 4.4.

4.3.2 Computational Screening using a Ligand Based Template

To identify ligand candidates which are related to the known ligands, a ligand-based screening template was generated from the six available AsnRS ligands: three asparagine molecules, two AMP molecules, and one S-adenosyl-asparagine (S-AMP-Asn) molecule. This template consisted of 13 points and was able to successfully dock the two AMP molecules and the S-AMP-Asn ligand, but was not able to dock any of the three Asn ligands from the crystallographic structures. This is quite likely due to the increased variability of placement of the asparagine moieties, such that when the template is generated by averaging the atom positions, the template points are placed outside of the favorable binding positions. Screening with this template against the database of conformers generated from the 16 HTS ligands yielded no dockings for any of the HTS ligands. One possible explanation for this result is that none of the HTS ligands bind in the active site, which was

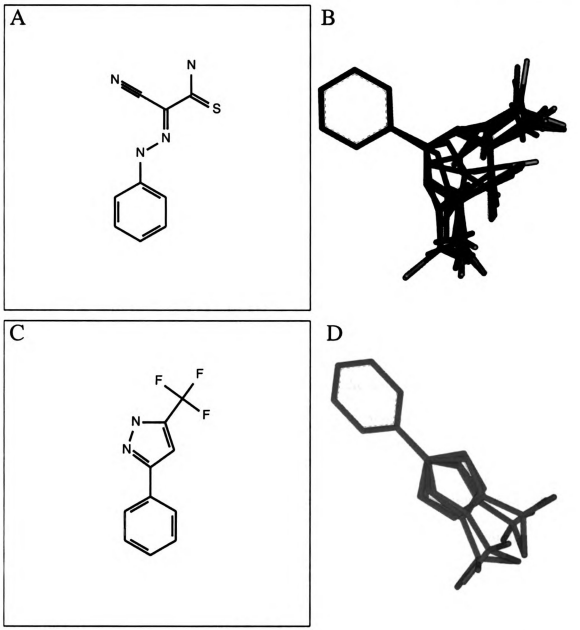


Figure 4.4: Conformers generated for HTS ligand 9 (A and B) and HTS ligand 15 (C and D). Shown in A and C are the 2-dimensional structures of the the HTS ligands; shown in B and D are the 3-dimensional structures of the generated conformers, each of which is superimposed onto the first conformer using the phenyl ring in the upper left as a reference. Ligand 15 is much less flexible and yields a small number of conformations (6) that are somewhat similar while ligand 9 yields a broader range of conformations (20) due to its increased flexibility.

screened against, which could be tested with additional activity assays to determine the nature of the inhibition. Another explanation based on the screening algorithm is that the HTS ligands do not strongly resemble the ligands from the crystallographic structure. One would not expect a template generated from these crystallographic ligands to be able to select for these different HTS ligands as they do not resemble the known ligands and are likely to take advantage of different sets of interaction sites within the binding site.

In addition to the screening against the databases of known ligands, either from the crystallographic structures or from the high-throughput screening tests, the ligand-based template was used to screen against the CSD for new ligands. A adenosyl compound, adenosine-5′-methylphosphonate (CSD code ADMPOT10; Barnes and Hawkinson 1979), a substrate analog, was the top scoring ligand using either the internal SLIDE scoring function or the DrugScore scoring function (Gohlke et al. 2000a; discussed briefly in Section 3.2.4). The docking of this CSD ligand closely resembles the docking orientation of the AMP portion of the S-AMP-Asn ligand from the *Brugia* AsnRS crystallographic structure. This confirms that the ligand-based template is able to select ligands that mimic the natural ligands' binding modes. Other adenosine analogs were also docked with high scores during the screening. Another ligand that was highly ranked by both the internal SLIDE scoring function and the DrugScore scoring function was variolin B (CSD code LEPWIM; Perry et al. 1994). Variolin B is a natural product derived from an Antarctic sponge and has been shown to have possible antiviral and antitumor activity. Other marine sponge products have been shown to have antihelminthic and antibiotic effects (Alvi et al., 1991). The two-dimensional structure of variolin B is presented in Figure 4.5 and the docking orientation

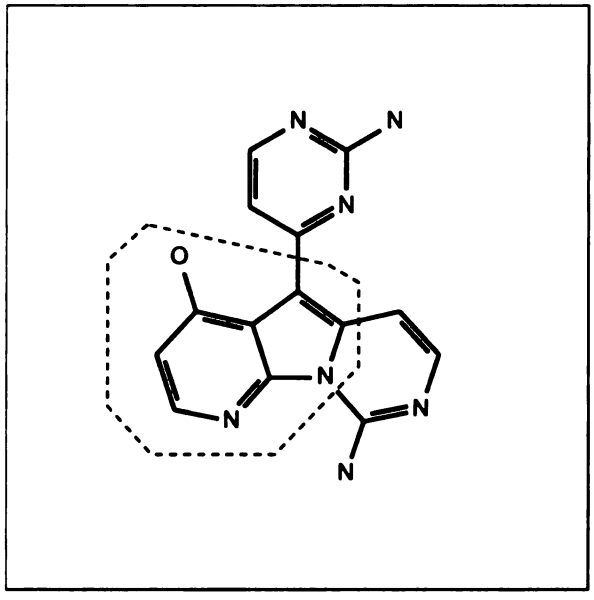


Figure 4.5: Two-dimensional structure of variolin B, CSD code LEPWIM. Circled is the ring structure mapped onto the purine ring of the AMP portion of the S-AMP-Asn ligand in the crystal structure of *Brugia* AsnRS to achieve a second docking of the variolin ligand.

as calculated by SLIDE is presented in Figure 4.6. An additional, manual docking was generated by mapping the variolin B ring circled in Figure 4.5 onto the purine ring of the adenosyl group, with subsequent rotation of rotatable side-chain and ligand bonds to alleviate intermolecular and intramolecular collisions in *InsightII* (Accelrys, San Diego, CA). This docking assessed whether is was possible for variolin B to bind such that it matched the binding of the adenosyl moiety in S-AMP-Asn. This docking, found to be feasible, also places a hydroxyl group of variolin B onto the adenosyl N_6 group, the hydroxyl oxygen can act as either a hydrogen bond acceptor, the "normal" hydrogen bond forming role for oxygen atoms, or a donor, thereby mimicking the adenosyl nitrogen's function. This docking, shown in Figure 4.7, scored higher than the docking in Figure 4.6 when using the internal SLIDE scoring function, but somewhat lower when using the DrugScore scoring function.

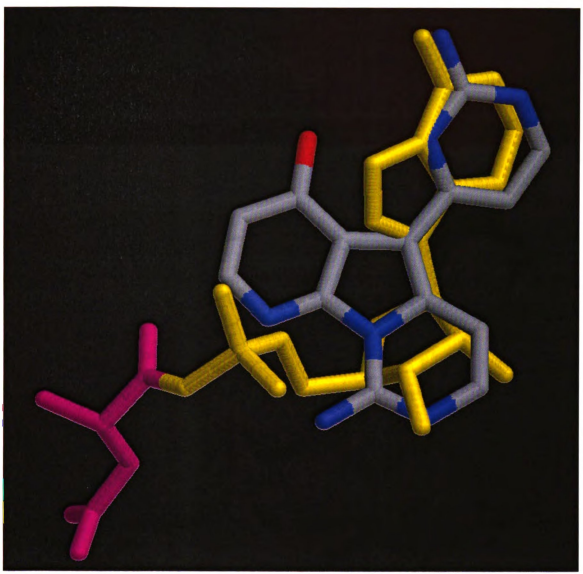


Figure 4.6: Variolin B docking as determined by SLIDE. The S-AMP-Asn ligand from the *Brugia* AsnRS crystallographic structure is shown in yellow (AMP portion) and magenta (Asn portion). Variolin B is colored by atom type: carbon, grey; oxygen, red; and nitrogen, blue. There is a clear match between the 6-membered ring of the purine group and the isolated ring of the variolin (upper right). The amine nitrogen of this ring coincides with N_6 of the adenosine, making the same hydrogen bond to the AsnRS protein.

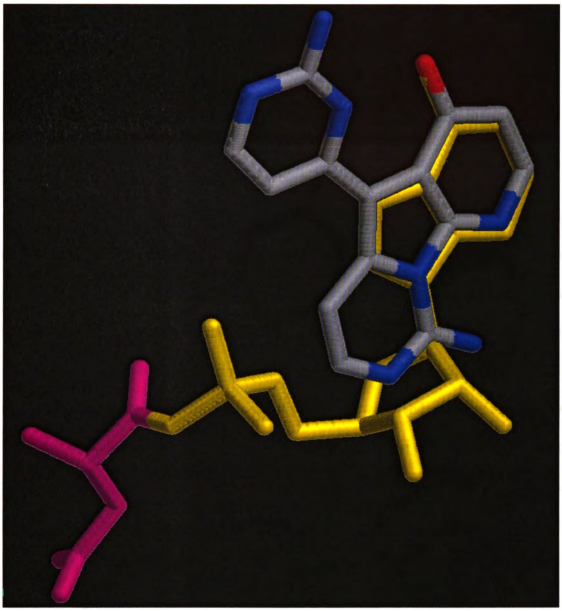


Figure 4.7: Variolin B docking assessed manually by superimposing the ring structure highlighted in Figure 4.5 onto the purine ring of the AMP moiety of the S-AMP-Asn ligand. The S-AMP-Asn ligand from the crystallographic structure is shown in yellow (AMP portion) and magenta (Asn portion). Variolin B is colored by atom type: carbon, grey; oxygen, red; and nitrogen, blue. The S-AMP-Asn ligand is in the same orientation as in Figure 4.6. This docking orientation yields a more filled AMP lobe of the binding pocket compared to the SLIDE orientation and scored somewhat higher with the internal SLIDE scoring function. Thus, the binding modes predicted in this figure and Figure 4.6 are both feasible and have favorable complementarity with AsnRS.

Given these two favorable binding modes and the known biological activities of variolin B, it was suggested as a potential ligand for further testing.

4.3.3 Computational Screening using an Unbiased Template

To identify ligand candidates that do not necessarily resemble the known ligands, an unbiased template of 140 points was generated from the *Brugia* AsnRS crystallographic structure. This template had 50 points which reside at the extreme ends of the S-AMP-Asn ligand marked as key points, meaning that all matches in SLIDE must include at least one of these points. Screening with this template against the database of six ligands derived from the crystallographic structures resulted in dockings of all six of these ligands, indicating the unbiased template can also correctly dock known ligands into the binding site. SLIDE was also able to dock all 16 of the HTS ligands in at least one, and generally more than one, orientation. The best scoring dockings, according to the DrugScore scoring function, for two of the HTS compounds are shown in Figures 4.8 and 4.9. This results in prediction of the mode of binding of the HTS inhibitors to AsnRS, which is not known experimentally.

In addition to predicting docking orientations of potential AsnRS inhibitors identified by high-throughput screening, SLIDE was used to identify additional novel potential inhibitors, without bias towards known ligands. This was done by screening with the unbiased template against the CSD as described above. Three of the CSD compounds were docked and ranked in the top 10 ligands using both the internal SLIDE scoring function and the DrugScore scoring function (Figure 4.10), indicating these potential ligands form very

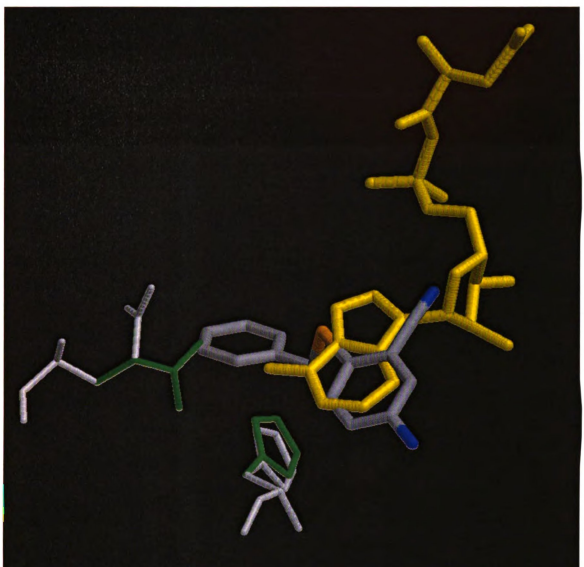


Figure 4.8: Best scoring SLIDE docking for high-throughput screening ligand number seven. The S-AMP-Asn ligand from the *Brugia* crystallographic structure is shown in yellow (AMP portion) and magenta (Asn portion) and the ligand is shown colored by atom (C, grey; O, red; N, blue). Side chains in the *Brugia* AsnRS that were rotated by SLIDE during the screening process are shown in green, with the crystallographic conformations shown in grey.

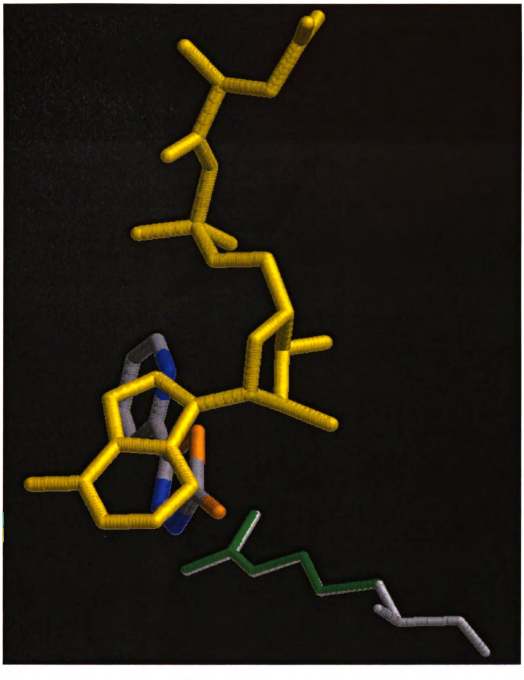


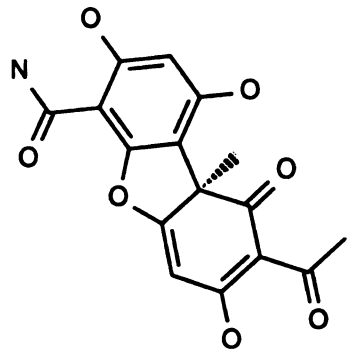
Figure 4.9: Best scoring SLIDE docking for high-throughput screening ligand number 13. The S-AMP-Asn ligand from the *Brugia* crystallographic structure is shown in yellow (AMP portion) and magenta (Asn portion) and the ligand is shown colored by atom (C, grey; O, red; N, blue). Side chains in the *Brugia* AsnRS that were rotated by SLIDE during the screening process are shown in green, with the crystallographic conformations shown in grey.

favorable interactions with AsnRS. One of these candidates, MSFURY, has a second crystal form entered in the CSD as entry code MSFURY01. This second form was also selected in the screening with rankings of 15 and 2 (SLIDE score and DrugScore respectively).

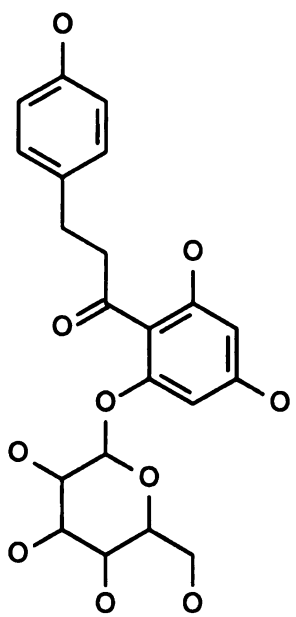
Cercosporamide, CSD code SIVXIE (Sugawara et al., 1991), was the top scoring candidate based on the SLIDE scoring function and was the 10th ranked ligand based on the DrugScore scoring function; it is shown in its docked orientation in Figure 4.11. This potential ligand fills the AMP lobe of the AsnRS binding pocket quite well, which is likely to be important for both binding affinity and specificity. Cercosporamide, a phytotoxin produced by a cassava pathogen, has been shown to have biological activity as a toxin to plant protoplasts as well as to a variety of fungi (Sugawara et al., 1991). Given the high scores and the known activities, this would be a good candidate to test for activity against both *Brugia* AsnRS and against the *Brugia* nematodes. It will also be important to test this and other potential ligands for inhibition of human AsnRS, which would be undesirable in therapeutical applications. The ultimate goal is to obtain compounds that specifically inhibit *Brugia* AsnRS, but not human AsnRS.

A second potential ligand of interest selected during computational screening was phlorizin, CSD code CEWWAC20 (Auf'mkolk et al., 1986). Phlorizin, a dihydrochalcone glycoside produced by apple trees, has been shown to inhibit Na^+ and glucose transport with high nanomolar concentration via direct interaction with the Na^+/glucose cotransporter (Hirayama et al., 2001). It has also been suggested that it may interact with the NADPH binding site in some mammalian catalases (Kitlar et al., 1994) and has been shown to be toxic to malarial pathogens (Loyevsky and Cabantchik, 1994). The docking orientation

SIVXIE



CEWWAC20



MSFURY

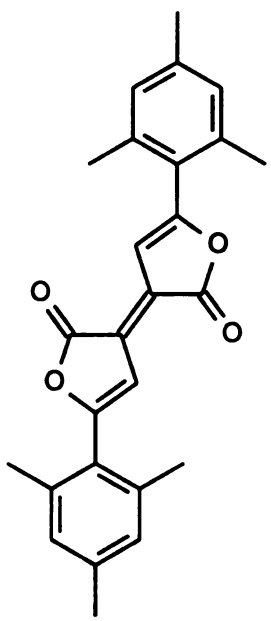


Figure 4.10: Molecules from the CSD selected as potential ligands by SLIDE screening with an unbiased template. These molecules (SIVXIE, cercosporamide; CEWWAC20, phlorizin; and MSFURY, (E)-4,4′-dimesityl-but-3-enolidylidene-but-3′-enolide) were ranked in the top 10 potential ligands using both the SLIDE scoring function and the DrugScore scoring function, indicating that they are highly favorable ligands to pursue further through *in vitro* and *in vivo* experimental testing.

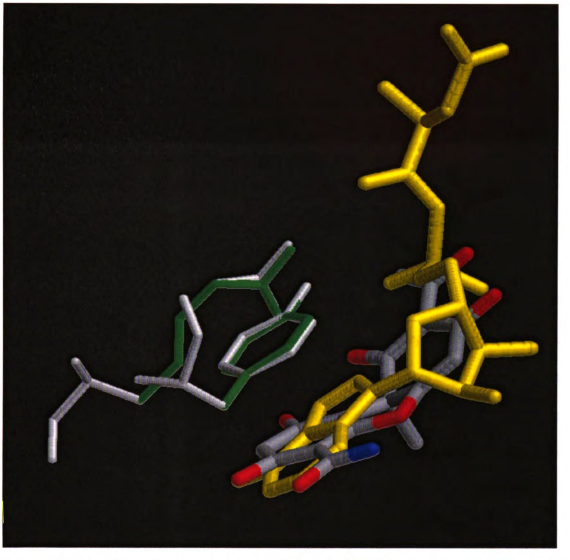


Figure 4.11: Docking of cercosporamide, CSD code SIVXIE, with *Brugia* asparaginyl-tRNA synthetase. The S-AMP-Asn ligand from the *Brugia* crystallographic structure is shown in yellow (AMP portion) and magenta (Asn portion) for comparison, and the ligand is shown colored by atom (C, grey; O, red; N, blue). Rotated side chains are shown in green with the native positions shown in grey. The docking orientation shown was achieved by screening with the AsnRS unbiased template against the CSD and was ranked as the best potential ligand using the internal SLIDE scoring function and the 10th best potential ligand using the DrugScore scoring function. The AMP pocket of AsnRS is well filled by the potential ligand, lower center of the figure.

determined by SLIDE is shown in Figure 4.12. These results suggest the compound is bioavailable, but also, particularly in the case of binding to mammalian catalases, suggest the possibility of toxicity or other side effects due to non-specific interactions of this potential inhibitor. Phlorizin was ranked as the 10th best potential ligand using the internal SLIDE scoring function and the ninth best potential ligand based on the DrugScore scoring function, indicating that this docking to *B. malayi* AsnRS is still favorable, but not as favorable as the interactions of the cercosporamide previously analyzed. The docking determined by SLIDE also fills the AMP lobe of the AsnRS binding pocket, but to a lesser extent. Also, more of the ligand extends away from the deepest portion of the binding site (the portion extending down on the right side of Figure 4.12). There are several side chain rotations in this docking, though most are still quite small, except for the movement of tyrosine 223, behind the ligands on the left side of the figure. This degree of movement is still not large and is comparable to what is seen between ligand-free and ligand-bound crystal complexes (Maria Zavodszky, unpublished results). In addition to the known activities, phlorizin is commercially available (Sigma-Aldrich) and would be a good candidate for *in vitro* testing against AsnRS and *in vivo* testing against *Brugia* larvae and nematodes.

The third molecule that was ranked as the 10th or better ligand with both the internal SLIDE scoring function and the DrugScore scoring function (ranked eighth with both scoring functions) was (E)-4,4′-dimesityl-but-3-enolidylidene-but-3′-enolide, CSD code MSFURY (Begley et al., 1981). Figure 4.13 shows the docking orientation determined by SLIDE for this potential ligand. Unfortunately, little biological work has been done with this molecule, so no information about potential biological activity exists. It is interesting to

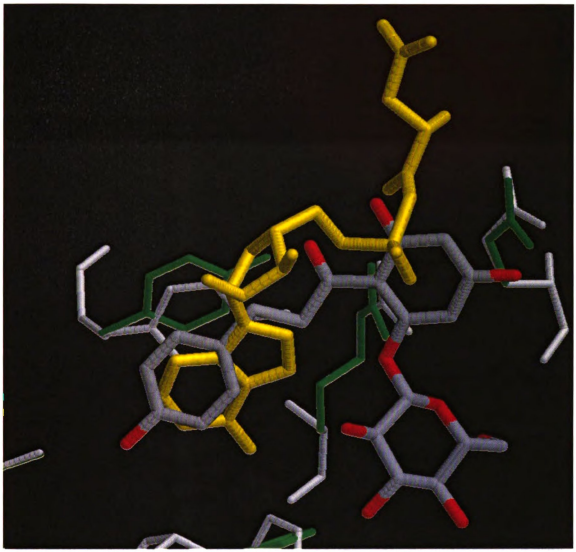


Figure 4.12: Docking of phlorizin, CSD code CEWWAC20, with *Brugia* asparaginyl-tRNA synthetase. The S-AMP-Asn ligand from the *Brugia* crystallographic structure is shown in yellow (AMP portion) and magenta (Asn portion) for comparison, and the ligand is shown colored by atom (C, grey; O, red; N, blue). Rotated side chains are shown in green, with the native positions shown in grey. The docking orientation shown was achieved by screening with the AsnRS unbiased template against the CSD and was ranked as the 10th best potential ligand using the internal SLIDE scoring function and the ninth best potential ligand using the DrugScore scoring function. Phlorizin fills a fair amount of the AMP binding lobe, lower left. This potential ligand extends further away from the Asn binding lobe compared to cercosporamide.

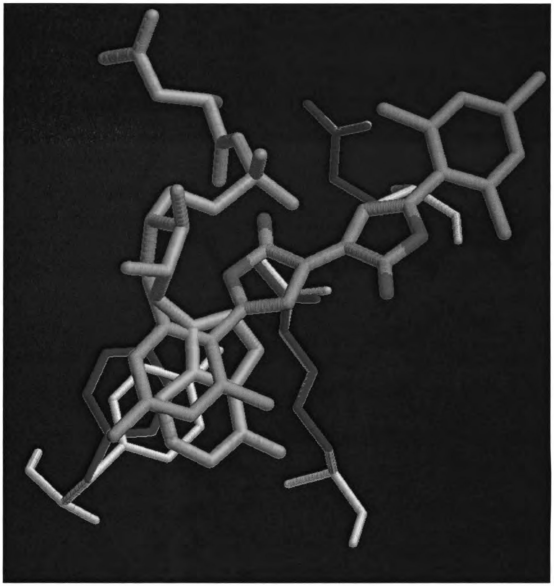


Figure 4.13: Docking of (E)-4,4′-dimesityl-but-3-enolidylidene-but-3′-enolide, CSD code MSFURY, with *Brugia* asparaginyl-tRNA synthetase. The S-AMP-Asn ligand from the *Brugia* crystallographic structure is shown in yellow (AMP portion) and magenta (Asn portion) for comparison, and the ligand is shown colored by atom (C, grey; O, red; N, blue). Rotated side chains are shown in green with the native positions shown in grey. The docking orientation shown was achieved by screening with the AsnRS unbiased template against the CSD database and was ranked as the 8th best potential ligand using both the internal SLIDE scoring function and the DrugScore scoring function. MSFURY also binds mostly in the AMP lobe of the binding pocket, but extends away from the binding site of AsnRS (towards the upper right of the figure).

note that a second crystal form of MSFURY exists in the CSD, code MSFURY01, in which the distal substituated phenyl rings are rotated approximately 90° from planar with respect to the central rings. (In the MSFURY entry, all four of the rings are roughly coplanar.) This conformation was also selected by SLIDE as a potential ligand, with SLIDE making conformational changes to the ligand to rotate the distal phenyl rings closer to being coplanar with the central rings. The resulting dockings are very similar, indicating the SLIDE has the ability to select different conformations of the same molecule as potential ligands. This potential ligand is a possibility for testing given its high scores and multiply docked conformations, however given its significant hydrophobic character, suggesting low solubility, and the absence of commercial availability, it could prove difficult to test.

4.4 Discussion – Analysis of Potential Ligands selected by SLIDE

As a summary of the above results, a computation screening study was performed using asparaginyl-tRNA synthetase (AsnRS) from *Brugia malayi* as a target. A template generated from six ligands taken from the crystallographic structures was generated and used to screen against a subset of the Cambridge Structural Database (CSD) and identified variolin B as a potential ligand. An unbiased template was generated and used to suggest binding orientations of the 16 inhibitors selected from *in vitro* high-throughput screening against AsnRS. This unbiased template was also used to screen the CSD subset and identified three potential ligands: cercosporamide, phlorizin, and (E)-4,4′-dimesityl-but-3-enolidylidene-

but-3′-enolide.

One of the difficulties encountered in virtual screening experiments is the selection of the best potential ligands. Generally, a large number of molecules pass through to the scoring step. The number of molecules reaching this step can be controlled by changing the stringency of the screening parameters, such the allowed interatomic overlap parameters in SLIDE. However, increasing the stringency of these steric aspects of, which decreases the number of potential ligands, also increases the chance of missing worthwhile potential ligands. The ideal case would be to use loose parameters, allowing many molecules to pass, and then have a good scoring function for protein:ligand complementarity with which to choose the most promising candidates for further work, such as *in vitro* inhibition assays followed by drug lead optimization for successful inhibitors. While there are several scoring functions and docking forcefields available (Böhm 1994; Jain 1996; Eldridge et al. 1997; Böhm 1998; Murray et al. 1998; Mügge and Martin 1999; reviewed in Section 1.4), all have shortcomings. One problem with many of the currently available forcefields is their sensitivity to minor changes in relative protein-ligand orientation. During the screening process, it is not possible to finely tune each orientation due to computational time costs, but the untuned orientation may have a significantly lower score, thereby seeming like a poor potential ligand when minor changes could make it rise much higher in the score rankings. Such problems could potentially be handled by doing subsequent refinement of the orientation, using such a forcefield coupled with a molecular dynamics algorithm, for the best scoring candidates.

The approach of consensus scoring, suggested by Charifson et al. (1999) and applied in

several cases (Bissantz et al., 2000; Stahl and Rarey, 2001; Tripos Associates, 2001), is to score each docking with several scoring functions and then choose the dockings that score well with several scoring functions. The work presented here used a simple consensus function in which potential ligands which score well with both the SLIDE scoring function and the DrugScore scoring function are examined in detail. Various methods have been suggested and evaluated with respect to weighting and combination of the scoring functions (Wang and Wang, 2001).

One advantage of the internal SLIDE scoring function over DrugScore is the clear separation of the score into the hydrophobic contribution and the hydrogen bonding contribution. This separation of score into components can be a very useful to increase the affinity of a particular ligand as it can point towards deficiencies in the current ligand, e.g., that it is too hydrophobic or has too high a positive charge. Several other scoring functions employ this separation, but many of the empirically based scoring functions do not make such distinctions. Unfortunately, since DrugScore does not include a separation of scores in component parts, it is not possible to examine the reasons why some potential ligands score quite highly with one scoring function, but quite poorly with another. In fact, in general, the scoring functions do not correlate well with each other; over seven various screening runs, the mean correlation was only -0.59 (standard deviation of 0.17), indicating why consensus scoring is important. The ability to disentangle the score into its component parts is an advantage to forcefield based approaches, but often empirical approaches yield better correlation with binding affinities. The advantages and disadvantages with each scoring function point directly to the problem that none of the currently available scoring func-

tions are extremely accurate and that additional research must be done to improve both the accuracy and computational complexity of current functions to yield functions which can accurately compute the binding affinity of a docked protein-ligand complex in a reasonable amount of time. However, development of an improved scoring function is beyond the scope of this work.

Another currently difficulty of analysis of potential ligands selected by computational screening techniques is that of multiplicity, i.e., the identification of multiple binding orientations of a single ligand. In cases where the desire is to simply obtain a list of potential ligands, this may not be a concern, but in cases when one wants to use the docking results to obtain insight into how a protein may function, it becomes a significant problem. This situation arises in SLIDE since each database molecule interaction point triplet is matched to each template point template. Two sets of pairings may orient the database molecule in virtually the same orientation with respect to the target protein. The simplest approach to resolve this would be to employ a clustering algorithm on the ligand positions, giving a set of most similar orientations. This issue becomes more prevalent when dealing with multiple conformers of a ligand. Each of the multiple conformers may dock in multiple orientations, leading to a large increase in the number of dockings to be analyzed. While the docking orientations of very different conformations cannot be the same, molecular conformers which start in reasonably close conformations can merge into a very similar docking when SLIDE rotates ligand bonds. However, even with varying final, docked conformations, it may be instructional to identify groups in the ligands which tend to bind in similar locations in the target binding site. Direct visual analysis for a few dockings is pos-

sible, but when many favorable orientations are identified, as in this study, where 3 to 5300 orientations were identified for each HTS ligand, visual analysis becomes impossible. In this case, one could generate distribution maps for functional groups of particular interest, such as those generated in Isostar (Bruno et al., 1997) for molecules in the CSD and Protein Databank (PDB) databases.

Yet a third question arises from analysis of a set of potential ligands: are there trends in the types of molecules selected? Techniques for clustering molecules using a one-dimensional chemical representation of the molecules (Barnard et al., 2000), i.e., the SMILES string (Weininger, 1988), using one-dimensional bitstring representations of the molecules (Matter, 1997), and using a compatibility based Tanimoto coefficient (Verkhivker et al., 2000), have been previously implemented and could be applied in SLIDE as a post-screening step. The identification of groupings of potential ligands is useful to determine general chemical characteristics of potential ligands (i.e., are there specific functional groups which all selected molecules share?) and to reduce the ligands to a set which is analyzable. Also, having such clusters would yield information about the diversity of molecules which can bind to the target by analysis of the molecule clusters produced, i.e, are there many clusters containing only a small number of ligands produced, indicating diverse binding, or are there only a few, well occupied clusters produced, indicating binding over only a narrow range of molecules. A practical result of clustering would be to provide information about chemically similar neighbors of potential ligands that are commercially available for unavailable selected molecules or may be easier to work with experimentally than selected molecules. A second application of such a technique would be to estimate

the promiscuity of a protein, i.e., the ability of a protein to bind to diverse types of ligands, and may help identify difficult targets for drug design work.

The use of computational screening tools is to provide a set of molecules with likely binding to a protein of interest for further testing. The ability to dock known ligands strongly suggests that ligands which reside in the database used for screening will be selected in the list of potential ligands. The scoring function can rank these potential ligands to give a set of more probable actual ligands, but knowledge of the researcher can play a significant role in what potential ligands are most likely to dock, or, in pharmaceutical companies, are promising candidates for drug development. One approach would be to limit the database to molecules that are "drug-like" molecules, such as those that follow Lipinski's "rule of fives" (Lipinski et al., 2001). Other options would be simply limit the molecular weight of the molecules or exclude molecules which have a certain functional group. Another problem that sometimes arises is the availability of selected compounds for further work. While this may be less of a concern to a large, pharmaceutical company with the ability to readily synthesize many compounds, small research entities may wish to limit the screening database to only available compounds, such as those on the Available Chemicals Database (ACD). All of this gets down to limiting the screening database to the most promising candidates, which is the main goal of computational screening techniques.

The use of computational screening is only likely to increase. One idea not yet widely considered is the idea of reverse screening, i.e., beginning with a ligand molecule and identifying which proteins in a database, e.g., the PDB, it could dock to. This would require a automated technique to identify potential binding sites on proteins. Such a method

could provide leads as to what targets compounds of unknown function may bind too, leading to elucidation of their role in the cell, as well as provide an estimate of cross-reactivity potential for a drug candidate. While impossible to do now due to inadequate algorithmic techniques and computational resources, perhaps a distant future screening application would be to do computational hybrid screening, i.e., screening a protein of interest for binding to a set of database proteins, which could prove especially applicable as the proteomics movement comes into its own. One can be assured that new and innovative screening techniques and applications will continue to arrive on the scene.

Appendix A

Summary of Publications Outside of the Scope of the Work Presented in this Dissertation

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

  Water-mediated ligand interactions are essential to biological processes, from product displacement in thymidylate synthase to DNA recognition by Trp repressor, yet the structural chemistry influencing whether bound water is displaced or participates in ligand binding is not well characterized. Consolv, employing a hybrid k-nearest-neighbors classifier/genetic algorithm, predicts bound water molecules conserved between free and ligand-bound protein structures by examining the environment of each water molecule in the free structure. Four environmental features are used:

the water molecule's crystallographic temperature factor, the number of hydrogen bonds between the water molecule and protein, and the density and hydrophilicity of neighboring protein atoms. After training on 13 non-homologous proteins, Consolv predicted the conservation of active-site water molecules upon ligand binding with 75% accuracy (Matthews coefficient C_m = 0.41) for seven new proteins. Mispredictions typically involved water molecules predicted to be conserved that were displaced by a polar ligand atom, indicating that Consolv correctly assesses polar binding sites; 90% accuracy (C_m = 0.78) was achieved for predicting conserved active-site water or polar ligand atom binding. Consolv thus provides an accurate means for optimizing ligand design by identifying sites favored to be occupied by either a mediating water molecule or a polar ligand atom, as well as water molecules likely to be displaced by the ligand. Accuracy for predicting first-shell water conservation between independently determined structures was 61% (C_m=0.23). The ability to predict water-mediated and polar interactions from the free protein structure indicates the surprising extent to which the conservation or displacement of active-site bound water is independent of the ligand, and shows that the protein micro-environment of each water molecule is the dominant influence.

- M. L. Raymer, W. F. Punch, E. D. Goodman, P. C. Sanschagrin, and L. A. Kuhn. Simultaneous feature scaling and selection using a genetic algorithm. In T. Bäck,

editor, *Proceedings of the Seventh International Conference on Genetic Algorithms*, pages 561–567. Morgan Kaufmann Publishers, 1997.

Statistical pattern recognition techniques classify objects in terms of a representative set of features. The selection of features to measure and include can have a significant effect on the cost and accuracy of an automated classifier. Our previous research has shown that a hybrid between a k-nearest-neighbors (knn) classifier and a genetic algorithm (GA) can achieve greater classification accuracy than a knn alone by weighting features during knn classification. Here we describe an extension to this approach which further enhances feature selection through the simultaneous optimization of feature weights and selection of key features by including a masking vector on the GA chromosome. We present the results of our GA/knn feature selection method on two important problems from biochemistry and medicine: identification of conserved water molecules bound to protein surfaces, and diagnosis of thyroid deficiency. By allowing the GA to explore the effect of eliminating a feature from the classification without losing the weight knowledge already learned, the feature masking technique allows the GA/knn to efficiently examine noisy, complex, and high-dimensionality datasets to find combinations of features which classify the data more accurately. In both biomedical applications, use of the feature masking technique resulted in equivalent or better accuracy than feature weighting alone, while using fewer features for the

classification.

- L. Craig, P. C. Sanschagrin, A. Rozek, S. Lackie, L. A. Kuhn, and J. K. Scott. The role of structure in antibody cross-reactivity between peptides and folded proteins. *J. Mol. Biol.*, 281:183–201, 1998.

  Peptides have the potential for targeting vaccines against pre-specified epitopes on folded proteins. When polyclonal antibodies against native proteins are used to screen peptide libraries, most of the peptides isolated align to linear epitopes on the proteins. The mechanism of cross-reactivity is unclear; both structural mimicry by the peptide and induced fit of the epitope may occur. The most effective peptide mimics of protein epitopes are likely to be those that best mimic both the chemistry and the structure of epitopes. Our goal in this work has been to establish a strategy for characterizing epitopes on a folded protein that are candidates for structural mimicry by peptides. We investigated the chemical and structural bases of peptide-protein cross-reactivity using phage-displayed peptide libraries in combination with computational structural analysis. Polyclonal antibodies against the well-characterized antigens, hen eggwhite lysozyme and worm myohemerythrin, were used to screen a panel of phage-displayed peptide libraries. Most of the selected peptide sequences aligned to linear epitopes on the corresponding protein; the critical binding sequence of each epitope was revealed from these alignments. The structures of the critical sequences as they occur in other non-homologous proteins were analyzed

using the Sequery and Superpositional Structural Assignment computer programs. These allowed us to evaluate the extent of conformational preference inherent in each sequence independent of its protein context, and thus to predict the peptides most likely to have structural preferences that match their protein epitopes. Evidence for sequences having a clear structural bias emerged for several epitopes, and synthetic peptides representing three of these epitopes bound antibody with sub-micromolar affinities. The strong preference for a type II beta-turn predicted for one peptide was confirmed by NMR and circular dichroism analyses. Our strategy for identifying conformationally biased epitope sequences provides a new approach to the design of epitope-targeted, peptide-based vaccines.

- L. Fan, P. C. Sanschagrin, L. S. Kaguni, and L. A. Kuhn. The accessory subunit of mtDNA polymerase shares structural homology with aminoacyl-tRNA synthetases: Implications for a dual role as a primer recognition factor and processivity clamp. *Proc. Natl. Acad. Sci. USA*, 96(17):9527–32, 1999.

  The accessory subunit of the heterodimeric mtDNA polymerase (polγ) from *Drosophila* embryos is required to maintain the structural integrity or catalytic efficiency of the holoenzyme. cDNAs for the accessory subunit from *Drosophila*, man, mouse, and rat have been identified, and comparative sequence alignment reveals that the C-terminal region of about 120 aa is the most conserved. Furthermore, we demonstrate that the accessory subunit of animal polγ has both sequence and structural similarity

with class IIa aminoacyl-tRNA synthetases. Based on sequence similarity and fold recognition followed by homology modeling, we have developed a model of the three-dimensional structure of the C-terminal region of the accessory subunit of polγ. The model reveals a rare five-stranded beta-sheet surrounded by four alpha-helices with structural homology to the anticodon-binding domain of class IIa aminoacyl-tRNA synthetases. We postulate that the accessory subunit plays a role in the recognition of RNA primers in mtDNA replication, to recruit polγ to the template-primer junction. A similar role is served by the γ-complex in *Escherichia coli* DNA polymerase III, and indeed our accessory subunit model shows structural similarity with the N-terminal domain of the δ' subunit of the γ-complex. Structural similarity is also found with *E. coli* thioredoxin, the accessory subunit and processivity factor in bacteriophage T7 DNA polymerase. Thus, we propose that the accessory subunit of polγ is involved both in primer recognition and in processive DNA strand elongation.

Bibliography

R. Abagyan, M.. Totrov, and D. Kuznetsov. ICM-a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.*, 15(5):488–506, 1994.

F. H. Allen and O. Kennard. 3D search and research using the Cambridge Structural Database. *Chemical Design Automation News*, 8:1 & 31–37, 1993.

K. A. Alvi, L. Tenenbaum, and P. Crews. Anthelmintic polyfunctional nitrogen-containing terpenoids from marine sponges. *J. Nat. Prod.*, 54(1):71–78, 1991.

J. G. Arnez and D. Moras. Structural and functional considerations of the aminoacylation reaction. *Trends Biochem. Sci.*, 22(6):211–216, 1997.

M. Auf'mkolk, J. Koehrle, R. D. Hesch, S. H. Ingbar, and V. Cody. Crystal structure of phlorizin and the iodothyronine deiodinase inhibitory activity of phloretin analogues. *Biochem. Pharmacol.*, 35(13):2221–2227, 1986.

G. Ausiello, G. Cesarani, and M. Helmer-Citterich. ESCHER: A new docking procedure applied to the reconstruction of protein tertiary structure. *Prots*, 28:556–567, 1997.

K. Awadzi. Research notes from the Onchocerciasis Chemotherapy Research Centre,Ghana. *Ann. Trop. Med. Parasitol.*, 91:703–711, 1997.

E. N. Baker and R. E. Hubbard. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.*, 44(2):97–179, 1984.

J. M. Barnard, G. M. Downs, A. von Scholley-Pfab, and R. D. Brown. Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries. *J. Mol. Graph. Model*, 18(4-5):452–463, 2000.

C. L. Barnes and S. W. Hawkinson. Structure of adenosine 5′-methylphosphonate hemihydrate. *Acta Crystallogr. B*, 35:1724–1727, 1979.

C. R. Beddell, P. J. Goodford, F. E. Norrington, S. Wilkinson, and R. Wootton. Compounds designed to fit a site of known structure in human haemoglobin. *Br. J. Pharmacol.*, 57 (2):201–209, 1976.

M. J. Begley, L. Crombie, G. L. Griffiths, R. C. F. Jones, and M. Rahmani. Charge transfer and noncharge transfer forms of (e)-5,5′-dimesitylbifuranylidenediones: An X-ray structural investigation. *J. Chem. Soc.,Chem. Commun.*, 16:823–825, 1981.

G. W. Bemis and M. A. Murcko. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.*, 39(15):2887–2893, 1996.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000. `http://www.rscb.org/pdb/`.

P. Bernard, A. Golbraikh, D. Kireev, J. R. Chretien, and N. Rozhkova. Comparison of chemical databases: Analysis of molecular diversity with self-organising maps (SOM). *Analysis*, 26:333–341, 1998.

F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.

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

C. Bissantz, G. Folkers, and D. Rognan. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.*, 43 (25):4759–4767, 2000.

R. A. Blevins and A. Tulinsky. Comparison of the independent solvent structures of dimeric alpha-chymotrypsin with themselves and with gamma-chymotrypsin. *J. Biol. Chem.*, 260 (15):8865–8872, 1985.

N. S. Blom and J. Sygusch. High resolution fast quantitative docking using Fourier domain correlation techniques. *Proteins*, 27(4):493–506, 1997.

D. M. Blow, J. J. Birktoft, and B. S. Hartley. Role of a buried acid group in the mechanism of action of chymotrypsin. *Nature*, 221:337–340, 1969.

H. J. Böhm. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput. Aided. Mol. Des.*, 6(1):61–78, 1992a.

H. J. Böhm. LUDI: Rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aided. Mol. Des.*, 6(6):593–606, 1992b.

H. J. Böhm. 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(3):243–256, 1994.

H. J. Böhm. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput. Aided. Mol. Des.*, 12(4):309–323, 1998.

J. Boström, P. O. Norrby, and T. Liljefors. Conformational energy penalties of protein-bound ligands. *J. Comput. Aided. Mol. Des.*, 12(4):383–396, 1998.

B. C. Braden, B. A. Fields, and R. J. Poljak. Conservation of water molecules in an antibody-antigen interaction. *J. Mol. Recognit.*, 8(5):317–325, 1995.

R. D. Brown and Y. C. Martin. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.*, 36(3):572–584, 1996.

R. M. Brunne, E. Liepinsh, G. Otting, K. Wüthrich, and W. F. van Gunsteren. Hydration of proteins. A comparison of experimental residence times of water molecules solvating the bovine pancreatic trypsin inhibitor with theoretical model calculations. *J. Mol. Biol.*, 231:1040–1048, 1993.

I. J. Bruno, J. C. Cole, J. P. Lommerse, R. S. Rowland, R. Taylor, and M. L. Verdonk. IsoStar: a library of information about nonbonded interactions. *J. Comput. Aided. Mol. Des.*, 11(6):525–537, 1997.

S. K. Burley and G. A. Petsko. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*, 229(4708):23–28, 1985.

M. W. Casewell and R. L. Hill. In-vitro activity of mupirocin ('pseudomonic acid') against clinical isolates of Staphylococcus aureus. *J. Antimicrob. Chemother.*, 15(5):523–531, 1985.

P. S. Charifson, J. J. Corkery, M. A. Murcko, and W. P. Walters. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.*, 42(25):5100–5109, 1999.

H. Claussen, C. Buning, M. Rarey, and T. Lengauer. FLEXE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.*, 308(2):377–395, 2001.

G. M. Clore, A. Bax, P. T. Wingfield, and A. M. Gronenborn. Identification and localization of bound internal water in the solution structure of interleukin 1 beta by heteronuclear three-dimensional 1H rotating-frame Overhauser 15N-1H multiple quantum coherence NMR spectroscopy. *Biochemistry*, 29(24):5671–5676, 1990.

M. L. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221: 709–713, 1983. `http://www.biohedron.com`.

L. Craig, P. C. Sanschagrin, A. Rozek, S. Lackie, L. A. Kuhn, and J. K. Scott. The role of structure in antibody cross-reactivity between peptides and folded proteins. *J. Mol. Biol.*, 281:183–201, 1998.

D. J. Cummins, C. W. Andrews, J. A. Bentley, and M. Cory. Molecular diversity in chemical databases: Comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci.*, 36(4):750–763, 1996.

S. Cusack. Eleven down and nine to go. *Nat. Struct. Biol.*, 10:824–831, 1995.

Q. D. Dang and E. Di Cera. Residue 225 determines the Na^+-induced allosteric regulation of catalytic activity in serine proteases. *Proc. Natl. Acad. Sci. USA*, 93(20):10653–10656, 1996.

V. P. Denisov, J. Peters, H. D. Hörlein, and B. Halle. Using buried water molecules to explore the energy landscape of proteins. *Nat. Struct. Biol.*, 3:505–509, 1996.

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

R. S. DeWitte and E. I. Schaknovich. SMoG: *de novo* design method based on simple, fast, and accurate free energy estimates. 1. methodology and supporting evidence. *J. Am. Chem. Soc.*, 118:1651–1663, 1996.

E. Di Cera, E. R. Guinto, A. Vindigni, Q. D. Dang, Y. M. Ayala, M. Wuyi, and A. Tulinsky. The Na^+ binding site of thrombin. *J. Biol. Chem.*, 270(38):22089–22092, 1995.

J. S. Dixon. Evaluation of the CASP2 docking section. *Proteins*, Suppl:198–204, 1997.

R. L. Dunbrack, Jr. and F. E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, 6(8):1661–1681, 1997.

T. Earnest, E. Fauman, C. S. Craik, and R. Stroud. 1.59 A structure of trypsin at 120 K: Comparison of low temperature and room temperature structures. *Proteins*, 10(3): 171–187, 1991.

M. B. Eisen, D. C. Wiley, M. Karplus, and R. E. Hubbard. HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins*, 19(3):199–221, 1994.

D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319(6050):199–203, 1986.

M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided. Mol. Des.*, 11(5):425–445, 1997.

C. T. Esmon. The protein C anticoagulant pathway. *Arterioscler. Thromb.*, 12(2):135–145, 1992.

T. Ewing and I. D. Kuntz. Critical evaluation of search algortihms for automated molecular docking and database screening. *J. Comput. Chem.*, 18:1175–1189, 1997.

C. H. Faerman and P. Andrew Karplus. Consensus preferred hydrtation sites in six FKBP12-drug complexes. *Proteins Struct. Funct. Genet.*, 23:1–11, 1995.

E. B. Fauman, E. E. Rutenber, G. F. Maley, F. Maley, and R. M. Stroud. Water-mediated substrate/product discrimination: The product complex of thymidylate synthase at 1.83 Å. *Biochemsitry*, 33(6):1502–1511, 1994.

A. V. Filikov, V. Mohan, T. A. Vickers, R. H. Griffey, P. D. Cook, R. A. Abagyan, and T. L. James. Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. *J. Comput. Aided. Mol. Des.*, 14(6):593–610, 2000.

J. S. Finer-Moore, A. A. Kossiakoff, J. H. Hurley, T. Earnest, and R. M. Stroud. Solvent structure in crystals of trypsin determined by X-ray and neutron diffraction. *Proteins Struct. Funct. Genet.*, 12:203–222, 1992.

D. Fischer, S. L. Lin, H. L. Wolfson, and R. Nussinov. A geometry-based suite of molecular docking processes. *J. Mol. Biol.*, 248(2):459–477, 1995.

E. Fischer. Einfluss der configuration auf die wirkung der enzyme. *Berichte Deutsche Chemische Gesellschaft*, 27:2985–2993, 1894.

P. A. Fitzpatrick, A. C. Steinmetz, D. Ringe, and A. M Klibanov. Enzyme crystal structure in a neat organic solvent. *Proc. Natl. Acad. Sci. USA*, 90(18):8653–8657, 1993.

J. D. Forman-Kay, A. M. Gronenborn, P. T. Wingfield, and G. M. Clore. Determination of the positions of bound water molecules in the solution structure of reduced human thioredoxin by heteronuclear three-dimensional nuclear magnetic resonance spectroscopy. *J. Mol. Biol.*, 220(2):209–216, 1991.

X. Fradera, R. M. Knegtel, and J. Mestres. Similarity-driven flexible ligand docking. *Proteins*, 40(4):623–636, 2000.

A. T. Fuller, G. Mellows, M. Woolford, G. T. Banks, K. D. Barrow, and E. B. Chain. Pseudomonic acid: an antibiotic produced by Pseudomonas fluorescens. *Nature*, 234 (5329):416–417, 1971.

B. Furie and B. C. Furie. The molecular basis of blood coagulation. *Cell*, 53(4):505–518, 1988.

D. K. Gehlhaar, K. E. Moerder, D. Zichi, C. J. Sherman, R. C. Ogden, and S. T. Freer. *De novo* design of enzyme inhibitors by Monte Carlo ligand generation. *J. Med. Chem.*, 38 (3):466–472, 1995a.

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

V. Gillet, A. P. Johnson, P. Mata, S. Sike, and P. Williams. SPROUT: a program for structure generation. *J. Comput. Aided. Mol. Des.*, 7(2):127–153, 1993.

V. J. Gillet, W. Newell, P. Mata, G. Myatt, S. Sike, Z. Zsoldos, and A. P. Johnson. SPROUT: Recent developments in the *de novo* design of molecules. *J. Chem. Inf. Comput. Sci.*, 34 (1):207–217, 1994.

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

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

P. J. Goodford. Drug design by the method of receptor fit. *J. Med. Chem.*, 27(5):558–564, 1984.

P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, 28(7):849–857, 1985.

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

J. Greer, J. W. Erickson, J. J. Baldwin, and M. D. Varney. Application of the three-dimensional structures of protein target molecules in structure-based drug design. *J. Med. Chem.*, 37(8):1035–1054, 1994.

S. Ha, R. Andreani, A. Robbins, and I. Muegge. Evaluation of docking/scoring approaches: a comparative study based on MMP3 inhibitors. *J. Comput. Aided. Mol. Des.*, 14(5): 435–448, 2000.

S. M. Habermann and K. P. Murphy. Energetics of hydrogen bonding in proteins: a model compound study. *Protein Sci.*, 5(7):1229–1239, 1996.

T. A. Halgren. Merck Molecular Forcefield. I. Basis, form, score, parametrization, and performance of MMFF94. *J. Comput. Chem.*, 17:490–519, 1996.

R. D. Head, M. L. Smythe, T. I. Oprea, C. L. Waller, S. M. Green, and G. R. Marshall. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.*, 118(16):3959–3969, 1996.

M. Helmer-Citterich and A. Tramontano. PUZZLE: a new method for automated protein docking based on surface shape complementarity. *J. Mol. Biol.*, 235(3):1021–1031, 1994.

B. A. Hirayama, A. Diez-Sampedro, and E. M. Wright. Common mechanisms of inhibition for the Na(+)/glucose (hSGLT1) and Na(+)/Cl(-)/GABA (hGAT1) cotransporters. *Br. J. Pharmacol.*, 134(3):484–495, 2001.

R. W. Hooft, C. Sander, and G. Vriend. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins*, 26(4):363–376, 1996.

J. Hughes and G. Mellows. Interaction of pseudomonic acid A with Escherichia coli B isoleucyl-tRNA synthetase. *Biochem. J.*, 191(1):209–219, 1980.

R. M. Jackson, H. A. Gabb, and M. J. E. Sternberg. Rapid refinement of protein interfaces incorporating solvation: Application to the docking problem. *J. Mol. Biol.*, 276:265–285, 1998.

A. K. Jain and B. Chandrasekaran. Dimensionality and sample size considerations in pattern recognition in practice. In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of Statistics*, volume 2, pages 835–855. North-Holland, 1982.

A. N. Jain. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput. Aided. Mol. Des.*, 10:427–440, 1996.

J. Janin and C. Chothia. The structure of protein-protein recognition sites. *J. Biol. Chem.*, 265(27):16027–16030, 1990.

A. Joachimiak, T. E. Haran, and P. B. Sigler. Mutagenesis supports water mediated recognition in the Trp repressor-operator system. *EMBO J.*, 245:43–53, 1994.

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

G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267(3):727–748, 1997.

P. A. Karplus and C. Faerman. Ordered water in macromolecular structure. *Curr. Opin. Struct. Biol.*, 4:770–776, 1994.

T. Kitlar, F. Doring, D. F. Diedrich, R. Frank, H. Wallmeier, R. K. Kinne, and J. Deutscher. Interaction of phlorizin, a potent inhibitor of the Na+/D-glucose cotransporter, with the NADPH-binding site of mammalian catalases. *Protein Sci.*, 3(4):696–700, 1994.

G. Klebe. The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands. *J. Mol. Biol.*, 237(2):212–235, 1994.

G. Klebe. Recent developments in structure-based drug design. *J. Mol. Med.*, 78(5):269–281, 2000.

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

R. M. Knegtel and M. Wagener. Efficacy and selectivity in flexible database docking. *Proteins*, 37(3):334–345, 1999.

P. Koehl and M. Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.*, 239:249–275, 1994.

P. Koehl and M. Delarue. Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.*, 6:222–226, 1996.

M. Konishi, M. Nishio, K. Saitoh, T. Miyaki, T. Oki, and H. Kawaguchi. Cispentacin, a new antifungal antibiotic. I. Production, isolation, physico-chemical properties and structure. *J. Antibiot. (Tokyo).*, 42(12):1749–1755, 1989.

I. Konrad and R. Roschenthaler. Inhibition of phenylalanine tRNA synthetase from Bacillus subtilis by ochratoxin A. *FEBS Lett.*, 83(2):341–347, 1977.

A. A. Kossiakoff, M. D. Sintchak, J. Shpungin, and L. G. Presta. Analysis of solvent structure in proteins using neutron D_2O-H_2O solvent maps: Pattern of primary and secondary hydration in trypsin. *Proteins Struct. Funct. Genet.*, 12:223–236, 1992.

B. Kramer, M. Rarey, and T. Lengauer. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins*, 37(2):228–241, 1999.

M. M. Krem and E. Di Cera. Conserved water molecules in the specificity pocket of serine proteases and the molecular mechanism of Na+ binding. *Proteins*, 30(1):34–42, 1998.

L. A. Kuhn, M. A. Siani, M. E. Pique, C. L. Fisher, E. D. Getzoff, and J. A. Tainer. The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. *J. Mol. Biol.*, 228:13–22, 1992a.

L. A. Kuhn, M. A. Siani, M. E. Pique, C. L. Fisher, E. D. Getzoff, and J. A. Tainer. The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. *J. Mol. Biol.*, 228:13–22, 1992b.

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

I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161(2):269–288, 1982.

I. D. Kuntz and W. Kauzmann. Hydration of proteins and polypeptides. *Adv. Protein Chem.*, 28:239–345, 1974.

J. E. Ladbury. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem. Biol.*, 3:973–980, 1996.

P. Y. Lam, P. K. Jadhav, C. J. Eyermann, C. N. Hodge, Y. Ru, L. T. Bacheler, J. L. Meek, M. J. Otto, M. M. Rayner, Y. N. Wong, and *et al.* Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science*, 263(5145):380–384, 1994.

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

A. R. Leach and I. D. Kuntz. Conformational analysis of flexible ligands in macromolecular receptor. *J. Comput. Chem.*, 13:730–748, 1992.

C. Lemmen, T. Lengauer, and G. Klebe. FLEXS: A method for fast flexible ligand superposition. *J. Med. Chem.*, 41(23):4502–4520, 1998.

M. Levitt and B. H. Park. Water: Now you see it, now you don't. *Structure*, 1:221–226, 1993.

C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 46(1-3):3–26, 2001.

S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The penultimate rotamer library. *Proteins*, 40(3):389–408, 2000.

M. Loyevsky and Z. I. Cabantchik. Antimalarial action of hydrophilic drugs: Involvement of aqueous access routes to intracellular parasites. *Mol. Pharmacol.*, 45(3):446–452, 1994.

M. De Maeyer, J. Desmet, and I. Lasters. All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold. Des.*, 2(1):53–66, 1997.

J. G. Mandell, V. A. Roberts, M. E. Pique, V. Kotlovyi, J. C. Mitchell, E. Nelson, I. Tsigelny, and L. F. Ten Eyck. Protein docking using continuum electrostatics and geometric fit. *Protein Eng.*, 14(2):105–113, 2001.

H. Matter. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.*, 40(8):1219–1229, 1997.

E. C. Meng, B. K. Shoichet, and I. D. Kuntz. Automated docking with grid-based energy evaluation. *J. Comput. Chem.*, 13(6):505–524, 1992.

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

E. Meyer. Internal water molecules and H-bonding in biological macromolecules: a review of structural features with functional implications. *Protein Sci.*, 1(12):1543–1562, 1992.

V. Mikol, C. Papageorgiou, and X. Borer. The role of water molecules in the structure-based design of (5-hydroxynorvaline)-2-cyclosporin: Synthesis, biological activity, and crystallographic analysis with cyclophilin A. *J. Med. Chem.*, 38(17):3361–3367, 1995.

M. D. Miller, S. K. Kearsley, D. J. Underwood, and R. P. Sheridan. FLOG: A system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J. Comput. Aided. Mol. Des.*, 8(2):153–174, 1994.

S. Miller, J. Janin, A. M. Lesk, and C. Chothia. Interior and surface of monomeric proteins. *J. Mol. Biol.*, 196(3):641–656, 1987.

G. W. A. Milne and J. A. Miller. The NCI Drug Information System. 1. system overview. *J. Chem. Inf. Comput. Sci.*, 26:154–159, 1986.

J. B. Mitchell, R. A. Laskowski, and J. M. Thornton. Non-randomness in side-chain packing: the distribution of interplanar angles. *Proteins*, 29(3):370–380, 1997.

J. B. Moon and W. J. Howe. Computer design of bioactive molecules: a method for receptor-based *de novo* ligand design. *Proteins*, 11(4):314–328, 1991.

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

G. M. Morris, D. S. Goodsell, R. Huey, and A. J. Olson. Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. *J. Comput. Aided. Mol. Des.*, 10:293–304, 1996.

J. Moult, T. Hubbard, S. H. Bryant, K. Fidelis, and J. T. Pedersen. Critical assessment of methods of protein structure prediction (CASP): Round II. *Proteins*, Supplement 1:2–6, 1997. `http://predictioncenter.llnl.gov/casp2/Casp2.html`.

I. Mügge and Y. C. Martin. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.*, 42(5):791–804, 1999.

C. W. Murray, T. R. Auton, and M. D. Eldridge. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comput. Aided. Mol. Des.*, 12(5):503–519, 1998.

T. Nakatsu, H. Kato, and J. Oda. Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. *Nat. Struct. Biol.*, 5(1): 15–19, 1998.

G. Nass, K. Poralla, and H. Zahner. Effect of the antibiotic Borrelidin on the regulation of threonine biosynthetic enzymes in E. coli. *Biochem. Biophys. Res. Commun.*, 34(1): 84–91, 1969.

Y. Nishibata and A. Itai. Confirmation of usefulness of a structure construction program based on three-dimensional receptor structure for rational lead generation. *J. Med. Chem.*, 36(20):2921–2928, 1993.

J. W. M. Nissink, M. L. Verdonk, and G. Klebe. Simple knowledge-based descriptors to predict protein-ligand interactions. methodology and validation. *J. Comput. Aided. Mol. Des.*, 14(8):787–803, 2000.

A. J. Oakley, M. L. Bello, A. Battistoni, G. Ricci, J. Rossjohn, H. O. Villar, and M. W. Parker. The structures of human glutathione transferase P1-1 in complex with glutathione and various inhibitors at high resolution. *J. Mol. Biol.*, 274(1):84–8100, 1997. PDB 12gs.

A. Ogilvie, K. Wiebauer, and W. Kersten. Inhibition of leucyl-transfer ribonucleic acid synthetasymol. *Biochem. J.*, 152(3):511–515, 1975.

T. Onda, Y. Hashimoto, M. Nagai, H. Kuramochi, S. Saito, H. Yamazaki, Y. Toya, I. Sakai, C. J. Homcy, K. Nishikawa, and Y. Ishikawa. Type-specific regulation of adenylyl cyclase; Selective pharmacological stimulation and inhibition of adenylyl cyclase isoforms. *J. Biol. Chem.*, 2001. in press.

C. M. Oshiro, I. D. Kuntz, and J. S. Dixon. Flexible ligand docking using a genetic algorithm. *J. Comput. Aided. Mol. Des.*, 9(2):113–130, 1995.

G. Otting, E. Liepinsh, and K. Wüthrich. Protein hydration in aqueous solution. *Science*, 254:974–980, 1991.

G. Otting and K. Wüthrich. Studies of protein hydration in aqueous solution by direct NMR observation of individual protein-bound water molecules. *J. Am. Chem. Soc.*, 111: 1871–1875, 1989.

Z. Otwinowski, R. W. Schevitz, R.-G. Zhang, C. L. Lawson, A. Joachimiak, R. Q. Marmorstein, B. F. Luisi, and P. B. Sigler. Crystal structure of *trp* repressor/operator complex at atomic resolution. *Nature*, 335:321–329, 1988.

W. Paetz and G. Nass. Biochemical and immunological characterization of threonyl-tRNA synthetase of two borrelidin-resistant mutants of Escherichia coli K12. *Eur. J. Biochem.*, 35(2):331–337, 1973.

D. A. Parrish and A. A. Pinkerton. Estradiol methanol hemihydrate. *Acta Crystallogr. C*, 55:IUC9900100, 1999.

D. A. Pearlman and M. A. Murcko. Concepts: New dynamic algorihm for *de novo* drug design. *J. Med. Chem.*, 38(3):466–472, 1995.

E. Perola, K. Xu, T. M. Kollmeyer, S. H. Kaufmann, F. G. Prendergast, and Y. P. Pang. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.*, 43(3):401–408, 2000.

J. J. Perona, C. S. Craik, and R. J. Fletterick. Locating the catalytic water molecule in serine proteases. *Science*, 261(5121):620–622, 1993.

N. B. Perry, L. Ettouati, M. Litaudon, J. W. Blunt, M. H. G. Munro, S. Parkin, and H. Hope. Alkaloids from the antarctic sponge *Kirkpatrickia varialosa*. Part 1: Variolin B, a new antitumour and antiviral compound. *Tetrahedron*, 50:3987–3992, 1994.

W. R. Pitt, J. Murray-Rust, and J. M. Goodfellow. AQUARIUS2: Knowledge-based modeling of solvent sites around proteins. *J. Comput. Chem.*, 14:1007–1018, 1993.

M. Rarey, B. Kramer, and T. Lengauer. The particle concept: Placing discrete water molecules during protein-ligand docking predictions. *Proteins*, 34(1):17–28, 1999.

M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261(3):470–489, 1996a.

M. Rarey, S. Wefing, and T. Lengauer. Placement of medium-sized molecular fragments into active sites of proteins. *J. Comput. Aided. Mol. Des.*, 10(1):41–54, 1996b.

A. A. Rashin, M. Iofin, and B. Honig. Internal cavities and buried waters in globular proteins. *Biochemistry*, 25:3619–3625, 1986.

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

D. C. Roe and I. D. Kuntz. BUILDER V.2: Improving the chemistry of a *de novo* design strategy. *J. Comput. Aided. Mol. Des.*, 9(3):269–282, 1995.

S. H. Rotstein and M. A. Murcko. GenStar: A method for *de novo* drug design. *J. Comput. Aided. Mol. Des.*, 7(1):23–43, 1993a.

S. H. Rotstein and M. A. Murcko. GroupBuild: a fragment-based method for *de novo* drug design. *J. Med. Chem.*, 36(12):1700–1710, 1993b.

J. A. Rupley and G. Careri. Protein hydration and function. *Adv. Protein Chem.*, 41:37–172, 1991.

J. S. Sack. CHAIN: A crystallographic modeling program. *J. Mol. Graph.*, 2:224–225, 1988.

B. Sandak, R. Nussinov, and H. J. Wolfson. A method for biomolecular structural recognition and docking allowing conformational flexibility. *J. Comput. Biol.*, 5(4):631–654, 1998a.

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

P. C. Sanschagrin and L. A. Kuhn. 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, 1998.

M. Schapira, B. M. Raaka, H. H. Samuels, and R. Abagyan. Rational discovery of novel nuclear hormone receptor antagonists. *Proc. Natl. Acad. Sci. USA*, 97(3):1008–1013, 2000.

C. A. Schiffer, R. Huber, K. Wuthrich, and W. F. van Gunsteren. Simultaneous refinement of the structure of BPTI against NMR data measured in solution and X-ray diffraction data measured in single crystals. *J. Mol. Biol.*, 241(4):588–599, 1994.

P. Schimmel, J. Tao, and J. Hill. Aminoacyl tRNA synthetases as targets for new anti-infectives. *FASEB J.*, 12:1599–1609, 1998.

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

V. Schnecke and L. A. Kuhn. Modeling induced fit and controlling molecular diversity during database screening for ligands. *Proteins Struct. Funct. Genet.*, 2000a. In press.

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

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

N. E. Shemetulskis, J. r. Dunbar JB, B. W. Dunbar, D. W. Moreland, and C. Humblet. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput. Aided. Mol. Des.*, 9(5):407–416, 1995.

B. K. Shoichet, D. L. Bodian, and I. D. Kuntz. Molecular docking using shape descriptors. *J. Comput. Chem.*, 13:380–397, 1992.

B. K. Shoichet and I. D. Kuntz. Matching chemistry and shape in molecular docking. *Protein Eng.*, 6(7):723–732, 1993.

B. K. Shoichet, A. R. Leach, and I. D. Kuntz. Ligand solvation in molecular docking. *Proteins*, 34(1):4–16, 1999.

P. T. Singer, A. Smalas, R. P. Carty, W. F. Mangel, and R. M. Sweet. The hydrolytic water molecule in trypsin, revealed by time-resolved Laue crystallography. *Science*, 259 (5095):669–673, 1993.

J. Singh, J. Saldanha, and J. M. Thornton. A novel method for the modelling of peptide ligands to their receptors. *Protein Eng.*, 4:251–261, 1991.

J. Singh and J. M. Thornton. SIRIUS. An automated method for the analysis of the preferred packing arrangements between protein groups. *J. Mol. Biol.*, 211(3):595–615, 1990.

V. Sobolev, R. C. Wade, G. Vriend, and M. Edelman. Molecular docking using surface complementarity. *Proteins*, 25(1):120–129, 1996.

U. Sreenivasan and P. H. Axelsen. Buried water in homologous serine proteases. *Biochemistry*, 31(51):12785–12791, 1992.

M. Stahl and H. J. Böhm. Development of filter functions for protein-ligand docking. *J. Mol. Graph. Model*, 16(3):121–132, 1998.

M. Stahl and M. Rarey. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.*, 44(7):1035–1042, 2001.

N. C. Strynadka, M. Eisenstein, E. Katchalski-Katzir, B. K. Shoichet, I. D. Kuntz, R. Abagyan, M. Totrov, J. Janin, J. Cherfils, F. Zimmerman, A. Olson, B. Duncan, M. Rao, R. Jackson, M. Sternberg, and M. N. James. Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. *Nat. Struct. Biol.*, 3(3):233–239, 1996.

A. I. Su, D. M. Lorber, G. S. Weston, W. A. Baase, B. W. Matthews, and B. K. Shoichet. Docking molecules by families to increase the diversity of hits in database screens: Computational strategy and experimental evaluation. *Proteins*, 42(2):279–293, 2001.

F. Sugawara, S. Strobel, G. Strobel, R. D. Larsen, D. L. Berglund, G. Gray, N. Takahashi, S. J. Coval, T. J. Stout, and J. Clardy. The structure and biological activity of cercosporamide from *Cercosporidium henningsii*. *J. Org. Chem.*, 56:909–910, 1991.

K. Tanaka, M. Tamaki, and S. Watanabe. Effect of furanomycin on the synthesis of isoleucyl-tRNA. *Biochim. Biophys. Acta.*, 195(1):244–245, 1969.

D. M. Tanenbaum, Y. Wang, S. P. Williams, and P. B. Sigler. Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. *Proc. Natl. Acad. Sci. USA*, 95(11):5998–6003, 1998.

M. Totrov and R. Abagyan. Detailed ab initio prediction of lysozyme-antibody complex with 1.6 A accuracy. *Nat. Struct. Biol.*, 1(4):259–263, 1994.

M. Totrov and R. Abagyan. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins Struct. Funct. Genet.*, Supplement 1:215–220, 1997.

J. Travis. Proteins and organic solvents make an eye-opening mix. *Science*, 262(5138): 1374, 1993.

Tripos Associates. *The SYBYL Software*. St. Louis, MO, 2001. `http://www.tripos.com`.

G. V. Trunk. A problem of dimensionality: A simple example. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 1:306–307, 1979.

I. A. Vakser and C. Aflalo. Hydrophobic docking: a proposed enhancement to molecular recognition techniques. *Proteins*, 20(4):320–329, 1994.

W. F. van Gunsteren, H. J. Berendsen, J. Hermans, W. G. Hol, and J. P. Postma. Computer simulation of the dynamics of hydrated protein crystals and its comparison with x-ray data. *Proc. Natl. Acad. Sci. USA*, 80(14):4315–4319, 1983.

G. M. Verkhivker, D. Bouzida, D. K. Gehlhaar, P. A. Rejto, S. Arthurs, A. B. Colson, S. T. Freer, V. Larson, B. A. Luty, T. Marrone, and P. W. Rose. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput. Aided. Mol. Des.*, 14(8): 731–751, 2000.

G. M. Verkhivker, P. A. Rejto, D. Bouzida, S. Arthurs, A. B. Colson, S. T. Freer, D. K. Gehlhaar, V. Larson, B. A. Luty, T. Marrone, and P. W. Rose. Towards understanding the mechanisms of molecular recognition by computer simulations of ligand-protein interactions. *J. Mol. Recognit.*, 12(6):371–389, 1999.

J. Vijayalakshmi, K. P. Padmanabhan, K. G. Mann, and A. Tulinsky. The isomorphous structures of prethrombin2, hirugen-, and PPACK-thrombin: Changes accompanying activation and exosite binding to thrombin. *Protein Sci.*, 3(12):2254–2271, 1994.

H. O. Villar and L. M. Kauvar. Amino acid preferences at protein binding sites. *FEBS Lett.*, 349(1):125–130, 1994.

J. H. Voigt, B. Bienfait, S. Wang, and M. C. Nicklaus. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.*, 41 (3):702–712, 2001.

W. P. Walters, M. T. Stahl, and M. A. Murcko. Virtual screening – an overview. *Drug Discov. Today*, 3(4):160–178, 1998.

H. Wang and A. Ben-Naim. A possible involvement of solvent-induced interactions in drug design. *J. Med. Chem.*, 39:1531–1539, 1996.

R. Wang and S. Wang. How does consensus scoring work for virtual library screening? an idealized computer experiment. *J. Chem. Inf. Comput. Sci.*, 41(5):1422–1426, 2001.

S. J. Weiner, P. A. Kollmann, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta Jr., and P. Weiner. A new forcefield for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784, 1984.

S. J. Weiner, P. A. Kollmann, D. T. Nguyen, and D. A. Case. An all atom forcefield for simulations of proteins and nucleic acid. *J. Comput. Chem.*, 7:230–252, 1986.

D. Weininger. SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28: 31, 1988.

W. Welch, J. Ruppert, and A. N. Jain. Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.*, 3:449–462, 1996.

R. G. Werner, L. F. Thorpe, W. Reuter, and K. H. Nierhaus. Indolmycin inhibits prokaryotic tryptophanyl-tRNA ligase. *Eur. J. Biochem.*, 68(1):1–3, 1976.

I. A. Wilson and D. H. Fremont. Structural analysis of MHC Class I molecules wirh bound peptide antigen. *Semin. Immunol.*, 5:75–80, 1993.

A. Wlodawer, M. Miller, M. Jaskolski, B. K. Sathyanarayana, E. Baldwin, I. T. Weber, L. M. Selk, L. Clawson, J. Schneider, and S. B. Kent. Conserved folding in retroviral proteases: Crystal structure of a synthetic HIV-1 protease. *Science*, 245(4918):616–621, 1989.

R. X. Xu, R. P. Meadows, and S. W. Fesik. Heteronuclear 3D NMR studies of water bound to an FK506 binding protein/immunosuppressant complex. *Biochemistry*, 32(10):2473–2480, 1993.

E. Zhang and A. Tulinsky. The molecular environment of the Na^+ binding site of thrombin. *Biophys. Chem.*, 63(2-3):185–200, 1997.

X.-J. Zhang and B. W. Matthews. Conservation of solvent-binding site in 10 crystal forms of T4 lysozyme. *Protein Sci.*, 3:1031–1039, 1994.