## Title

The design of a ligand discovery and optimization system for structured-based drug design

## Author

Clark, Kevin Patrick

## Publication Date

1997

# The Design of a Ligand Discovery and Optimization System for Structure-based Drug Design

by
Kevin Patrick Clark
M.S., EECS, Massachusetts Institute of Technology, 1990
B.S., EECS, University of California at Berkeley, 1987

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISIONS

of the

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

and

UNIVERSITY OF CALIFORNIA BERKELEY

[redacted]

Date University Librarian

Degree Conferred: ..........................................................................

To my mother and father,

Mary and Wesley Clark

# Acknowledgements

I thank the Computer Graphics Laboratory for their support and encouragement in this project. I begin by thanking Professors Langridge and Ferrin for the computing facilities, their guidance, and their support. I thank Norma Belfer, Willa Crowell, and Al Conde for making the Computer Graphics Laboratory a great working environment. I thank Elaine Meng and Teri Klein for their introducing me to molecular docking and for their suggestions. I thank Eric Pettersen, Greg Couch, and Conrad Huang for our discussions and for answering my numerous questions. Finally, I thank Ajay with whom I collaborated during most of this work. This collaboration began while he was a postdoc in the Computer Graphics Laboratory and has continued since he began working at Vertex Pharmaceuticals.

Chapter 4 of this thesis is a reprint of the material as it appeared in: Clark, K. P. and Ajay, Journal of Computational Chemistry, 16(10):1210–1226, 1995.

I thank my dissertation committee: Professors Robert Langridge, Thomas Ferrin, Fred Cohen, and Boris Rubinsky for their comments and suggestions.

I thank the UCSF/UCB Joint Bioengineering Graduate Group for their support. Debra Harris has always been there to help me get over some of the hurdles of the joint campus program. I thank my fellow bioengineering students for our discussions, happy hours, and other activities.

I thank the NIH Division of Research Resources (RR-1081) for providing the facilities and for funding this research.

Finally, I thank my parents, my sisters, and my brother for their love and encouragement.

# Abstract

Unprecedented opportunities in drug design now exist because the basis of disease is being understood on a more molecular level. With the increasing number of known protein structures, it is likely that the structure of the target enzyme will be known or can be determined using X-ray crystallography, NMR techniques, or homology modeling. The structure of the enzyme can be used to design new therapeutic agents that are complementary to the receptor and bind with high affinity. In this dissertation I have developed a new system for the discovery and optimization of new lead compounds in the drug design process. The system contains a molecular graphics component as well as integrated molecular docking methods. Molecular graphics are important in the drug design process. They identify receptor-ligand interactions and suggest modifications to the lead compound that may improve binding. The system uses the most recent developments in computer graphics to interactively compute and display molecular surfaces and volumes of interest. Texture mapping is also used to allow the chemist to more clearly render the information. Molecular docking methods are invaluable for drug design. They are used to screen for new lead compounds and identify chemically plausible structures for the intermolecular complex. Once a binding mode is found, chemical modifications to the lead compound can be suggested. In this dissertation I developed methods which use genetic algorithms to dock flexible ligands to rigid receptors. I studied the use of molecular mechanics force fields and empirical estimates of the binding affinity as scoring functions. I found that the latter work much better within the context of a genetic algorithm optimization. I performed three different types simulations on seven different receptor-ligand complexes. With an initially docked fragment, my system performed as well as other incremental construction approaches. I also demonstrate the my system is robust to errors in the orientation of the base fragment. Furthermore, in all cases, when no information about the binding was provided, my system found solutions very close to the crystal structure. This system should prove to be useful for the design of new therapeutic agents.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Unprecedented opportunities in drug design now exist because the basis of disease is being understood on an increasingly molecular level. Infectious agents create enzymes that are necessary for their replication and survival, and knowledge of the three-dimensional structure of these targets offers a direct route to the development of therapeutic agents. Traditionally, most lead compounds have been found through high volume random screening of corporate databases or natural products [1]. Families of related compounds are then synthesized and tested. Subsequent developments are often assisted by quantitative structure activity relationships (QSAR) [2] or other related techniques [3, 4]. An alternative approach has been to use an understanding of the biological or biochemical mechanisms to direct the design of a new drug [5, 6]. Structure-based drug design is a relatively new approach that has been brought about largely by improvements in molecular structure determination and new computational tools. Using this technique, new inhibitors for thymidylate synthase [7], purine nucleoside phosphorylase [8], and HIV-1 protease [9, 10] have been designed. Some of these are now in clinical use.

## 1.1 Structure-based Drug Design

The goal of drug design is to apply the principles of molecular recognition to the development of novel ligands. Seymour Cohen proposed a general paradigm for

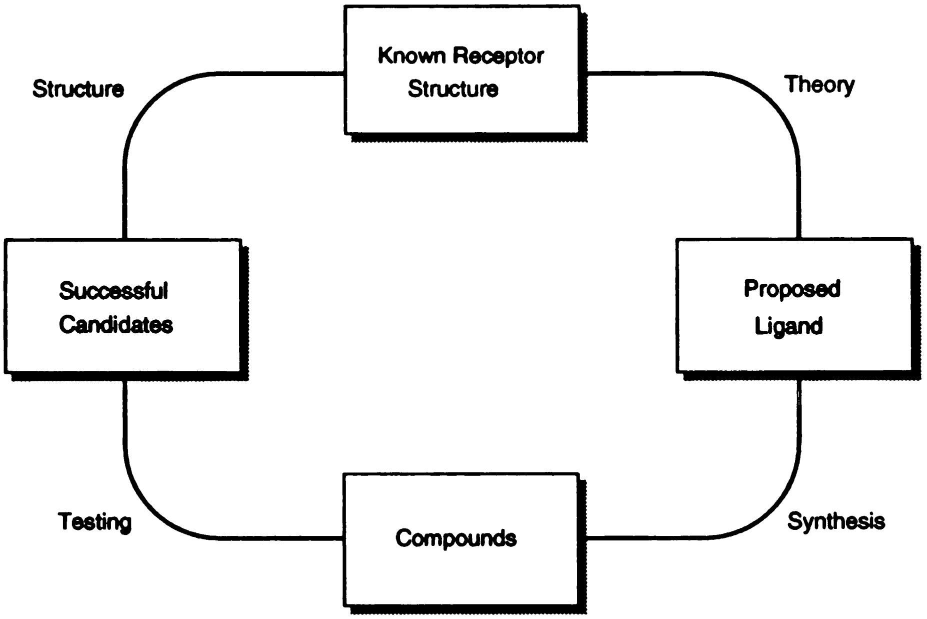


Figure 1.1: A four step strategy for the structure-based design of new therapeutic agents.

the design of inhibitors of infectious diseases in 1977 [11]. Now, nearly two decades later, the computational and experimental techniques have advanced to the state that the approach shown in Figure 1.1 is now practical. This method depends upon the existence of structural information on the receptor site. With recent advances in methods for protein structure determination, it is likely that the structure of the target will be available in the Brookhaven Protein Data Bank [12] or can be determined through either X-ray crystallography, NMR techniques, or homology modeling. Model structures of proteases from the schistosome and malaria parasites have recently been used as drug design targets [13]. Given structural information about the receptor site, ligands can be designed, synthesized, and tested. The information about the structure of the ligand-receptor complex can be used to further optimize the ligand. The procedure can be iterated until a potential drug has been designed. Although not all targets are enzymes, I only consider these in this dissertation.

One technique for discovering new biologically active compounds is to search databases of small molecules. There are two common approaches. The first involves pharmacophore matching where a database is searched for structures that have similar arrangements of atoms or group of atoms [14]. Molecular docking, on the other hand, searches for molecules that fit into a specific binding site. These algorithms orient molecules into the receptor and by considering specific interactions with the receptor and ranks the individual compounds in terms of their interaction energies. The docking problem can be stated as follows: Given the structures of the free ligand and receptor, predict the structure and binding free energy of the complex.

Computational approaches to predicting the binding geometries of ligands to receptors are of interest because they yield insight into the mechanisms of molecular recognition and for their utility in designing new therapeutic agents. Such techniques are useful for the optimization of lead compounds. Once a compound has been found and its preferred mode of binding identified, structural modifications of the compound can be made to optimize the interactions with the receptor. The underlying assumption of structure-based design is that ligands must be structurally and chemically complementary to their targets. This idea is not new to medicinal chemistry. Emil Fischer introduced this "lock and key" concept [15, 16] to describe the interactions between ligands and receptors over a century ago. Most molecular docking system make use of this idea.

Ideally, a structure-based design system should be able to 1) screen databases of small molecules for lead compounds, 2) rank a set of molecules according to their binding affinities, 3) elucidate all the chemically plausible binding modes for a given ligand and receptor, and finally, 4) provide methods for at least interactively optimizing lead compounds. The DOCK system, published in 1982, was one of the first docking methods [17]. Since that time Kuntz and coworkers have continued to develop and improve various aspects of the algorithm [18, 19, 20]. DOCK is one of the most successful molecular docking systems. Its strongest feature is its ability to rapidly screen compounds and identify potential leads, or alternatively, rule out those that will most likely not have any chance of binding [13]. It has been used to discover lead compounds for inhibitors of a variety of different macromolecular targets (see

Kuntz [1] and references therein).

One area where most molecular docking procedures, including DOCK, fall short of the ideal is in the ranking of a set of compounds according to their relative binding affinities. Docking simulations often find other binding geometries that are sterically and chemically plausible, but cannot be distinguished by the particular scoring function used. This has been seen in both protein-protein [21, 22] and protein-ligand [23] interactions. The quantity of interest for the design of new ligands is the free energy of binding, $\Delta G_{bind}$, in aqueous solution. Free energy perturbation techniques [24, 25, 26] in the best cases can have accuracies of ±1 kcal/mol. Although these calculations are perhaps the most promising for the purpose of ranking compounds according to their binding affinities, they are not without their limitations. First, they are computationally very costly because they require good sampling of the conformational and configurational states of both the ligand and the receptor. Second, they require an accurate model of the protein-ligand structure [27] which is usually not available since the goal of the calculation is to identify the binding mode of the ligand. Furthermore, the results are accurate only for closely related ligands and cannot be applied to different molecules [28]. Quantitative estimates of binding affinity require relative accuracies within 1 kcal/mol in free energy. An order of magnitude change in the binding affinity is equivalent to a change in free energy of 1.4 kcal/mol. Force field calculations rarely achieve accuracies better than ±2 kcal/mol. For computational efficiency, however, most evaluation schemes use much simpler scoring functions. These are reviewed along with the molecular docking approaches in Chapter 2.

Another important characteristic of any molecular docking system is that it provide a means of rapidly determining chemically plausible structures for the protein-ligand complex. This will assist the medicinal chemist to identify unforeseen binding modes which can provide alternative design routes. The DOCK system generates a list of possible binding orientations of a compound. There are cases where subsequent crystallography has shown that DOCK incorrectly predicted the ligand's actual binding mode [29, 30]. One reason for this is that it is unlikely during a database search to find a compound with a very high binding affinity. DOCK generally finds compounds with micromolar affinities, and there may be many orientations of a molecule that would yield a similar affinity and be indistinguishable in terms of the scoring function. A more important reason, however, is that DOCK is based on Fischer's "lock and key" concept. Because ligands and receptors are flexible, changes in their conformations should also be included in the docking simulations. These "induced fit" effects seriously undermine the "lock and key" model [31, 32].

Typical drug molecules are small organic molecules that may contain several rotatable bonds. Recently, there have been many studies comparing the changes in conformation [33, 34, 35, 36] and conformational energy [37, 38, 36] of a ligand in its free and bound states. The evidence supports the "induced fit" model of ligand-receptor interactions; ligand and receptors do change their conformations when they form a complex. One might propose that a possible solution to flexible ligand docking might be to include several low energy conformations of a compound in the docking simulation. The obvious drawback to this approach is determining how many and which conformations of the molecule to use. If a favorable binding geometry is not included, then a potentially potent inhibitor or binding mode may be missed. For example, methotrexate, a potent inhibitor of dihydrofolate reductase and an anti-tumor drug, has two different conformations in the Cambridge Structure Database (CSD) [39], but both are very different from the complexed conformation [40]. Furthermore, Nicklaus, *et al.* [36] suggest that ligands can "use up" a substantial part of the energy gained by forming hydrogen bonds with the receptor in order to deform its conformation. Inhibitors do not always bind in their minimum energy conformations [41]. Therefore, including molecular flexibility is an important component to a molecular docking system.

## 1.2 Outline of Dissertation

The first two stages in the development of a new drug are the discovery of a lead compound and its optimization. New lead compounds can be discovered using a variety of techniques including *de novo* design [42, 43, 44, 45], computational screening of databases of compounds [1], and high volume screening. With the development of combinatorial chemistry techniques and high throughput screening with robotics, the latter approach has become the most popular approach in the pharmaceutical industry. Current methods for computational screening are limited to rigid ligands and receptors. One of the goals of this dissertation is to develop techniques for computationally screening databases of flexible ligands.

Once a lead compound has been discovered the next step is to chemically modify the ligand so as to optimize its binding affinity to its target enzyme. In order to address this step of drug design, I have developed a molecular graphical ligand design and optimization system which also includes methods for flexible ligand docking. Interactive molecular graphics are invaluable for drug design [46, 47, 48]. They enable the chemist to visualize various properties of the receptor — the van der Waals surface [49], solvent-accessible surface [50], molecular surface [50, 51], and different molecular interactions. With recent improvements in computer graphics hardware and software, it is now possible to convey much more information. Because most molecular docking systems are based on force fields or receptor surface properties (see Chapter 2), molecular graphics can illustrate regions on the ligand where new functional groups can be added to enhance the binding affinity. The system designed in this dissertation has also been invaluable in analyzing the results of molecular docking simulations.

Once a ligand has been optimized, it is important that the different binding modes of the molecule be enumerated. A common mistake is to assume that closely related compounds will bind nearly identically. Structural data have shown that even small changes in a ligand can completely alter its binding geometry [38]. Therefore, a flexible ligand docking system is important for enumerating other chemically plausible structures. These structures illustrate the states that a ligand can occupy and may lead to new directions in developing a new therapeutic agent. One of the most important limitations of most molecular docking systems is that ligand flexibility up until recently has been excluded. Much of this disseration describes the development of a flexible ligand docking system. As mentioned above, there are two major components to a molecular docking system, predicting the structure for the intermolecular complex and estimating its binding affinity. My approach has been to use genetic algorithms [52], an optimization technique that combines a nondeterministic search procedure with a "survival of the fittest" strategy. In my first system, I showed that genetic algorithms could be used for both rigid and flexible ligand docking [23] (see Chapter 4). There were a number of problems with the techinique. These are reviewed along with other docking approaches in the next chapter. Several of the limitations of my previous system and other docking methods are addressed in my design of a new flexible ligand docking system. The remainder of the dissertation is organized as follows:

- Chapter 2 contains a description of the molecular docking problem with its mathematical complexities and a survey of the approaches that have been used to try to solve it. Particular emphasis is placed on the limitations of other receptor-ligand docking systems and how my research has addressed these issues.

- Chapter 3 contains a description of the molecular graphics portion of this project. In this chapter I illustrate how new state of the art molecular graphics can be used to allow the chemist to visualize much more information about the receptor and how this can be used in lead optimization.

- Chapter 4 contains a discussion of my flexible ligand docking system which is based on a molecular mechanics force field. These results have been published [23].

- Chapter 5 contains a description of a new flexible ligand docking system that uses genetic algorithms with a binding free energy estimate as the scoring function. I discuss a number of limitations of other methods, and I show how my system addresses them.

- Chapter 6 contains the conclusions of my research along with some areas for future work.

# Bibliography

[1] I. D. Kuntz. Structure-based strategies for drug design and discovery. *Science*, 257(5073):1078–82, 1992.

[2] C. Hansch. A quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.*, 2:232, 1969.

[3] R. D. Cramer, D. E. Patterson, and J. D. Bunce. Recent advances in comparative molecular field analysis (CoMFA). *Prog. Clin. Biol. Res.*, 291(5073):161–5, 1989.

[4] G. M. Crippen. Quantitative structure-activity relationships by distance geometry: systematic analysis of dihydrofolate reductase inhibitors. *J. Med. Chem.*, 23(6):599–606, 1980.

[5] J. W. Black, W. A. Duncan, C. J. Durant, C. R. Ganellin, and E. M. Parsons. Definition and antagonism of histamine H 2-receptors. *Nature*, 236(5347):385–90, 1972.

[6] M. A. Ondetti, B. Rubin, and D. W. Cushman. Design of specific inhibitors of angiotensin-converting enzyme: new class of orally active antihypertensive agents. *Science*, 196(4288):441–4, 1977.

[7] S. H. Reich, M. A. Fuhry, D. Nguyen, M. J. Pino, K. M. Welsh, S. Webber, C. A. Janson, S. R. Jordan, D. A. Matthews, W. W. Smith, C. A Bartlett, C. L. J. Booth, S. M. Herrmann, E. F. Howland, C. A. Morse, R. W. Ward, and J. White. Design and synthesis of novel 6,7-imidazotetrahydroquinoline in-

hibitors of thymidylate synthase using iterative protein crystal structure analysis. *J. Med. Chem.*, 35(5):847–58, 1992.

[8] J. A. Montgomery, S. Niwas, J. D. Rose, J. A. Secrist, Y. S. Babu, C. E. Bugg, M. D. Erion, W. C. Guida, and S. E. Ealick. Structure-based design of inhibitors of purine nucleoside phosphorylase. 1. 9-(arylmethyl) derivatives of 9-deazaguanine. *J. Med. Chem.*, 36(1):55–69, 1993.

[9] A. K. Ghosh, W. J. Thompson, H. Y. Lee, S. P. McKee, P. M. Munson, T. T. Duong, P. L. Darke, J. A. Zugay, E. A. Emini, W. A. Schleif, J. R. Huff, and P. S. Anderson. Cyclic sulfolanes as novel and high affinity P2 ligands for HIV-1 protease inhibitors. *J. Med. Chem.*, 36(7):924–7, 1993.

[10] 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, C-H. Chang, P. C. Weber, D. A. Jackson, T. R. Sharpe, and S. Erickson-Viitanen. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science*, 263(5145):380–4, 1994.

[11] S. S. Cohen. A strategy for the chemotherapy of infectious disease. *Science*, 197(4302):431–2, 1977.

[12] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer, 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(3):535–42, 1977.

[13] C. S. Ring, E. Sun, J. H. McKerrow, G. K. Lee, P. J. Rosenthal, I. D. Kuntz, and F. E. Cohen. Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. U. S. A.*, 90(8):3583–7, 1993.

[14] Y. C. Martin. 3d database searching in drug design. *J. Med. Chem.*, 35(12):2145–54, 1992.

[15] E. Fischer and H. Thierfelden. Verhalten der verschiedenen zucher gegen reine hefen. *Chemische Berichte*, 27:2031–2037, 1894.

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

[17] 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):269–88, 1982.

[18] 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):722–9, 1988.

[19] B. K. Shoichet, D. L. Bodian, and I. D. Kuntz. Molecular docking using shape descriptors. *J. Comp. Chem.*, 13(3):380–397, 1992.

[20] E. C. Meng, B. K. Shoichet, and I. D. Kuntz. Automated docking with grid-based energy evaluation. *J. Comp. Chem.*, 13(4):505–524, 1992.

[21] B. K. Shoichet and I. D. Kuntz. Protein docking and complementarity. *J. Mol. Biol.*, 221(1):327–46, 1991.

[22] J. Cherfils, S. Duquerroy, and J. Janin. Protein-protein recognition analyzed by docking simulation. *Proteins*, 11(4):271–80, 1991.

[23] K. P. Clark and Ajay. Flexible ligand docking without parameter adjustment across four ligand-receptor complexes. *J. Comp. Chem.*, 16(10):1210–1226, 1995.

[24] K. M. Merz and P. A. Kollman. Free energy perturbation simulations of the inhibition of thermolysin- prediction of the free energy of binding of a new inhibitor. *J. Am. Chem. Soc.*, 111(15):5649–5658, 1989.

[25] D. L. Beveridge and F. M. Dicapua. Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem.*, 18(4):431–92, 1989.

[26] W. F. van Gunsteren and H. J. C. Berendsen. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. *Angew. Chem. Int. Ed. Engl.*, 29:992–1023, 1990.

[27] T. P. Straatsma and J. A. McCammon. Computational alchemy. *Ann. Rev. Phys. Chem.*, 43(V43):407–435, 1992.

[28] P. R. Gerber, A. E. Mark, and W. F. Van Gunsteren. An approximate but efficient method to calculate free energy trends by computer simulation: application to dihydrofolate reductase-inhibitor complexes. *J. Comput. Aided. Mol. Des.*, 7(3):305–23, 1993.

[29] B. K. Shoichet, R. M. Stroud, D. V. Santi, I. D. Kuntz, and K. M. Perry. Structure-based discovery of inhibitors of thymidylate synthase. *Science*, 259(5100):1445–50, 1993.

[30] E. Rutenber, E. B. Fauman, R. J. Keenan, S. Fong, P. S. Furth, D. E. Ortiz, E. Meng, I. D. Kuntz, D. L. Decamp, R. Salto, J. R. Rosé, C. S. Craik, and R. M. Stroud. Structure of a non-peptide inhibitor complexed with HIV-1 protease. developing a cycle of structure-based drug design. *J. Biol. Chem.*, 268(21):15343–6, 1993.

[31] D. E. Koshland. Molecular basis of enzyme catalysis and control. *Pure. Appl. Chem.*, 25(1):119–33, 1971.

[32] W. L. Jorgensen. Rusting of the lock and key model for protein-ligand binding. *Science*, 254(5034):954–5, 1991.

[33] E. M. Ricketts, J. Bradshaw, M. Hann, F. Hayes, N. Tanna, and D. M. Ricketts. Comparison of conformations of small molecule structures from the protein data

bank with those generated by CONCORD, COBRA, CHEMDBS-3D, and converter and those extracted from the cambridge structural database. *J. Chem. Inf. Comput. Sci.*, 33(6):905–925, 1993.

[34] K. G. Rice, P. G. Wu, L. Brand, and Y. C. Lee. Experimental determination of oligosaccharide 3-dimensional structure. *Curr. Opin. Struct. Biol.*, 3(5):669–674, 1993.

[35] S. L. Moodie and J. M. Thornton. A study into the effects of protein binding on nucleotide conformation. *Nucleic Acid Research*, 21(6):1369–1380, 1993.

[36] M. C. Nicklaus, S. Wang, J. S. Driscoll, and G. W. Milne. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.*, 3(4):411–428, 1995.

[37] P. R. Andrews, D. J. Craik, and J. L. Martin. Functional group contributions to drug-receptor interactions. *J. Med. Chem.*, 27(12):1648–57, 1984.

[38] D. Ringe. Binding by design. *Nature*, 351:185–186, 1995.

[39] F. H. Allen, S. Bellard, M. D. Brice, B. A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B. G. Hummelink-Peters, O. Kennard, W. D. S. Motherwell, J. R. Rodgers, and D. G. Watson. The Cambridge Crystal Data Centre: Computer-based search, retrieval, analysis, and display of information. *Acta. Crystallogr.*, B35:2331–2339, 1979.

[40] J. T. Bolin, D. J. Filman, D. A. Matthews, R. C. Hamlin, and J. Kraut. Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 åresolution. I. general features and binding of methotrexate. *J. Biol. Chem.*, 257(22):13650–13662, 1982.

[41] D. Ringe, B. A. Seaton, M. H. Gelb, and R. H. Abeles. Inactivation of chymotrypsin by 5-benzyl-6-chloro-2-pyrone: 13C NMR and X-ray diffraction analyses of the inactivator-enzyme complex. *Biochemistry*, 24(1):64–8, 1985.

[42] J. B. Moon and W. J. Howe. Computer design of bioactive molecules: a method for receptor-based de novo ligand design. *Proteins: Struct. Funct. Genet.*, 11:314–328, 1991.

[43] Y. Nishibata and A. Itai. Automatic creation of drug candidate structures based on receptor structure starting point for artificial lead generation. *Tetrahedron*, pages 8985–8990, 91.

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

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

[46] R. Langridge, T. E. Ferrin, I. D. Kuntz, and M. L. Connolly. Real-time color graphics in studies of molecular interactions. *Science*, 211(4483):661–6, Feb 1981.

[47] N. C. Cohen, J. M. Blaney, C. Humblet, P. Gund, and D. C. Barry. Molecular modeling software and methods for medicinal chemistry. *J. Med. Chem.*, 33(3):883–94, Mar 1990.

[48] R. S. Bohacek and C. McMartin. Definition and display of steric, hydrophobic, and hydrogen-bonding properties of ligand binding sites in proteins using Lee and Richards accessible surface: validation of a high-resolution graphical tool for drug design. *J. Med. Chem.*, 35(10):1671–84, May 1992.

[49] P. A. Bash, N. Pattabiraman, C. Huang, T. E. Ferrin, and R. Langridge. Van der Waals surfaces in molecular modeling: Implementation with real-time computer graphics. *Science*, 222:1325–1327, 1983.

[50] F. M. Richards. Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.*, 6(4612):151–76, 1977.

[51] M. L. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–13, Aug 1983.

[52] D. E. Goldberg. *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley Pub. Co., Reading, Mass., 1989.

# Chapter 2

# Molecular Docking: A Review

Molecular docking is the prediction of the structure and binding affinity of a protein-ligand complex given only the structures of the free ligand and receptor. This is a complex problem that can be divided into two separate, but related tasks. The first is a method to "dock" ligands to proteins. This involves searching the conformational and configurational space of the ligand within the receptor for chemically plausible binding geometries. The second is a scoring function that can at least predict which binding modes are favorable and which are not. Ideally, one would like the scoring function to rank the set of compounds by their binding affinities. Both of these components are essential for a successful docking system. The docking algorithm must sample the conformational and orientational space sufficiently so that the favorable binding modes can be found, and the scoring function must be able to recognize these structures when they are found.

In the case of rigid molecular docking where both the ligand and the receptor are rigid and assumed to be in their crystal structure conformations, a direct search of the six-dimensional space is not practical. Kuhl *et al.* [1], for the docking problem defined as finding the orientation of the ligand that maximizes the number of pairwise contacts made between the ligand and the receptor, showed that the brute force method of finding all maximal matchings runs in time $O(m^3n^3\min(m,n))$ where $m$ and $n$ are number of atoms in the ligand and the receptor, respectively. Kuhl *et al.* propose an algorithm that is based on a combinatorial graph of pairwise distances.

For this problem, finding a maximal matching reduces to finding maximal cliques in a graph, a problem known to belong to the class of "nondeterministic polynomial time NP-complete" computational problems. Thus, any progress will depend on the quality of the approximations used to search the conformational space of the ligand and the scoring function used to evaluate the intermolecular complexes. In the remainder of this chapter, I review some of the approaches to solve the molecular docking problem, and I illustrate the relevance of my dissertation research in this field. For convenience, the different docking approaches have been divided into several categories. Furthermore, I focus only on the techniques that are applicable to docking small molecules. I do not review the protein-protein docking literature.

## 2.1 Energy Minimization Methods

Energy minimization techniques use various methods to minimize the intermolecular conformation using a molecular mechanics force field. These methods include ligand flexibility. Goodsell and Olsen [2] use simulated annealing to find the low energy conformation of ligands within the receptor site. Caflisch *et al.* [3] used a Monte Carlo method to dock a peptide inhibitor of HIV-1 protease. Yamada and Itai [4, 5] developed ADAM, another energy minimization technique. Their system includes flexibility in the ligand through a systematic search of the torsion space. An initially docked structure is minimized with AMBER [6]. DiNola *et al.* [7] describe a molecular dynamics approach to the docking of phosphocholine into McPC603. These techniques consider flexibility in the receptor as well as the ligand. They can also account for desolvation effects with discrete solvent molecules or with continuum solvation models [8, 9, 10]. Because of the complex topology of the energy landscape, these techniques often get trapped in local minima, have long run times, and are computationally too expensive.

## 2.2 Descriptor Methods

Descriptor methods characterize the receptor in terms of various physico-chemical descriptors and search for ligand orientations and conformations that maximize the complementarity. The DOCK suite of programs [11] was one of the first methods available for molecular docking. Although both the ligand and the receptor are considered to be rigid, DOCK is still used to screen databases of small molecules. The method consists of several steps. First, a negative image of the receptor site is created using spheres. A set of ligand atoms, typically four or five, are matched to the sphere centers to generate various orientations of the ligand. Finally, the ligand orientations are scored. Recently, Meng *et al.* [12] used the grid-based approximation to the molecular mechanics force field first proposed by Pattabiraman *et al.* [13] to score docked structures. CLIX [14] is another rigid docking algorithm. It uses Goodford's GRID [15] program which characterizes the binding site using twenty-three different types of probe atoms. The ligand orientation is optimized with molecular mechanics. Bacon and Moult [16] describe a rigid docking algorithm that is based on a web representation of the solvent accessible surface [17]. They apply a least squares fitting procedure to map the ligand onto the surface. The orientation is scored based on electrostatics and steric interactions. Another class of approaches use graph theoretical methods. Kasinos *et al.* [18] represent the problem of finding the largest number of ligand/receptor matches as the problem of finding the maximal common subgraph. The authors used potential hydrogen bonding sites as match points for the algorithm. Smellie *et al.* [19] applied a similar approach to dock flexible ligands to rigid receptors. The major drawback of their algorithm is that the scoring function was too simple, and many of the generated binding modes were sterically impossible. FLOG [20] uses an approach similar to the approach in DOCK, but the conformational space of the ligand is represented by a number of low energy ligand structures. The methods described in this section all try to summarize the important features of the receptor site and will most likely be of importance in any successful flexible ligand docking system. The descriptors of the receptor provide targets which can guide the placement of complementary groups of the ligand and thereby limit the search of orientational and conformational space.

## 2.3 Fragment-based Methods

Fragment-based methods dock a ligand by docking the individual fragments of a ligand into the receptor. There are two approaches taken with the fragment-based approaches. First, the individual fragments of the ligand can be docked separately and then joined to form the complete ligand. One of the most challenging aspects of this technique is to connect the fragments in a synthetically accessible way. The other approach is the "grow" the ligand from an initially docked fragment. The fragment joining approach forms the basis of many *de novo* design systems. Some examples of these include GROW [21], LEGEND [22], LUDI [23, 24], GroupBuild [25], and GenStar [26]. A number of flexible ligand docking systems are also based on this idea. DesJarlais *et al.* [27] used DOCK to place an individual fragment of the ligand, and the other fragments were subsequently added. This methods includes only partial ligand flexibility. Hart and Read's [28] Multi-start Monte Carlo system used Monte Carlo techniques to dock the individual fragments. Leach and Kuntz [29] describe an incremental fragment construction algorithm. In this case a variant of DOCK is used to place an "anchor" fragment of a ligand into the receptor, and a systematic search algorithm is used to place the remainder of the ligand. Rarey *et al.* [30] recently described their incremental ligand construction algorithm called FLEXX. In their algorithm a base fragment is chosen and docked using pose clustering, a pattern recognition technique [31]. The ligand is constructed using a ligand conformation generation program MIMUMBA [32], and the intermolecular conformation is scored using an estimate of the binding free energy which is based on the work of Böhm [33]. Welch *et al.* [34] developed Hammerhead, another fragment-based approach ligand docking system. The method starts with a fragment docked in placed, and the complex is scored using a variation of the same free energy estimation [33] procedure. Their implementation of the scoring function, however, is differentiable and the ligand's orientation can be minimized. The limitations of these techniques are covered in the last section of the chapter.

## 2.4 Genetic Algorithm-based Systems

Perhaps the newest approach to the problem of flexible ligand docking is the use of a genetic algorithm (GA) [35] to search the conformational and orientational space of the ligand. Genetic algorithms are non-deterministic optimization procedures that are based on a "survival of the fittest" strategy. In 1994 Judson *et al.* [36] used genetic algorithms to dock Cbz-GlyP-Leu-Leu (ZGLL) to thermolysin. This is the first published account of the use of genetic algorithms for flexible ligand docking. The authors used a molecular mechanics scoring function and the ligand was "grown" from an initially docked fragment. Oshiro *et al.* [37] describe a flexible ligand docking extension to DOCK. The genetic algorithm encodes a mapping between the ligand atoms and the sphere centers of the negative image of the receptor in addition to the torsion angles of the ligand. The receptor-ligand conformations are scored using a molecular mechanics force field. In my first flexible ligand docking system [38], I used a genetic algorithm search procedure with a grid-based molecular mechanics force field [12, 13] (see Chapter 4). The purpose of the work was to study the applicability of genetic algorithms to rigid and flexible ligand docking. Two different approaches were taken with flexible ligand docking. The first was the combination of a systematic search within the context of a genetic algorithm. The second was a fragment-based approach where an initial fragment of the ligand was docked approximately in place. The primary focus of this work was to determine how much information about the binding mode might be necessary in order to find the crystal structure binding mode consistently (at least 50% of the time). Jones *et al.* [39] describe a flexible ligand docking system which encodes putative mappings between hydrogen bond donors and acceptors as well as the torsion angles of the ligand. The scoring function is a weighted sum of the van der Waals energy and the hydrogen bond energy. The authors point out that the method is suitable for incorporating receptor flexibility into the receptor side chains. Leach [40], however, has described the most detailed treatment of receptor flexibility in molecular docking simulations. His approach uses the dead-end elimination algorithm [41] to enumerate the receptor conformations and the A* algorithm, a well-known artificial intelligence algorithm for finding the optimal, least cost solution to a search problem. All of the genetic algorithm methods described above have similar performance in that they all converge to local minima and often take several runs to find the solutions.

## 2.5 Comparisons and Future Directions

The two most promising approaches to flexible ligand docking to date are the fragment-based and the genetic algorithm-based approaches. Rarey *et al.* [30] have described the most comprehensive system FLEXX to date for the fragment-based approach. There are a number of limitations to their approach. First, the base fragment for the docking procedure was chosen interactively. In fact, the choice of the base fragment plays an important role in determining whether the system will find the solution or not. If a fragment that has no clearly predominant directionality with the receptor is selected, then their procedure will most likely not find the correct binding mode. Second, the actual placement of the base fragment is context dependent. Ligands are known to deform when they bind to receptors [42]. There is no guarantee that the orientation or conformation of the base fragment will be the same in the ligand as it is when the base fragment is docked independently. Rotstein and Murcko [25] have shown that it is difficult to dock ligands if the fragment is not oriented properly. Uncertainties in the placement of the base fragment can be particularly important if one is trying to design a drug from a model structure of the target enzyme or if receptor flexibility is included. Third, there are many potential sites in which the base fragment can be docked. Rarey *et al.* found that orientation of the base fragment in the crystal structure does not necessarily correspond to the best position of the base fragment docked independently. This brings up a very interesting question. How many different orientations of the base fragment must be considered to find micromolar or better inhibitors? Finally, the authors use a "greedy" construction algorithm to "grow" the ligand. This approach begins with an initially docked fragment and adds subsequent fragments in the orientation that maximizes the fragment's contribution to the score. It is not necessarily the case that the lowest energy conformation of the entire ligand is the one where the fragments are added in their lowest energy orientation. This approach is most suited for docking highly optimized ligands where each functional group is designed to maximize the complementarity with the receptor. The purpose of this dissertation is to develop a molecular docking system for ligand discovery and optimization so one most likely will not be docking a highly optimized ligand to the receptor. The goal is to elucidate all the plausible binding modes of a ligand to discover different strategies for the design of new therapeutic agents that optimize the complementarity between the ligand and its target. Furthermore, the performance of the "greedy construction" approach will be affected the other uncertainties in the base fragment that were mentioned above.

Methods that rely on genetic algorithms to search the conformation space of the ligand have not been studied as well. The approaches described in the previous section all had similar results. The genetic algorithms converged prematurely to local minima. Many runs were needed to find the crystal structure binding mode, and alternate solutions with scores similar to the crystal structures were found. Some classes of problems are known as "GA deceptive" [35]. These problems are difficult to solve using genetic algorithms because good partial solutions lead to poor global solutions. In formulating a problem within the context of a genetic algorithm solution, one must carefully choose the scoring function and the representation of the optimization parameters. For molecular docking, the representation determines how the parameter space is searched. It helps direct the search to areas where good global solutions are likely to be found. For example, a representation that uses a mapping of putative hydrogen bonding sites guides the search in the regions where these contacts can be made. The scoring function, on the other hand, is linked directly to the dynamics of the genetic algorithm through the selection operator. It determines which partial solutions are favorable and are thereby explored further. If a conformation does not score well even though it is close to the crystal structure, then it will most likely not survive to the next iteration of the genetic algorithm. The genetic algorithm-based molecular docking systems described in the previous section use different representations but very similar scoring functions, an electrostatic or hydrogen bonding term and a van der Waals score. In Chapter 5, I describe a molecular docking system which uses genetic algorithms to search the ligand's conformational space. In this system, I use a scoring function similar to the Böhm free energy estimate [33]. Furthermore, I use the system to address some of the limitations of the fragment-based docking procedures. Finally, in the last chapter, I discuss some alternate representations that address the other limitations of the fragment-based approaches and describe an approach that may be more suitable for molecular docking.

# Bibliography

[1] F. S. Kuhl, G. M. Crippen, and D. K. Friesen. A combinatorial algorithm for calculating ligand binding. *J. Comp. Chem.*, 5(1):24–34, 1984.

[2] D. S. Goodsell and A. J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins*, 8(3):195–202, 1990.

[3] A. Caflisch, P. Niederer, and M. Anliker. Monte Carlo docking of oligopeptides to proteins. *Proteins*, 13(3):223–230, 1992.

[4] M Yamada and A. Itai. Development of an efficient automated docking method. *Chem. & Pharm. Bul.*, 41:1200–1202, 1993.

[5] M Yamada and A. Itai. Application and evaluation of the automated docking method. *Chem. & Pharm. Bul.*, 41:1203–1205, 1993.

[6] S.J. Weiner, P. A. Kollman, D.A. Case, U. C. Singh, C Ghio, G. Alagona, S. Profeta Jr., and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765, 1984.

[7] A. DiNola, D. Roccatano, and H. J. C. Berendsen. Molecular dynamics simulation of the docking of substrates to proteins. *Proteins*, 19(3):174–182, 1994.

[8] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. S.*, 112(16):6127–6129, 1990.

[9] L. Wesson and D. Eisenberg. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein. Sci.*, 1(2):227–235, 1992.

[10] C. A. Schiffer, J. W. Caldwell, R. M. Stroud, and P. A. Kollman. Inclusion of solvation free energy with molecular mechanics energy: alanyl dipeptide as a test case. *Protein. Sci.*, 1(3):396–400, 1992.

[11] 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):269–88, 1982.

[12] E. C. Meng, B. K. Shoichet, and I. D. Kuntz. Automated docking with grid-based energy evaluation. *J. Comp. Chem.*, 13(4):505–524, 1992.

[13] N. Pattabiraman, M. Levitt, T. E. Ferrin, and R. Langridge. Computer graphics in real-time docking with energy calculation and minimization. *J. Comp. Chem.*, 6(5):432–436, 1985.

[14] M. C. Lawrence and P. C. Davis. CLIX: a search algorithm for finding novel ligands capable of binding proteins of known three-dimensional structure. *Proteins*, 12(1):31–41, 1992.

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

[16] D. J. Bacon and J. Moult. Docking by least-squares fitting of molecular surface patterns. *J. Mol. Biol.*, 225(3):849–858, 1992.

[17] M. L. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–13, Aug 1983.

[18] N. Kasinos, G. A. Lilley, N. Subbarao, and I. Haneef. A robust and efficient automated docking algorithm for molecular recognition. *Protein. Eng.*, 5(1):69–75, 1992.

[19] A. S. Smellie, G. M. Crippen, and W. G. Richards. Fast drug-receptor mapping by site-directed distances: a novel method of predicting new pharmacological leads. *J. Chem. Inf. Comput. Sci.*, 31(3):386–392, 1991.

[20] 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.

[21] J. B. Moon and W. J. Howe. Computer design of bioactive molecules: a method for receptor-based de novo ligand design. *Proteins: Struct. Funct. Genet.*, 11:314–328, 1991.

[22] Y. Nishibata and A. Itai. Automatic creation of drug candidate structures based on receptor structure starting point for artificial lead generation. *Tetrahedron*, pages 8985–8990, 91.

[23] 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, 1992.

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

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

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

[27] R. L. Desjarlais, R. P. Sheridan, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.*, 29(11):2149–53, 1986.

[28] T. N. Hart and R. J. Read. A multiple-start Monte Carlo docking method. *Proteins*, 13(3):206–222, 1992.

[29] A. R. Leach and I. D. Kuntz. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comp. Ch.*, 13(6):730–748, 1992.

[30] 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, 1996.

[31] 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, 1996.

[32] 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.

[33] 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.

[34] W. Welch, J. Ruppert, and A. N. Jain. Hammerhead - fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.*, 3(6):449–462, 1996.

[35] D. E. Goldberg. *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley Pub. Co., Reading, Mass., 1989.

[36] R. S. Judson, E. P. Jaeger, and A. M. Treasurywala. A genetic algorithm based method for docking flexible molecules. *J. Mol. Struct.*, 308:191–206, 1994.

[37] 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.

[38] K. P. Clark and Ajay. Flexible ligand docking without parameter adjustment across four ligand-receptor complexes. *J. Comp. Chem.*, 16(10):1210–1226, 1995.

[39] 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(1):43–53, 1995.

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

[41] J. Desmet, M. DeMaeyer, M. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature (London)*, 356:539–542, 1992.

[42] M. C. Nicklaus, S. Wang, J. S. Driscoll, and G. W. Milne. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.*, 3(4):411–428, 1995.

# Chapter 3

# A Ligand Discovery and Optimization System

Most of the drugs in clinical use today were developed through high volume random screening of natural products or corporate databases of compounds. This is just the first stage in the lengthy process of drug development. According to Table 3, the ligand discovery and optimization process can take from two to four years. The purpose of this drug design system is to help further our understanding of protein-ligand interactions, to assist structure-based drug design, and to decrease the development time during the first two steps in the drug design process.

## 3.1 Drug Discovery and Optimization

As mentioned above, high volume screening is one of the most useful techniques for ligand discovery. Other techniques include *de novo* drug design or computational screening [1]. The latter, however, is currently practical only if the ligand and receptor are rigid. Recently, with the development of combinatorial chemistry libraries and high throughput screening techniques that use robotics to carry out standardized biochemical assays, the high volume screening approach is being used more and more in the pharmaceutical industry. This technology enables more rapid synthesis and testing of compounds and identification of potential lead compounds. The challenge

| Drug development step | Years |
|---|---|
| Discovery and lead generation | 1 – 2 |
| Lead optimization | 1 – 2 |
| In vitro and in vivo assays | 1 – 2 |
| Toxicology trials | 1 – 3 |
| Human safety trials | 1 |
| Human efficacy trials | 1 – 2 |
| Total development time | 6 – 12 |

Table 3.1: This table, which appears in Kuntz[1], describes the steps required to get a drug to the market. This chapter of my dissertation focuses on the first two steps in the table.

once a lead compound is discovered is to determine how the compound binds and to suggest chemical modifications that will increase its binding affinity.

The underlying assumption of structure-based drug design is that ligands and receptors must have significant structural and chemical complementarity. In optimizing a lead compound, a chemist exploits information about the receptor to suggest chemical modifications that will increase the binding affinity to the target enzyme. Even when the structure of the receptor is known, this can be a difficult task. Interactions between the ligand and receptor involve many non-bonded contacts. In many cases the true binding mode may not be known [2, 3], and there may be several chemically plausible structures. Each of these may provide a starting point for lead optimization. Furthermore, small changes in the ligand can have a profound effect on the binding mode [4]. Therefore, two important components of any ligand discovery and optimization system are a method for molecular docking and a molecular graphical tool which highlights the potential interaction sites on the receptor. The remainder of this chapter describes the details of the molecular graphics system, and the flexible ligand docking methods are discussed in the next two chapters.

## 3.2 A Ligand Discovery and Optimization System

Molecular graphics are an invaluable tool for providing structural insights for directing the design of new therapeutic agents [5]. By studying how different inhibitors can be accommodated within the same active site of a receptor, we can develop a better understanding of the protein-ligand interactions that are applicable to structure-based drug design. Bohacek and McMartin [6] describe a molecular graphical tool for drug design. They investigated several methods for displaying key interaction sites on the receptor and illustrated how molecular surfaces are important for rendering this information. Although their system used parallel contour lines to represent the molecular surface, they demonstrated the value of displaying it interactively. Molecular surfaces illustrate the overall topology of the binding site and provide detailed information about the which pockets are unoccupied. Functional groups can be proposed to extend the ligand to that it can interact with more sites on the receptor. Bohacek and McMartin quantified the relationship between surface complementarity and binding affinity. Their approach involves dividing the molecular surface into three different regions — hydrophobic, hydrogen bond donor, and hydrogen bond acceptor. For a set of nine thermolysin inhibitors, they found a good correlation ($r^2 = 0.99$) between $\log 1/K_i$ and the number of nonpolar carbon and complementary hydrogen bond contacts. In the five years since Bohacek and McMartin developed their system, there have been many advances in computer graphics hardware and software. It is now possible to display much more information to the chemist.

This drug design system has been developed as an extension to Chimera [7], a new molecular modeling system that is being developed in the Computer Graphics Laboratory at UCSF to eventually replace MidasPlus [8]. The key features of the system include:

- Extensible and user customizable
- Comprehensive graphical primitives
- Texture Mapping

Extensibility is one of the primary features of the new system. Chimera, in fact, will contain only the core functionality of the molecular modeling system — the graphical display, the user interface, and several extension mechanisms. All other functionality and applications will be implemented as extensions to this core functionality. One of the extension mechanisms is the use of Python [9], a complete programming language with control flow constructs, as the Chimera command language. Through Python application programmers will have access to the data in Chimera. Chimera's user interface is also customizable, and extensions can be readily integrated into menus or placed on Chimera's tool bar. More elaborate extensions to the user interface can be made using X-windows *Motif*, *Tk*, or *HTML* libraries of Python. Thus, entire systems such as a molecular dynamics or drug design systems can be developed as extensions to Chimera's core functionality.

Another key feature of Chimera is that the graphics are based on OpenInventor [10], an object oriented toolkit of objects and methods for creating interactive 3D graphics applications. OpenInventor greatly extends the graphics capability of the system beyond that of MidasPlus. Chimera can interactively display wire-frame, ball-and-stick, and ribbon style representations of molecules. Chimera can also display non-molecular graphical objects such as points, lines, spheres, and polygons. The orientation, color, and translucency of these objects can be changed interactively. OpenInventor also supports texture mapping applications. Texture mapping is not a new technique in computer graphics. It has been used for many years for creating realistic computer generated images where software-based rendering systems were fast enough. Recently, however, advanced graphics workstations with special hardware for texture mapping have appeared on the market. Three components are required. One must have a vertex-based three dimensional object, a texture or set of parameters associated with each vertex, and a mapping function which relates the texture to the display of the object. Because texture mapping modifies pixel information interactively during the rendering procedure, it provides a new framework from which to display and analyze scientific data. One important application of texture mapping is volume visualization.

My ligand discovery and optimization system is implemented as an extension to

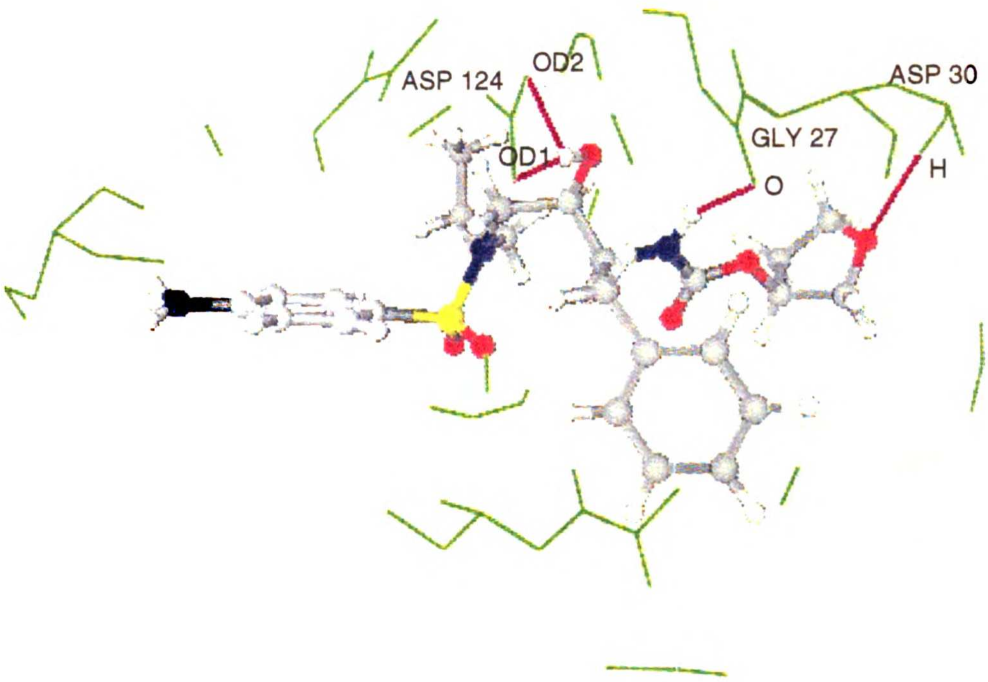


Figure 3.1: The hydrogen bonds that are formed between VX-478 and the HIV-1 protease. The ligand is shown with a ball-and-stick representation, the neighboring receptor atoms are shown in green, and the hydrogen bonds are shown in magenta.

Chimera. It consists of a set of applications written in C, C++, or Python for extending the core functionality of Chimera so that it can be used as a drug design system. Many of these applications are for visualizing various physico-chemical properties of the active site of the receptor. Other extensions provide user interfaces for computing and displaying these quantities or for browsing the results of a flexible ligand docking simulation. The applications described here have been integrated into Chimera and are available from Chimera's user interface. In the remainder of this chapter, I present examples of these features using the real world example of an HIV-1 protease inhibitor VX-478 that was designed at Vertex Pharmaceuticals. Phase I/II clinical trials for this drug began in 1995. The docking simulations of this system are considered in Chapter 5.

In designing a therapeutic agent, it is important to be able to identify interactions that are being made or can be made with the receptor. A number of applications are available in this system for this purpose. Hydrogen bonds, for example, are one of the most important interactions that can be formed between a ligand and receptor. During a molecular recognition event, hydrogen bonds with the solvent are replaced with hydrogen bonds between the ligand and receptor. In some cases, water molecules mediate hydrogen bonds between the ligand and receptor. Because these interactions are so important in molecular docking studies, I have written a C++ application to identify and display the hydrogen bonds within a complex or a given molecule. The molecules are read into the program, and sets of atoms that satisfy various hydrogen bond geometries are identified. Currently, the system searches for approximately linear hydrogen bonds as described in Böhm [11] or those that satisfy the criteria specified in LUDI [12]. Other geometries can be used simply by changing the input file. The output of the program is an OpenInventor file containing a graphical representation of the hydrogen bonds. This file is rendered in the graphical display using a core utility of Chimera for displaying OpenInventor files. The hydrogen bond geometries are also written to Chimera's reply window using a Python script so that the user can identify the atoms involved and evaluate the geometry. Figure 3.1 shows the hydrogen bonds being made between VX-478 and the HIV-1 protease. Other utilities are available for highlighting hydrogen bonding opportunities.

Bohacek and McMartin [6] illustrated the utility of a molecular surface in drug design. My ligand design system has an application for interactively computing and displaying a molecular surface. A number of programs now exist for the calculation of a triangulated molecular surface [13, 14]. I have compared these programs and found that the latter produces better molecular surfaces. Both systems generate the surface in real-time, but the Varshney *et al.* surface had problems with degenerate triangles. My system uses the MSMS [14] program to compute the surface which is represented as a list of vertices and triangles that represent the molecular surface. I have written a Python script to convert the the output of MSMS, a list of triangles, into OpenInventor format so that it can be displayed within Chimera. The surface can be colored using standard techniques or texture mapping. Furthermore, translucency can be used to improve the interpretability of the data. Figure 3.2 shows the molecular surface of the HIV-1 protease that has been colored by the protein's chain identification

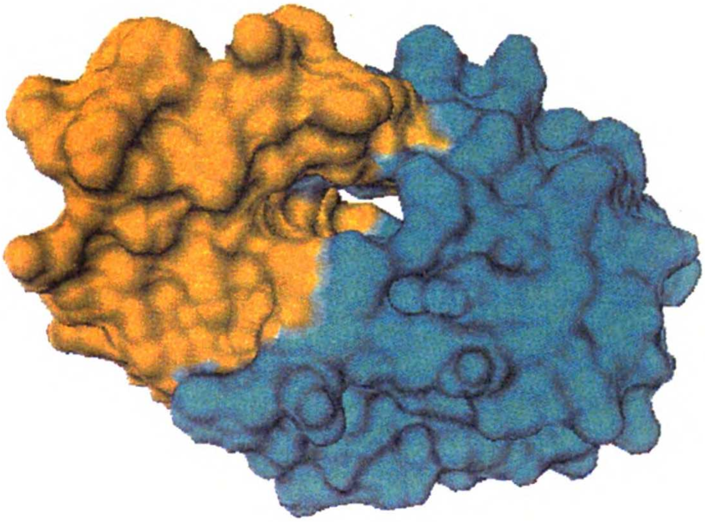


Figure 3.2: The symmetric nature of the HIV-1 protease. The two dimers are shown in gold and cyan.

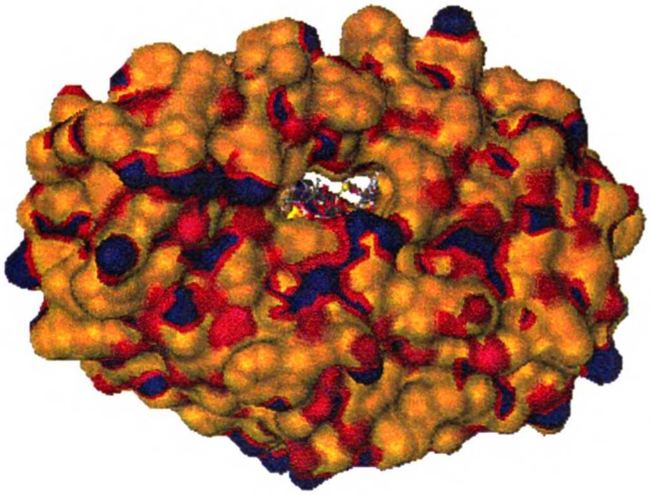


Figure 3.3: The molecular surface of the HIV-1 protease colored according the the hydrogen bond donors or acceptor sites on the receptor. Hydrogen bond acceptors and donors are shown in red and blue, respectively. The yellow regions on the receptor correspond to areas that are neither hydrogen bond donors or acceptors.

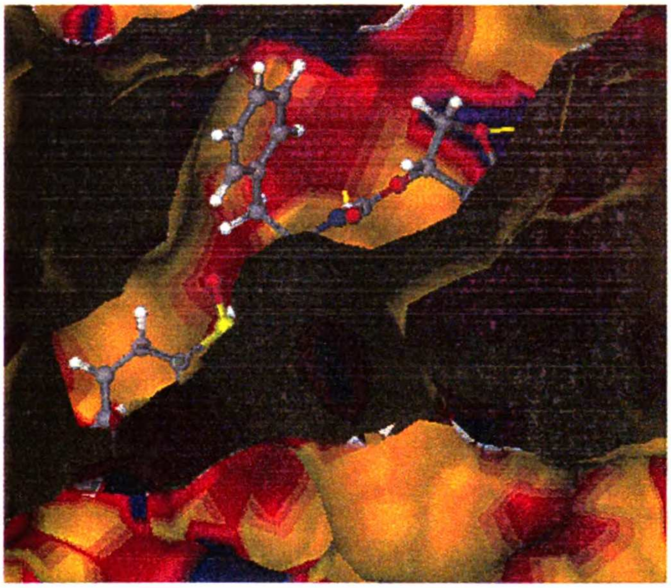


Figure 3.4: The interactions of the inhibitor with the protease as in Figure 3.3 after part of the surface has been clipped away using transparency as one of the components of the texture space.

number. The figure illustrates the symmetric nature of the protein. Figure 3.3 shows the surface of the HIV-1 protease that has been colored according to the presence of hydrogen bond donor or acceptor groups on the receptor along with a ball-and-stick representation of the ligand. By implementing the coloring using texture mapping, one can use transparency to interactively clip away part of surface and display those features of interest. This is shown in Figure 3.4. In this case, I wrote an application in C which computes the distance from a given vertex to a clipping plane. The output of this program is stored as the second texture component in the OpenInventor file. Vertices above the plane and further than a distance threshold from it are made transparent. The distance threshold can be changed using the texture editor, which was written by Conrad Huang as part of Chimera. Notice how the different functional groups interact with the different subgroups of the protease. The figure shows the hydrophobic groups are docked into hydrophobic pockets (yellow), and the hydrogen bonds are shown with yellow lines. This, however, is better illustrated interactively

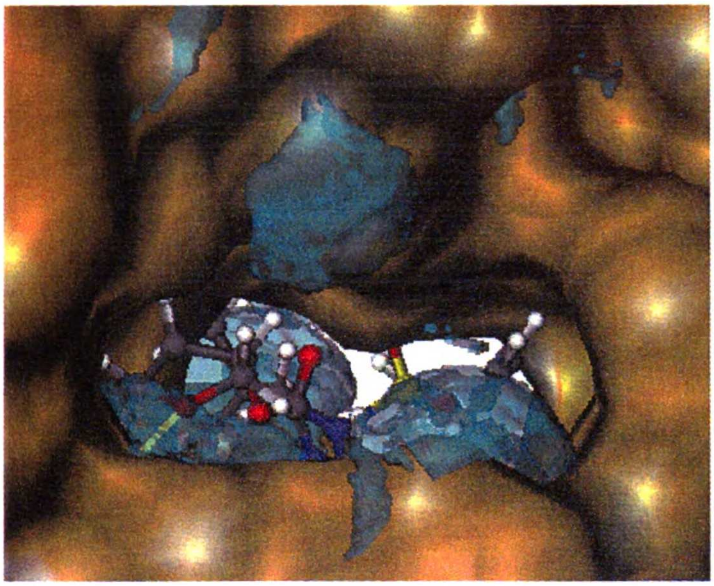


Figure 3.5: A hydrogen bond acceptor isosurface (cyan). This surface highlights where acceptor groups on the ligand can be placed to form a hydrogen bond with the receptor.

on a graphics screen. Because molecular surfaces are invaluable for visualizing a receptor's binding site and identifying regions where functional groups can be added to increase a lead compounds binding affinity, I have written a graphical user interface to the surface calculation application in *Tk* and Python. This utility facilitates the calculation of the molecular surface and allows the user to color the surface either a solid color, the color of atoms closest to the vertex, or by using texture mapping. The texture fields can be set to any program that can be executed from the Unix command shell. The vertices are read into the program from the standard input, the output is written to the appropriate texture field of the OpenInventor output file. The molecular surface is displayed using a core utility in Chimera.

The examples considered so far have displayed lines or surfaces, but molecular interactions are governed by quantities in the volume of the active site of the receptor. Line segments can illustrate hydrogen bonds that have been made between the ligand and the receptor. This is particularly useful when analyzing the results from docking

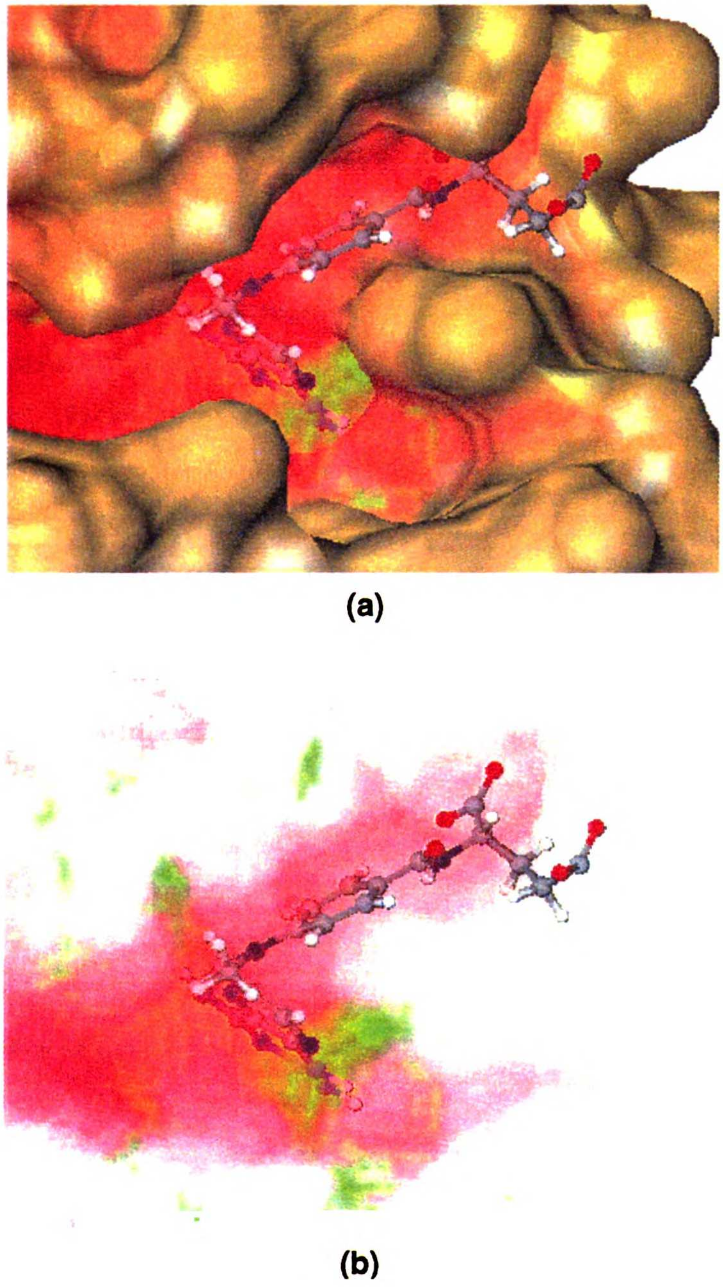
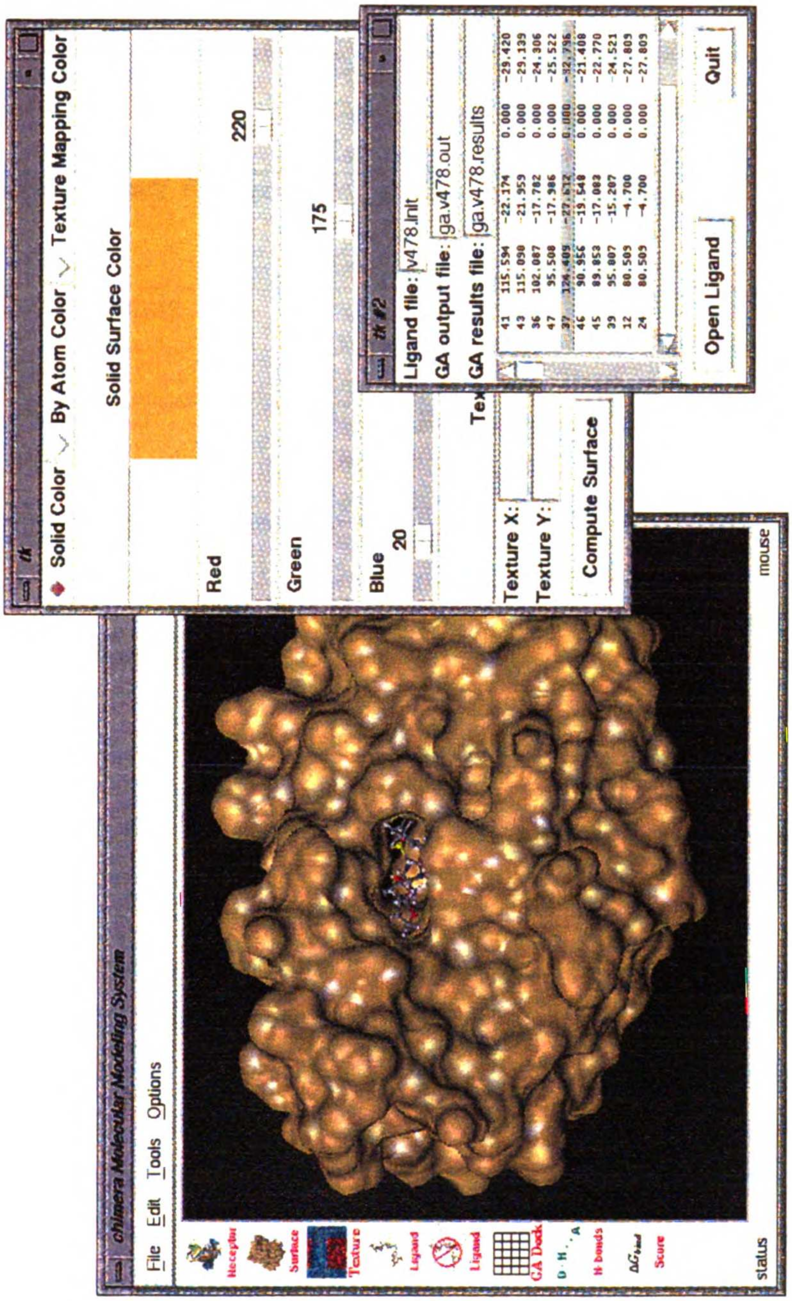


Figure 3.6: a) The volume visualization of the electrostatic potential within the active site of dihydrofolate reductase. b) The surface has been removed to show that the pteridine ring lies in the region of higher electrostatic potential. The green corresponds to the regions of higher electrostatic potential.

simulations. In designing a new ligand, however, one is interested in identifying the properties in the volume of space of the active site. For example, one might be interested in identifying those regions where new hydrogen bonds can be formed. For this purpose, I have computed on a three dimensional grid the location of hydrogen bond donor and acceptor potential and contoured the region using the isosurface routine in the Molecular Inventor [15] software package, a chemistry visualization toolkit that is based on OpenInventor. Figure 3.5 shows the regions of space where hydrogen bond acceptor groups on the ligand can be placed to form hydrogen bonds with the receptor. Notice how the ligand's acceptor atom lies in the cyan region and forms a hydrogen bond with a donor group on the receptor surface. This utility of this functionality is better illustrated with interactive molecular graphics instead of an image. As another example of displaying volumetric data, I have computed the electrostatic potential of a receptor's active site. The potential is computed on a three dimensional grid as described in Chapter 4 and is displayed using a volume visualization technique that is implemented using texture mapping. Figure 3.6 shows the electrostatic potential for dihydrofolate reductase which has a strongly polarized active site with a positive charge at one end and a negative charge at the other. The potential that is stored on the grid is converted into a texture mapping object and is displayed in Chimera. I have considered only two specific examples for visualizing volumetric data, but the two C++ applications can be applied to any volumetric data that is stored on a three dimensional grid. A graphical user interface is needed for this application.

Lastly, the system has an integrated flexible ligand docking module. Although flexible ligand docking, in general, cannot be done interactively, there are some simulations, particularly with an initially docked fragments, where this is the case. Eventually, as computers become faster and our docking methods improve, this may be more useful. My ligand design system has a browser for organizing and displaying the results of a docking simulation. It is implemented in *Tk* and displays information about the simulation, the score of the docked structure, and its rms deviation from the crystal structure if it is known. Selecting an entry in the table transforms the initial ligand conformation with the particular translations and rotations which are



● Solid Color

Red 220

Green 175

Blue 20

Texture X:

Texture Y:

Ligand file: v478.init

GA output file: ga.v478.out

GA results file: ga.v478.results

```
41  115.594  -22.174  0.000  -29.420
43  115.090  -21.959  0.000  -29.139
36  102.087  -17.782  0.000  -24.306
47   95.508  -17.986  0.000  -25.522
37  124.409  -27.612  0.000  -32.796
46   90.956  -19.548  0.000  -21.408
45   89.853  -17.083  0.000  -22.770
39   95.807  -15.207  0.000  -24.521
12   80.509   -4.700  0.000  -27.809
24   80.509   -4.700  0.000  -27.809
```

Figure 3.7: This figure shows Chimera with my ligand discovery and optimization extensions. My applications are included in the toolbar on the left hand edge of Chimera. The figure also shows my surface calculation and molecular docking browser extensions.

| Extension | Implementation |
|---|---|
| Molecular Surface | *Input:* PDB file<br>*Output:* OpenInventor<br>*Language:* Python, Tk |
| Molecular Dock Browser | *Input:* GA output information<br>*Output:* PDB file<br>*Language:* C, Python, Tk |
| Hydrogen Bonds | *Input:* PDB file(s)<br>*Output:* OpenInventor, ascii text<br>*Language:* C++, Python |
| Dock Score | *Input:* Receptor (pdb), ligand (mol2)<br>*Output:* ascii text (see Chapter 5)<br>*Language:* C, Python |
| Isosurface | *Input:* 3D grid<br>*Output:* OpenInventor<br>*Language:* C++, Molecular Inventor |
| Volume Visualization | *Input:* 3D grid<br>*Output:* OpenInventor<br>*Language:* C++ |

Table 3.2: This table list the extensions that are part of my ligand discovery and optimization system.

read from the genetic algorithm docking output file (see Chapters 4 and 5). The docked structure is then displayed within Chimera. Figure 3.7 shows Chimera with my drug design extensions. My system is contained within the icons of the toolbar on the left hand edge of the Chimera window. The "Surface" icon executes my surface calculation application described above. The "GA Dock" icon brings up the browser for the flexible ligand docking results. This application proved to be most useful in the analysis of molecular docking simulations.

In conclusion, a ligand discovery and optimization system has been implemented as an extension to the core functionality of Chimera. Table 3.2 shows a list of the extensions. This system is useful for analyzing and visualizing data that can be important for the design of new drugs. It can help us further our understanding of the protein-ligand interactions that are applicable to structure-based drug design. It can be used to identify interaction sites on the receptor that may be used to improve the binding affinity of a ligand to its target. The molecular docking browser facilitates the analysis of the results. With basic methods for adding and replacing groups on a molecule, this system will be useful for structure-based drug design. As new features are added to Chimera, more extensions can be added which will greatly enhance the discovery and optimization of ligands in the drug design process.

# Bibliography

[1] I. D. Kuntz. Structure-based strategies for drug design and discovery. *Science*, 257(5073):1078–82, 1992.

[2] B. K. Shoichet, R. M. Stroud, D. V. Santi, I. D. Kuntz, and K. M. Perry. Structure-based discovery of inhibitors of thymidylate synthase. *Science*, 259(5100):1445–50, 1993.

[3] E. Rutenber, E. B. Fauman, R. J. Keenan, S. Fong, P. S. Furth, D. E. Ortiz, E. Meng, I. D. Kuntz, D. L. Decamp, R. Salto, J. R. Rosé, C. S. Craik, and R. M. Stroud. Structure of a non-peptide inhibitor complexed with HIV-1 protease. developing a cycle of structure-based drug design. *J. Biol. Chem.*, 268(21):15343–6, 1993.

[4] D. Ringe. Binding by design. *Nature*, 351:185–186, 1995.

[5] N. C. Cohen, J. M. Blaney, C. Humblet, P. Gund, and D. C. Barry. Molecular modeling software and methods for medicinal chemistry. *J. Med. Chem.*, 33(3):883–94, Mar 1990.

[6] R. S. Bohacek and C. McMartin. Definition and display of steric, hydrophobic, and hydrogen-bonding properties of ligand binding sites in proteins using Lee and Richards accessible surface: validation of a high-resolution graphical tool for drug design. *J. Med. Chem.*, 35(10):1671–84, May 1992.

[7] C. C. Huang, G. S. Couch, E. F. Pettersen, and T. E. Ferrin. Chimera: An extensible molecular modelling application constructed using standard components.

Singapore, 1996. Biocomputing: Proceedings of the 1996 Pacific Symposium, World Scientific Publishing Co.

[8] T. E. Ferrin, C. C. Huang, L. E. Jarvis, and R. Langridge. The MIDAS display system. *J. Mol. Graphics*, 6(1):13–27, 1988.

[9] M. Lutz. *Programming Python.* O'Reilly & Associates, Inc., Sebastopol, CA, 1996.

[10] J. Wernecke. *The Inventor Mentor.* Addison Wesley, Reading, Massachusetts, 1995.

[11] 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.

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

[13] A. Varshney, F. P. Brooks, Jr., and W. V. Wright. Linearly scalable computation of smooth molecular surfaces. IEEE Computer Graphics and Applications, Sept 1994.

[14] M. F. Sanner, A. J. Olson, and J. C. Spehner. Reduced surface - an efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, 1996.

[15] Silicon Graphics, Inc. http://www-europe/sgi.com/ChemBio/MolInventor, 1996.

# Chapter 4

# Flexible Ligand Docking with a Molecular Mechanics Force Field

Many of the functions performed by biological molecules depend on appropriate interactions with each other. Molecular recognition is the process by which intermolecular forces act to bring about a productive collision between molecules; it is inherently a dynamic and stochastic process. One application where a better understanding of the mechanisms of molecular recognition would be particularly important is in the design of new therapeutic agents that bind to target receptors with high affinity and specificity.

Molecular docking is one way to computationally formulate the problem of molecular recognition. Docking is simply the collision of the ligand with the binding site in the correct relative conformation and orientation to form a complex. Within the context of molecular docking, the problem is to identify the "best" conformation of a ligand in the binding site of the receptor. There are two essential parts to any docking algorithm — a scoring function and an efficient algorithm for searching conformational space.

Different ligands can bind to a receptor in many different conformations with varying affinities. A scoring function is used to rank these affinities. An important criteria for the form of the scoring function requires a compromise between the complexity of the function to maintain reasonable computational efficiency and its ability to correctly rank the bindings. A variety of scoring methods have been used in molecular docking. For example, Kuntz and coworkers [1] use shape-based descriptors. Pattabiraman, *et al.* [2], Goodford [3], and Meng, *et al.* [4] use three-dimensional grids to evaluate energies (nonbonded, electrostatics, and H-bonds). Bohm [5] uses empirical scoring functions based on experimental binding data.

Algorithms for efficiently searching conformational space are another integral part of any docking algorithm. In the simplest formulation of the molecular docking problem, both the ligand and the receptor are assumed to be rigid. Searching conformational space is reduced to a search of the much smaller orientational space. DOCK [1] was one of the first algorithms developed to address this problem. While most algorithms consider the receptor to be fixed, a few algorithms [6, 7] allow side-chains to take on values from a limited set of well-defined conformations. Despite the simplifications that are made, a brute force search of conformational space is not practical. For example, in docking a rigid ligand in a 20x20x20Å box with 5 degrees and 0.5Å rotational and translational resolution, respectively, there are more than $3 \times 10^{10}$ orientations to be tested. Methods based on distance geometry [8], Monte Carlo [9, 10], graph theory and tree search algorithms [11] have been used. Kuntz, *et al.* [12] provide an overview of these algorithms as they are used for molecular docking.

A useful molecular docking algorithm must satisfy at least the following three conditions: First, it should be computationally efficient while simultaneously providing a reasonably thorough search of conformation space. Second, it should identify the "best" binding modes consistently as brute force exhaustive searches are not feasible. Without an exhaustive search, it is not possible to guarantee that a given minimum is global. Third, the algorithm's parameters should not be too sensitive to the particular ligand-receptor complex in question. Algorithm parameter searches should be avoided especially since there is no independent way to verify a docking result in a real-world application.

In this paper we explore the application of genetic algorithms (GAs) to the problem of rigid and flexible ligand docking to fixed receptors. Genetic algorithms are used because they are computationally efficient and easily parallelizable. They have also been proven to robustly search complex spaces [13]. For our scoring function, we use an AMBER-type potential function [14, 15] to evaluate the "binding affinities" of the different ligand conformations. To improve the algorithm's efficiency we employ a simplification used by others, namely, a grid based energy evaluation [2, 3]. Lastly, we introduce a masking operator, which provides a convenient mechanism by which different binding hypothesis can be tested. This operator also improves the consistency with which solutions are found. We report the results on four different ligand-receptor complexes (in this paper we are concerned only proteins) using the system, DIVALI (Docking wIth eVolutionary ALgorIthms).

## 4.1 Methods

### Scoring Function

We adopt an AMBER-type potential function [14, 15] to score the different ligand orientations with the underlying assumption that the correct ligand binding conformation corresponds to the minimum of this function. In general, our scoring function is

$$\text{Score} = E_{inter} + E_{intra} + C \tag{4.1}$$

where $E_{inter}$ and $E_{intra}$ are the intermolecular and intramolecular energies, respectively, and $C$ is a positive constant whose value is chosen to prevent negative scores which would violate the assumptions with the selection operator. We use $C = 100$ kcal/mol.

In general, the energy of interaction between the ligand and the receptor is governed by electrostatics, van der Waals, and hydrophobic interactions. This intermolecular energy can be approximated by

$$E_{inter} = \sum_{l}^{lig} \sum_{r}^{recep} \left( \frac{A_{lr}}{r_{lr}^{12}} - \frac{B_{lr}}{r_{lr}^{6}} + 332.0 \frac{q_l q_r}{\epsilon(r) r_{lr}} \right), \tag{4.2}$$

where $A_{lr}$ and $B_{lr}$ are the non-bonded parameters, $\epsilon(r)$ is the dielectric constant, $q_r$ and $q_l$ are the partial charges on the receptor and ligand atoms, respectively. The factor 332.0 converts the electrostatic energy into kilocalories/mole. A distance-dependent dielectric function, $\epsilon(r) = 4r$, is used to model the effect of the solvent.

As mentioned above, in order to improve the algorithm efficiency a grid-based energy evaluation procedure is used. Following the procedure outlined by Pattabiraman *et al.* [2], the intermolecular energy can be written as

$$E_{inter} = \sum_{l} \left( q_l \sum_{r} \frac{332 q_r}{\epsilon(r) r_{lr}} + \sqrt{A_l} \sum_{r} \frac{\sqrt{A_r}}{r_{lr}^{12}} - \sqrt{B_l} \sum_{r} \frac{\sqrt{B_r}}{r_{lr}^{6}} \right). \qquad (4.3)$$

Thus, the electrostatics and van der Waals contributions due to the receptor can be pre-calculated and stored on a grid. We interpolate between the eight neighboring grid points to calculate the receptor contributions on each ligand atom. The AMBER all atom parameter set is used for both the receptor and the ligand atoms. The partial charges on the ligands (except methotrexate) were obtained from the Gasteiger-Marsili option within SYBYL [16].

The location of the grid is determined so that it encloses the active site or a large enough cavity on the surface of a receptor. Potential binding sites can be identified by using Connolly's MS algorithm [17, 18], for example. The spacing between the grid points is 0.3 Å. Clearly, the grid calculation is an approximation, but experience has demonstrated that the ranking of different orientations is maintained [2, 4].

For rigid-body docking the scoring function is simply the intermolecular energy added to a constant. For flexible ligand docking, the intramolecular energy of the ligand must also be included. This energy is

$$E_{intra} = \sum_{l<l'}^{ligand} \left( \frac{A_{ll'}}{r_{ll'}^{12}} - \frac{B_{ll'}}{r_{ll'}^{6}} + 332.0 \frac{q_l q_{l'}}{\epsilon(r) r_{ll'}} \right) + \sum_{diheds} \frac{V_n}{2} \left[ 1 + \cos\left\{ \eta_n (\phi_n - \gamma_n) \right\} \right] \qquad (4.4)$$

where the van der Waals and electrostatic sums are over the one-four and higher interactions. The last term is the torsional angle component of the intermolecular energy of the ligand. We have assumed the bond lengths, bond angles, and ring conformations found in the crystal structures. These are not allowed to vary. Unlike the intermolecular energy which is precalculated and stored on a grid, the intramolecular energy must be calculated for each conformation. In AMBER [14, 15] certain atom pairs are modeled using explicit H-bonds to fine tune the H-bond distances and energies. This prevents unrealistically short H-bonds. In our implementation we do not use a specific H-bond term. We assume that contributions due to H-bonding are included in the electrostatic term. We also remove all crystallographic water molecules from the receptor.

## Genetic Algorithms

A genetic algorithm is an optimization procedure that combines a probabilistic search algorithm with a directed search strategy based on the fitness function. They have been shown to search complex spaces robustly [13]. Genetic algorithms differ from other optimization techniques in a number of ways. First, they explore the different regions of the solution space with a population of points. Second, they only need the value of the function being optimized. They are not limited by the details of the function such as continuity or the existence of derivatives. Third, they do not manipulate the parameters individually. The parameters are instead coded as a continuous string, and the algorithm modifies this string without any direct knowledge of individual parameters. A genetic algorithm can simultaneously process local and non-local information within the string. This feature is known as implicit-parallelism [19] and is one of the most interesting strengths of genetic algorithms. It is important to note, however, that implicit-parallelism is independent of the particular choice of representation.

A genetic algorithm maintains a population of individuals with an associated fitness. Each individual represents a possible solution to the problem. A genetic algorithm alters the individuals of the population and thereby searches different regions of the solution space in two separate stages. During the selection stage a new population is created by proportionally selecting the more fit individuals from the previous population. The members of the populations are then transformed in the alteration step. The mutation operator creates a new individual from a single individual by randomly changing an element in its representation. The crossover operator, on the other hand, exchanges information between two (occasionally more) members of the population. We use a two-point crossover operator where two points are chosen at random, and the information between the two points is exchanged. We chose this implementation because there is evidence that it works better than a single point

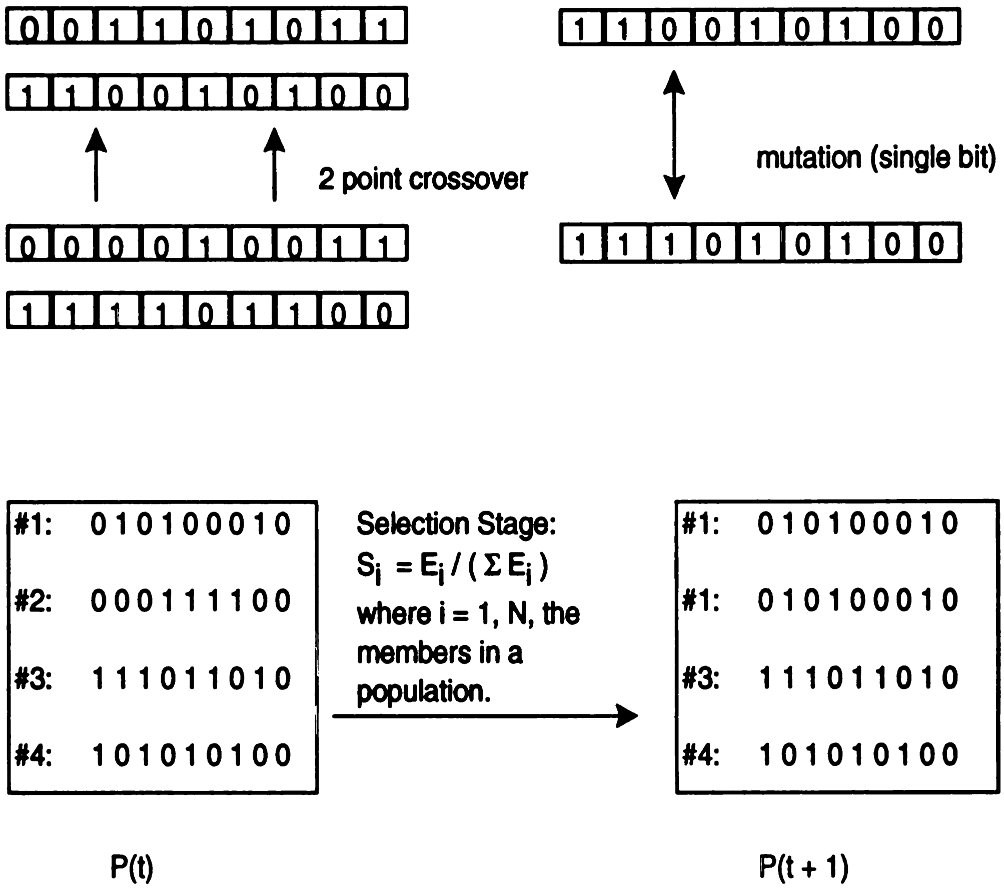


Figure 4.1: Shows the basic operations of a genetic algorithm that uses bit strings. We show the mutation and crossover operators (top figures) and the selection procedure (bottom figure).

| PDB | Center | Dimension | Bonds | Translation | Bits |
|---|---|---|---|---|---|
| 2gbp | $43.1 \times 31.8 \times 51.5$ | $20 \times 28 \times 20$ | 6 | 8, 8, 8 | 114 |
| 3cpa | $-1.1 \times 31.7 \times -7.4$ | $18 \times 18 \times 22$ | 8 | 7, 7, 8 | 132 |
| 4dfr | $25.0 \times 68.0 \times 46.0$ | $25 \times 15 \times 25$ | 11 | 9, 8, 9 | 166 |
| 6rsa | $33.2 \times 11.8 \times 11.1$ | $28 \times 18 \times 24$ | 6 | 8, 7, 8 | 113 |

**Table 4.1: Summarizes the data representation issues for the four complexes studied. The first column corresponds to the PDB descriptor for the complex. The box center and its dimensions are given in the next to columns. The remaining entries are the number of rotatable bonds, the number of bits to represent each of the translation components and the total number of bits used to represent the solution for each complex.**

crossover [13]. The selection step is based on a simple heuristic idea that we would expect to find the best solution in that region of space which contains a large number of good solutions. Mutation and crossover operators extend this region to the space of potential solutions. Figure 4.1 illustrates these operators. Our implementation of the genetic algorithm is based on the GAucsd package [20]. We also use the so-called elitist mode while running the GA. In this mode the best member of the population in each generation is copied into the next.

In our formulation of the molecular docking problem, the binding mode is described by three translations ($T_x$, $T_y$, $T_z$), three rotations ($R_x$, $R_y$, $R_z$), and a number of bond rotations. The translational components usually specify the position of the centroid of the molecule, but they can also specify the transformed coordinates of a given atom. The rotational components are the Euler angles of rotation. As in most applications of genetic algorithms we chose to use a binary representation. In particular, we use Gray-coded binary strings throughout because the genetic algorithm literature has amassed substantial empirical evidence that Gray-coding performs better than a simple binary representation. With Gray-scale coding the closest values within the representation differ by at most one bit. Furthermore, because this representation is periodic, it is well suited to representing rotations. The number of bits for each translation component $b_i$ is given such that $2^{b_i} \geq 2L_i/0.3$ where $L_i$ is the dimension of the box in the *i*th direction, and 0.3 is the grid spacing. All rotations whether Euler angles or bond rotations are represented with 10 bits. Table 4.1 summarizes this information for each of the complexes studied.

## Receptor-Ligand Complexes

Four widely differing well-resolved crystallographic complexes from the Protein Data Bank [21, 22] are used to demonstrate our method. These are periplasmic binding protein-glucose (2gbp), carboxypeptidase A-glycyl-L-tyrosine (3cpa), dihydrofolate reductase-methotrexate (4dfr), and ribonuclease A-uridine vanadate (6rsa).

The periplasmic binding protein (2gbp) binds the sugar with about thirteen hydrogen bonds, and the structure is known to a resolution of 1.9 Å. The bound sugar is completely engulfed in the cleft between the two domains of the protein. The X-ray results indicate that the final geometry of the binding site depends both on the final folded protein and any conformational changes induced by the ligand [23]. The glucose binding is not strongly controlled either by charge (net charge zero) or shape (a rough ellipsoid).

The structure of carboxypeptidase A glycyl-L-tyrosine complex (3cpa) is known to a resolution of 2 Å [24] at low temperatures. The hydroxybenzyl group of tyrosine resides in the hydrophobic pocket of the $S_1'$ subsite. The dipeptide appears to bind with the hydrolytically important zinc-bound water displaced and excluded from the active site.

The structure of methotrexate bound to dihydrofoloate reductase (4dfr) is known to a resolution of 1.7 Å [25]. There are two independent protein molecules in the PDB file with a methotrexate inhibitor bound to each. We use the molecule designated "B" in the PDB file. The p-aminobenzamide is bound to the enzyme by hydrophobic and van der Waals interactions. There are bound water molecules in the active site.

The ribonuclease A and uridine vanadate (6rsa) structure [26] is known to a resolution of 2 Å. The uridine vanadate appears to be a transition state analog. The position and orientation of the uridine base is completely clear in the electron density map, but the structure of the sugar ring is less obvious. The reported structure deviates from ideal geometry, and some difficulties were encountered in interpreting the electron density map in the neighborhood of the vanadium atom. To make it possible to use AMBER based parameters, we replace the vanadium atom by phosphorus [4]. The X-ray structure of uridine vanadate (phosphate) is restrained energy minimized after replacing vanadium with phosphorus (see figure 4.2). The ligand conformation can, therefore, no longer be thought of as a transition state analog.

## 4.2 Results

### Rigid Ligand Docking

We began our studies in the simplest possible way. Given a ligand in the crystal structure conformation, we wanted to test whether the genetic algorithm could find the correct orientation of the ligand within the receptor. That is, we wanted to find the correct translations and rotations. The GA was initialized with a population of random $\vec{T}'s$ and $\vec{R}'s$. The masking operation maintained the centroid of the ligand within the box. Ligand atoms outside the box were assigned large positive energies ($10^9$ kcal/mole) for both rigid and flexible simulations. Assigning a large positive number to these atoms does not bias the search as the box in all four cases lies well outside the region of interest of the receptor.

Different population sizes, number of generations, and mutation and crossover rates were considered in many different combinations. For each combination of parameters twenty runs with different initial populations were attempted. The results were not encouraging. Occasionally, we would find the solution, often we would not. No solution was found for the rigid body docking of glucose to the periplasmic binding protein after hundreds of different runs. The GA consistantly found solutions about 8–10 kcal/mole higher than the crystal structure. Similarly, the GA was unable to find the binding orientation of glycyl-L-tyrosine complexed with carboxypeptidase A. We did find solutions for rigid docking of methotrexate to dihydrofolate reductase and uridine vanadate to ribonuclease A. The results, however, were not consistent and involved different mutation and crossover rates. The only reasonable conclusion is that the standard GA recipe is inappropriate.

There are a number modifications that could be made to this system. Recall from the discussion in the introduction that we would ideally like to have an efficient algorithm that consistently converges to reasonable binding modes without system dependent parameter adjustments. We could, for example, use a different representation, provide a method to search space more thoroughly, introduce new genetic operators, or even alter the selection procedure. The simplicity of the binary representation and the fact that a binary coding offers the maximum number of schemata per bit [13] led us to retain the binary representation. We decided to introduce a new operator to the GA which promotes a more thorough search of the orientational or conformational space. We address the other possible modifications in the discussion.

The idea behind the new operator is to introduce the notion of precisely-defined and well-maintained sub-populations. In a binary representation the most significant bit of the representation divides the region into two halves. The next significant bit divides the region into fourths and so on. In the context of molecular docking by fixing the most significant bit of each of the translation components, we create eight different sub-populations, each of which explores one of the mutually exclusive and collectively exhaustive regions of the box. Deciding to search in any one of the different regions is equivalent to testing a different binding hypothesis. We could equally well have defined eight different regions in rotation space, but using sub-populations based on translation space is more intuitive. This operator is called a masking operator, and because it constrains the search to subpopulations, it ensures a wider and more thorough search in the space of possible conformations.

As far as the GA parameters are concerned, we decided to fix them for all other simulations. We chose the crossover rate to be 0.6 because we wanted the crossover rate to be greater that half while simulataneously avoiding a GA that was completely dominated by crossover. We tried a variety of mutation rates before deciding to use a mutation rate of one bit in approximately 1000. The population size $n$ and the number of generations were determined in an ad hoc fashion. The population size was determined with the following considerations: (1) it should not be too small so as to avoid pre-mature convergence, (2) it should not be too large as it increases the

| 2gbp | | 3cpa | | 4dfr | | 6rsa | |
|---|---|---|---|---|---|---|---|
| $E$ | rms (Å) | $E$ | rms (Å) | $E$ | rms (Å) | $E$ | rms (Å) |
| -23.9 | 0.35 | -47.8 | 0.07 | -63.0 | 0.25 | -61.6 | 0.98 |
| -23.4 | 0.36 | -47.7 | 0.12 | -49.3 | 4.49 | -61.2 | 0.65 |
| -23.6 | 0.35 | -47.6 | 0.14 | -54.6 | 1.98 | -60.8 | 1.00 |
| -23.8 | 0.30 | -47.9 | 0.19 | -27.4 | 9.04 | -61.4 | 0.74 |
| -23.9 | 0.33 | -47.7 | 0.12 | -48.2 | 2.02 | -61.5 | 1.00 |
| -23.9 | 0.38 | -47.6 | 0.18 | -49.3 | 4.68 | -61.5 | 0.84 |
| -24.1 | 0.33 | -47.6 | 0.10 | -65.6 | 0.37 | -61.0 | 0.91 |
| -24.0 | 0.36 | -47.7 | 0.09 | -66.3 | 0.26 | -61.0 | 1.05 |
| -23.3 | 0.41 | -47.9 | 0.07 | -65.7 | 0.21 | -60.1 | 1.06 |
| -20.3 | 0.53 | -47.5 | 0.24 | -62.0 | 0.76 | -61.4 | 0.75 |
| $4s$ | | $4s$ | | $17s$ | | $5s$ | |

Table 4.2: Shows the results for rigid ligand docking for all four complexes. The energies (in kcal/mole) and the RMS deviations are given for each of the ten different GA runs. The last row summaries the times that are required for a single GA run each system.

computational burden, and (3) larger ligands may require larger $n$. We chose the population size to be 64 and 256 for rigid and flexible docking, respectively. Three of the four systems (not 4dfr) satisfied the condition that the solution be found five out of ten times with a population of 32 for rigid docking. Finally, the number of generations $N_g$ should depend on the size of the problem or number of parameters that are being optimized. This implies a certain minimum number of generations that the GA needs to be run. For the upper bound we tried to find the number of generations that would be necessary to find the solution in at least five out of tens tries across the four systems. For rigid docking, $N_g = 500$, and for flexible docking we chose $N_g = 1000$. One reason why *a priori* bounds on $N_g$ is useful in docking studies is because not all initial populations converge within reasonable CPU time to the solution (minimum energy conformation), and in our experience it was always better to restart a run with a new population. The robustness of this methodology is demonstrated by the fact that it works consistently across four different ligand-receptor complexes.

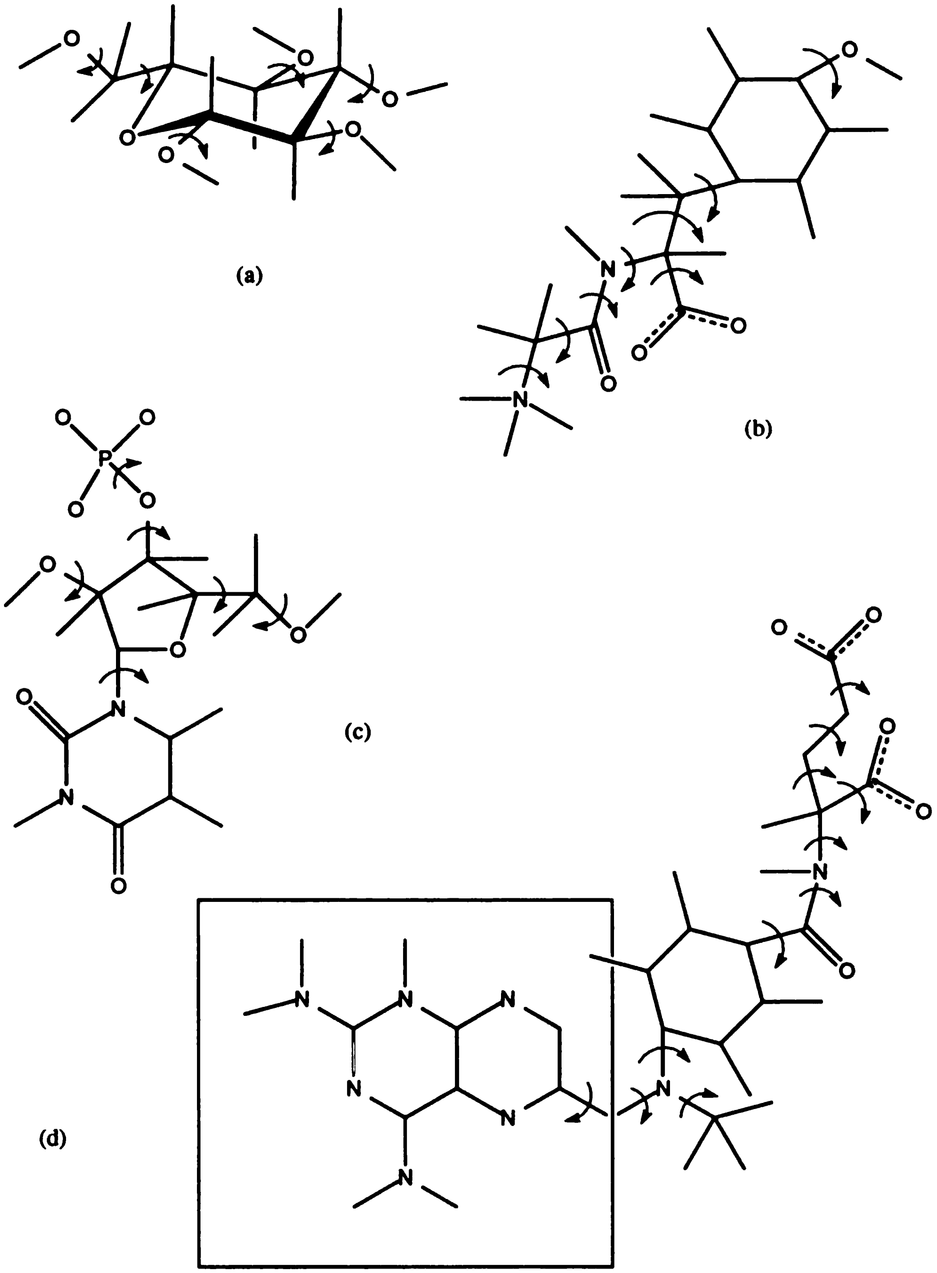


Figure 4.2: Shows the ligands with the dihedrals that change for each of the four complexes studied. (a) $\beta$-D-glucose, (b) glycyl-L-tyrosine, (c) methotrexate, (d) uridinephosphate. The rectangular box around a part of methotrexate is the rigid-part used in the anchoring simulations.

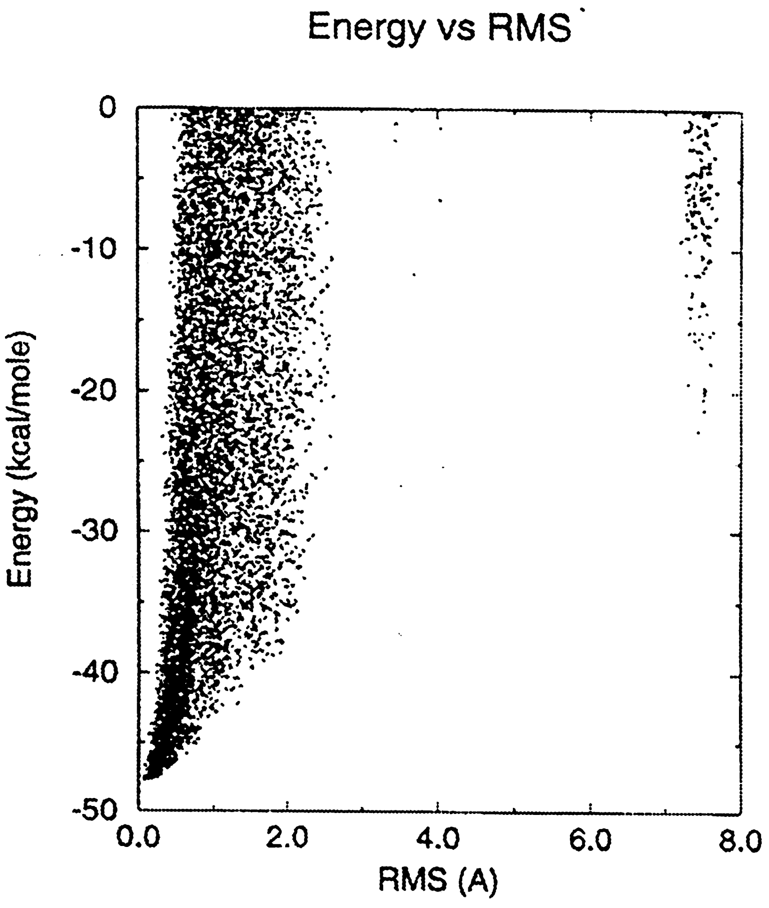


Figure 4.3: Shows the energy versus RMS of the rigid docked conformations for carboxypeptidase A-glycyl-L-tyrosine. The results are shown for the sub-population that yields the lowest energies.

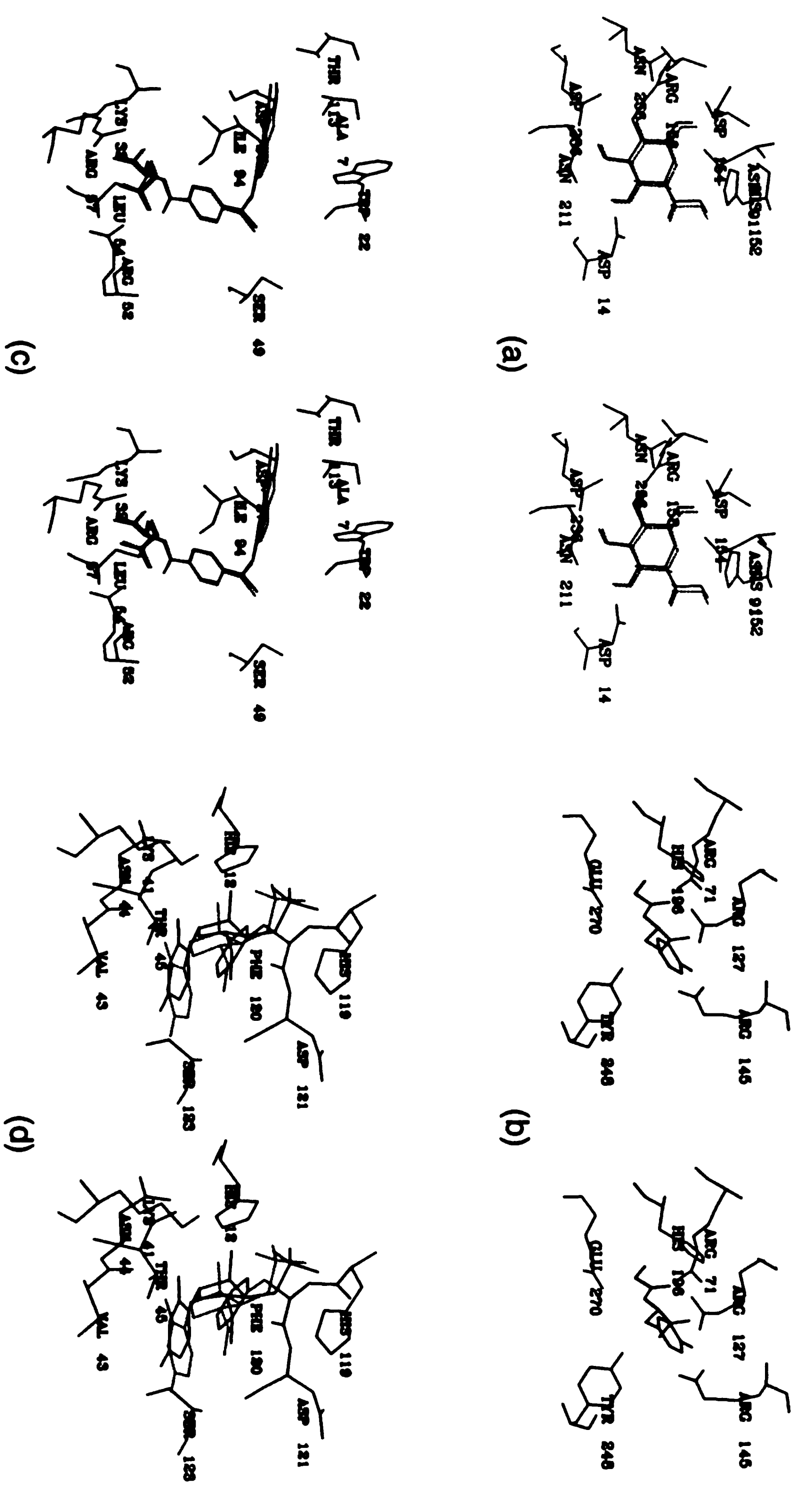


Figure 4.4: Best rigid docked (dashed) and crystal conformations for all four complexes. (a) periplasmic binding protein-glucose (2gbp), (b) carboxypeptidase A (3cpa), (c) dihyrofolate reductase (4dfr), and (d) ribonuclease A (6rsa).

With this modification we have been able to satisfy the requirements of a successful docking algorithm described above, namely, computational efficiency. Table 4.2 gives the RMS error and energies of the docked conformation. In all the rigid docking systems we were able to find the crystal structure conformation at least six out of the ten runs with differently initialized populations. No genetic algorithm parameter adjustments were necessary. Our rigid body docking runs converged very quickly for all four systems including the pteridine part alone and the complete methotrexate (see figure 4.2). The crystal and rigid docked conformations of all four complexes are shown in figure 4.4 for comparison. The scoring function appears to work very well. We have not been able to find any conformations with $E \leq E_{min} + 0.5$ kcal/mole that does not correspond closely to the X-ray structure. Also all the structures within 0.5 kcal/mole of the lowest energy structure have pair-wise RMS errors of 0.5 Å or less. Furthermore, we do not require a final energy minimization step like many other procedures [4, 27]. Considering the accuracy of the scoring function used it appears that the computational efficiency of the method is sufficient to perform rigid database searches. As figure 4.3 shows, there are many structures within 5 kcal/mole of $E_{min}$ that have an RMS deviation within 1 Å of the crystal structure. This is not too surprising due to the rugged nature of the energy function. No other generalizations seem possible across systems from this (and similar) figure(s). Table 4.2 also gives the timing details for each complex. For an exhaustive analysis the total time required will be eighty times (eight sub-populations and ten runs for each population) the numbers in the last row if no other information about the receptor is used.

## Flexible Ligand Docking

The flexible docking procedure follows the one used for rigid docking. The GA parameters are generated in the same fashion. That is, $p_c = 0.6$, and $p_m$ is as calculated as above. Clearly, allowing the flexible ligands greatly increases the size of conformational space. Therefore, larger populations ($n = 256$) and longer GA runs ($N_g = 1000$) are used. Masks on the translation vectors are also applied in order to create and maintain different sub-populations. Table 4.1 gives the details of the

| Complex | PDB | $E_{inter}$ | $E_t$ |
|---|---|---|---|
| periplasmic binding protein/glucose | 2gbp | -24.0 | -22.2 |
| carboxypeptidase A/glycyl-L-tyrosine | 3cpa | -47.9 | -63.2 |
| dihydrofolate reductase/methotrexate | 4dfr | -72.0 | -87.7 |
| ribonuclease A/uridine vanadate | 6rsa | -59.8 | -66.8 |

Table 4.3: Lists the four complexes studied with the intermolecular energy and the variable part of the intramolecular energy.

number of rotatable bonds and the total number of bits used in each case to represent the problem.

A genetic algorithm is also capable of simple energy minimization. We seed the initial population with the crystal structure, and within 50 generations we obtain a local minimum structure with a lower energy. Table 4.3 shows the energies after this minimization. It would be interesting to compare this local minimum structure to the local minimum structure obtained using a simple gradient descent procedure. The RMS deviation after this minimization step is less than 1 Å in all four cases. However the RMS deviations reported in the tables uses the original crystal structure and not the one obtained after this minimization.

There are many questions that arise in comparing the flexibly docked ligand with the X-ray structure. First, it is possible to maintain important ligand-receptor interactions with differences in the two structures. This is especially true when parts of the ligand do not take part in explicit interactions with the receptor. Second, other differences can also appear because of the particularities of our implementation. For example, crystallographic waters including the ones that mediate H-bonding interactions between the ligand and the receptor were removed. In addition, hydrophobic interactions are not accounted for in the energy function. Lastly, other differences in two structures arise because the X-ray structure is an average over different conformations and not an energy minimum structure.

Comparing the docked and X-ray ligands is not necessarily straightforward. If we find structures with small RMS deviations, then obviously the agreement must be quite good, and it would be expected that most of the important interactions are maintained. On the other hand, a structure with a higher RMS deviation does not necessarily mean that the docked structure is incorrect. It is quite possible that the some ligand conformations may satisfy the important specific interactions while other parts of the ligand which do not take place in such interactions can exhibit torsional freedom. Thus, structures with varying RMS deviations can actually be similar structures because they participate in the same interactions. Therefore, we adopt two methods by which the comparisons are made. First, total and flexible RMS deviations are calculated. These two measures differ in that in calculating the flexible RMS, which we shall refer to as simply the RMS, we subtract out the contributions due to the centroids not overlapping. Stereo images of the ligand within the active site are also provided so that the specific interactions can be inspected. Another method to compare the structures is to perform a detailed examination of the important interactions that contribute to the energy ($E_{inter}$) of the structure in both the docked and X-ray structures. However, no systematic conclusions could be drawn from the complexes studied except for 2gbp.

## 2GBP

The flexible torsions are shown in figure 4.2. Table 4.4 shows the energy and RMS deviations of some of the best structures found. Notice that the structure with the smallest energy (-22.1 kcal/mole) corresponds to the structure with the smallest RMS error. Figure 4.5 shows a stereo picture of the best conformation and the crystal structure of the ligand in the active site. The results indicate the the docked structure is quite close to the crystal structure. One reason for such small RMS deviations is that most dihedral changes affect only the location of the hydrogen atoms. As crystallographic water molecules in the active site take part in the glucose binding to the protein, work is currently in progress examining how docking results will change on including them. Because ring geometries are fixed in our implementation, we have not examined whether our procedure would reproduce the preference for $\beta$-D-glucose in the ${}^4C_1$ conformation.

| 2gbp | | 3cpa | | 4dfr | | 6rsa | |
|---|---|---|---|---|---|---|---|
| *E* | rms (Å) | *E* | rms (Å) | *E* | rms (Å) | *E* | rms (Å) |
| -13.0 | 0.6 / 0.3 | -57.2 | 7.4 / 1.8 | -40.5 | 2.8 / 2.3 | -70.6 | 1.6 / 1.0 |
| 8.2 | 4.1 / 0.4 | -63.6 | 1.4 / 0.9 | -71.2 | 1.9 / 1.2 | -67.8 | 4.9 / 1.3 |
| -14.9 | 0.7 / 0.3 | -63.9 | 2.3 / 1.8 | -85.2 | 2.3 / 2.0 | -62.4 | 3.3 / 1.0 |
| -22.1 | 0.4 / 0.3 | -60.5 | 1.7 / 1.1 | 2.1 | 8.1 / 3.0 | -68.9 | 2.0 / 1.2 |
| -17.3 | 0.7 / 0.5 | -64.7 | 1.7 / 1.1 | 0.2 | 8.8 / 2.8 | -69.8 | 4.8 / 1.0 |
| -13.4 | 0.9 / 0.7 | -58.0 | 7.5 / 1.1 | -82.1 | 1.5 / 1.1 | -71.0 | 2.2 / 1.1 |
| -19.5 | 0.5 / 0.4 | -58.4 | 2.0 / 1.0 | -0.9 | 8.4 / 2.9 | -67.0 | 4.9 / 1.3 |
| -19.8 | 0.9 / 0.8 | -55.1 | 7.4 / 1.2 | -77.1 | 2.3 / 1.9 | -71.7 | 1.6 / 1.2 |
| 5.1 | 3.3 / 1.1 | -61.8 | 2.3 / 2.0 | 19.8 | 8.1 / 3.3 | -69.3 | 1.3 / 1.0 |
| -4.3 | 3.5 / 0.4 | -61.8 | 1.4 / 1.0 | -74.1 | 2.0 / 1.9 | -64.1 | 5.1 / 1.0 |
| 120*s* | | 540*s* | | 950*s* | | 370*s* | |

Table 4.4: Shows the results for flexible ligand docking for all four complexes. The energies (in kcal/mole) are given along with two different measures of the RMS deviations. The first number is the total RMS between the docked structure and the crystal structure. The second number refers to the flexible RMS values which reflects the RMS error with the orientational component removed. As with the rigid results, the last row refers to the times for a single GA run of each system.

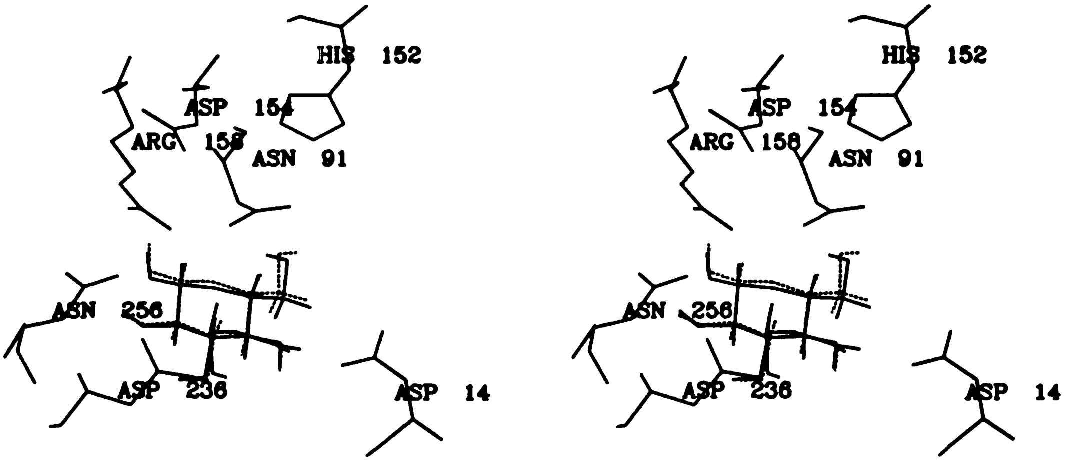


Figure 4.5: Best flexible docked and crystal conformations for periplasmic binding protein-glucose (2gbp). The docked structure is in gray.

## 3CPA

The flexible torsions are shown in figure 4.2. The energy and corresponding RMS deviations of the docked structures are in table 4.4. The energies of the docked structures vary from −55.1 to −64.7 kcal/mole with total RMS deviations from 7.5Å to 1.4Å. The structure with the lowest energy (-64.7 kcal/mole) is quite close to the crystal structure. Figure 4.6 shows two different binding orientations with energy differences of 4 kcal/mole. Interestingly, the two docked structures are rotated by almost 180 degrees but still have energies quite close to the crystal structure energy and to each other. The flipped docked structure has the tyrosine ring towards the surface of the protein. We expect that including a solvation term in the energy function will weed out the $180^o$ rotated structure. In addition several water molecules are also displaced upon binding of GY. Therefore, the entropic effects of the release of the these water molecules will play a major role in its binding. The overall agreement is quite good. Higher RMS deviations compared with 2gbp arise because the location of a larger number of heavy atoms are unknown and also due to the larger number of atoms in the ligand.

## 4DFR

Figure 4.2 shows the 11 rotatable bonds in methotrexate, the largest ligand among the four systems considered. Docking methotrexate to dihydrofolate reductase turned out to be quite difficult using the masking operation described above. We do find solutions close to the crystal structure occasionally, but it is more by chance. We do not consistently find the solution. The GA exhibits "premature convergence" in these runs. This is equivalent to getting stuck in a local minimum (see section 4.3). To verify that our system could, in fact, find the crystal structure conformation, we narrowed the region of our search. We assumed that the centroid of the ligand in the crystal structure conformation must lie in a 6 × 4 × 6 box about the true centroid and that the rotations are within 90 degrees of the actual rotations. In practice, it might be possible to formulate such a hypothesis with knowledge about the structure of the receptor. The energies and RMS deviations for the ten runs is

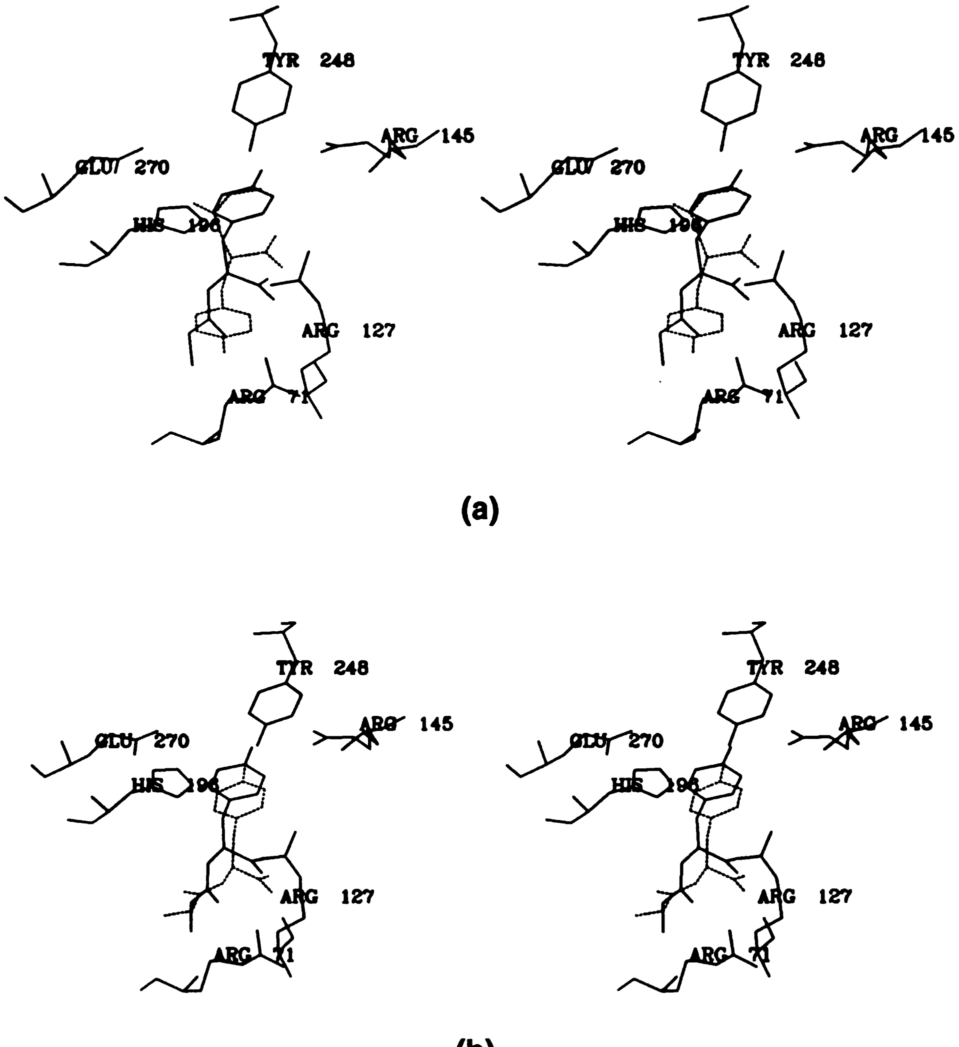


Figure 4.6: (a) A good flexible docked solution and the crystal conformations for carboxypeptidase A with glycine-L-tyrosine and (b) the best flexible docked and crystal conformations for carboxypeptidase A with glycine-L-tyrosine. The docked structure is dashed.

| $E$ | total rms (Å) | rms (Å) |
|---|---|---|
| -81.2 | 3.48 | 2.56 |
| -79.6 | 9.90 | 2.80 |
| -89.3 | 4.73 | 2.68 |
| -80.0 | 6.41 | 3.20 |
| -90.0 | 4.11 | 1.81 |
| -84.2 | 4.88 | 2.73 |
| -85.4 | 4.68 | 3.28 |
| -83.2 | 3.90 | 1.91 |
| -85.9 | 4.93 | 2.73 |
| -81.2 | 3.48 | 2.56 |

Table 4.5: Shows the results for the flexible docking of methotrexate to dihydrofolate reductase using the masking operator to keep the pteridine ring approximately in place.

given in table 4.4. The best energy structure (-85.2 kcal/mole) that the GA found had an RMS deviation of 2.3Å. Three solutions had an RMS deviation of less than or equal to 2.0Å. The best energy of these structures is −82.1 kcal/mole while the crystal energy (after minimization) is -87.5 kcal/mole. Thus, with the region of space narrowed as described above, the GA is able to find solutions within 2.0Å of the crystal structure.

We also tried a more "intelligent" masking operation, one that incorporates more specific information about the binding mode. It is known that the pteridine ring of methotrexate is rigid and binds to the receptor in roughly the same site [25]. We, therefore, docked the pteridine moiety as a rigid unit using the methodology described in the rigid docking section (see also [4]). As with the rigid docking for 2gbp, 3cpa, 4dfr and 6rsa, this was also a simple problem for the genetic algorithm. A structure very close to the crystal structure was obtained. We created a mask that would keep the pteridine ring of the methotrexate in the vicinity of the rigid-docked structure. The mask allowed each translation and rotation component to vary the position of the ring by approximately 1Å and 22.5 degrees. These results are shown in table 4.5. Unlike any of the previous runs, all the runs converge. That is, all runs result in

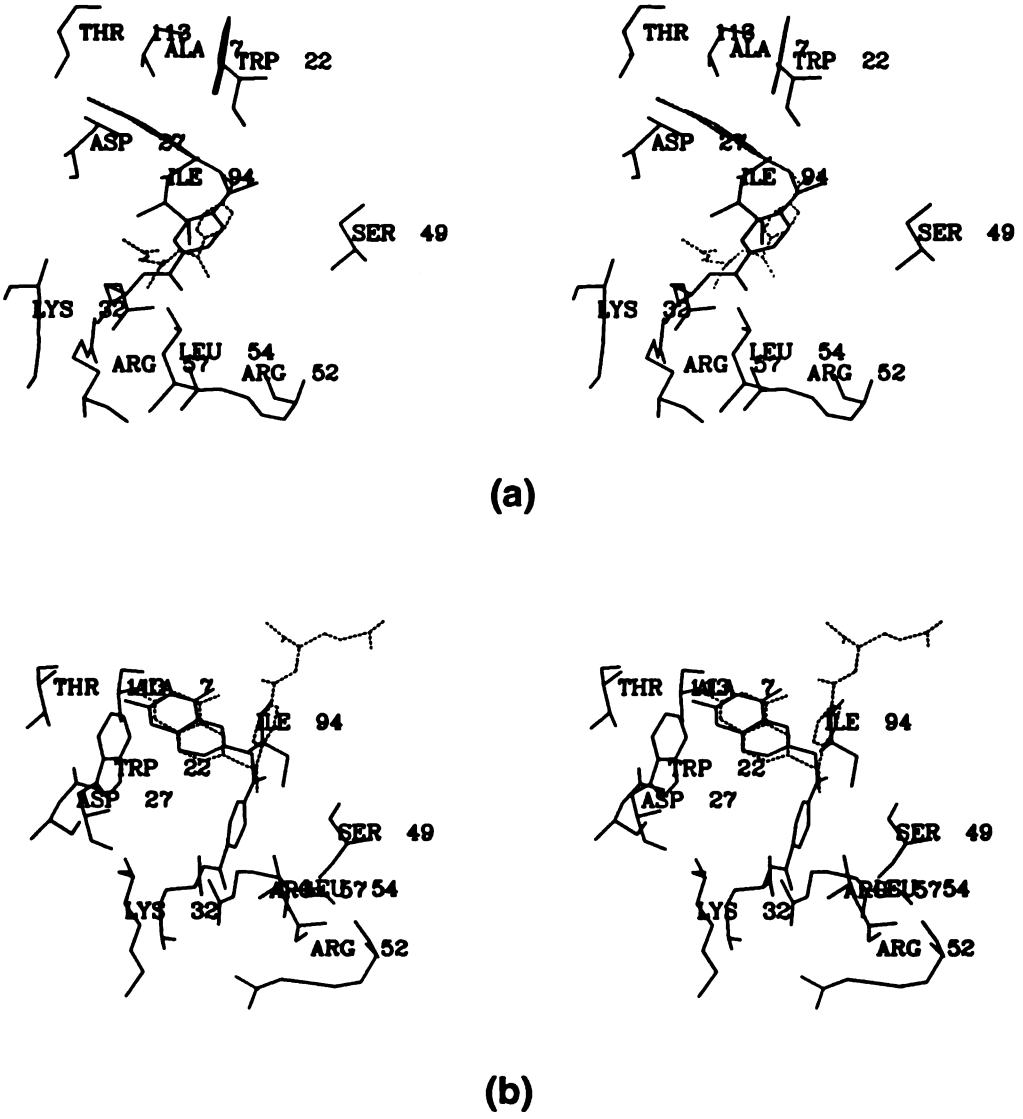


Figure 4.7: The crystal structure and two docked conformations of dihydrofolate reductase with methotrexate. The top figure, (a) shows the docked structure with $E_t$ = $-85.9$ kcal/mol; and the bottom figure, (b) shows a very different conformation with comparable energy ($E_t = -86.2$ kcal/mol to the crystal structure ($E_t = -87.7$ kcal/mol). The docked structures are dashed.

energies very close to the minimized crystal structure energy. The most interesting part of this result is that widely differing structures have energies close to the crystal structure energy. In fact, the 10 Å structure follows the NADPH binding site. Some of these conformations along with the crystal structure are shown in figure 4.7.

The placement of the benzyl moiety of MTX in a hydrophobic pocket formed by Leu28, Phe31 and Ile50 is a major source of binding free energy, but this is not accounted for in our energy function and contributes primarily to the alternative binding modes found in the pteridine docked structures. There are many interesting aspects to the methotrexate docking results, *e.g.,* the differences between the many ways of masking (centroid, one atom, pteridine ring), the differences between charged and uncharged methotrexate.

## 6RSA

Recall that the vanadium is replaced by phosphorus, and the structure is energy minimized to obtain the "crystal" structure to which the docked structures are compared. Figure 4.2 shows the variable torsions. The vanadium atom occupies the center of a distorted trigonal bipyramid with the ribose O2' at the apical position. There are two water molecules in the active site [26], and one of the water molecules exists close to O3' and O7. Replacing the vanadium with phosphorus and minimizing the structure with no explicit water molecules alters the structure of the ligand. It is this minimized structure that we consider to be the "experimental" structure against which RMS deviations are calculated. Table 4.3 shows the total and intermolecular energy of this structure. Notice that the intramolecular energy is positive.

The systematic masking operation is quite successful in finding good solutions. Table 4.4 shows the energy and RMS deviations of the docked structures. The energies of the docked structures vary from $-64.1$ to $-71.7$ kcal/mole with total RMS deviations from 5.1Å to 1.3Å. The minimum energy structure (-71.7 kcal/mole) has a total RMS error very close to the minimum RMS. Furthermore, there is a structure with an total RMS of 4.8Å with an energy of $-69.8$ kcal/mole. The difference in energy between this structure and the structure closest to the crystal structure

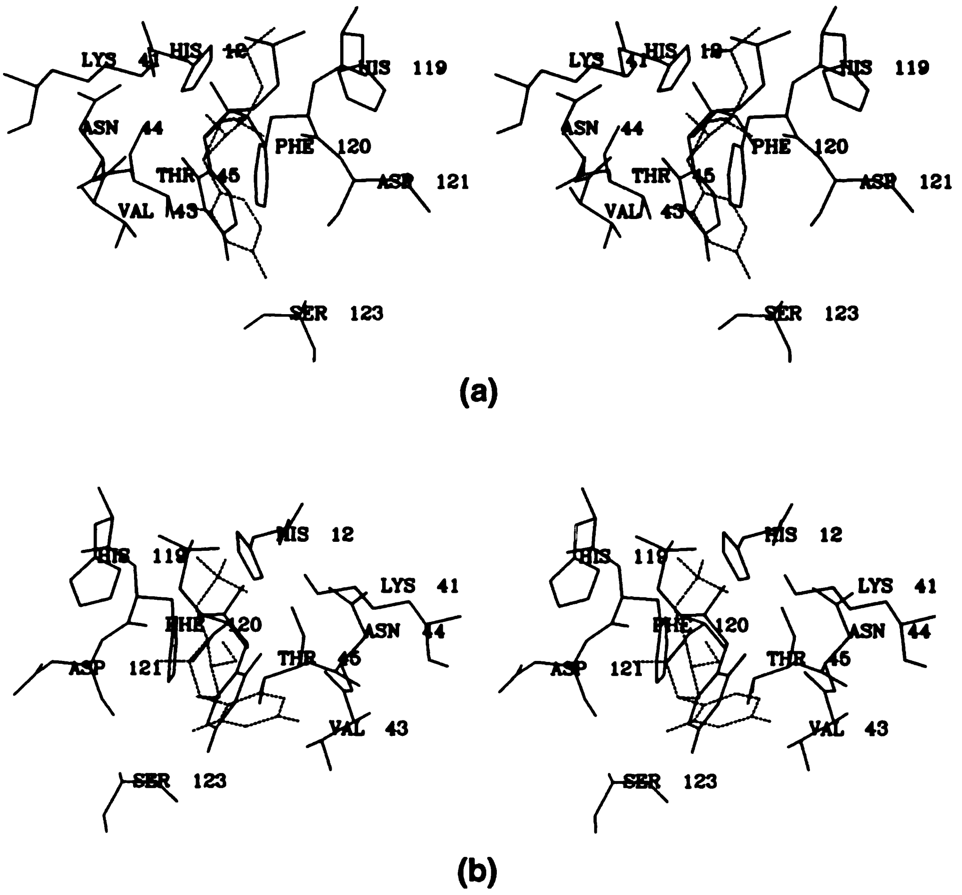


Figure 4.8: (a) The best flexible docked solution and the crystal conformations for ribonuclease A-uridine vanadate and (b) another solution which is very close in energy to the crystal conformation. The docked structure is dashed.

is only 0.5 kcal/mole. It appears that uridine-phosphate and uridine-vanadate bind in similar orientations. Figure 4.8 shows the best energy structure and the smallest RMS deviation structure.

### Timings

The timing results for the flexible docking simulations are also included in table 4.4. All simulations were run on a DEC Alpha running OSF/1. The times in the table indicate the time for one GA run of 1000 generations. In our method we propose that the GA should not be run until a majority of the bits have converged. Instead, it is computationally more efficient to run a number of shorter GAs in parallel. For all the complexes studied we found around five conformations close to the crystal structure in ten runs of the GA. The actual times are ten times the numbers shown in the table. Furthermore, if we wanted to automate the system or if none of the sub-populations could be eliminated by experimental (SAR) information, the GA would have to be run in the other eight regions as well.

We would like to emphasize that these timings guarantee that DIVALI finds structures with energies very close to the crystal structure energy. DIVALI has not optimized for speed, hence, the timings can be improved. For example, the results show that the time required by the GA scales approximately with the number of atoms in the ligand. This is expected since most of the time is spent in calculating the intramolecular energy. In the present implementation of the system, the variable component of the intramolecular energy is computed for every individual for every generation of the genetic algorithm. We could realize significant savings in run times by aborting the calculation after bad contacts are detected. Future versions of DIVALI will include this enhancement.

## 4.3 Discussion

The formation of a ligand-receptor complex results from a balance of contributions from direct binding, desolvation, entropic and and environmental effects. In addition protein ligand complexes exhibit fluctuations at many rates. These structural fluctuations help to determine the specificity of binding. It is well known that closely related molecules can bind quite differently, and sometimes the same molecule can bind with more than one conformation or binding mode with close binding energies. This implies that molecular recognition is inherently a dynamic phenomena. These features provide the major theoretical hurdle in our understanding of molecular recognition. We approximate this problem using an optimization procedure for computational tractability.

All molecular mechanics methods depend fundamentally on the quality of the energy function used. Clearly, molecular recognition is not equivalent to energy minimization, it is much more likely that the bound conformation is the global free energy minimum. Because binding is directed by electrostatic, van der Waals, and hydrophobic interactions, a well-established force field like AMBER is used in our studies [14, 15]. Hydrophobic interactions between two molecules depend on the existence of water. Solvation effects, for the most part, have been neglected. Nevertheless, in both the rigid and flexible ligand docking simulations, the energy of the crystal structure appears to be close in magnitude to the global minimum. Verifying this, however, would most likely require an exhaustive search of conformational space. In the case of rigid ligand docking, the lowest energy orientations were consistently within 0.6 Å of the crystal structure in all four complexes. Although the crystal structure appears to be close to the energy minimum in the flexible ligand docking simulations, a number of different conformations with similar energies can be found. With glycyl-L-tyrosine docking to carboxypeptidase A, for example, a low energy binding mode within 4 kcal/mole of the lowest energy structure in which the tyrosine ring points towards the surface of the protein is found. With uridine vanadate binding to ribonuclease A, there are many ligand conformations and orientations that are very close in energy but vary in RMS from 1.3 to 5.1Å from the crystal structure. This may be due to the peculiarities of substituting the vanadium with a phosphorus. The simulations of methotrexate binding to dihydrofolate reductase with the pteridine ring approximately fixed in the active site demonstrated that it is possible to find widely differing conformations with scores similar to the crystal structure's. Although the score of the crystal structure may be close to the global minimum of this particular scoring function, it is not the only such structure. This scoring function, therefore, lacks specificity. The implication of this is that there exists a limit to how well the energy can be calculated, and all structures within this limit should be included in the list of possible binding modes. A reasonably safe limit that many people have proposed seems to be about 5–7% of the crystal structure energy [12]. Each of the contributing terms to binding energy (electrostatics, hydrogen bonding, hydrophobic, loss of conformational, translational and rotational flexibility) may be as large as any other. Therefore, it is important to calculate each to within an accuracy of 2–3 kcal/mole.

Water molecules play an important role in molecular recognition, but this role is often underrepresented in theoretical studies. There are two related but different aspects to including water molecules. First is the role of the crystallographically known water molecules. Our simulations do not include any waters. An interesting question is whether we can simultaneously optimize the position of the ligand and water molecules known to bridge the interactions between the ligand and the protein. In all four of the complexes that we studied, we were able to find excellent structural agreement between the crystal structure and at least a few docked ligand conformations without including the bound waters. Interestingly, Guida, *et al.* [28], on the other hand, found that they had to include a continuum solvation model to obtain good agreement between the docked and crystal structures of thermolysin and its inhibitors once the crystallographic waters were removed. Second, we must account for the hydrophobic interactions that govern ligand binding. A solvation term might improve the rank ordering of the different conformations. For example, it might eliminate the flipped glycyl-L-tyrosine structure in the 3cpa complex. Historically, estimating solvation energies has been the most difficult part of binding mode energy evaluation. A number of continuum solvation models have recently appeared. Two such models include the Wesson and Eisenberg model [29] and the Clark Still model [30]. We are currently exploring different solvation models.

Given the computational requirements of molecular docking, our GA performed surprisingly well. For all four complexes, our GA with our new masking operator was able to find structures with energies close to the crystal structure energy. For all our rigid and flexible docking simulations except the docking of flexible methotrexate, it was necessary to mask the most significant bit of each of the translation components. The set of all such masks is equivalent to specifying eight mutually exclusive, collectively exhaustive binding mode hypotheses. For docking flexible methotrexate to dihydrofolate reductase, we were able to find structures with energies similar to crystal structure and RMS deviations less than 2Å with the hypothesis that the centroid of methotrexate is located within a region approximately 6 × 4 × 6Å and the angular rotations are within 90 degrees of the actual angles of rotation. Without masking the GA converges prematurely to a local minimum. As an alternative hypothesis for docking methotrexate to dihydrofolate reductase, we used the docked pteridine ring structure and allowed it a translational movement of about 1Å in each direction and an angular rotation of 22.5 degrees. With this docking hypothesis the GA found good solutions (energies close to crystal structure energy) for all the ten runs. These conformations vary in RMS error from 3 to 10Å from the crystal structure. This implies that a variety of binding modes are possible even with a good potential energy function. We would like to point out that DIVALI reproduces the conformations (less than 1.5 Å error) of the ligands in all four systems once the translational and rotational degrees of freedom have been removed.

There are a number of improvements that are being explored. For instance, to address the premature convergence of the GA, one could introduce other genetic operators that are designed to maintain the diversity of the population. We are presently investigating the utility of these operators. Convergence of the population is one important characteristic of most GA applications. Our results do not show any convergence of the population, this is because of the extremely rugged nature of the the van der Waals function. In our experience this non-convergence of the population has not been a negative feature. In molecular docking with the AMBER [14] potential function, the energy landscape is extremely rugged. Because of the van der Waals interactions, a slight displacement of an atom can result in a several orders of magnitude change in energy. Consider, for example, structures where just one atom forms a bad contact with the receptor and all other atoms are in the desired positions for optimum binding. These structures will result in a large energy, and because of our selection procedure, they will most likely not survive to subsequent generations. Furthermore, once a low energy structure is found, it tends to dominate the population with a corresponding loss in population diversity. This can result in premature convergence, longer runs, and a decreased likelihood of finding the correct binding mode. We are presently trying different selection procedures and multiple optimization functions to address this problem. Our new selection procedure allows individuals with larger energies, but fewer bad contacts to survive to the next generation. Similarly, the multiple optimization function tries to minimize the energy and the number of bad contacts simultaneously.

Much of the success of our system is due to our novel masking operator which by fixing certain schemata, provides a means for maintaining different subpopulations. This operator provides a convenient mechanism for trying out different binding hypotheses. Experimental data on the binding mode could be used to constrain the population. A 1-D NMR spectra of the complex, if a well-resolved structure of the receptor is known, can be used for this purpose. If the ligand is capable of forming a large number of specific interactions with the receptor, then QSAR studies can constrain the possible binding sites for the ligand. In all of our docking simulations we found that it was necessary to provide additional information to the genetic algorithm. We had to decompose the full docking problem into a smaller, more manageable problem for the GA. In the rigid docking of glucose to the periplasmic binding protein, for instance, the crystal structure orientation was only found after localizing the search to a particular region of the receptor with the masking operator. The procedure used to dock MTX can be used as part of a fragment based design strategy to computationally create novel ligands. Thus, experimental data about the actual binding mode can be incorporated easily using the masking operation.

The masking operation provided another advantage also. It greatly improved the performance of our system and made it more like the ideal system that we described in the introduction. By decomposing a docking problem into a set of mutually exclusive and collectively exhaustive smaller docking problems, the masking operator was used to improve both the thoroughness of the search as well as the consistency of obtaining the solution. For all the systems studied we found the solution on average about five times out of ten with the masking operator. The masking operator is easily automated. Therefore, since GAs are easily parallelizable, it is possible to extend the present system so that each processor explores a different binding hypotheses.

# Bibliography

[1] I. D. Kuntz. Structure-based strategies for drug design and discovery. *Science*, 257(5073):1078–82, 1992.

[2] N. Pattabiraman, M. Levitt, T. E. Ferrin, and R. Langridge. Computer graphics in real-time docking with energy calculation and minimization. *J. Comp. Chem.*, 6(5):432–436, 1985.

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

[4] E. C. Meng, B. K. Shoichet, and I. D. Kuntz. Automated docking with grid-based energy evaluation. *J. Comp. Chem.*, 13(4):505–524, 1992.

[5] 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.

[6] F. Jiang and S. H. Kim. "Soft docking": matching of molecular surface cubes. *J. Mol. Biol.*, 219(1):79–102, May 1991.

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

[8] A. K. Ghose and G. M. Crippen. Geometrically feasible binding modes of a flexible ligand molecule at the receptor-site. *J. Comp. Chem.*, 6:350–359, 1985.

[9] D. S. Goodsell and A. J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins*, 8(3):195–202, 1990.

[10] T. N. Hart and R. J. Read. A multiple-start Monte Carlo docking method. *Proteins*, 13(3):206–222, 1992.

[11] A. R. Leach and I. D. Kuntz. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comp. Ch.*, 13(6):730–748, 1992.

[12] I. D. Kuntz, E. C. Meng, and B. K. Shoichet. Structure-based molecular design. *Acc. Chem. Re.*, 27(5):117–123, 1994.

[13] D. E. Goldberg. *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley Pub. Co., Reading, Mass., 1989.

[14] S.J. Weiner, P. A. Kollman, D.A. Case, U. C. Singh, C Ghio, G. Alagona, S. Profeta Jr., and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765, 1984.

[15] S.J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. *J. Comp. Chem.*, 16:548, 1986.

[16] Molecular Modeling System SYBYL. Version 5.4 Tripos Associates inc. 1991.

[17] M. L. Connolly. *J. Appl. Crystallogr.*, 16:548, 1983.

[18] M. L. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–13, Aug 1983.

[19] J. H. Holland. *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, Ann Arbor, MI, 1975.

[20] Nicol N Schraudolph. Tech. Rep. CS92-249. UC San Diego, 1990.

[21] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer, 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(3):535–542, May 1977.

[22] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, 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. *Arch. Biochem. Biophys.*, 185(2):584–591, Jan 1978.

[23] N. K. Vyas, M. N. Vyas, and F. A. Quiocho. Sugar and signal-transducer binding sites of the Escherichia coli galactose chemoreceptor protein. *Science*, 242(4883):1290–1295, 1988.

[24] D. W. Christianson and W. N. Lipscomb. X-ray crystallographic investigation of substrate binding to carboxypeptidase A at subzero temperature. *Proc. Natl. Acad. Sci. U. S. A.*, 83(20):7568–7572, 1986.

[25] J. T. Bolin, D. J. Filman, D. A. Matthews, R. C. Hamlin, and J. Kraut. Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 åresolution. I. general features and binding of methotrexate. *J. Biol. Chem.*, 257(22):13650–13662, 1982.

[26] B. Borah, C. W. Chen, W. Egan, M. Miller, A. Wlodawer, and J. S. Cohen. Nuclear magnetic resonance and neutron diffraction studies of the complex of ribonuclease A with uridine vanadate, a transition-state analogue. *Biochemistry*, 24(8):2058–2067, 1985.

[27] 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.

[28] W. C. Guida, R. S. Bohacek, and M. D. Erion. Probing the conformational space available to inhibitors in the thermolysin active site using Monte-Carlo energy minimization techniques. *J. Comp. Chem.*, 13(2):214–228, 1992.

[29] L. Wesson and D. Eisenberg. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein. Sci.*, 1(2):227–235, 1992.

[30] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. S.*, 112(16):6127–6129, 1990.

# Chapter 5

# Flexible Ligand Docking Using Genetic Algorithms and Binding Free Energy Estimates

Many of the functions performed by biological molecules depend on appropriate interactions with other molecules. Small organic molecules, for example, can bind specifically to biological macromolecules and modulate their function. The goal of computer-aided drug design is the ability to design putative ligands that bind to a particular therapeutically relevant target protein. If the three dimensional structure of the target molecule is known, then the problem can be formulated as the molecular docking problem. Within this context, the problem is to find the best conformation for the intermolecular complex. This requires an efficient algorithm for searching the ligand's conformational space within the receptor and a scoring function for ranking the binding mode. Molecular docking systems provide a means for identifying the chemically plausible structures of a ligand-receptor complex and may ultimately allow for the screening of large databases of flexible ligands for new lead compounds.

Automated ligand docking has been the subject of intense research. DOCK [1], one of the first molecular docking systems, is still widely used even though it considers both the ligand and the receptor to be rigid. It can rapidly screen databases of small molecules for new lead compounds. Computational screening of flexible ligands, however, is still not practical. Other methods for screening databases of rigid ligands include CLIX [2] and FLOG [3]. The conformational space of each ligand is represented by a number of low energy conformations in the latter. A more detailed review of these techniques is included in Kuntz *et al.* [4]. In Chapter 2, I discussed some of the different approaches to flexible ligand docking. A brief review of this discussion is described below in order to highlight the issues addressed in this chapter.

One approach to flexible ligand docking is to use energy minimization techniques. Goodsell and Olson [5], for example, used simulated annealing to dock a flexible ligand into a receptor. GRID [6] interaction energy maps were used to more efficiently compute the energy of the intermolecular complex. Hart and Read [7] describe a multi-start Monte Carlo system for flexible ligand docking. Yamada and Itai [8, 9] developed ADAM, another energy minimization technique. Their system includes flexibility in the ligand through a systematic search of the torsion space once an initially docked structure is minimized with AMBER [10]. DiNola *et al.* [11] describe a molecular dynamics approach to the docking of phosphocholine into McPC603. Because of the complex topology of the energy landscape, these techniques have long run times and are computationally too expensive for docking simulations.

Another approach is to use fragment joining techniques. These systems either dock the individual fragments of a ligand into a receptor and then join them together or, more commonly, start with an initially docked fragment and "grow" the complete ligand. This technique forms the basis of many *de novo* design systems. Some of these include GROW [12], LEGEND [13], LUDI [14, 15], and GroupBuild [16]. One of the first fragment-based ligand docking systems is described by DesJarlais *et al.* [17]. Their system included only partial ligand flexibility. Leach and Kuntz [18] describe an incremental construction approach that begins with an initially docked structure. Rarey *et al.* [19] developed FLEXX, another incremental construction technique. In their system a base fragment is chosen interactively and docked using a pose clustering algorithm [20]. For scoring the positions of the individual fragments and the entire intermolecular complex, a modified version of Böhm's binding free energy prediction function [21] is used. Welch *et al.* [22] describe a similar system. The docking procedure begins with an initially docked fragment and "grows" the ligand.

The scoring function is also based on Böhm's binding free energy estimate, but in their approach it is differentiable, and the ligand's conformation is optimized using a gradient descent algorithm. Although the fragment-based methods perform very well, there are a number of limitations with these approaches. First, the base fragment for the docking procedure must be chosen interactively. In fact, the choice of the initial fragment can have a significant impact on whether the ligand's true binding mode will be found. Second, the actual placement of the base fragment is context dependent. Ligands are known to deform when they bind to a receptor [23]. There is no guarantee that the orientation of the base fragment will be the same in the ligand as it is when the base fragment is docked independently. Rotstein and Murcko [16] have shown that it is difficult to dock ligands if the initial fragment is not oriented properly. Third, Rarey *et al.* [19] found that there are many potential sites in which the base fragment can dock. Each of these initial fragments could be tried separately, but how many different initial fragment locations have to be tried to identify micromolar or better inhibitors? Finally, Rarey *et al.* use a "greedy" construction algorithm to add subsequent fragments to the ligand. As described in Chapter 2, "greedy algorithms" are best applied to problems where the optimal solution to the problem is the solution that is obtained by choosing the best orientation of the fragment at each step of the construction procedure. This technique may perform very well for highly optimized ligand- receptor interactions, but in general, we will not have such a structure. This is the purposes for this flexible ligand docking system.

Finally, another approach is to use a genetic algorithm (GA) [24] to search the conformational space of the ligand. Judson *et al.* [25] published the first account of the use of genetic algorithms for flexible ligand docking in 1994. In their system an initial fragment is docked, and the ligand is grown. Oshiro *et al.* [26] describe a flexible ligand docking extension to DOCK. The algorithm encodes a mapping between the ligand atoms and sphere centers which form the negative image of the receptor site as well as the torsion angles of the ligand. In earlier work, we described our system, DIVALI [27]. We developed a method which combines a systematic search option with a genetic algorithm search procedure by introducing a new genetic algorithm operator. The primary focus of this work was to investigate how much information would be required to consistently find the crystal structure binding mode. We also explored the approach of docking with an initial base fragment docked only approximately in the correct orientation. Jones *et al.* [28] describe a flexible ligand docking system where the genetic algorithm encodes the putative mapping between hydrogen bond donors and acceptors on the ligand and the receptor in addition to the torsion angles. All of these systems employ very similar scoring functions which consist of a van der Waals term and either an electrostatic term or hydrogen bond energy. The results are also very similar. The systems converge to local minima and often require many runs to find the best solution. Furthermore, solutions with energies similar to the crystal structures are often found.

Despite the limited success of the genetic algorithm-based methods described above, we believe that these approaches have a great potential for docking flexible ligands to rigid, and ultimately, flexible receptors. Some problems are known to be "GA deceptive [24]" and are difficult to solve using genetic algorithms because good partial solutions can lead to poor global ones. This can occur when van der Waals penalties in the binding site can cause the genetic algorithm to search elsewhere for a solution. Therefore, in formulating a problem for a genetic algorithm, one must carefully choose the scoring function and the representation of the optimization parameters (see the next section for a detailed description of genetic algorithms). The representation determines how the conformation space is searched. It provides a way to specify which interactions are important. Jones *et al.* [28], for example, used putative mappings between hydrogen bond donors and acceptors. The scoring function, on the other hand, is linked directly to the dynamics of the genetic algorithm through the selection operator. If a conformation does not score well even though it is close to the crystal structure, then it may not survive to the next generation. The issues of representation and scoring function are very closely linked, and both have to be considered before applying a genetic algorithm to the flexible ligand docking problem. The approaches above have employed very similar scoring functions with different representations. The goal of this work is to develop a new genetic algorithm-based flexible ligand docking system that is based on a more GA conducive scoring function. We leave the issue of representation to future work, and we encode the ligand translations, euler angles, and bond rotations in the genetic algorithm. Although this is a rather naive representation, it represents an intermediate step between other representations and the transformation of the ligand. There are applications where this would be useful. For example, this provides an excellent framework for the chemist to introduce hypotheses about the ligand's binding mode.

In this paper we describe a new flexible ligand docking system that uses genetic algorithms. We introduce a smoother scoring function, we change the basic genetic algorithm operators to more efficiently search the conformational space of the ligand, and finally, we show that this system can be applied to a variety of different problems in docking flexible ligands to rigid receptors. In particular, we describe three different simulations. First, we show that this approach can be used to dock ligands with initially docked base fragments with results similar to the other methods. We also assume that the initial fragment is docked incorrectly and show that the system can dock the base fragment in context with the entire ligand. The centroid of the fragment is allowed to move up to 2Å along the $x$, $y$, and $z$ directions, and the Euler angles are allowed to vary over a range of 45 degrees. Finally, we also dock the ligands using the method described in our previous work [27]. The primary purpose of the last simulation is to evaluate the scoring function. The centroid of the ligand is allowed to vary within a box of approximately 10Å on a side. These simulations are applied to seven different ligand/receptor systems which have differing sizes and bind through a variety of interactions.

## 5.1 Methods

### Genetic Algorithms

Reviewing briefly the discussion of genetic algorithms from Chapter 4, a genetic algorithm is an optimization procedure that combines a probabilistic search algorithm with a directed search strategy based on the fitness function. Genetic algorithms have been shown to search complex spaces robustly [24] and differ from other optimization techniques in a number of ways. First, they explore the different regions of the

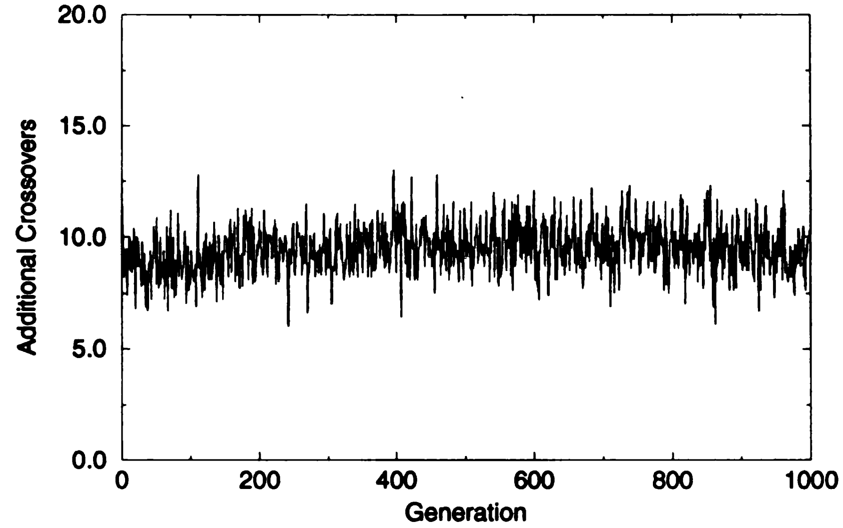


Figure 5.1: The effect of the new crossover operation on the genetic algorithm performance. This graph show that on average fewer than 10 additional crossover operations are required to generate 128 conformations of methotrexate with no bad internal contacts.

solution space with a population of points. Second, they only need to compute the value of the function being optimized. They are not limited by the details of the function such as continuity or the existence of derivatives. Third, they do not manipulate the optimization parameters individually. The parameters are instead coded as a continuous string (usually binary digits), and the algorithm modifies this string without any direct knowledge of individual parameters. A genetic algorithm can simultaneously process local and non-local information within the string. This feature is known as implicit-parallelism [29] and is one of the most interesting strengths of genetic algorithms.

A genetic algorithm maintains a population of individuals with an associated fitness. Each individual represents a possible solution to the problem, and the algorithm alters the individuals of the population and thereby searches different regions of the solution space in two separate stages. During the selection stage a new population is created by proportionally selecting the more fit individuals from the previous population. The members of the populations are then transformed in the alteration step.

The mutation operator creates a new individual from a single individual by randomly changing an element in its representation. The crossover operator, on the other hand, exchanges information between two (occasionally more) members of the population. We use a two-point crossover operator where two points are chosen at random, and the information between the two points is exchanged. We chose this implementation because there is evidence that it works better than a single point crossover [24]. The selection step is based on a simple heuristic idea that we would expect to find the best solution in that region of space which contains a large number of good solutions. Mutation and crossover operators extend the region to the space of potential solutions. Figure 4.1 illustrates these operators. In our earlier work [27], we observed that many of the individuals within the population contained bad intramolecular van der Waals interactions and that the conformational space of the ligand was not being search efficiently. Therefore, we have changed the basic crossover and mutation operator to avoid this. In each case, before the conformation is included in the population a bump check between the ligand atoms is performed. For all atoms $A_i$ and $A_j$ that are greater than four bonds apart, we compare the distance between the two atoms $r_{ij}$ with the sum of the van der Waals radii, $R_i$ and $R_j$. If $r_{ij} < R_i + R_j - 0.3$, then the structure is not included in the population. The atoms are allowed to overlap by 0.3Å. The mutation or crossover operation is performed again. This continues until the population is complete or the appropriate number of mutations have been performed. Figure 5.1 shows the performance cost of the modified crossover operator. The simulation described in the figure is the docking of methotrexate to dihydrofolate reductase (4dfr) for a population size of 256 with a crossover rate of 0.5. The genetic algorithm was run for 1000 generations. In order to generate 128 new members of the population from crossover, fewer than 10 additional crossovers must be performed at each generation. This approach not only assures us that our population does not contain bad intramolecular van der Waals contacts, but also improves the overall efficiency of the algorithm since we do not bother to evaluate individuals which would ultimately score poorly. Our implementation of the genetic algorithm is based on the GAucsd package [30].

A number of parameters are associated with a genetic algorithm simulation. One must choose the population size, the number of generations, the crossover rate, and the mutation rate. One of the primary focuses of this work was to try to determine a set of parameters that would work across all of the receptor-ligand complexes. This, however, must be determined within the limitations of the other components of the system such as the scoring function. We chose the population size and the number of generations so that the genetic algorithm would perform almost as well for all the systems. For the simulations with the base fragment docked only approximately in place and with no initially docked fragment, we chose the the population size and number of generations to be 256 and 1500, respectively. For a number of systems this was far more computations than were necessary to solve the problem. These parameters have a dramatic effect on the length of the simulation. The population size controls the extent to which the solution space is searched in each generation. The number of generations determines how long the genetic algorithm is run. As shown in Chapter 4, for some formulations of the molecular docking problem many runs have to be performed to find the solution. Therefore, the simulations do not converge to the same solution, and they must be terminated. Doubling either one will approximately double the simulation time. For the simulations with an initial fragment docked in the crystal structure orientation, the genetic algorithm has less space to search and these parameters can be changed to reflect this. For these simulations, the population size and the number of generations were chosen to be 128 and 1000, respectively. The crossover and mutation rates determine the convergence of the population. The two most important operators for the genetic algorithm are the selection procedure and the crossover operator. These are the primary operators for the convergence of the optimization procedure. The mutation rates are traditionally chosen to be small, and their purpose is to introduce a little diversity into the population. For our simulations we chose the crossover and mutation rates to be 0.5 and 0.005, respectively. This choice of parameters slows down the converge of the genetic algorithm, and allows the genetic algorithm to explore more of the solution space.

The genetic algorithm encodes three translations ($T_x$, $T_y$, $T_z$), three rotations ($R_x$, $R_y$, $R_z$), and a number of bond rotations. The translational components usually specify the position of the centroid of the molecule, but they can also specify the transformed coordinates of a given atom. The rotational components are the Euler angles of rotation. As in most applications of genetic algorithms we chose to use a Gray-coded binary representation for these parameters. With Gray-scale coding the closest values within the representation differ by at most one bit [24]. Because this representation is periodic, it is well suited to representing rotations. The number of bits for each translation component $b_i$ is given such that $2^{b_i} \geq 2L_i/0.3$ where $L_i$ is the dimension of the box in the $i$th direction. All rotations of 360 degrees whether Euler angles or bond rotations are represented with 10 bits of precision.

## Scoring Function

One of the major requirements of the scoring function is that it should be "smooth" so that good local solutions lead to good global ones. We choose a scoring function that is based on the one developed by Böhm [21]. The scoring function estimates the binding free energy of a ligand, and it consists of primarily of terms that measure how different types of interactions enhance the binding affinity. There are no large penalty terms that will cause the genetic algorithm to converge to regions far from the actual binding site. Böhm represented the binding free energy as a sum of several types of interactions and performed a linear regression using several crystal structures. In its functional form, the Böhm binding free energy estimate is

$$\Delta G_{binding} = \Delta G_0 + \Delta G_{hbond} + \Delta G_{ionic} + \Delta G_{hydration} + \Delta G_{rotation} \tag{5.1}$$

where

$$\begin{aligned}
\Delta G_{hbond} &= k_{hbond} \sum_{hbonds} w(r, \alpha) \\
\Delta G_{ionic} &= k_{ionic} \sum_{ionic} w(r, \alpha) \\
\Delta G_{hydration} &= k_{hydration} A_{hydration} \\
\Delta G_{rotation} &= k_{rotation} N_{rotation}
\end{aligned}$$

$\Delta G_0$ represents the loss of overall translational and rotational degrees of freedom. $\Delta G_{hbond}$ and $\Delta G_{ionic}$ are the free energies of hydrogen bonded and ionic interactions, respectively. $\Delta G_{hydration}$ is the free energy due to burying hydrophobic surface area $A_{hydration}$. $\Delta G_{rotation}$ is the free energy lost from freezing the $N_{rotation}$ rotatable bonds in the ligand. The weighting function $w(r,\alpha)$ penalizes deviations from the ideal hydrogen bond or ionic interaction geometries. It has a distance dependence $w_1(r)$ where $r$ represents the deviation from the ideal hydrogen bond or ionic interaction distance of 1.9 Å and 2.0 Å, respectively. The angular dependence $w_2(\alpha)$ represents the deviation from the ideal hydrogen bond angle of 180°.

$$w(r,\alpha) = w_1(r)w_2(\alpha) \tag{5.2}$$

where

$$\begin{aligned} w_1(r) &= \begin{cases} 1 & r \leq 0.2\text{Å} \\ 1-(r-0.2)/0.4 & 0.2 \leq r \leq 0.6\text{Å} \\ 0 & r > 0.6\mathring{A} \end{cases} \\ w_2(\alpha) &= \begin{cases} 1 & \alpha \leq 30^o \\ 1-(\alpha-30)/50 & 30 \leq \alpha \leq 80^o \\ 0 & \alpha > 80^o \end{cases} \end{aligned}$$

Bohm determined the values of the regression constants, $k_{hbond}$, $k_{ionic}$, and $k_{hydration}$, to be -4.7 kJ/mol, -8.3 kJ/mol, and -0.17 kJ/mol, respectively. Because our docking studies which involve comparisons between different conformations of the same ligand, we assume that $\Delta G_0$ and $\Delta G_{rotation}$ do not vary much and are, therefore, assumed to be constant.

In order for the above formulation to be useful within the context of optimization with a genetic algorithm, a number of changes were made. First, the estimation of the area of hydration $A_{hydration}$ is computationally too expensive to calculate at each step of the procedure. This would involve $p \times g$ calculations of the area of hydration where $p$ is the population size and $g$ is the number of generations. The purpose of this term is to model the burial of hydrophobic groups on the ligand away from the solvent and into hydrophobic pockets on the receptor. A simpler form is used during the search phase of the algorithm.

$$\Delta G'_{hydration} = k'_{hydration} \sum_{atoms} w_h(r) \tag{5.3}$$

where

$$w_h(r) = \begin{cases} 0 & r \leq -0.3\text{Å} \\ 1 & -0.3 < r \leq 0.7\text{Å} \\ 1 - (r - 0.7)/0.5 & 0.7 < r < 1.2\text{Å} \\ 0 & r \geq 1.2\text{Å} \end{cases}$$

$r$ is the difference between the distance between the ligand and receptor atoms minus the sum of their respective van der Waals radii. This term represents the number of lipophilic atoms within 5.0 Å of the ligand atom, but those atoms that are closer are weighted more heavily. This is similar to the idea described by Bohacek and McMartin [31] for characterizing hydrophobic pockets in receptors. The summation above is taken over the ligand atoms as well as the receptor atoms, and this allows us to consider intramolecular lipophilic interactions. In these simulations $k'_{hydration}$ is -0.25. During the first half of the simulation, however, this term is not computed so that the genetic algorithm can identify the surface and those regions where other interactions can be formed. The primary reason for this approach is that hydrogen and ionic bonding terms have some directionality that can be used to effectively place the ligand.

The Böhm estimate of the binding free energy does not have a van der Waals term. Instead of penalizing atoms that make van der Waals contact with the receptor, we reward those that do not. Other docking techniques [1, 18] have used similar rewarding functions, but unlike those approaches, we do not penalize bad contacts. This results in van der Waals interactions which are much smoother. It allows the genetic algorithm to find the conformation and orientation that places the ligand atoms at the surface where they can make other favorable hydrogen bond, ionic, or lipophilic interactions. The contact score is

$$f_{contact} = k_{contact} w_{contact}(r) \tag{5.4}$$

where

$$w_{contact}(r) = \begin{cases} 1 & 2.5\AA \leq r \leq 4.8\AA \\ 0 & \text{otherwise} \end{cases}$$

and $k_{contact}$ was chosen to be 2.0. The genetic algorithm keeps track of the contact score for each member of the population so that the solutions with good contact scores can be identified, and the contact contribution to the score can be subtracted. A good contact score implies that most of the atoms lie adjacent to the receptor surface, and few form bad van der Waals interactions with the receptor. The scoring function is

$$\text{Score} = \Delta G_{hbond} + \Delta G_{ionic} + \Delta G'_{hydration} + f_{contact} \tag{5.5}$$

There is no intramolecular energy calculation for the ligand. We want our algorithm to be able to find the crystal structure when a ligand undergoes some deformation to form the intermolecular complex. We have instead changed the basic crossover and mutation operators of the genetic algorithm so that the population contains only conformations that do not have bad intramolecular contacts. This improves the efficiency of the conformational search of the ligand without too large a computational penalty. In essense, the receptor is represented by a shell and the genetic algorithm searches through the conformational and orientational space of the ligand and tries to find orientations that place most of the atoms close to the surface in such a way that favorable hydrogen bonding, ionic, or lipophilic interactions can be made. Because the scoring function is based primarily on distances between ligand and receptor atoms, it can potentially be used in conjunction with flexible receptor docking simulations. Here, however, we consider the receptor to be rigid. Therefore, for convenience of implementation and efficiency, we have implemented much of the scoring function on a grid with a spacing of 0.25 Å.

## Receptor-ligand Complexes

A set of seven receptor-ligand complexes were used to test the flexible ligand docking system. Six of these are available from the Protein Data Bank [32, 33], and the last one is an HIV-1 protease inhibitor that was developed at Vertex Pharmaceuticals

| PDB file | Receptor/ligand complex |
|---|---|
| 1stp | Streptavidin/biotin |
| 3cpa | Carboxypeptidase-A/glycyl-L-tyrosine |
| 121p | H-RAS P21/5′-B,G-methylene-triphosphate |
| 4dfr | Dihydrofolate reductase/methotrexate |
| 3dfr | Dihydrofolate reductase with NADPH/methotrexate |
| p478 | HIV-1 protease/VX-478 |
| 1dwc | Thrombin-argatroban |

Table 5.1: A list of the complexes used to test this system.

and is now in the clinical trials stage of development. These systems were chosen because of the differing sizes and degrees of flexibility in the ligands and the types of interactions that are involved in binding. These complexes are a subset of the systems analyzed by Rarey *et al.* [19]. Jones *et al.* [28] chose systems where the hydrogen bonds were the primary interactions involved in the formation of the complex. Many of the system have major hydrophobic regions in addition to hydrogen bonding and ionic interactions. Table 5.1 shows a list of these complexes. In the remainder of this section, we describe the steps taken to prepare the ligands and receptors for the docking simulations.

The ligands were extracted from the PDB file and converted into SYBYL "mol2" file format. Using the SYBYL system, hydrogen atoms were added and the correct atom and bond types were assigned to the molecule. Rotatable bonds were rotated at random, and the molecule was randomly oriented at the origin of the crystal structure coordinate system. For the receptor, all water molecules were removed, and polar hydrogens were added using the "protonate" program of the AMBER [34] system. No optimization of the ligand nor receptor was performed. The physico-chemical properties of the atoms within the active site were assigned using software and template files that we developed. The scoring function described above was evaluated on a scoring grid with a spacing of 0.25 Å. The locations of hydrogen bond acceptors on the receptor were stored on the grid since the score depends on the orientation of the ligand's donor group and cannot be precalculated.

## 5.2 Results

In this section we describe the results of three different simulations on seven different receptor-ligand complexes. First, we dock a flexible ligand given that we know the position of a base fragment. Then, we assume that the base fragment is positioned incorrectly and dock the flexible ligand while correcting the base fragment orientation. The error in the position of the initial fragment can be as much as 2 Å in the $x$, $y$, and $z$ directions, and the rotations about each of the axes can be off by as much as 45 degrees. Finally, the last simulation performed on each of the systems assumes only that the centroid of the ligand must lie within a box with a dimension on each side of approximately 10Å. This simulation allows us to test the usefulness of the new scoring function and to compare our results with those that we have published previously [27]. Fifty runs were performed in each case. In the remainder of this chapter, the results for each complex are discussed in the following subsections. Table /reftbl:FlexDockResults summarizes the performance of the docking procedure on each of the different types of simulations.

### 1stp

Biotin binds to streptavidin with a very high affinity due to the formation of specific hydrogen bonds and overall steric fit. The structure of the complex has been solved to 1.55 Å resolution [35]. Biotin contains a fused ring system with a ureido group that is linked to a carboxylate group through a four member alkyl chain. The dominant effect contributing to biotin binding is the polarization of the ureido group so that the negative charge is located on the ureido oxygen which binds in an oxyanion "hole" on the protein. Biotin contains five rotatable bonds, and the fused ring system is used for the fragment-based docking simulations.

For the simulations with the initial fragment docked in the crystal structure orientation, the correct binding mode was found in all runs. The solutions varied from 0.16 to 1.02 Å rms deviation from the crystal structure, and all solutions were within 5 kJ/mol. Most of the variations in the solutions were due to the alkyl chain. When the

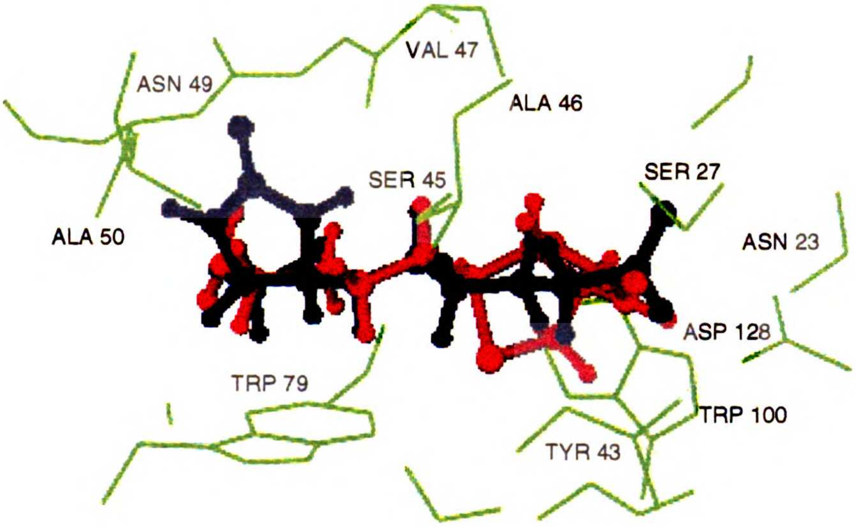


Figure 5.2: The crystal structure binding mode of biotin to streptavidin (red) and another solution that is 5.94Å rms deviations away (blue).

initial fragment was only approximately placed in the crystal structure orientation, the performance was still very good. Most of the solutions were between 0.67 and 1.53 Å rms from the crystal structure. In these simulations the genetic algorithm quickly converged to solutions with good contact score. Most of the atoms were placed near the receptor surface, and no bad van der Waals contacts were made with the receptor.

Without the initially docked fragment, there were a variety of other solutions found. In this case the centroid of the ligand was allowed to vary over a box with dimensions 9 × 9 × 12Å. Most of the simulations converged to solutions close to the crystal structure. The closest solution was at 0.88 Å rms deviation from the crystal structure. Another potential binding mode was found at 5.94 Å from the crystal structure and had a score that was approximately 5 kJ/mol less than the score of the crystal structure. This difference in energy is slightly more than the energy of one hydrogen bond. This solution is shown with the crystal structure in Figure 5.2. The proximity of these two solutions in energy, however, may be due to the fact that water molecules and the polarization of the ureido group have not been included.

The carboxylate group on the biotin is known to form additional hydrogen bonds with water molecules. In nearly all simulations the solutions had good contact scores.

## 3cpa

Carboxypeptidase A is a zinc protease. The structure of the complex with glycyl-L-tyrosine is known to 2.0 Å at low temperatures [36]. The hydroxybenzyl group of the tyrosine resides in the hydrophobic pocket of the $S_1'$ site. The dipeptide has five rotatable bonds and appears to bind with the hydrolytically important zinc-bound water displaced from the active site. As in Rarey *et al.* [19], the carboxylate group of the glycine residue is chosen as the base fragment.

With the initial fragment docked in the crystal structure orientation, the genetic algorithm quickly found the crystal structure conformation. All solutions were within 3 kJ/mol of each other, and most of the variation in the rms deviation from the crystal structure was in the placement of the hydroxybenzyl group. These results are not surprising given that the binding site is a very deep pocket, and a fragment of the ligand is initially docked. When error is introduced into the base fragment, more variability is seen in the solutions found by the docking procedure. The most commonly found solution is very close to the crystal structure, but other solutions are found between 2.2 and 2.6 Å rms deviation from the crystal structure. In all simulations, the final solutions had good contact scores. Without the carboxylate group initially docked, the docking procedure again found a variety of solutions including structures close to the crystal structure. It is important to point out that the genetic algorithm did converge prematurely to the solution that is approximately 6 Å rms from the crystal structure as in our previous work [27]. However, in these simulations this inverted solution is not found as often, and the energy difference between this structure and the crystal structure is greater than 14 kJ/mol. Thus, with carboxypeptidase-A the new scoring function improves the results of the simulations (see Table 5.2).

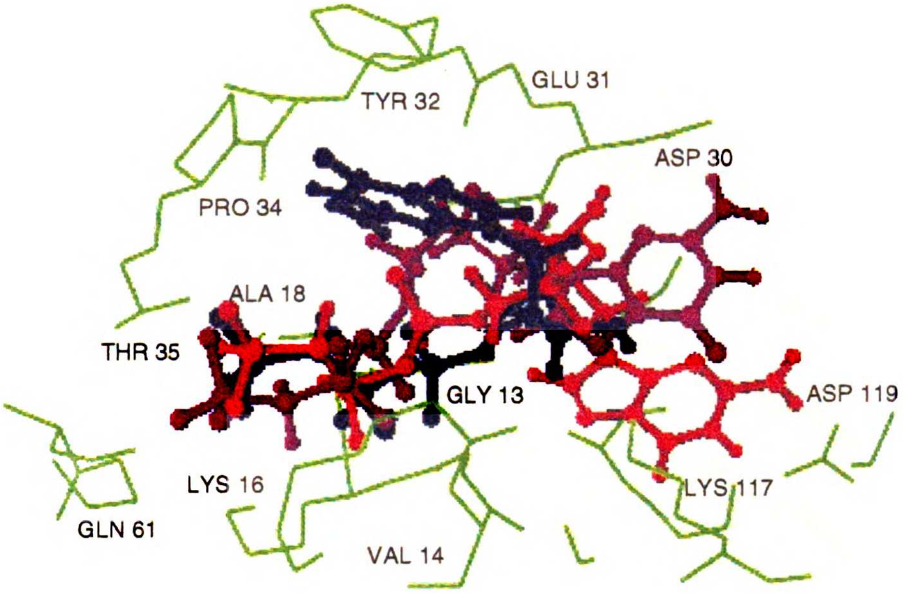


Figure 5.3: A couple of solutions from the 121p docking simulations. The crystal structure is shown in red. The 3.35 and 5.08Å solutions are shown in magenta and blue, respectively.

## 121p

The structure of the oncogene protein H-RAS P21 and bound guanosine 5′-B,G-methylene-triphosphate has been solved to a resolution of 1.54 Å [37]. The $Mg^{2+}$ is included in the active site. The inhibitor has seven rotatable bonds, and the guanosine group is chosen as the initial fragment.

For the docking simulations with the base fragment fixed, the crystal structure binding mode was found in every run. The solutions varied from 0.27 to 1.05 Å. When the initial fragment is docked incorrectly, and the docking algorithm has to reposition it within the context of the entire ligand, most of the simulations generated solutions that were within 1.4 Å rms deviations from the crystal structure. These structures had the best scores. The best solution in terms of score and proximity to the crystal structure had an rms deviation of 0.81 Å. A few simulations that did not converge to a conformation close to the crystal structure did not have near optimum contact scores. Many atoms were positioned away from the receptor surface. Longer run

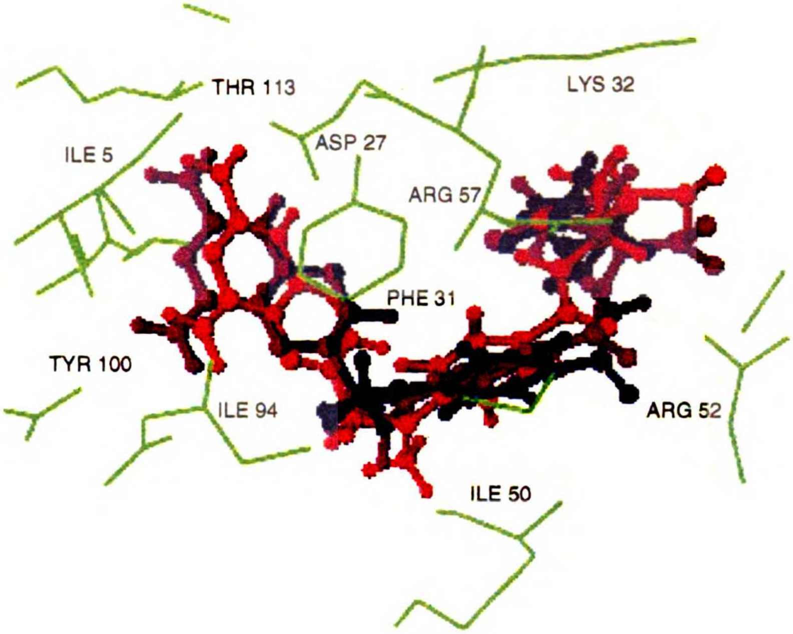


Figure 5.4: Results from the 4dfr docking simulations. The crystal structure binding mode is shown in magenta. The 1.64 and 2.5Å solutions are shown in blue and red, respectively.

times may help solve this.

When the centroid of the ligand is allowed to vary within a box with dimensions 15 × 10 × 10Å, three distinct binding modes appear. The solution closest to the crystal structure has an rms deviation of 1.68 Å. This solution also had the best score which turns out to be about 3.6 kJ/mol better than the crystal structure solution without any optimization of the crystal structure. The other binding modes are 3.35 and 5.08 Å from the crystal structure. The scores for these structures are 3.3 kJ/mol and 4.5 kJ/mol worse than the best scoring solution. These solutions are shown in Figure 5.3. As above longer simulations may be required to get better contact scores. The rms deviations in these structures are due to the placement of the guanosine group.

## 4dfr

Dihydrofolate reductase is an enzyme which is vital for the replication of DNA and has thus been the target for anti-bacterial and anti-cancer drugs. The structure of methotrexate bound to dihydrofolate reductase is known to 1.7 Å resolution [38]. There are two independent protein molecules in the PDB file; we chose the molecule designated "B". Dihydrofolate reductase has a rather large active site which is strongly polarized. One end has a positive charge while the other end has a negative charge. Methotrexate has ten rotatable bonds which divide the molecule into three different regions — a pteridine ring, a hydrophobic *p*-aminobenzoyl group, and a flexible chain with two carboxylate groups. In our simulations the pteridine ring was chosen to be the initial fragment.

With the pteridine ring docked in place, most of the simulations converged to solutions less than 1.64 Å rms deviations from the crystal structure. The closest solution was 0.51 Å away. The best scoring solution, however, was 1.64 Å from the crystal structure and scored 4.3 kJ/mol higher in energy. This solution is shown with the crystal structure binding mode in Figure 5.4. The most commonly found solutions were approximately 1.5 Å rms deviations away from the crystal structure. These score approximately 2.5 kJ/mol worse than the best scoring solution. When the pteridine ring is initially docked incorrectly, many different solutions are found. Most simulations found solutions with rms deviations between 1.1 and 1.8 Å. Two simulations converged to solutions that were approximately 2.5 Å away. One of these is shown in Figure 5.4. Most of the rms deviations resulted from trying to position the last carboxylate group so that it could hydrogen bond with the receptor. This is due to the absense of the solvent molecules and the attempts of the genetic algorithm to satisfy the hydrogen bonding potential in the flexible tail of the methotrexate molecule.

In the docking simulations without the pteridine ring initially docked, the centroid of methotrexate is allowed to vary over a box with dimensions 12.5 Å on each side. In this case over half of the simulations converged to structures with good contact scores. Six solutions were found within 2.0 Å rms deviation of the crystal structure. The closest solution was 1.24 Å away, and the best scoring solution again was the 1.64 Å solution. This is much better than results we reported previously [27] and demonstrates that this scoring function is better suited to flexible ligand docking with genetic algorithms (see Table 5.2).

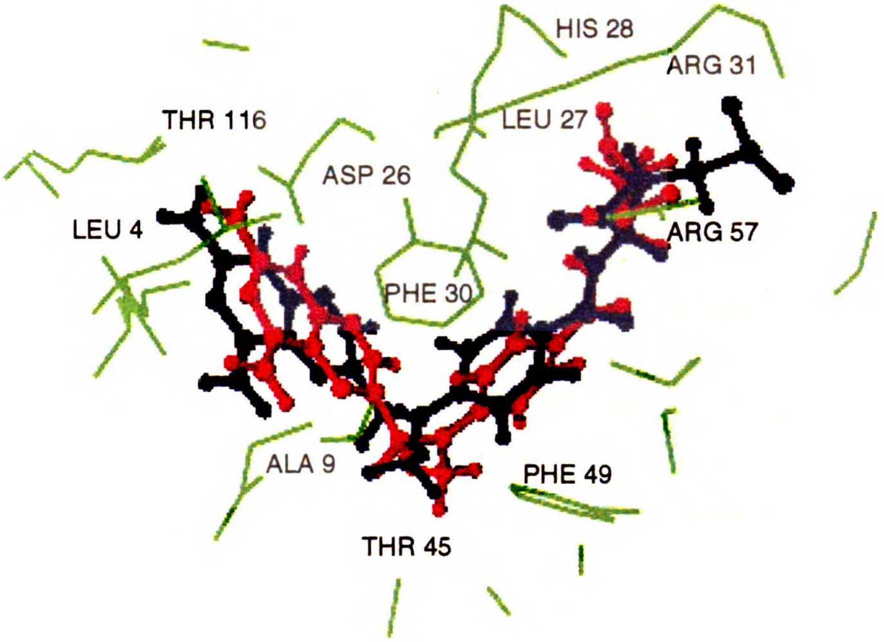


Figure 5.5: Results from the 3dfr docking simulations. A 1.93Å solution (blue) is shown with the crystal structure binding mode (red).

## 3dfr

In this complex methotrexate is bound with dihydrofolate reductase and NADPH. The structure has been determined to 1.7 Å resolution [38]. An interesting feature of this system is that the amide bond on methotrexate appears to deviate markedly from being planar. This illustrates the importance of being able to consider strain in the ligand as suggested by Nicklaus *et al.* [23].

With the pteridine ring docked in the crystal structure orientation, most of the simulations find solutions within 1.3 Å rms deviations from the crystal structure. The closest solution found is 0.80 Å from the crystal structure, and the best scoring solution is 1.11 Å away. When the orientation of the pteridine ring is also allowed to vary, the overall rms deviation of the solutions from the crystal structure increases, but all solutions lie within 1.97 Å from the crystal structure and have good contact scores.

In the centroid-based docking strategy, the centroid of the methotrexate molecule can be placed anywhere within a box with dimensions of 10 Å on a side. The two

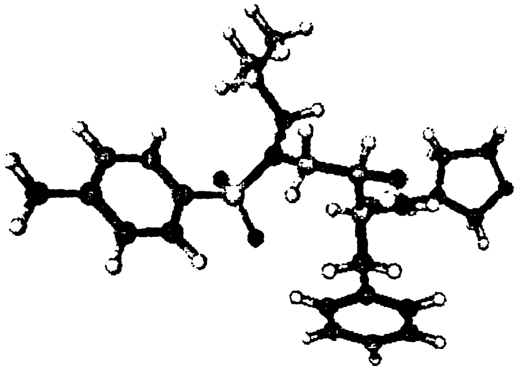


Figure 5.6: The HIV-1 protease inhibitor, VX-478, that was designed at Vertex Pharmaceuticals. Nitrogen atoms are colored blue, oxygens are blue, sulfurs are yellow, carbons are gray, and hydrogens are white.

closest solutions are 1.93 and 2.03 Å rms deviations away from the crystal structure. The first of these is shown with the crystal structure in Figure 5.5. Only about half of the simulations had contact scores near optimum, and this suggests that longer simulation times may be needed. Again, much of the error may be due to the fact that water has been excluded, and the genetic algorithm is trying to form hydrogen bonds between the last carboxylate group and the receptor.

## p478

Inhibitors of HIV-1 protease have recently been introduced to help in the treatment of AIDS. Here we describe the docking simulations of VX-478, a drug that has been designed at Vertex Pharmaceuticals. Phase I/II clinical trials of this new HIV-1 protease inhibitor began in 1995. VX-478 is the largest inhibitor that we consider in this paper. It has thirteen rotatable bonds and is shown in Figure 5.6. For the base fragment we chose the ring system with the nitrogen atom.

With the initial fragment docked in its crystal structure orientation, two different solutions were found. Several simulations converged to solutions within approximately 3Å rms deviations from the crystal structure. Almost all of the other simulations, however, converged to conformations within 1.0 and 1.3Å rms deviations from the

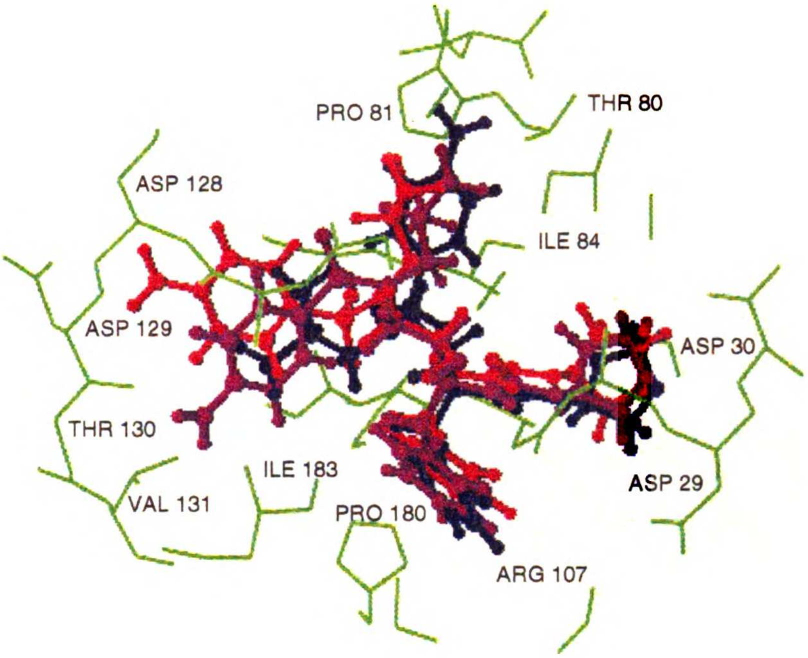


Figure 5.7: Results from the HIV-1 protease inhibitor docking simulations. The crystal structure binding mode is shown in red. The 1.27 and 3.02Å solutions are shown in magenta and blue, respectively.

crystal structure. The best scoring solution is also the closest to the crystal structure (1.01Å). This solution is approximately 9kJ/mol better than the score of the 3Å solution. When the orientation of the initial fragment is allowed to vary, other possible binding modes were found. The best scoring solution is 3.02Å rms deviation from the crystal structure and scores about 2kJ/mol worse than the 1.01Å solution. The closest solution is 1.27Å from the crystal structure. These solutions are shown in Figure 5.7. The difference between these two solutions is that in the 3.02Å solution the benzyl group and hydrophobic chain are docked in the opposite hydrophobic pockets. The receptor also seems to be able to accommodate some variability in the placement of these groups. In the simulations with the incorrectly docked initial fragment, many of the solutions placed many of the atoms away from the receptor surface. Furthermore, many of the solutions have large rms deviations from the crystal structure. This demonstrates the sensitivity that docking simulations can have to the placement of the initial base fragment.

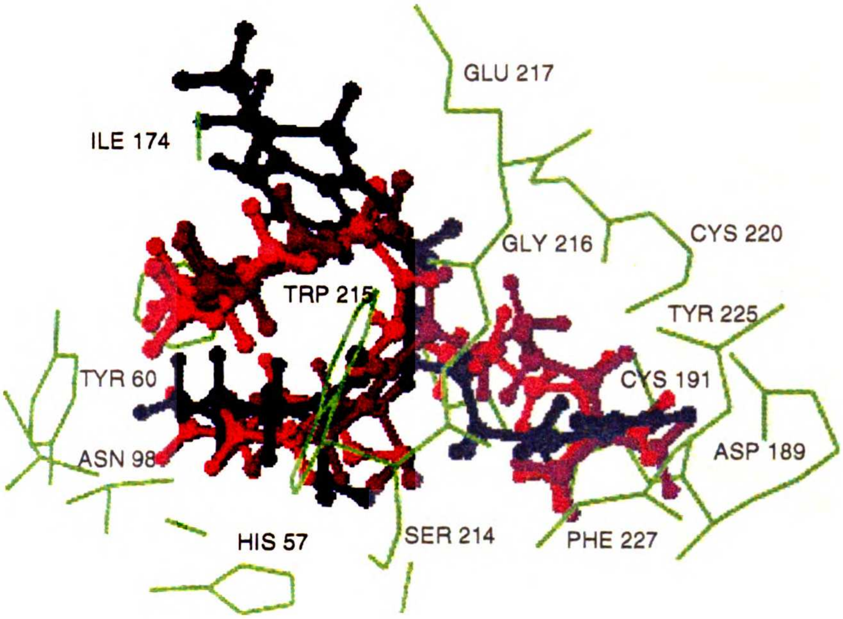


Figure 5.8: The results for the 1dwc-argatroban binding simulations. The crystal structure (red) is shown with 1.14 and 2.29Å solutions in magenta and blue, respectively.

Because the protease has $C_2$-symmetry, we allowed one of the Euler angles to vary over $\pi$ radians, and the position of the inhibitor's centroid was allowed to vary over a 10Å box. The closest solution was 1.35Å rms deviations away from the crystal structure. Other good solutions were found at 3, 7.4, and 9.3Å. All of these structures have scores approximately 9kJ/mol worse than the best solution. Most of the solutions had good contact scores.

## 1dwc

The structure of the thrombin-argatroban complex has been solved to a resolution of 3.0 Å [39]. Argatroban consists of two hydrophobic ring systems (a piperidine and quinoline group) and a guanidinium group which is connected *via* an alkyl chain. According to Banner and Hadvary [39], the piperidine group binds in the proximal(P) pocket, and the quinoline group binds in the distal(D) pocket and packs tightly against the piperidine group. An important feature of binding in this case is the

| Complex | Fixed Fragment | | Variable Fragment | | No Fragment | |
|---|---|---|---|---|---|---|
| | rms (Å) | $N_s$ | rms (Å) | $N_s$ | rms (Å) | $N_s$ |
| 1stp | 0.16 | 50 | 0.67 | 12 | 0.88 | 19 |
| 3cpa | 0.38 | 50 | 0.82 | 28 | 0.79 | 22 |
| 121p | 0.27 | 50 | 0.81 | 31 | 1.68 | 4 |
| 4dfr | 0.51 | 50 | 1.10 | 34 | 1.64 | 6 |
| 3dfr | 0.80 | 50 | 1.03 | 50 | 1.93 | 2 |
| p478 | 1.01 | 37 | 1.27 | 3 | 1.35 | 1 |
| 1dwc | 0.69 | 23 | 1.14 | 2 | 2.29 | 0 |

Table 5.2: A summary of the fixed fragment, variable fragment, and no initial fragment docking simulations. For each, the best rms structure and the number of structures with rms deviations less than 2.0 Å ($N_s$) are reported.

strong interaction between the two hydrophobic groups on the ligand. There is also an oxyanion "hole" in the receptor that is not used directly by argatroban. The guanidinium group was chosen to be the base fragment.

With the initial fragment docked in place, almost half the simulations had solutions within 2.0 Å rms deviations from the crystal structure. The best solution was 0.69 Å away, and the best scoring solution had an rms deviation of 0.73 Å. The performance of the docking system worsened when the base fragment orientation was allowed to vary from the crystal structure orientation. Only two solutions with rms deviations less than 2.0 Å were found. These solutions are 1.14 and 1.71 Å from the crystal structure. The latter solution had the best score. Other solutions were found at 3.3 and 5.2 Å. Many of the simulations did not had good contact scores. Longer simulations may improve the results, but the major limitation in this case appears to be with the scoring function.

In the docking procedure with no initially docked fragment, the crystal structure solution was not found. The closest solutions had rms deviations of 2.29 and 2.77 Å. A number of solutions were found near 4.0 Å, but their scores were approximately 12kJ/mol worse than the best docked structure, the 2.29 Å solution. This solution is shown in Figure 5.8. The docking procedure had trouble docking the quinoline group, but it did well with the rest of the ligand.

## 5.3 Discussion

Molecular docking has been the focus of intense research for a number of years. The goals of this research are to be able to predict the conformation for a protein-ligand complex and ultimately, to screen databases of flexible ligands for new lead compounds. The formation of a receptor-ligand complex results from a balance of contributions from direct binding, desolvation, entropy, and environmental effects. Therefore, the key to any molecular docking system are the underlying approximations. Two different types of approximations are often made. The first have to do with the scoring function used to evaluate the complexes. Ideally, this function would rank the different conformations by their binding affinities. Two commonly used scoring functions are molecular mechanics force fields and empirical estimates of the binding free energy [21]. Another set of approximations deal with the formulation of the docking problem. A number of systems, for example, consider the ligand to be rigid or only partially flexible. Almost all docking methods except for the energy minimization techniques treat the receptor as rigid. Even in the case of flexible ligand docking to rigid receptors, a number of approximations are often made. A common one is that ligands can be docked from an initially placed fragment, but as pointed out above, there are many limitations to this approach.

In this paper, we have described a new flexible ligand docking system which uses a genetic algorithm to search the conformational space of the ligand. The scoring function is based on the binding free energy estimate developed by Böhm [21]. Variations of this function have been used in some fragment joining docking systems. In order for this empirical scoring function to be useful within the context of docking with a genetic algorithm, a number of modifications were made. The most notable of these is the inclusion of a "good contact" score with no penalty for bad van der Waals interactions with the receptor. Our results show that the docking simulations generally converge to solutions with near optimal contact scores with most atoms being placed near the receptor surface and few overlapping with the receptor. Three different simulations were performed on seven receptor-ligand complexes.

The first two simulations were designed to address some of the limitations of fragment-based flexible ligand docking procedures and to evaluate the performance of the docking method if some information about the binding mode were given. In all cases where the base fragment is docked in the crystal structure orientation, our docking system converged to solutions close to the crystal structure in terms of both score and rms deviations. By using a genetic algorithm to search the ligand's conformational space, we have avoided the problems associated with using a "greedy" algorithm as in the incremental construction procedure [19]. Another limitation of the fragment-based docking approach is that the position of the base fragment must be context dependent; its position must depend on the conformation of the entire ligand. Our docking system allows errors in the placement of the initial fragment to be corrected during the docking of the ligand. In all complexes considered, our docking algorithm found solutions close to the crystal structure when the base fragment was initially docked incorrectly. These solutions were found consistently in every case except for p478 and 1dwc, the two largest and most hydrophobic ligands. Our approach offers an alternative to the incremental construction procedure. Another limitation of the fragment-based approach is that the initial fragments often dock in many different locations on the receptor. One solution could be to make many runs, possibly in parallel, with each of the initial locations. Another solution, which is not presented here, would be to use the representation of the genetic algorithm parameters to encode mappings to the different initial fragment docking positions. This approach can also be used to address the final problem of the fragment-based approach, the choice of the initial base fragment. Instead of choosing one initial fragment, the ligand can be divided into a number of different "hot spots", and the genetic algorithm can encode mappings between the different parts of the ligand and their positions on the receptor surface. Work is currently in progress to address these issues of optimization parameter representation.

The last set of simulations were designed for the purpose of evaluating the performance of the scoring function and to compare our approach with other methods that dock flexible ligands without an initially docked base fragment. By using the naive translation and Euler angle representation for the ligand's orientation, we could better separate the performance of the scoring function from that of the representation and compare our present results with those that we published earlier. In all our simulations, our docking procedure found solutions that were close to the crystal structure. In most cases, the solutions had good contact scores. For 3cpa and 4dfr, our results with the new scoring function are much better than the results that we reported previously [27]; we found the crystal structure solution more consistently. The two most difficult complexes for our system were the two largest and most hydrophobic ligands, the HIV-1 protease inhibitor and argatroban. These systems highlight the need for improvements in the scoring function. This is particularly important in the modeling of hydrophobic interactions. For example, in the simulations with the HIV-1 protease inhibitor a number of solutions were found where the benzyl group and hydrophobic chain were docked in the incorrect hydrophobic pockets. With argatroban, the piperidine and quinoline hydrophobic groups interact strongly with each other. These intramolecular interactions must be better characterized. Another limitation of the current scoring function is that effects of water molecules are not included. This is particularly noticeable in the streptavidin and dihydrofolate reductase docking simulations. In each case, carboxylate groups form hydrogen bonds with water molecules in the crystal structure. The high binding affinity of biotin to streptavidin results from the additional hydrogen bonds that form with the solvent. Because the water molecules have been removed, the docking procedure greatly underestimates the binding affinity. In the methotrexate-dihydrofolate reductase simulations, much of the rms deviations from the crystal structure resulted from changes in the conformation of the flexible tail of methotrexate to form hydrogen bonds between the carboxylate groups and the receptor. Discrete water molecules can also play an important role in complex formation by mediating hydrogen bonds between the ligand and the receptor. Guida *et al.* [40] found that they had to include the effects of solvent to obtain good agreement between the docked and crystal structures of thermolysin and its inhibitors once the crystallographic water molecules were removed. Despite these shortcomings in the scoring function, our docking procedure performed very well on a number of complexes with a variety of different types of interactions.

We have developed a flexible ligand docking system that is versatile and can address different formulations of the molecular docking problem. In this paper we have focused on the issues involved with the scoring function that is used with a genetic algorithm since this is directly related to the dynamics of the genetic algorithm. In future work we will address the issue of parameter representation. Jones *et al.* [28] considered one possible representation with putative mappings between hydrogen bonding groups on the ligand and receptor. This approach may perform well with ligands that bind primarily through hydrogen bonds, but as we have seen, there are many complexes of interest that do not fall into this category. We are currently developing a docking procedure that combines a genetic algorithm search procedure with features of fragment-based docking methods. In particular, the genetic algorithm encodes mappings between different "hot spots" on the ligand with complementary regions on the receptor. As we have seen in Chapter 3, molecular docking is important for the optimization of lead compounds in the drug design procedure. These simulations provide information about the possible binding modes of the ligand so that modifications can be suggested to improve the binding affinity of the ligand to its target.

# Bibliography

[1] 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):269–88, 1982.

[2] M. C. Lawrence and P. C. Davis. CLIX: a search algorithm for finding novel ligands capable of binding proteins of known three-dimensional structure. *Proteins*, 12(1):31–41, 1992.

[3] 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.

[4] I. D. Kuntz, E. C. Meng, and B. K. Shoichet. Structure-based molecular design. *Acc. Chem. Re.*, 27(5):117–123, 1994.

[5] D. S. Goodsell and A. J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins*, 8(3):195–202, 1990.

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

[7] T. N. Hart and R. J. Read. A multiple-start Monte Carlo docking method. *Proteins*, 13(3):206–222, 1992.

[8] M Yamada and A. Itai. Development of an efficient automated docking method. *Chem. & Pharm. Bul.*, 41:1200–1202, 1993.

[9] M Yamada and A. Itai. Application and evaluation of the automated docking method. *Chem. & Pharm. Bul.*, 41:1203–1205, 1993.

[10] S.J. Weiner, P. A. Kollman, D.A. Case, U. C. Singh, C Ghio, G. Alagona, S. Profeta Jr., and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765, 1984.

[11] A. DiNola, D. Roccatano, and H. J. C. Berendsen. Molecular dynamics simulation of the docking of substrates to proteins. *Proteins*, 19(3):174–182, 1994.

[12] J. B. Moon and W. J. Howe. Computer design of bioactive molecules: a method for receptor-based de novo ligand design. *Proteins: Struct. Funct. Genet.*, 11:314–328, 1991.

[13] Y. Nishibata and A. Itai. Automatic creation of drug candidate structures based on receptor structure starting point for artificial lead generation. *Tetrahedron*, pages 8985–8990, 91.

[14] 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, 1992.

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

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

[17] R. L. Desjarlais, R. P. Sheridan, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.*, 29(11):2149–53, 1986.

[18] A. R. Leach and I. D. Kuntz. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comp. Ch.*, 13(6):730–748, 1992.

[19] 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, 1996.

[20] 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, 1996.

[21] 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.

[22] W. Welch, J. Ruppert, and A. N. Jain. Hammerhead - fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.*, 3(6):449–462, 1996.

[23] M. C. Nicklaus, S. Wang, J. S. Driscoll, and G. W. Milne. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.*, 3(4):411–428, 1995.

[24] D. E. Goldberg. *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley Pub. Co., Reading, Mass., 1989.

[25] R. S. Judson, E. P. Jaeger, and A. M. Treasurywala. A genetic algorithm based method for docking flexible molecules. *J. Mol. Struct.*, 308:191–206, 1994.

[26] 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.

[27] K. P. Clark and Ajay. Flexible ligand docking without parameter adjustment across four ligand-receptor complexes. *J. Comp. Chem.*, 16(10):1210–1226, 1995.

[28] 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(1):43–53, 1995.

[29] J. H. Holland. *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, Ann Arbor, MI, 1975.

[30] Nicol N Schraudolph. Tech. Rep. CS92-249. UC San Diego, 1990.

[31] R. S. Bohacek and C. McMartin. Definition and display of steric, hydrophobic, and hydrogen-bonding properties of ligand binding sites in proteins using Lee and Richards accessible surface: validation of a high-resolution graphical tool for drug design. *J. Med. Chem.*, 35(10):1671–84, May 1992.

[32] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer, 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(3):535–542, May 1977.

[33] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, 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. *Arch. Biochem. Biophys.*, 185(2):584–591, Jan 1978.

[34] S.J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. *J. Comp. Chem.*, 16:548, 1986.

[35] P. C. Weber, J. J. Wendoloski, M. W. Pantoliano, and F. R. Salemme. Crystallographic and thermodynamic comparison of natural and synthetic ligands to streptavidin. *J. Am. Chem. Soc.*, 114:3197–3200, 1992.

[36] D. W. Christianson and W. N. Lipscomb. X-ray crystallographic investigation of substrate binding to carboxypeptidase A at subzero temperature. *Proc. Natl. Acad. Sci. U. S. A.*, 83(20):7568–7572, 1986.

[37] U Krengel. *Struktur und guanosintriphosphat-hydrolysemechanismus des c-terminal verkuerzten menschlichen krebsproteins.* PhD thesis, Heidelberg, 1991.

[38] J. T. Bolin, D. J. Filman, D. A. Matthews, R. C. Hamlin, and J. Kraut. Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 åresolution. I. general features and binding of methotrexate. *J. Biol. Chem.*, 257(22):13650–13662, 1982.

[39] D. W. Banner and P. Hadvary. Crystallographic analysis at 3.0-A resolution of the binding to human thrombin of four active site-directed inhibitors. *J. Biol. Chem.*, 266(30):20085–93, 1991.

[40] W. C. Guida, R. S. Bohacek, and M. D. Erion. Probing the conformational space available to inhibitors in the thermolysin active site using Monte-Carlo energy minimization techniques. *J. Comp. Chem.*, 13(2):214–228, 1992.

# Chapter 6

# Conclusions and Future Directions

In this dissertation I have designed a ligand discovery and optimization system for structure-based drug design. There are two key components to this system — an interactive molecular graphics system and methods for docking flexible ligands to rigid receptors. The molecular graphics system is implemented as an extension to Chimera, a new molecular modeling system under development at the Computer Graphics Laboratory at UCSF. Currently, the system has methods for finding and displaying hydrogen bonds, computing and displaying molecular surfaces, and for interactively visualizing any volumetric data that is represented on a grid. Texture mapping is available and provides the user with a means for further controlling how the data are rendered. For example, the user can interactively change the color or transparency of an object to highlight regions where new functional groups can be added to increase the binding affinity of the ligand to the target receptor. The system has not been used to design a new therapeutic agent, but this is a goal of future work. Nevertheless, the system has been very useful for analyzing the results of docking simulations. It has also been used to visualize how different ligands are accommodated in a common binding site. This has been invaluable for understanding the principles that govern protein-ligand interactions.

The other component of the system is a flexible ligand docking module. Molecular docking is essential for a drug design system because in order to propose chemical modifications to a lead compound, one has to know the ligand's binding mode. Furthermore, once a change has been made to a lead compound, it is good practice to verify that the molecule binds in the expected conformation and orientation. A common mistake is to assume that the molecule does. Even small changes in a ligand can change how it binds to the target enzyme. In this dissertation, I have chosen flexible ligand docking methods that use a genetic algorithm to search the conformational space of the ligand. I have compared the molecular mechanics force fields and simple, empirical binding free energy estimates as potential scoring functions for docking flexible ligands to rigid receptors. I have shown not only that the simple binding free energy scoring function can be used with genetic algorithms, but I have developed a docking system that performed very well with several different receptor-ligand complexes. The system also addresses a number of limitations of other docking approaches. My docking system can be used as an alternative to the incremental construction approach. A genetic algorithm is used to search the conformational space of the ligand within the receptor as an alternative to a "greedy" procedure. My system also allows the initial base fragment to be docked within the context of the entire ligand. The focus of my research has been on developing better scoring functions for genetic algorithm-based docking procedures, but I have also described ways in which the choice and representation of the parameters can be used to address other aspects of flexible ligand docking. Two areas of future work are developing new representations for molecular docking and improving various aspects of the scoring function. Another goal is to ultimately develop methods for computationally screening databases of flexible ligands for new lead compounds.