

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Exploration of macromolecular interactions

**Permalink**

<https://escholarship.org/uc/item/1nh486mk>

**Author**

Hendrix, Donna K.

**Publication Date**

1999

Peer reviewed|Thesis/dissertation

**Exploration of Macromolecular Interactions:  
Development and Implementation  
of Descriptors for Macromolecular Docking**

by  
**Donna K. Hendrix**

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

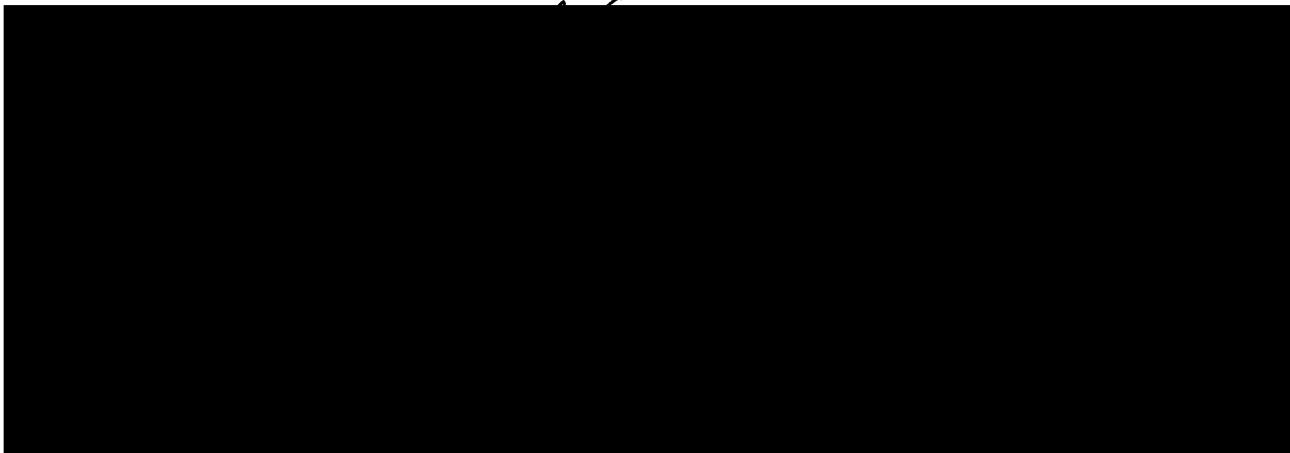
in  
**Biophysics**

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA SAN FRANCISCO



Date

University Librarian

Degree Conferred: .....

copyright 1999  
by  
Donna K. Hendrix

**It doesn't get any better than this.**

- Paul Albrecht, Ph.D., lecturer  
ca. 11pm, March, 1988  
Building 20, MIT, Cambridge, Massachusetts  
to Donna Hendrix and Voula Georgopolous, Course 6 graduate students



**These are the best days of your life.**

- Chancellor Joseph Martin  
ca. April, 1996  
Medical Sciences elevators, UCSF, San Francisco, California  
to Kevin Clark and Donna Hendrix, graduate students,  
as they commiserated about writing dissertations and  
preparing for oral qualifying exams.



**In graduate school, I never worry acutely about any one thing, but I do feel a general kind of *cosmic* stress.**

- Tom Defay, graduate student in Biophysics  
ca. October, 1995  
between San Francisco and Tiburon, California  
en route to the Biophysics Graduate Group Retreat,  
to Rachel Brem, graduate student in Biophysics.



**Biophysics students seem to enjoy themselves more than most graduate students I know.**

- Donna Hendrix  
ca. February, 1993  
Millberry Union, UCSF, San Francisco, California  
To Professor Peter Kollman, while interviewing for the  
Graduate Group in Biophysics.

years after I had left their labs. My friends and mentors at MIT were also helpful, particularly my master's thesis advisor, Tom Weiss, who was kind when he encouraged me to stay in graduate school, and wise when he told me to have no fear when I decided to try something else. My managers and co-workers at Oracle Corporation have kept in touch and occasionally bought me breakfast or lunch, and still seem to be surprised that I left the world of high-tech.

At UCSF, I have felt a part of the larger community, within my lab, my graduate program, student organizations, and the community as a whole. I have worked hard to be a part of that community by serving as the Women's Advocacy Officer in the Graduate Students' Association, and as an active participant in Women In Life Sciences. My affiliation with these groups enriched my experience in graduate school. I am honored to have known the women who founded WILS, Theresa Gamble, Renee Williard, and Tina Settineri, who offered their guidance and friendship. I have worked with numerous other incredible people, too numerous to mention, who also gave their time and efforts to WILS, and I am proud to have had that experience.

The Kuntz Lab, both past and present, has been a fantastic place to work. Tack sets a tone of friendly cooperation, and the lab follows. I have had the tremendous fortune of sharing an office with Connie Oshiro and Geoff Skillman. Connie has been supremely supportive and helpful, offering a realistic analysis of scientific (and system) proposals, gentle nudges, and thoughtful advice. Geoff always has a kind, useful and appropriate response to my pesky questions and complaints. Our "fourth office mates" have all terrific additions: Barbara Chapman, Ron Knegtel, Shinichi Katakura and Seth Hopkins. I have benefited from interacting with many Kuntz Lab members, especially alumnae Cindy Cor-

win, Dan Gschwend, Elaine Meng, Diana Roe, Todd Ewing, Yax Sun and Rob Cerpa, and current members Malin Young (the meeting buddy extraordinaire!), Keith Burdick, David Sullivan, Xiaoqin Zou, Michelle Lamb, and Ken Brameld.

Graduate school would have been much more difficult without the assistance of the Computer Graphics Lab. Al Conde came to my rescue an uncountable number of times while I was working on the Kuntz Lab system. Greg Couch helped with Midas and programming questions, and let me bend his ear on thousands of occasions. Heidi Houtkooper generously provided assistance with figures for publications and for my dissertation. Conrad Huang gave valuable advice on the presentation of my data, and he built viewdock in a day, just for us! Eric Petterson was always helpful with any system or Midas questions, and Tom Ferrin was very generous to me and the Kuntz Lab with CGL resources.

The Biophysics Graduate Group is an amazing group of people, and it is one of my best decisions to have come to UCSF and be a part of it. Julie Ransom is the hub and heart, and is a large and underacknowledged part of what makes the Group great. The faculty are incomparable, and I am amazed that I have had the opportunity to chat with and take classes from Peter Kollman, Ken Dill, Fred Cohen, and David Agard. The students make the program what and who it is, and my class -- Shane Atwell, Keith Burdick, Julian Chen, Cathy Collins, Jeremy Gollub, Mark Kaplan, Dave Miller and Laleh Shayestah -- was no exception.

My closest friends and family offered support throughout these last years. Good friends are rare and wonderful, and I have been blessed with many while I have been here, especially Linda Brinen, Lisa Kim-Shapiro, Lisa Uyechi and Tracy Ware. Outside of

UCSF, Chris Ring, Craig Lee and Joanna Muench gave me a valuable perspective of life beyond graduate school, and pep talks on the many occasions they were needed. My parents, Sidney and Kyung Bok Hendrix, were frequent visitors to San Francisco, so that we could spend time together, and were patient with delayed calls home and postponed visits. They always encouraged my interests in math, science, music, service and education, and always helped me in thousands of ways. My sister and her husband, Sharon and Steve Kramer, offered an ear when I needed to vent frustrations, and a place to go (and a big adventure for my dog) when I needed to get out of town. Finally, my partner, Mike Machado, maintained great tolerance while I chose and forged a difficult path to my dreams. Perhaps nothing in life worth having is easy, but hard choices and paths can be made easier with help from a supportive partner. I thank Mike for being there in good and bad times.

**DISSERTATION ABSTRACT**

**EXPLORATION OF MACROMOLECULAR INTERACTIONS:  
DEVELOPMENT AND IMPLEMENTATION  
OF DESCRIPTORS FOR MACROMOLECULAR DOCKING**

by

**Donna K. Hendrix**

Structure-based design offers insight that cannot be made with new methods that are in use for drug discovery and development, such as genomics, high-throughput screening and combinatorial chemistry. By advancing structure-based methods, it may be possible to make predictions about the structural aspects of macromolecular interactions from sequence information. These predictions may be used to identify new targets and to optimize existing targets.

This dissertation is a first step toward the long-term goal of using macromolecular docking to predict interactions and function. Chapter 2 details a method that describes the shape of complex macromolecular surfaces for docking. In chapter 3, these descriptors are used to explore the interactions of known systems, and predict a geometry of human growth hormone receptor that is not seen in x-ray crystal structures. Chapter 4 describes the discovery of small molecule inhibitors of HIV-1 integrase by using molecular docking methods to target a site at the dimer interface, more than 15Å from the active site with no known function. Chapter 5 is a preliminary study of the use of DOCK to screen a database of small molecules with the eventual goal of building novel, larger molecules from the small molecules. Our model is FK506 binding protein, and the inherent difficulties of working with this type of site are revealed in the work.





# Table of Contents

## Chapter 1

<b>Introduction</b> .....	<b>1</b>
Background .....	2
Goals .....	4
Strategies and Results .....	5
References .....	6

## Chapter 2

<b>Surface Solid Angle-based Site Points for Molecular Docking</b> .....	<b>8</b>
Abstract .....	9
Introduction .....	10
Methods .....	11
Determining site points: building regions .....	13
Docking with shape-based site points .....	14
Results .....	16
Regions .....	16
Macromolecular Docking .....	17
Discussion .....	23
Acknowledgments .....	24
References .....	25

## Chapter 3

<b>Macromolecular Docking of a Three-Body System: the Recognition of Human Growth Hormone by its Receptor</b> .....	<b>27</b>
Abstract .....	28
Introduction .....	29
Results .....	32
Docking with surface shape-based site points .....	32
Comparison of spheres to surface shape-based site points .....	35
Docking human growth hormone to its receptor .....	37
Docking mutational sites on hGH to the hGHR .....	41
Discussion .....	43
Shape-based site points .....	43
Docking hGH to the hGHR .....	45
Conclusions .....	48
Methods .....	49
Building site points .....	49
Docking macromolecules .....	52
Scoring docked orientations .....	53
Generating spheres .....	54
Test cases .....	55

Macromolecular docking techniques: hGH to the hGHR .....	55
Acknowledgments .....	57
References .....	58

## **Chapter 4**

<b>Discovery of Inhibitors of HIV-1 Integrase by Docking to an Uncharacterized Site at the Dimer Interface .....</b>	<b>71</b>
Abstract .....	72
Introduction .....	73
Methods .....	74
Computational methods .....	74
Assays .....	76
Results .....	79
RT compounds .....	79
DOCK compounds .....	82
Discussion .....	86
Possible mechanism of inhibition .....	87
Ongoing work and future directions .....	89
Acknowledgments .....	91
References .....	91

## **Chapter 5**

<b>Computational Approaches to SAR by NMR by DOCK .....</b>	<b>107</b>
Abstract .....	108
Introduction .....	109
Methods .....	111
Part I: Searching the ACD .....	112
Part II: Ranking a database .....	114
Results .....	116
Part I: Searching the ACD .....	116
Part II: Ranking a database .....	119
Selecting site points and examining sampling methods .....	119
Scoring .....	121
Conclusions .....	122
Acknowledgments .....	125
References .....	126

## **Chapter 6**

<b>Epilogue .....</b>	<b>145</b>
-----------------------	------------

# List of Tables

## Chapter 2

### Surface Solid Angle-based Site Points for Molecular Docking

Performance of DOCK runs with and without shapelite points .....	18
Performance of DOCK runs with shapelite points .....	19
Results of DOCK runs with shapelite points .....	20

## Chapter 3

### Macromolecular Docking of a Three-Body System: the Recognition of Human Growth Hormone by its Receptor

Test set for macromolecular docking .....	34
Results of macromolecular docking .....	36
Sites for hGH docking to the hGHR .....	38
Calculations for docking hGH to the hGHR .....	39
Results of docking binding sites of hGH to the hGH receptor .....	41
Results of docking mutational sites on hGH to mutational sites on hGH receptor .....	42

## Chapter 4

### Discovery of Inhibitors of HIV-1 Integrase by Docking to an Uncharacterized Site at the Dimer Interface

Inhibition of HIV IN activity. ....	80
Inhibition and toxicity data, IN DOCK compounds .....	84
Crystals of HIV-1 IN catalytic core domain .....	90

## Chapter 5

### Computational Approaches to SAR by NMR by DOCK

Clustering of 1,010 highest-scoring molecules .....	116
Results of docking the ACD to FK506 binding protein .....	118

# List of Figures

## Chapter 2

### Surface Solid Angle-based Site Points for Molecular Docking

- Figure 1. A description of the surface solid angle. . . . . 12
- Figure 2. Trypsin inhibitor docked into trypsin (2ptc) using  
shapelite points from the entire inhibitor surface. . . . . 22

## Chapter 3

### Macromolecular Docking of a Three-Body System: the Recognition of Human Growth Hormone by its Receptor

- Figure 1. Docking sites of human growth hormone binding  
to its receptor. . . . . 63
- Figure 2. Orientation of receptor monomers docked by receptor  
stem sites. . . . . 66
- Figure 3. Description of the surface solid angle. . . . . 67
- Figure 4. Building site points from surface points. . . . . 68
- Figure 5. Site points on trypsin inhibitor, colored by relative value. . . . . 69
- Figure 6. DOCK score vs. RMS deviation of DOCK orientations  
for chymotrypsin and turkey ovomucoid third domain. . . . . 70

## Chapter 4

### Discovery of Inhibitors of HIV-1 Integrase by Docking to an Uncharacterized Site at the Dimer Interface

- Figure 1. Structural alignment of the catalytic core domains of  
HIV-1 IN and ASV IN. . . . . 99
- Figure 2. Alignment of the sequences of the catalytic cores of  
ASV IN and HIV-1 IN. . . . . 100
- Figure 3. Computational approach. . . . . 101
- Figure 4. Docking sites on ASV IN catalytic core domain. . . . . 102
- Figure 5. Docking sites on ASV IN catalytic core domain,  
side view. . . . . 103
- Figure 6. ASV IN and HIV-1 IN with MFCD0070629 docked into  
the dimer site. . . . . 104
- Figure 7. MFCD0070629 in the ASV IN dimer site. . . . . 105
- Figure 8. MFCD0070629 in HIV-1 IN dimer site. . . . . 106

## Chapter 5

### Computational Approaches to SAR by NMR by DOCK 107

- Figure 1. FK506 binding protein. . . . . 129
- Figure 2. SPHGEN spheres and SURFSPH spheres in FK506  
binding site of FKBP. . . . . 130
- Figure 3. DOCK ligand in predicted orientation in FKBP site. . . . . 131

Figure 4. DOCK scores of top 1,000-scoring molecules from docking ACD95.2 to FKBP. ....	132
Figure 5. Determining the appropriate amount of sampling. ....	133
Figure 6. Results from docking with SURFSPH spheres compared to results from SPHGEN spheres with contact score. ....	135
Figure 7. Results from docking with SURFSPH spheres compared to results from SPHGEN spheres with energy score. ....	136
Figure 8. Results from docking with SURFSPH spheres compared to results from SPHGEN spheres with contact score. ....	137
Figure 9. Results from docking with SURFSPH spheres compared to results from SPHGEN spheres with energy score. ....	138
Figure 10. Ideal and random enrichment curves. ....	139
Figure 11. Enrichment curve. ....	141
Figure 12. Enrichment curve, first 25% of the database. ....	142
Figure 13. Weighting factors to correct bias. ....	143
Figure 14. Weighting factors to correct bias, first 25% of the database. ....	144

## **Chapter 1**



# **EXPLORATION OF MACROMOLECULAR INTERACTIONS: DEVELOPMENT AND IMPLEMENTATION OF DESCRIPTORS FOR MACROMOLECULAR DOCKING**

## **INTRODUCTION**



## **Background**

Structure-based design has come of age, having proven itself in the last decade as a valuable tool. The success of HIV-1 protease inhibitors has shown that computational tools can be used for drug discovery -- to find lead compounds, to optimize leads, and to help elucidate the energy of binding (Wlodawer and Vondrasek 1998). The maturation of structure-based design is concurrent with the emergence of high through-put screening and combinatorial chemistry, and the availability of numerous sequences from genome projects and microarray screening. These techniques have changed drug design by enormously increasing the volume of compounds that can be screened, the number of molecules that can be synthesized and the identification of targets.

These methods appear to pose a challenge to structure-based methods. Genomics, high-volume assays and combinatorial chemistry do not directly take advantage of structural information of the target molecules, but structural information and structure-based techniques remain useful. Small-molecule structure methods characterize the diversity of libraries. Predictions from structural data narrow the scope of combinatorial libraries. Secondary and tertiary structure prediction can help assign function to sequence data. Structures will continue to reveal how ligands and proteins interact, and those data will continue to contribute to the understanding of the interactions, the identification of targets for drug design and the optimization of lead compounds.

Molecular docking is one of the aforementioned structure-based technique. It has been used to screen databases of molecules to search for drug discovery and to predict the binding orientation of known and predicted ligands. In the world of high-volume drug

design, it has been used to design combinatorial libraries. Macromolecular docking may predict the interactions of proteins with proteins and DNA. With the addition of more sequence and structural data, macromolecular docking may be used to identify new targets and characterize macromolecular function.

Macromolecules have many roles: as enzymes, hormones, channels, links in signaling cascade, transcription factors, cell and tissue infrastructure. (Additionally, many macromolecules play more than one role, for example, thrombin is a protease and binds to protease-activated receptor, a seven membrane-spanning G-protein coupled receptor, initializing a signal (Liu, et al. 1991; Vu, et al. 1991).) Many successful drug designs have been made to enzymatic targets, but as more sequences are determined more non-enzymatic macromolecular targets will be identified. Their roles and interactions can be revealed with methods such as the yeast two-hybrid assay, but structural data will increase the understanding of complex biological systems and unveil possible new targets. In the absence of complete structural information, macromolecular docking can predict interactions. Additionally, the interfaces of such macromolecules are targets for drugs. A drug may increase or decrease a macromolecular interaction by binding in or near the interface, and thus alter the resulting biological function or signal.

Cell surface receptors, and signalling proteins -- macromolecules that have evolved large binding surfaces to identify other macromolecules -- tend to have large, solvent-exposed and complex surfaces. This feature is in contrast to the solvent-shielded active sites of proteases. For a drug to bind to these types of molecules, it must recognize the site and disrupt the interaction of the target macromolecule with other macromolecules.



The manner that a small molecule disrupts a macromolecular system is to change the structure of a macromolecule or to block the binding site of the molecule. For example, the inhibitor may prevent or enhance a conformational change, as proposed for HA (Hoffman, et al. 1997). It may alter the quaternary structure of a macromolecule and reduce or increase the molecule's ability to form a heterodimer or oligomerize, as taxol binds to tubulin and stabilizes microtubules (Schiff, et al. 1979). It may bind to a recognition site and hinder recognition by having a higher affinity to the site than the native molecule, thus blocking recognition, as seen in FK506 (Griffith, et al. 1995). It may mimic a larger molecule that enhances recognition, as has been seen in the small-molecule mimic of granulocyte-colony stimulating factor (Tian, et al. 1998) and the peptide mimic of erythropoietin (Livnah, et al. 1998), and, in some respects, tamoxifen binding to estrogen receptor (Shiau, et al. 1998).

## **Goals**

The ultimate goal is to begin with sequence data, use structure prediction methods to determine structure, and then dock macromolecules to unveil their interactions. This goal includes the prediction of interactions of macromolecules that are known to bind, but for which no structural data of the macromolecules in complex exists, and characterization and use of the interfaces of these macromolecules as targets for drug design. This dissertation takes initial steps in that direction. The objective of this dissertation is to explore the interactions of macromolecules of known complexes within the context of identifying targets for structure-based design.

## Strategies and Results

There are still many challenges in the world of structure-based design. An excellent ab initio protein fold has a root-mean-square distance (RMSd) of 4.0Å to the true structure. Docking can frequently identify a small-molecule, but it is considered a success if one among ten to twenty selected molecules binds with better than 100 μM affinity. It is not realistic at this time to begin with sequence data and then use folding and docking to identify targets.

As a beginning, well-characterized structures can be used to explore molecular docking. A first step toward using macromolecular docking to explore unknown complexes is to dock known complexes. To determine whether it is possible to design a drug that disrupts oligomerization, one can dock to well-characterized interfaces of macromolecules.

This dissertation is a first step toward the goal of using macromolecular docking to predict interactions and function. In chapter 2, I developed a method to describe the complex surfaces of macromolecules for macromolecular docking. In chapter 3, I used these descriptors to explore the interactions of known systems, and predicted a geometry of human growth hormone receptor that is not seen in x-ray crystal structures. In chapter 4, I used molecular docking methods to discover small molecule inhibitors of HIV-1 integrase by targeting a site at the dimer interface, more than 15Å from the active site with no known function. In chapter 5, I used DOCK in a preliminary study to screen a database of small molecules with the eventual goal of building novel, larger molecules from the small molecules. Our model is FK506 binding protein, and the inherent difficulties of working with this type of site are revealed in the work.

## References

- Griffith JP, JL Kim, EE Kim, MD Sintchak, JA Thomson, MJ Fitzgibbon, MA Fleming, PR Caron, K Hsiao and MA Navia. 1995. X-ray structure of calcineurin inhibited by the immunophilin-immunosuppressant FKBP12-FK506 complex. *Cell* 82:507-22.
- Hoffman LR, ID Kuntz and JM White. 1997. Structure-based identification of an inducer of the low-pH conformational change in the influenza virus hemagglutinin: irreversible inhibition of infectivity. *Journal of Virology* 71:8808-8820.
- Livnah O, DL Johnson, EA Stura, FX Farrell, FP Barbone, Y You, KD Liu, MA Goldsmith, W He, CD Krause, S Pestka, LK Jolliffe and IA Wilson. 1998. An antagonist peptide EPO receptor complex suggests that receptor dimerization is not sufficient for activation. *Nature Structural Biology* 5:993-1004.
- Liu LW, TK Vu, CT Esmon, SR Coughlin. 1991. The region of the thrombin receptor resembling hirudin binds to thrombin and alters enzyme specificity. *Journal of Biological Chemistry* 266:16977-80.
- Schiff PB, J Fant, SB Horwitz. 1979. Promotion of microtubule assembly in vitro by taxol. *Nature* 277:665-7.
- Shiau AK, D Barstad, PM Loria, L Cheng, PJ Kushner, DA Agard and GL Greene. 1998. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* 95:927-937.
- Tian S, P Lamb, AG King, SG Miller, L Kessler, JI Luengo, L Averill, RK Johnson, JG Gleason, LM Pelus, SB Dillon and J Rosen. 1998. A small, nonpeptidyl mimic of granulocyte-colony stimulating factor. *Science* 281:257-159.

Vu TK, DT Hung, VI Wheaton, SR Coughlin. 1991. Molecular cloning of a functional thrombin receptor reveals a novel proteolytic mechanism of receptor activation. *Cell* 64:1057-68.

Wlodawer A and J Vondrasek. 1998. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu Rev Biophys Biomol Struct* 27:249-284.

## Chapter 2



### **SURFACE SOLID ANGLE-BASED SITE POINTS FOR MOLECULAR DOCKING**

by

**Donna K. Hendrix and Irwin D. Kuntz**



*This has been published in Pacific Symposium on Biocomputing 1998:317-326. It is reprinted with the permission of the Pacific Symposium on Biocomputing.*

## Abstract

We are developing a new site descriptor for the DOCK molecular modeling program suite. Sphgen, the current site description program for the DOCK suite, describes the pockets of a macromolecule by filling a volume with intersecting spheres. DOCK then identifies possible ligand orientations in the pocket by overlapping the atoms of proposed ligands with the sphere centers. Sphgen limits use of the DOCK program to concave binding regions, but macromolecular binding regions can be solvent-exposed rather than buried pockets. We present a more general site descriptor, based on the surface solid angle, which generates site points by determining the solid angle of exposure for points on the surface of the molecule, then identifying patches of surface with similar solid angle values which are then built into site points. We find possible ligand orientations by matching shape-based site points on the ligand and protein and demanding complementary solid angle values. Orientations are evaluated using the DOCK's force field-based score, which evaluates the Coulombic and van der Waals energy. The surface solid angle descriptor displays the complementary characteristics of the interfaces of our test systems: trypsin/trypsin inhibitor, chymotrypsin/turkey ovomucoid third domain, and subtilisin/chymotrypsin inhibitor. The solid angle site points can be used by DOCK to generate orientations within 1.5Å RMSd of the crystal structure orientation.

## Introduction

The interactions of proteins with other proteins and with DNA perform many of the signaling, recognition and catalytic functions within cells. The specificity of macromolecular interactions is due to a matching of complementary features in the interface of the complexed molecules. These features are both chemical in nature (e.g., salt bridges, hydrogen bonds, hydrophobic interactions) and geometric.

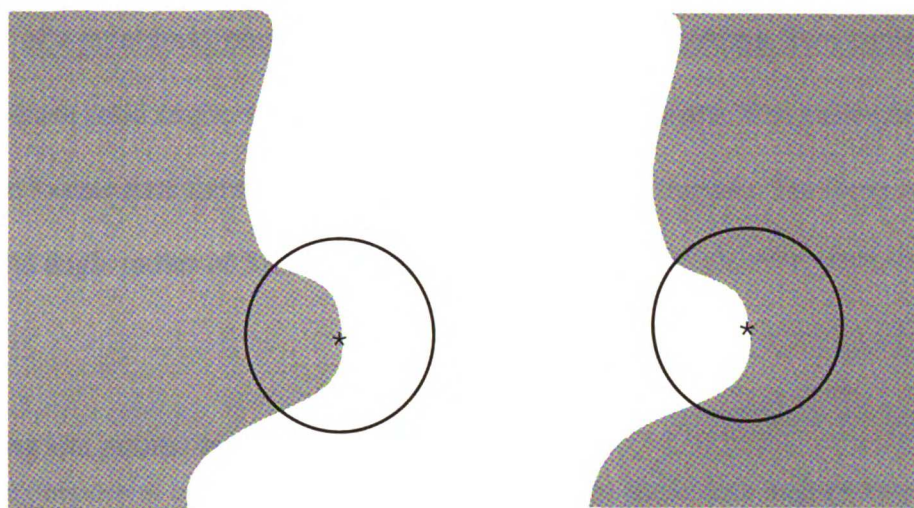
Solutions to the molecular docking problem have used approaches based upon the chemistry and geometry of macromolecules to reduce the solution space of the problem. (Shoichet and Kuntz 1991)(Connolly 1986b, Connolly 1992)(Fischer, et al. 1995) Lin et al. (Lin, et al. 1994) define geometric "critical points" on the molecule, based upon the Connolly molecular surface. Each critical point also has an associated surface normal, and a defined character based upon the type of surface from which it was generated: cap, pit or belt. Several groups (Kuntz, et al. 1982)(Connolly 1986b) describe the complementary nature of protein-protein and protein-ligand docking by describing the geometric interactions as protrusions which fit into invaginations, or knobs-into-holes.

Macromolecular interactions do not always have a knobs-into-holes character, but can have large, smooth interfaces. In order to take advantage of these types of interactions with an existing docking algorithm, we have developed a site descriptor for docking based on the surface solid angle. This descriptor describes the local shape of the surface regardless of the features of the surface. We use these site descriptors as site points for DOCK, which generates orientations and evaluates them based upon electrostatic interactions.

## Methods

Defining the surface and the local surface shape The surface of proteins and ligands are described with Connolly's molecular surface (MS) program (Connolly 1983). To calculate the solid angle, we require surface normals and associated areas in addition to coordinates for the surface points. The solid angle of each surface point is calculated using Connolly's solid angle algorithm (Connolly 1986a), which places a test sphere center on a point, then determines the area of the test sphere which lies within the protein. The area within the protein is calculated by sampling the surface of the test sphere and calculating whether sampled points lie inside or outside the surface of the protein. The solid angle is then the portion of the surface area of the sphere that lies inside the protein, multiplied by  $4\pi$  (see Figure 1). The solid angle is measured in steradians. The result of these calculations is a set of points in 3-space, each with an associated surface solid angle value.





**Figure 1. A description of the surface solid angle.**

The gray shaded area represents the interior of the protein. On the left, measuring the surface solid angle at the asterisk, approximately 1/4 of the test sphere lies inside the protein, therefore its solid angle is  $\frac{1}{4} \times 4\pi = \pi$ . On the right, measuring the surface solid angle at the asterisk, approximately 3/4 of the test sphere lies inside the protein, therefore its solid angle is  $\frac{3}{4} \times 4\pi = 3\pi$ . The surface at these two points complements, and the sum of their solid angles is  $4\pi$ .

WU  
LIBRARY

The solid angle of a point lying well outside of a surface is 0 steradians, while the solid angle of a point lying entirely within the surface is  $4\pi$  steradians. Two complementary points have solid angles which sum to a value of  $4\pi$  steradians. The radius of the test sphere used to calculate the solid angle is variable and set by the user. For these calculations, a solid angle radius of  $5\text{\AA}$  was used.

***Determining site points: building regions***

The purpose of calculating the solid angle is to use these data to dock two molecules together. Our docking algorithm grows geometrically with the number of site points. In order to reduce the docking time, the program, shapesite, reduces the number of site points by amalgamating them into shape regions. Shapesite examines near neighbor points and defines shape regions as clusters of adjoining points with similar solid angle values.

Neighbor lists are determined by a simple point-by-point search. The near neighbors of a point are defined as points within a distance of the square root of the density of surface points multiplied by two. For example, for these studies, a surface density of 1 dot/ $\text{\AA}^2$ , and a neighbor search radius of  $1.4\text{\AA}$  is used. Up to 8 near neighbors are found.

Regions begin as a seed point. The near neighbor list of the seed point is evaluated, and the neighborhood forms a region if all neighbors have a solid angle value within  $\pi/8$  steradians. Once a point is placed into a region, it is removed from all neighbor lists; a point can be assigned to one region, only. The region can grow larger by accumulating more near neighbor points if their surface solid angle is within  $\pi/8$  steradians of the seed point. Regions have a minimum size of  $5\text{\AA}^2$  and a maximum size of  $15\text{\AA}^2$ , with the average size of a region on a protein surface varying from 7 to  $8.5\text{\AA}^2$ . Each region has an asso-

ciated solid angle value, and is represented as a site point in the DOCK algorithm by its center of mass and solid angle value. Regions may span several atoms.

Because a point, once selected to be part of a region, is taken off the list of points searched to form regions, the resulting site points formed from regions can vary with the order in which the surface points are searched. To determine the effect of the order of the search on the derived site points, we decoupled trypsin/trypsin inhibitor, derived site points and for the binding site first by the default order of points, which is ordered by residue in the protein from the N-terminus to the C-terminus. We then “shuffled” the residues in the data file so they were no longer in the N-to-C order (yet surface points from the same residues remain together). We re-assembled the complex and examined the complementarity of the site points from the unshuffled and shuffled data sets.

#### ***Docking with shape-based site points***

Molecules are docked using version 4 of DOCK (Ewing and Kuntz 1997) and implementing the solid angle values as a shape-based filter. DOCK determines orientations by searching for distances between pairs of site points on the ligand that also exist between pairs of site points on the receptor. With shape-based site points, we additionally demand that the matched distances align such that the resulting adjacent ligand and protein site points which determine the match have complementary solid angles, with

$$\text{solid angle}(\textit{site point 1}) + \text{solid angle}(\textit{site point 2}) > 3\pi$$

This shape filter is implemented as a chemical matching filter in DOCK (Shoichet and Kuntz 1993).

We have selected three proteinase-inhibitor complexes for this study: chymotrypsin/turkey ovomucoid third domain (1cho) (Fujinaga, et al. 1987); trypsin/trypsin inhibitor (2ptc) (Marquart, et al. 1983); subtilisin/chymotrypsin inhibitor (2sni) (McPhalen and James 1988). These structures were selected based on their resolution, which ranges from 1.9Å to 2.1Å, and for comparison to a previous study (Shoichet and Kuntz 1991). For each structure, the complexes were decoupled and shapesite generated shape-based site points for the entire inhibitor and proteinase surfaces. Two docking studies were done, one with an area on the proteinase and the inhibitor which covered the binding region plus an additional 5Å in all directions beyond the site, and a second study with the entire inhibitor surface, with the binding region plus additional 5Å of the proteinase site. (See figure 2 for an example of a proteinase site for docking trypsin/trypsin-inhibitor.)

Input parameters to version 4 of DOCK include the minimum distance between site points, which is the shortest distance that will be compared between pairs of site points on the protein and ligand, and a distance tolerance. Two distances whose lengths differ by the distance tolerance are considered equal distances. For these studies we selected a large minimum distance tolerance, 4Å, because of the large ligand molecules. We selected distance tolerances of 0.65Å.

After determining possible orientations, DOCK places the ligand molecule into each orientation and scores it with the force field. Once the ligand molecule is positioned, DOCK uses a rigid-body simplex minimization to find a local minimum. This minimization step is the most CPU-intensive step of the docking algorithm. In order to reduce the number of orientations minimized, we first use a “bump” filter. The bump filter evaluates

the orientation and determines if there will be a significant overlap between ligand and receptor atoms.

## Results

### *Regions*

Shapesite quickly defines regions from solid angle data. For the entire surface of trypsin, a structure with 223 residues, the calculation requires less than 5 CPU seconds, and less than 2 CPU seconds for trypsin inhibitor, with 56 residues, on an SGI Octane (single processor MIPS R10000 CPU). Because the solid angle algorithm requires a comparison of all surface points to all other surface points (Connolly 1986a), it requires significantly more CPU time. For example, for trypsin inhibitor calculating the solid angle on the SGI Octane requires 2,178 CPU seconds.

The formation of regions from individual points is dependent upon the order in which points are searched during the calculation; however, regardless of the order, derived site points display the complementary nature of known interfaces. For the case of trypsin-trypsin inhibitor, we compare the site points from the default ordering with site points derived from “shuffling” the residues in the input file. We examined the re-assembled trypsin-trypsin inhibitor interface for site points which lie within 2Å of one another across the interface. For the first, unshuffled run, there are 24 site points on trypsin within 2Å of a site point on trypsin inhibitor. These adjacent regions display complementarity: when the surface solid angles of the adjacent site points are summed, their average value is  $3.3\pi$  steradians, and their standard deviation is  $0.40\pi$ . For the shuffled data, there are 25 adjacent regions on the trypsin-trypsin interface, with an average value of  $3.2\pi$  and standard deviation of  $0.36\pi$ .

### ***Macromolecular Docking***

Docking studies are summarized in Tables 1, 2 and 3. For each of the test cases, the top-scoring orientation also has the lowest root-mean-square distance (RMSd) from the crystal complex orientation, and it is always less than 1.5Å RMSd. For comparison purposes, we report the DOCK force field score of the crystal complex, and the score of the complex minimized against the DOCK force field. For each proteinase-inhibitor complex in both sets of studies, the DOCK orientation is within 3.0 DOCK score units from the minimized crystal complex structure, and the DOCK orientation has a more favorable score than the non-force field-minimized orientation.

The use of the shape-based filter vastly reduces the number of orientations searched, and therefore the computational time for docking these molecules. The reduction in both computer time and number of orientations searched approaches 50-fold, as shown in Table 1. When examining the entire inhibitor surface, the improvement for trypsin/trypsin inhibitor was nearly 100-fold, with 11,655,935 orientations generated without shape-based filtering and 124,517 orientations generated with shape filtering.

**Table 1: Performance of DOCK runs with and without shapelite points**

Complex	Site points, inhibitor	Site points, proteinase	Orientations without shape	CPU (min.)	Orientations with shape	CPU (min.)
2ptc	50	67	145,163	2.0	4,601	0.23
1cho	49	61	88,541	55.4	2,707	1.6
2sni	56	79	745,934	593.4	20,803	13.8

These studies were performed with a subset of points on the protein and inhibitor, where the inhibitor sites are within 5Å of the proteinase in the crystal complex, and the proteinase sites are within 5Å of the inhibitor. We report the number of site points on the proteinase and the inhibitor, as well as the number of orientations generated by DOCK with and without the use of shape filtering. We also report the CPU time, in minutes, to perform the DOCK runs on a Silicon Graphics Octane (R10000).

**Table 2: Performance of DOCK runs with shapelite points**

Complex	Site points, Inhibitor	Site points, proteinase	Orientations	CPU (in minutes)
2ptc	213	67	124,517	5.1
1cho	189	61	18,375	35.2
2sni	226	79	321,591	625.0

These studies were performed using the entire inhibitor surface and the active site of the inhibitor surface, as defined by surface atoms which lie within 5Å of the inhibitor in the crystal complex. We report the number of site points on the proteinase and the inhibitor, as well as the number of orientations generated by DOCK with shape filtering. We also report the CPU time, in minutes, to perform the DOCK runs on a Silicon Graphics Octane (R10000).



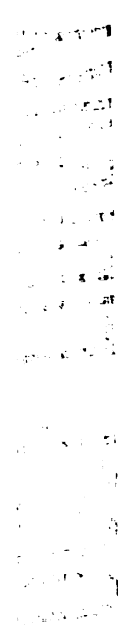
**Table 3: Results of DOCK runs with shapelite points**

Complex	Score, complex	Minimized score, complex	Score, DOCK/ shape	RMSd DOCK/ shape
2ptc	-72.26	-87.53	-85.71	0.85Å
1cho	-21.05	-75.94	-72.24	1.46Å
2sni	-50.02	-70.04	-68.65	0.37Å

Reported are the DOCK force field score of the proteinase-inhibitor complex before and after minimization, the top-scoring orientation from the DOCK runs with shape-based site points from the entire inhibitor surface, and the RMSd from the unminimized complex of the top-scoring orientation.

CPU times varied with the number of orientations generated and the number of orientations minimized. For subtilisin/turkey ovomucoid third domain, fewer orientations were generated than for trypsin/trypsin inhibitor, however, more of those orientations passed the bump filter, so more orientations were minimized, and therefore significantly more CPU time was required to search the orientation space.

An example of the docked conformation and native crystal complex conformation of 2ptc can be found in Figure 2.





**Figure 2. Trypsin inhibitor docked into trypsin (2ptc) using shapesite points from the entire inhibitor surface.**

The crystal structure orientation of trypsin inhibitor is shown in black and the docked structure is in gray. The RMSd between the two structures is 0.85Å. The docking site of the trypsin surface is shown in light gray.

## Discussion

Like Connolly's (Connolly 1986b)(Connolly 1992) and Lin et al.'s (Lin, et al. 1994) molecular shape descriptors, these shape-based site points are derived from the Connolly molecular surface. Unlike Connolly's earlier attempts, we do not describe the geometric fit of proteins and ligands as strictly knobs-into-holes, but allow for a range of shape. Like Lin et al., our site descriptor is closely tied to our docking algorithm. Their algorithm makes use of critical points classified as a cap, pit or belt, and normal vector. Our method allows for a range of shape but does not use a projected normal.

Earlier efforts from this group focused on the complexes examined in this study and uncomplexed forms of the same molecules (Shoichet and Kuntz 1991). In that study, the inhibitor was partitioned into several smaller groups of 40 to 60 spheres, or site points, and the proteinase active site was represented by 40 to 90 site points. The selection and reduction of site points was a highly interactive process. Some 1-2 million orientations were generated for the complexed sites in several separate DOCK2 runs which took several days to run.

For this study, a similar number of site points were generated for the proteinases and inhibitors as were for the previous study. Our site points were generated by their shape criteria, and required no further clustering efforts. While the region-building algorithm is dependent upon the order in which the points are searched, the resulting number and complementary nature of the site points varied little with the ordering of site points.

Unlike the earlier study, we were able to simultaneously examine the entire inhibitor surface in one DOCK run. The number of orientations generated with DOCK is dependent upon the site points, a minimum distance between site points, and a distance

tolerance set by the user. For these runs, the minimum distance between site points was 4Å and the distance tolerance was 0.65Å. When searching the whole surface of the inhibitor, the number of orientations generated varied from a tens of thousands with shape filtering to nearly 12 million without shape filtering. The differences and variations in the number of orientations generated and the significant change in time required to perform these runs (less than 10 hours) is due to the improvements in both software and hardware technologies.

With filtering based upon shape-based site points, we generated from 4,500 to 325,000 orientations for a protein-protein complex, depending upon the test case. The improvement with shape-based site points approached 100-fold when examining the entire inhibitor surface. With shape filtering, we quickly reached the same, or better, orientation than without shape filtering.

### **Acknowledgments**

This work was supported by NIH Training Grants GM08284, GM08388 and GM31497 from the Institute of General Medical Sciences, National Institutes of Health.

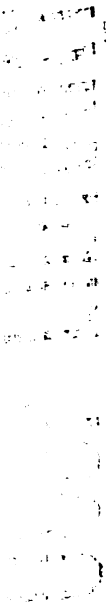
## References

- Connolly ML. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221:709-13.
- Connolly ML. 1986a. Measurement of protein surface shape by solid angles. *Journal of Molecular Graphics* 4:4-6.
- Connolly ML. 1986b. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. *Biopolymers* 25:1229-47.
- Connolly ML. 1992. Shape distributions of protein topography. *Biopolymers* 32:1215-36.
- Ewing TJA and ID Kuntz. 1997. Critical evaluation of search algorithms for automated molecular docking and database screening. *Journal of Computational Chemistry* 18:1175-89.
- Fischer D, SL Lin, HL Wolfson and R Nussinov. 1995. A geometry-based suite of molecular docking processes. *Journal of Molecular Biology* 248:459-77.
- Fujinaga M, AR Sielecki, RJ Read, W Ardelt, M Laskowski, Jr. and MN James. 1987. Crystal and molecular structures of the complex of alpha-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8 A resolution. *Journal of Molecular Biology* 195:397-418.
- Kuntz ID, JM Blaney, SJ Oatley, R Langridge and TE Ferrin. 1982. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology* 161:269-88.
- Lin SL, R Nussinov, D Fischer and HJ Wolfson. 1994. Molecular surface representations by sparse critical points. *Proteins* 18:94-101.
- Marquart M, J Walter, J Deisenhofer, WB Bode and R Huber. 1983. The geometry of the reactive site and of the peptide groups in trypsin, tyrypsinogen and its complexes with inhibitors. *Acta Crystallographica B* 39:480-490.

McPhalen CA and MN James. 1988. Structural comparison of two serine proteinase-protein inhibitor complexes: eglin-c-subtilisin Carlsberg and CI-2-subtilisin Novo. *Biochemistry* 27:6582-98.

Shoichet BK and ID Kuntz. 1991. Protein docking and complementarity. *Journal of Molecular Biology* 221:327-346.

Shoichet BK and ID Kuntz. 1993. Matching chemistry and shape in molecular docking. *Protein Engineering* 6:723-32.



## Chapter 3



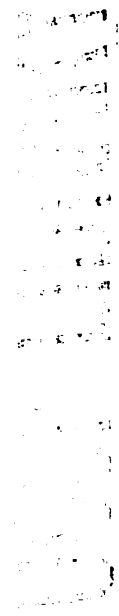
### **MACROMOLECULAR DOCKING OF A THREE-BODY SYSTEM: THE RECOGNITION OF HUMAN GROWTH HORMONE BY ITS RECEPTOR**

by

**Donna K. Hendrix, Teri E. Klein and Irwin D. Kuntz**



*This chapter was accepted for publication by Protein Science in January,  
1999. It is reprinted with the permission of Cambridge University Press.*





## Abstract

Human growth hormone (hGH) binds to its receptor (hGHR) in a three-body interaction: one molecule of the hormone and two identical monomers of the receptor form a trimer (Cunningham, et al. 1991). Curiously, the hormone-receptor interactions in the trimer are not equivalent and the formation of the complex occurs in a specific kinetic order (Cunningham, et al. 1991). In this paper, we model the recognition of hGH to the hGHR using shape complementarity of the three-dimensional structures and macromolecular docking to explore possible binding modes between the receptor and hormone. The method, reported previously (Hendrix and Kuntz 1998), is based upon matching complementary-shaped strategic sites on the molecular surface. We modify the procedure to examine three-body systems. We find that the order of binding seen experimentally is also essential to our model. We explore the use of mutational data available for hGH to guide our model. In addition to docking hGH to the hGHR, we further test our methodology by successfully reproducing sixteen macromolecular complexes from X-ray crystal structures, including enzyme-inhibitor, antibody-antigen, protein dimer and protein-DNA complexes.

## Introduction

Many cellular and physiological processes, from signaling to transcription, are performed and regulated by the interactions of macromolecules. Macromolecules must be able to recognize their binding partners in a specific manner to drive cellular events.

There are many facets to understanding macromolecular recognition. The assembly of a macromolecular complex is a multi-step event: the initial association of the molecules via random collision or a direct mechanism, followed by the formation of the non-covalent interactions that build a stable interface between the molecules that allow them to discern among many molecules, or recognize, a particular ligand. From available crystal structures, we know that the macromolecular surfaces that come together in a complex are highly complementary not only in their chemical features, but also in their shapes. These interfaces have been described as “lock-and-key”, or “hand-and-glove” interactions.

While these recognition interfaces, termed structural epitopes, tend to be large, mutational studies suggest that the energy of the interactions can be limited to the contribution of a small number of residues, the functional epitope (Cunningham and Wells 1993).

We can explore the possible binding modes, or orientations, of two or more molecular structures by docking. Docking programs position molecules to form complementary interfaces. When the functional epitope is known, it can be used to direct docking, for example by considering only the orientations where those residues are buried. A question we can address is whether the functional epitope provides sufficient information to identify the correct binding modes. More information may be required to find the most stable conformation. In this study, we use macromolecular docking of the functional and structural epitopes of human growth hormone and its receptor to examine recognition.

The cytokine superfamily of hormones and their receptors, which include prolactin, tissue factor, interleukins 1-7 and erythropoietin, regulates numerous physiological processes, such as growth and differentiation of blood cells, muscle, bone and cartilage (Nicola 1994). These proteins initiate signaling by forming a complex with their extracellular receptors. Human growth hormone (hGH) and erythropoietin (EPO) are well-characterized members of this superfamily. Both hGH and EPO initiate downstream effects by binding two monomers of their receptors to form a trimeric complex (Wells and de Vos 1993) (Matthews, et al. 1996).

Due to the availability of structural and mutational data for hGH and hGHR, we chose to focus on this complex rather than on EPO or other cytokine/cytokine receptor systems. Monumental efforts have been made to design EPO agonists and antagonists (Wrighton, et al. 1996) (Wrighton, et al. 1997) (Johnson, et al. 1997) (Johnson, et al. 1998). Peptide EPO agonists were discovered (Wrighton, et al. 1996), and a small molecule agonist was found to activate the granulocyte-colony stimulating factor (Tian, et al. 1998), but small-molecule or peptide hGH agonists have not been reported. There are X-ray crystal structures of the EPO receptor and its peptide agonist (Livnah, et al. 1996) and a recent structure of EPO receptor and a peptide antagonist (Livnah, et al. 1998); however, only recently has a structure of EPO bound to the EPO receptor been published (Syed, et al. 1998), while a structure of hGH bound to its receptor (de Vos, et al. 1992) is obtainable through the Protein Data Bank.

The recognition of hGH by its receptor has been examined by biochemical, mutational and structural analyses. These studies have revealed that the hormone binding to the receptor molecules sets off the signaling cascade through a specific sequence of

events (Cunningham, et al. 1991). The hormone first binds to one receptor molecule at site 1, then the hormone-receptor dimer complex binds to a second receptor molecule at a second site on the hormone, site 2. This is the signaling trimer: the receptor dimer plus the hormone. Crystallographic coordinates of hGH bound to the extracellular domain of its receptor show that the two receptor molecules form a large interface with one another (de Vos, et al. 1992). Mutational studies reveal “hot spots” (Clackson and Wells 1995) -- the functional epitope — that correlate with site 1. While the energy of binding appears to be localized to a small patch of surface, the recognition of hormone by receptor requires the formation of the trimeric complex that launches the signaling cascade. The structure of hGH bound to the hGHR reveals a structural epitope much larger in surface area than the functional epitope mapped by mutational analysis. The erythropoietin complex also has been shown to have hot spots by mutational analysis (Matthews, et al. 1996) (Middleton, et al. 1996) (Barbone, et al. 1997) (Syed, et al. 1998).

To explore the association of macromolecules, we use a docking method to match local regions of complementary shape on the macromolecular surfaces. Active sites of enzymes tend to be concave, shielding the active site residues from solvent; however, receptor systems, with extracellular domains extended into solvent, have convex and flat regions as well as concave sections. We have developed a method to describe these surfaces, and we use these descriptors to guide the docking of receptor systems and other macromolecular complexes.

Our goal is to predict the binding modes of macromolecular complexes from structures or models of individual molecules. We take a first step toward this goal by examining known complexes of x-ray crystal structures. To establish our method for receptor

systems, we first dock sixteen determined macromolecular complexes. These complexes are classified as enzyme-inhibitor (ten), protein-protein (three), and protein-DNA (three). We then use this method to explore the recognition of hGH binding to its receptor.

For the hGH/hGHR complex, we have adapted our general approach of examining two-body systems to three-body systems by a divide-and-conquer protocol: we first modeled the binding of the hormone to each receptor molecule; we next modeled the binding of receptor monomer to receptor monomer; we then modeled hormone binding to the dimerized receptor; and finally we modeled the binding of a second receptor molecule to the ligand-monomer complex. While the search for complementary docking surfaces can be done over the entire ligand and receptor, we also explore the usefulness of mutational and biochemical data to guide the search for surfaces known to contribute to the binding energy.

## **Results**

We present the results of docking our test suite of sixteen bimolecular complexes with shape-based site points, followed by the modeling of the hGH/hGHR complex. For bimolecular complexes, we refer to the smaller molecule in the complex as the ligand and larger molecule as the receptor.

### **Docking with surface shape-based site points**

We limited the search to known binding sites of the receptor molecules for this study. This strategy permits more intensive sampling of possible orientations within a fixed

amount of computer time. The sampling is defined by the input parameters to the UCSF DOCK program (Ewing and Kuntz 1997). Please see Methods for a more detailed discussion of the generation of shape-based site points, the selection of site points, the selection of docking parameters and the scoring of orientations.

Orientations of complexes were generated using shape-based site points and version 4 of UCSF DOCK. The site points used for docking included site points for the entire ligand surface of the inhibitors in enzyme-inhibitor complexes and for lysozyme in the lysozyme-antibody complex. For the receptor, site points were selected based upon previous knowledge of the crystal complex. A similar selection criteria, based upon the known geometry, was used for the both monomers when docking the protein dimers, a/b hemoglobin and HIV-1 protease. The selection criteria for the DNA molecules were based upon structural and biochemical data. The number of site points for each molecule in the test suite is listed in table 1.

The orientations generated by DOCK were scored with the force field score available with the DOCK suite of programs. The force field score selected an orientation from a cluster of orientations closest to the X-ray crystal structure as the best-scoring orientation for each of the sixteen test cases. For each of the enzyme-inhibitor test cases, the force field score, measured in DOCK units, favorably scored orientations similar to the native binding mode, and gave poorer scores for orientations farther from the native binding mode. It was common to see a small cluster of 3-5 orientations with scores within 5 DOCK units and RMSDs within 1.5Å of the native orientation. In three cases, serine proteinase B/potato inhibitor (4sgb), FAB/lysozyme (3hfm), and human TATA-box binding protein/DNA (1cdw), a favorably-scoring orientation could be found distant in

**Table 1: Test set for macromolecular docking**

Class	PDB Identifier <sup>a</sup>	Resolution (Å) <sup>b</sup>	# atoms, ligand <sup>b</sup>	# atoms, receptor <sup>b</sup>	# site points, ligand <sup>c</sup>	# site points, receptor
<i>proteinase-proteinase inhibitor</i>	2cpk	2.7	157	2665	132	114
	4cpa	2.5	285	2443	206	69
	4sgb	2.1	380	1310	263	74
	1cho	1.8	407	1757	189	62
	1tgs	1.8	416	1646	289	96
	2ptc	1.9	456	1631	213	67
	4tpi	2.2	471	1629	326	95
	2sni	2.1	515	1940	226	79
	1tec	2.2	522	1881	312	96
	2sec	2.8	530	1923	352	87
<i>protein-protein</i>	3hfm	3.0	1001	2114	582	83
	1hpy	1.9	758	758	163	290
	2mhb	2.0	1072	1137	230	110
<i>protein-DNA</i>	1aay	1.6	444	734	158	162
	1ytb	1.8	602	1421	179	183
	1cdw	1.9	652	1421	216	184

<sup>a</sup>Ten proteinase-proteinase inhibitor complexes, one antibody/protein complex, two protein dimer, and three protein-DNA complexes are included in this test set. Complexes are: 2cpk:c-AMP dependent protein kinase/protein kinase inhibitor; 4cpa:carboxypeptidase A/potato carboxypeptidase A inhibitor; 4sgb:serine proteinase B/potato inhibitor; 1cho:alpha-chymotrypsin/turkey ovomucoid third domain; 2ptc:beta-trypsin/trypsin inhibitor; 1tgs:trypsinogen/porcine pancreatic sensory trypsin inhibitor; 4tpi:trypsinogen/pancreatic trypsin inhibitor; 2sni:subtilisin novo/chymotrypsin inhibitor; 1tec:thermitase/eglin-c; 2sec:subtilisin carlsberg/N-acetyl eglin-c; 3hfm:IgG1 FAB fragment/lysozyme; 1hpy:HIV-1 protease; 2mhb:hemoglobin a/b; 1aay:ZIF268 zinc finger/DNA; 1ytb:yeast TATA-box binding protein/DNA; 1cdw:human TATA-box binding protein/DNA.

<sup>b</sup>The resolution of the crystal structure, the number of non-hydrogen atoms in the ligand and receptor molecules, and the number of site points representing these molecules is shown.

<sup>c</sup>Note that only the binding surface of the receptor is used for the dockings, not the entire surface of the receptor.

RMSd, but with scores within 10 DOCK units of the native orientation. For the 4sgb and 3hfm, several orientations within 3Å of the best-scoring orientations were found, each with scores more favorable than the alternate orientations. For 1cdw, the alternate orientation generated by DOCK was 26Å RMSd from the native complex, rotated 180 degrees about an axis through the DNA and protein relative to the native orientation. The protein-DNA contacts of this alternate orientation were primarily along the backbone. The score was only 1.5 DOCK units less favorable than the best-scoring orientation.

For all the test cases, the score of the docked orientation was within 4 DOCK units of the minimized score of the native orientation, except for one test case where the score favorably exceeded that of the minimized native orientation by more than 8 DOCK units. Scores, relative RMSds between all non-hydrogen atoms of the ligand, and CPU times are listed in table 2.

### **Comparison of spheres to surface shape-based site points**

In earlier docking studies we have used negative images of sites, primarily the overlapping sphere set created by the program SPHGEN (Kuntz, et al. 1982). This approach has been used for both small molecule and macromolecular docking (Shoichet and Kuntz 1991). We compared the surface shape-based procedure with our original method by generating spheres for the trypsin/trypsin inhibitor complex, 2ptc, and using the same docking parameters and scoring methods to dock the inhibitor to the enzyme. We used SPHGEN to build spheres for both the trypsin binding site and, in SPHGEN's ligand mode, for the interior of the trypsin inhibitor surface. Spheres were selected to be similar in number and distribution to the shape-based site points. 270 spheres filled the interior of



**Table 2: Results of macromolecular docking**

Class	PDB Identifier	Energy Score Minimized Crystal Complex <sup>a</sup>	Energy Score of DOCKed Complex, using Shape-based Site Points <sup>b</sup>	Relative RMSd (Å) from Minimized Crystal Complex	CPU time, in minutes (R10000)
<i>proteinase-proteinase inhibitor</i>	2cpk	-115.3	-112.9	0.45	44
	4cpa	-53.9	-53.8	1.10	36
	4sgb	-64.3	-61.9	1.62	116
	1cho	-75.9	-72.2	0.85	35
	1tgs	-95.4	-93.6	0.74	241
	2ptc	-87.5	-88.8	0.89	7.1
	4tpi	-91.4	-94.4	0.64	95
	2sni	-70.0	-69.8	0.85	339
	1tec	-77.4	-78.0	0.52	113
	2sec	-75.0	-74.4	0.85	130
<i>protein-protein</i>	3hfm	-67.2	-66.2	0.84	946
	1hpv	-179.0	-178.1	0.25	1129
	2mhb	-73.1	-76.1	1.59	201
<i>protein-DNA</i>	1aay	-127.3	-136.5	1.11	442
	1ytb	-183.3	-187.5	0.52	180
	1cdw	-97.4	-94.9	1.41	438

<sup>a</sup>The DOCK energy scores were calculated for the X-ray crystal complex.

<sup>b</sup>The entire ligand surfaces were sampled, while only the receptor binding sites, as defined by protein surface within 5Å of the ligand in the native complex, were used.

the trypsin inhibitor molecule, and 104 spheres described the volume of the inhibitor binding site on the trypsin molecule.

The trypsin inhibitor spheres were docked to the trypsin spheres. The number of orientations generated exceeded 660 million, in contrast to the 117,319 orientations

generated with the shape-based site points. The best orientation found with spheres had a score of -91.9 DOCK units, more favorable than the score of the minimized native complex of -87.5 DOCK units, and was 0.76 Å RMSd from the minimized native complex. The DOCK program required 4,464 minutes of CPU time on a SGI R10000 to generate this orientation. The best orientation found with shape-based procedure had a score of -88.8 DOCK units, and was 0.89 Å RMSd from the minimized native complex. The DOCK program ran for 7.1 CPU minutes to generate this orientation. The shape-based site points allowed DOCK to focus its matching on those orientations whose shapes complemented, while the spheres sampled all possible matches, eventually finding a match closer to the native complex, at the cost of 1000-fold greater sampling.

In a previous study directly comparable to the approach taken here (Shoichet and Kuntz 1991), the receptor was represented by clusters of 25 to 60 spheres, and the ligand molecule was partitioned and docked piecewise rather than docking the whole molecule at once. In that work,  $4 \times 10^5$  orientations were sampled to generate a near-native orientations for 2ptc, while the shape-based method of this study required  $1 \times 10^5$ ;  $1.6 \times 10^6$  orientations were examined for 1cho, while for these results we examined  $1.8 \times 10^4$ ;  $1.5 \times 10^6$  orientations were generated for 2sni, while  $1.5 \times 10^5$  were examined for this study. Thus, shape-based method provided a more than 3- to 20-fold improvement in sampling. A similar comparison was reported previously (Hendrix and Kuntz 1998).

### **Docking human growth hormone to its receptor**

We asked three questions:

Can the ternary complex be accurately reassembled using the DOCK program?

Does the order of docking individual components affect the final assembled structure?

Is knowledge of the mutational “hot spots” necessary and sufficient to achieve an accurate ternary structure?

To answer these questions, we examined the binding of human growth hormone to its receptor by docking, using the published structure from the Brookhaven Protein Databank, 3hhr (de Vos, et al. 1992). The docking sites are described in table 3 and shown in figure 1.

**Table 3: Sites for hGH docking to the hGHr**

Site	Number of site points	Selection method <sup>a</sup>
Hormone, full surface ( <b>H</b> )	581	
Hormone side 1 ( <b>H1</b> )	90	Binding surface, from structure
Hormone side 2 ( <b>H2</b> )	60	Binding surface, from structure
Hormone sides 1 and 2 ( <b>H12</b> )	150	Binding surface, from structure
Receptor, side 1, full surface ( <b>R1F</b> )	604	
Receptor, side 2, full surface ( <b>R2F</b> )	563	
Receptor, side 1 ( <b>R1</b> )	92	Binding surface, from structure
Receptor, side 2 ( <b>R2</b> )	58	Binding surface, from structure
Receptor, sides 1 and 2 ( <b>R1R2</b> )	150	Binding surface, from structure
Receptor, side 1, stem region ( <b>RS1</b> )	34	Binding surface, from structure
Receptor, side 2, stem region ( <b>RS2</b> )	38	Binding surface, from structure
Hormone side 1, mutants ( <b>H1M</b> )	57	Mutational data
Receptor, side 1, mutants ( <b>R1M</b> )	39	Mutational data
Receptor, side 2, mutants ( <b>R2M</b> )	45	Mutational data

<sup>a</sup>Site points were generated for the entire surfaces of each molecule, and docking sites were selected from the surfaces. Docking sites were selected based on two methods: one method based on the known structure, and a second method based on the mutational data. The sites based on mutational data are from residues identified as contributing a large portion of the binding energy (Wells 1996). There are no mutational site residues on site 2 of the hormone surface.

We executed twelve calculations, listed in table 4. Of these calculations, eight were carried out on sites defined by the structural data, and four were defined by mutational data, described in the following section. The results are shown in table 5.

**Table 4: Calculations for docking hGH to the hGHR**

Receptor sites <sup>a</sup>	Hormone sites <sup>a</sup>	
<b>R1</b>	<b>H</b>	Receptor 1 binding site/Whole hormone
<b>R1</b>	<b>H1</b>	Receptor 1 binding site/Hormone binding site 1
<b>R2</b>	<b>H</b>	Receptor 2 binding site/Whole hormone
<b>R2</b>	<b>H2</b>	Receptor 2 binding site/Hormone binding site 2
<b>RS1</b>	<b>RS2</b>	Receptor stem sites
<b>H2&amp;RS1</b>	<b>R2&amp;RS2</b>	Dimerized receptor 1 and hormone/Receptor 2 binding and stem sites
<b>R1R2</b>	<b>H</b>	Dimerized receptors/Whole hormone
<b>R1R2</b>	<b>H12</b>	Dimerized receptors/Hormone binding sites 1 and 2
<i>Docking with mutational sites</i>		
<b>R1M</b>	<b>H</b>	Receptor 1 mutational sites/whole hormone
<b>R1M</b>	<b>H1M</b>	Receptor 1 mutational sites/hormone mutational sites
<b>R2M</b>	<b>H</b>	Receptor 2 mutational sites/whole hormone
<b>R2M</b>	<b>H2M</b>	Receptor 2 mutational sites/Hormone binding site 2

<sup>a</sup>Sites are represented as described in table 3. Results are in tables 5 and 6.

We can reassemble the two body receptor-hormone complexes, R1+H and R2+H, and can dock the hormone accurately to the dimerized receptor, R1R2+H. Further, we can dock the hormone-receptor 1 complex, represented by the sites (H2&RS1), to the second receptor (R2&RS2). In all but one of these examples, the final energies and geometries

correspond closely with the crystallographic result. The exception is docking the hormone to the second receptor site where the best score is not as favorable as that for the crystal complex (-69.6 versus -91.3 DOCK units). A search using just the hormone and receptor sites is more successful (-87.7 versus -91.3 DOCK units), implying that the problem is sampling rather than a scoring issue. **Thus, we conclude that the extension of docking to multimeric systems is feasible.**

We turn to the question of whether the structure depends strongly on the order in which the components are docked. Here we find an interesting result. While the hormone can be added to either receptor first without difficulty, if the receptors are docked in the absence of the hormone, a pivoted receptor dimer structure is the best-scoring orientation. Since this structure, RS1+RS2, is some 20Å RMSD from the crystal structure, it is impossible to use it to obtain the crystallographic ternary complex. This rotated structure might represent a biologically interesting “off” state of the receptor.

**Table 5: Results of docking binding sites of hGH to the hGH receptor**

Sites docked <sup>a</sup>	Native score (DOCK units) <sup>b</sup>	Docked score (DOCK units)	Relative RMSd (Å)	Orientations searched
<b>R1 + H</b>	-148.6	-147.6	0.28	279,617
<b>R1 + H1</b>	-148.6	-144.8	0.21	2,767
<b>R2 + H</b>	-91.3	-69.6	2.00	126,133
<b>R2 + H2</b>	-91.3	-87.7	1.33	2,971
<b>RS1 + RS2</b>	-40.1	-46.5	19.20	2,010
<b>(H2&amp;RS1) + (R2&amp;RS2)</b>	-130.4	-130.1	0.47	2,440
<b>R1R2 + H</b>	-225.2	-225.2	0.04	2,401,568
<b>R1R2 + H12</b>	-225.2	-225.1	0.05	56,980

<sup>a</sup>Sites are represented as described in tables 3 and 4. The docking of sites is designated by “+”, while combined sites are designated by “&”. For example, **(H2&RS1) + (R2&RS2)** identifies the docking of the combined sites **H2** and **RS1** to the sites represented by **R2** and **RS2** combined.

<sup>b</sup>Five different boxes enclosing the scoring grid were used: the site 1 box, for all **R1** dockings; the site 2 box, for **R2** dockings; the stem box for docking **RS1 + RS2**; the stem plus site 2 box for docking **(H2&RS1) + (R2&RS2)**, and the site 1 plus site 2 box to dock **(R1&R2)+H**. Because of the different boundaries, scores are not additive, i.e., scores from the site 1 box plus scores from the site 2 box are not equivalent to scores from the site 1 plus site 2 box.

### Docking mutational sites on hGH to the hGHR

We next turn to mutational data (Wells 1996) to help select site points for docking.

Mutationally-derived sites on the receptor, R1M and R2M, and on the hormone surface, H1M, are shown in figure 1 and listed in table 3.

To determine whether these sites provide sufficient information to achieve a ternary structure, we docked the hormone mutational sites to the each receptor mutational sites. The results are shown in table 6. We found that the mutational sites do provide sufficient information since we were able to reassemble the two-body complexes, R1M+H and

R2M+H, However, we had to search 3-fold more orientations, and the scores of the docked orientations of R1M+H and R2M+H complexes were less favorable than those of the native orientations. We also docked the mutational sites of the hormone, H1M, to the mutational sites of receptor site 1, R1M. Similar to our results with the whole hormone, we were able to find an orientation close in RMSd, but with a less favorable score than the native complex. For R2M+H2, we were able to determine an orientation close in score and RMSd to the native orientation.

**Table 6: Results of docking mutational sites on hGH to mutational sites on hGH receptor**

Sites docked	Native score (DOCK units) <sup>a</sup>	Docked score (DOCK units)	Relative RMSd (Å)	Orientations searched
<b>R1M + H</b>	-148.6	-137.4	0.87	458,055
<b>R1M + H1M</b>	-148.6	-129.8	1.08	11,780
<b>R2M + H</b>	-91.3	-82.6	0.71	384,645
<b>R2M + H2</b>	-91.3	-90.0	1.28	75,650

<sup>a</sup>Two boxes enclosing the scoring grid were used: the site 1 box, for all **R1**, **R1M** dockings, and the site 2 box, for **R2** and **R2M** dockings.

It was curious that significantly more orientations were generated with the same matching parameters and similar number of site points for R2M+H2 than R1M+H1M. The R2M and H2 sites generated 75,650 orientations vs. 11,780 orientations on the R1M and H1M sites. The likely cause of this effect is the shape character of R2 and H2, which allowed for less shape filtering and generated more matches. The R1M and H1M sites contained greater variation of shape character over the patch, so fewer matches were generated.

We conclude that site points selected by mutational data can be used to dock the hGH complex, but additional characteristics in the structural epitope make the docking effort more computationally efficient.

## **Discussion**

We have introduced the use of shape-based site points for macromolecular docking and have shown that these site points may be used to explore a variety of complexes. Additionally, we have used shape-based site points to explore the binding of the hGH/hGHR complex, finding that the model mimics the experimental data, and discovering a novel orientation of the receptor.

### ***Shape-based site points***

Sixteen macromolecular complexes were re-assembled using shape-based site points. For each of these cases, we found the best-scoring orientation to also be within 1.6Å of the native orientation. The general trend of decreasing score with increasing RMSd was seen (figure 6), and it was common to see a small cluster of 3-5 orientations with scores within 5 DOCK units and RMSDs within 1.5Å of the native orientation. Alternate orientations were found in three cases, and for two of the three cases the orientations could be suspected as false positives due to the distribution of scores, with a cluster of orientations around the best-scoring orientation with scores within 10 DOCK units. For 1cdw, the alternate orientation would be much more difficult to detect in a blind docking experiment. Such alternate orientations have been commented on previously (Shoichet and Kuntz 1991) and might represent alternative binding modes of the ligand.



While the entire surface of ligand molecules was examined, a smaller portion of the receptor surface was searched. Because active site residues, mutational or biochemical data are frequently available, it was reasonable to localize the search to a region of one surface. However, it was not determined whether this method could find correct binding sites and binding modes without biasing the search.

In addition to docking the set of macromolecules, we compared docking with shape-based site points to docking with spheres from the program SPHGEN. SPHGEN creates an inverse image by filling the pocket or volume with intersecting spheres. The SPHGEN method requires that the molecules have concave or convex sites so that it might build spheres into the volume. These features are not always found in the binding sites of larger molecules, such as cellular receptors. Shape-based site points are derived by searching for patches of like shape character regardless of whether it is concave, convex, or flat.

A significant advantage to using shape-based site points is a reduction of the search space offered by shape-based filtering. While CPU times for docking are highly dependent on the features of the sites and the docking parameters, our docking algorithm is based on a comparison of all inter-site distances on the ligand to all inter-site distances on the receptor. The matching of distances begins by building a table whose elements are all distances (Ewing and Kuntz 1997). The rows and columns of the table are the pairs of possible overlapping site points on the receptor. This table is size  $(L \times R) \times (L \times R)$ , where L and R are the number of site points on the ligand and receptor. If all non-unique possibilities are compared, then the search is a function of the number of sites on the ligand and the number of sites on the receptor,  $(\sum L \times R) - (L \times R)$ . By shape-based filtering, we only compare distances on the ligand and receptor if the site points on the

ligand complement the site points on the receptor. This reduces the size of the table and the size of the search. At worst, all site points complement, and the shape-based filtering offers no benefit. At best, a vast reduction in the number of compared distances is seen. For most macromolecular systems, shape-based filtering reduces the number of sites examined in the search, so larger systems with more site points are docked in less CPU time.

For a direct comparison between shape-based site points and spheres, we used trypsin and pancreatic trypsin inhibitor. The sphere method found a better-scoring orientation than the shape-based method, but at greater computational cost. Both methods found orientations whose scores were better than the minimized native orientation.

### ***Docking hGH to the hGHR***

Cytokines function as signal transducers by recognizing molecules on cell surfaces and initiating signaling cascades within the cell, and they are known for pleiotropy and redundancy. Early studies showed that a cytokine may initiate a multitude of effects, and yet the removal of some cytokines or their receptors from a genome may have no visible effect (Nicola 1994). This feature can be explained, in part, by the ability of cytokines to bind multiple receptors, and receptors to bind multiple cytokines (Nicola 1994). Cytokine receptors have been optimized to recognize multiple cytokines, and cytokines have been optimized to recognize multiple receptors. For example, hGH binds and activates the prolactin receptor, using overlapping regions to bind both the prolactin and hGH receptors (Somers, et al. 1994).

In this study, we have used structural and biochemical data to model and recapitulate the binding behavior of a cytokine-receptor system. Experiments have shown the necessity of binding site 1 of hGH to the hGHR before formation of the receptor dimer and signaling can take effect (Cunningham, et al. 1991). We see a similar preference in our docking experiment. Site 1 of the receptor can readily select the proper face of the hormone from the surface of the entire hormone. When examining site 2, we must increase sampling and focus on one face of the hormone to produce an orientation close in score to that seen in the crystal complex.

Echoing what is seen in experimental data, we can dock the pre-formed dimer of hGH bound to one molecule of hGHR with relative ease, sampling only a few thousand orientations, but we do not easily form a native-like orientation with hormone binding to site 2, in the absence of receptor 1, or with the receptor molecules, in the absence of hormone. The receptor stem alone does not contribute significantly, as shown by the docking of the stem regions by themselves. We were able to generate orientations with scores more favorable than that of the native orientation, but distant in RMSd, when docking the stem sites. It may be that the receptor molecules “sample” this position in the cell membrane, but that this orientation does not activate the signal. Thus, the receptor molecules might lie in a nonproductive mode in close proximity, waiting for the hormone molecule to place the receptor molecules in proper position for activation. This observation is consistent with the data for EPO, where it is known that the relative orientation of the receptor dimer is key to the efficacy of activation (Syed, et al. 1998). Additionally, antagonists can form non-productive dimers of erythropoietin receptor

(Livnah, et al. 1998), and dimers of erythropoietin receptor exist in the cell in the absence of ligand (Miura and Ihle 1993).

Mutational data suggest that the binding determinants of hGH and hGHR are located within a handful of residues, the functional epitope (Clackson and Wells 1995). We found that by docking only sites associated with those amino acids, H1M, R1M, and R2M, we were able to generate near-native orientations with acceptable, favorable scores; however, in order to find these orientations, we had to increase sampling significantly, generating four-fold more orientations for H1M docking to R1M than for docking their structurally derived counterparts. We conclude that the geometric information found within the functional epitope is necessary and sufficient for the recognition event, but that using other structural features increased the speed of the docking calculation.

We have demonstrated another approach to a well-studied problem: docking recognition sites that are strongly complementary. As those who have preceded our efforts (Connolly 1986b) (Connolly 1992) (Shoichet and Kuntz 1991) (Fischer 1995) (Gabb, et al. 1997) (Jiang and Kim 1991), we were able to find orientations very close to the native orientation in RMSd space. We have consistently found that the best-scoring orientation was also in a cluster of orientations closest in RMSd to the minimized native complex. This observation was not always true in earlier work.

A novel feature of this work has been the examination of a tri-molecular complex, hGH and hGHR. The divide-and-conquer approach has worked well for these calculations, and we have generated results which compare favorably with the native structures.

Future work shall focus on determining a force field for previously uncomplexed molecules. We have modeled all molecules as rigid objects, but they are dynamic molecules which will change and shift upon binding, as illustrated in hGH's binding to prolactin receptor and hGHR. The challenge will be in determining how to account for these changes both in the generation of orientations and in the scoring of orientations. Macromolecules change structurally upon binding, and these adaptations must be identified and accommodated in docking models and scoring schemes. With more accurate modeling tools, it may be possible use docking methods and structure to explore the function of macromolecules. In addition to cytokine/receptor systems such as hGH/hGHR and prolactin/hGHR, many other intriguing multi-macromolecular systems exist: proteins from signaling pathways such as G $\beta$  subunits from G-proteins, or ras/raf from tyrosine kinase receptor pathways; protein-nucleic acid systems, like small nuclear ribonucleoprotein particles; and structural and motor systems, such as microtubules or actin-myosin complexes.

## **Conclusions**

We have developed a new approach to identify sites for macromolecular docking based upon the local shape of the molecular surface. This method reduces the sampling requirement by a factor of 1,000 when compared to the previous site representation method. We tested our method on sixteen protein-protein and protein-DNA complexes. Focusing the search on the binding surface of the larger molecule in the complex, we

accurately reproduced the crystal structure geometries of these complexes to within 1.7Å RMSd.

We extended the two-body method to explore the three-body binding of hGH to the hGHR. First, we examined the structural epitope. As seen in the experimental binding process, we found that site 1 of the receptor can select the native binding mode of the hormone, but site 2 cannot, unless the hormone is previously bound to site 1. Next, we examined the docking of the receptor molecules of the hGHR, finding an alternate binding mode 19.2Å RMSd from the native complex. Finally, we explored a representation of the functional epitope of hGH and hGHR. We found that we were able to reproduce the crystal structure orientation, but using the functionally important residues greater sampling was required than docking of the structural epitope.

## **Methods**

We report general approaches for docking our sixteen macromolecular test cases, including a description of the method to generate spheres for comparing the performance of spheres to that of the shape-based site points. We follow this discussion with a more detailed description of the methods of the hGH/hGHR calculations.

### **Building site points**

Our docking method consists of three steps: defining the shape-based site points, docking with the site points, and scoring the docked complexes. Details of the site point

generation algorithm have been described in a previous publication (Hendrix and Kuntz 1998). A brief description is given here.

Site points are derived from the molecular surface, as defined by the Connolly MS program (Connolly 1983b). The MS program result is a set of points which lie on the surface of the molecule, as well as a surface normal and associated area for each point. The solid angle of each point of the molecular surface is determined using the Connolly algorithm (Connolly 1986a). We use the solid angle to describe the shape of the surface. A probe sphere is placed with its center on a surface point, and the surface area of the sphere, normalized by the square of the radius of the sphere, which lies inside of the molecular surface is the solid angle. A two-dimensional description of the solid angle is shown in figure 3. A maximum solid angle value is  $4\pi$  steradians, for a point which is completely buried, while a point which lies entirely outside the molecular surface has a solid angle value of 0 steradians. For this study, the probe sphere has a radius of  $5.0\text{\AA}$ . The solid angle is determined for each point on the molecular surface.

We derive shape-based site points from the molecular surface and solid angle measurements. Near-neighbor lists are defined for each molecular surface point, and the neighbor lists are searched for points of similar solid angle value. Thus, a regional clustering is performed (figure 4), where regions of surface with similar, unchanging solid angle values are identified. These regions vary in size from  $5\text{\AA}^2$  to  $15\text{\AA}^2$ .

For the enzyme-inhibitor test cases, site points were generated for the entire inhibitor (ligand) surface, while for the enzyme (receptor) site points were generated only for the binding surface, as defined by surface within  $5\text{\AA}$  of the inhibitor when in complex with the enzyme. Similarly, for the antibody-lysozyme complex (3hfm), site points were generated

for the entire surface of lysozyme, and the binding surface of the antibody. For a/b hemoglobin (2mhb), for the a monomer, a 5Å binding surface was defined, while for the b monomer, site points within 10Å of the a monomer when in the dimer form were selected. The same method was used to select points on each monomer of the HIV-1 protease dimer.

For the human TATA-box binding protein structure (1cdw) and the ZIF268 zinc-finger (1aay) site points were defined on the entire surface of the DNA molecule except for the two most distal base pairs. Site points on the proteins came from 5Å binding surface. The yeast TATA-box DNA molecule was much larger (29 base pairs, vs. 16 base pairs), so a slightly different approach was used to reduce the number of site points for the yeast TATA-box DNA molecule. For the yeast structure (1ytb), site points on the DNA molecule associated with the TATA box and 2 base pairs 3' and 5' of the box defined the binding region. Again, site points on the protein came from the 5Å binding surface.

Surface shape-based site points are generated to represent the local shape of the molecule. Each site point is representative of a region of the surface and the shape of that region of surface. Site points are generated for both ligand and receptor surfaces, and are well-distributed over the surface of each molecule. The number of site points varies primarily with the size of the molecule, but the overall character of the molecular shape and the size of the probe sphere can influence the number of site points as well. If a molecule has a protrusion that is approximately the size of the probe sphere radius, the surface one probe sphere radius from the protrusion will at first capture the local shape of the protrusion, then show a discontinuity in the shape measurement when the protrusion is no longer within the probe sphere. This type of discontinuity in shape measurement can increase the number of site points.



For most macromolecules, the solid angle values of the site points span from  $0.2\pi$  steradians for points on very extended portions of the molecule, for example, a lysine side chain exposed to solvent, to nearly  $4\pi$  steradians for buried water molecules.

We have generated site points for a test set of sixteen macromolecular complexes. These complexes are enzyme-inhibitor, protein-protein, and protein-DNA complexes, with ligands varying in size from 157 to 1001 non-hydrogen atoms. The complexes and the number of site points are listed in table 1.

### **Docking macromolecules**

To dock the macromolecules, we used the most recent release of UCSF DOCK (Ewing and Kuntz 1997). DOCK generates orientations by identifying distances between site points on the ligand that are identical to distances between site points on the receptor. DOCK has several user-defined parameters, including the minimum distance compared, a distance tolerance, which is the allowable difference in distances that can generate a match, and the minimum and maximum number of nodes, the equivalent distances, to generate a match. These values were generously set for initial runs, but optimized to improve CPU times for the reported runs. The original parameters which generated near-native orientations for all of the test runs were: minimum distance of  $4\text{\AA}$ , distance tolerance of  $0.7\text{\AA}$ , 4 nodes minimum, and 10 nodes maximum.

DOCK finds equivalent distances by defining a table of all possible ligand-receptor site point matches (Ewing and Kuntz 1997). For these studies, we also demanded that the site points were matched only if they were of complementary shape measurement. We considered two points complementary if their shape measurements summed to a value

equal to or greater than approximately  $3\pi$  steradians. This was accomplished using the coloring feature that is one option of version 4 of UCSF DOCK. Site points were assigned a shape color which was the integer value of the solid angle measurement, thus losing some detail, but allowing for quicker implementation. We allowed some matches such that the sums of the solid angle values were less than the target  $3\pi$  steradians, based upon the observed matched site points in known complexes. We did not allow matches with solid angle sums to exceed  $4\pi$  steradians, as these matches did not complement.

Prior to scoring, the DOCK orientations are checked to determine if there is significant overlap between ligand and receptor. We refer to this step as a bump filter. A bump is defined as an overlap of the van der Waals radii. For initial runs, the bump filters were set to values which depended on the size of the ligand: larger ligands were allowed more bumps. Initial runs for the enzyme-inhibitor systems allowed 8 bumps. For the protein-protein and protein-DNA systems, initial runs allowed from 12 to 20 bumps, varying with the size of the ligand.

### **Scoring docked orientations**

Docked orientations of macromolecules were scored for fitness using the DOCK force field score generated by GRID, a revision of CHEMGRID (Meng, et al. 1992) available with UCSF DOCK version 4. The receptor binding region was placed on a grid with a resolution of  $0.3\text{\AA}$ . DOCK force field scoring represented the van der Waals and electrostatic components of intermolecular energies:

$$\text{force field score} = \sum_{i=1}^{lig} \left[ \sqrt{A_{ii}} \cdot \sum_{j=1}^{rec} \frac{\sqrt{A_{jj}}}{r_{ij}^{12}} - \sqrt{B_{ii}} \cdot \sum_{j=1}^{rec} \frac{\sqrt{B_{jj}}}{r_{ij}^6} + 332q_i \cdot \sum_{j=1}^{rec} \frac{q_j}{Dr_{ij}} \right]$$

*van der Waals* *electrostatic*

Charges were AMBER united-atom charges for proteins and nucleic acids (Weiner, et al. 1984), and were pre-calculated on the grid for the receptor. Scores for each orientation of ligand are calculated on the grid, interpolating from nearest grid points. Scores are reported as DOCK units.

An example of scoring 200 orientations is shown in figure 6. The energy well that is seen in figure 6 is representative of the enzyme-inhibitor and protein-protein complexes but is less well-defined for the protein-DNA test cases.

### **Generating spheres**

Spheres were generated for the test case of trypsin and trypsin inhibitor (2ptc) with the SPHGEN (Kuntz, et al. 1982), part of the DOCK suite of programs. SPHGEN places spheres tangent to two points on the molecular surface. We ran SPHGEN with default parameters for trypsin, but for trypsin inhibitor it was run in ligand mode, placing spheres in the interior of the molecular surface rather than in cavities outside of the molecule. For trypsin, we selected a spheres in the binding site, removing only those spheres which were not within the 5Å binding site as defined for the shape-based site points. For trypsin inhibitor, we began our sphere selection with all available spheres, then thinned them by placing them on a 1.35Å grid and requiring that only one sphere lie in any grid cell. Using this method, we reduced the number of spheres from 974 to 270.

### **Test cases**

The X-ray crystal structures of sixteen complexes were selected from the Protein Data Bank (Bernstein, et al. 1977)(Abola, et al. 1987) (Couch, et al. 1995). These complexes consist of enzyme-inhibitor complexes, dimers, one antibody-antigen complex and three protein-DNA complexes. The complexes, PDB identifiers and numbers of atoms are listed in table 1. The structure of human growth hormone bound to its receptor is available from the Protein Data Bank, with the identifier 3hrh (de Vos, et al. 1992).

### **Macromolecular docking techniques: hGH to the hGHR**

Site points were selected using two different methods, one based solely on structural information, and the other based upon mutational data. The structural method used the known complex structure, selecting shape-based site points on the hormone within 5Å of the receptor, and on the receptor molecules within 5Å of the hormone. These sites are shown in figure 1. Orientations were scored on a grid, built with GRID, as described previously for the macromolecular docking test suite. The grid dimensions were determined by the position of the site points, to include the site points used for a calculation and an additional 10Å along both directions of  $x$ -,  $y$ -, and  $z$ -axes. For the 12 calculations, five grids were created, based upon the sites R1R2, R1, R2, RS1 and (RS1&H2). For each grid, the score of the minimized crystal complex is reported in table 5 as the native score. All grids had a resolution of 0.3Å.

Matching parameters for the DOCK program dramatically affect the results of a docking calculation, particularly the minimum distance compared, the distance tolerance,

which determines how similar two distances must be to be considered equal, and the number of nodes, or distances, matched to generate an orientation. The selection of minimum distance was based on the size of the region, and the number of nodes was based on the number of site points. For docking to the R1 and R2 sites, the minimum distance was 4-4.5Å. For docking larger sites such as R1R2 and (H2&RS1), the minimum distance was set much larger because the sites were much larger: using 15Å for R1R2 and 12Å for (H2&RS1). This reduced the CPU time of the search by focusing the search on orientations that had distant complementary features. A match consisted of 5 nodes for the R1 and R2 sites. Distance tolerances for all runs varied from 0.55Å to 0.7Å.

The second method for selecting site points was based on mutational analysis. We selected site points on the receptor molecule surfaces associated with the residues identified as the hot spots, or functional epitopes, by mutational analysis (Clackson and Wells 1995). We refer to these sites as the mutational sites. These residues are reported to provide most of the binding affinity (Wells 1996) and include R43, E44, I103, W104, I105, P106, D164, I165, W169. There is a mutational site on each molecule of the receptor. We used the same criteria for the hormone surface. The mutational site residues on the hormone surface are on one face of the hormone, and thus primarily affect the binding to one of the two receptor molecules. The mutational site residues on the hormone that account for nearly 85% of the binding affinity include K41, L45, P61, R64, K172, T175, F176 and R178. These sites map to the binding region of hormone to receptor 1. None of these sites lie near the hormone binding site with receptor 2 (figure 1). The mutational sites are nearly a subset of the structural sites, with a few points from the mutational sites outside but near the structure-based sites.

As with the other receptor sites, for docking to the R1M site the minimum distance was 4-4.5Å. The number of nodes required for a match was 4 nodes for matching mutational sites on the receptor to the whole hormone, and 3 nodes to match R1M to H1M and R2M to H2. Distance tolerances varied from 0.55Å to 0.7Å.

### **Acknowledgments**

We thank Heidi Houtkooper for her assistance and guidance with the artwork, Dr. Elaine Meng for her many helpful comments and suggestions, and the UCSF Computer Graphics Lab for scientific and computational support. This work was supported by the National Institutes of Health GM31497 (I. Kuntz, principal investigator), P41 RR-01081 for T.E.K. and H.H. (T. Ferrin, Principal Investigator), and Training Grants GM08284, GM08388 for D.K.H.

## References

- Abola EF, FC Bernstein, SH Bryant, TF Koetzle and J Weng. 1987. Protein Data Bank: *Crystallographic Databases -- Information Content, Software Systems, Scientific Applications*. F. H. Allen, B. Bergerhoff and R. Sievers. Bonn/Cambridge/Chester, Data Commission of the International Union of Crystallography: 107-132.
- Barbone FP, SA Middleton, DL Johnson, FJ McMahon, J Tullai, RH Gruninger, AE Schilling, LK Jolliffe and LS Mulcahy. 1997. Mutagenesis studies of the human erythropoietin receptor. Establishment of structure-function relationships. *J Biol Chem* 272:4985-92.
- Bernstein FC, TF Koetzle, GJB Williams, EF Meyer, Jr., MD Brice, JR Rodgers, O Kennard, T Shimanouchi and M Tasumi. 1977. The Protein Data Bank: a computer-based archiveal file for macromolecular structures. *J Mol Biol* 112:535-542.
- Clackson T and JA Wells. 1995. A hot spot of binding energy in a hormone-receptor interface. *Science* 267:383-6.
- Connolly ML. 1983a. Analytical molecular surface calculations. *J Appl Crystallogr* 16:548-558.
- Connolly ML. 1983b. Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221:709-713.
- Connolly ML. 1986a. Measurement of protein surface shape by solid angles. *J Mol Graph* 4:4-6.
- Connolly ML. 1986b. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit Interface. *Biopolymers* 25:1229-1257.
- Connolly ML. 1992. Shape distributions of protein topology. *Biopolymers* 32:1215-1236.

Couch GS, EF Petterson, CC Huang and TE Ferrin. 1995. Annotating PDB files with scene information. *J Mol Graph* 13:153-158.

Cunningham BC, M Ultsch, AM De Vos, MG Mulkerrin, KR Clauser and JA Wells. 1991. Dimerization of the extracellular domain of the human growth hormone receptor by a single hormone molecule. *Science* 254:821-825.

Cunningham BC and JA Wells. 1993. Comparison of a structural and a functional epitope [published erratum appears in *J Mol Biol* 1994 Apr 8;237(4):513]. *J Mol Biol* 234:554-63.

de Vos AM, M Ultsch and AA Kossiakoff. 1992. Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science* 255:306-12.

Ewing TA and ID Kuntz. 1997. Critical evaluation of search algorithms for automated molecular docking and database screening. *J Comput Chem* 18:1175-1189.

Ferrin TE, CC Huang, LE Jarvis and R Langridge. 1988. The MIDAS display system. *J Mol Graphics* 6:13-27.

Fischer D, Lin, S.L., Wolfson, H.L., Nussinov, R. 1995. A geometry-based suite of molecular docking processes. *J Mol Biol* 248:94-101.

Gabb HA, RM Jackson and MJ Sternberg. 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* 272:106-20.

Hendrix DK and ID Kuntz. 1998. Solid angle-based site points for molecular docking. *Pacific Symposium on Biocomputing 1998*:1234-1244.

Huang CC, EF Pettersen, TE Klein, TE Ferrin and R Langridge. 1991. Conic: A fast renderer for space-filling molecules with shadows. *J Mol Graph* 9:230-236.

Jiang F and SH Kim. 1991. 'Soft docking': matching of molecular surface cubes. *J Mol Biol* 219:79-102.



Johnson DL, FX Farrell, FP Barbone, FJ McMahon, J Tullai, K Hoey, O Livnah, NC Wrighton, SA Middleton, DA Loughney, EA Stura, WJ Dower, LS Mulcahy, IA Wilson and LK Jolliffe. 1998. Identification of a 13 amino acid peptide mimetic of erythropoietin and description of amino acids critical for the mimetic activity of EMP1. *Biochemistry* 37:3699-3710.

Johnson DL, FX Farrell, FP Barbone, FJ McMahon, J Tullai, D Kroon, J Freedy, RA Zivin, LS Mulcahy and LK Jolliffe. 1997. Amino-terminal dimerization of an erythropoietin mimetic peptide results in increased erythropoietic activity. *Chem Biol* 4:939-50.

Kuntz ID, JM Blaney, SJ Oatley, R Langridge and TE Ferrin. 1982. A geometric approach to macromolecule-ligand interactions. *J Mol Biol* 161:269-88.

Livnah O, DL Johnson, EA Stura, FX Farrell, FP Barbone, Y You, KD Liu, MA Goldsmith, W He, CD Krause, S Pestka, LK Jolliffe and IA Wilson. 1998. An antagonist peptide EPO receptor complex suggests that receptor dimerization is not sufficient for activation. *Nature Structural Biology* 11:993-1004.

Livnah O, EA Stura, DL Johnson, SA Middleton, LS Mulcahy, NC Wrighton, WJ Dower, LK Jolliffe and IA Wilson. 1996. Functional mimicry of a protein hormone by a peptide agonist: the EPO receptor complex at 2.8 Å. *Science* 273:464-71.

Matthews DJ, RS Topping, RT Cass and LB Giebel. 1996. A sequential dimerization mechanism for erythropoietin receptor activation. *Proc Natl Acad Sci USA* 93:9471-6.

Meng EC, BK Shoichet and ID Kuntz. 1992. Automated docking with grid-based energy evaluation. *J Comput Chem* 13:505-524.

Middleton SA, DL Johnson, R Jin, FJ McMahon, A Collins, J Tullai, RH Gruninger, LK Jolliffe and LS Mulcahy. 1996. Identification of a critical ligand binding determinant of

the human erythropoietin receptor. Evidence for common ligand binding motifs in the cytokine receptor family. *J Biol Chem* 271:14045-54.

Miura O and JN Ihle. 1993. Dimer- and oligomerization of the erythropoietin receptor by disulfide bond formation and significance of the region near the WSXWS motif in intracellular transport. *Arch of biochem and biophys* 306:200-208.

Nicola NA, Ed. 1994. Guidebook to cytokines and their receptors. Oxford, UK, Oxford University Press, 1994.

Shoichet BK and ID Kuntz. 1991. Protein docking and complementarity. *J Mol Biol* 221:327-346.

Somers W, M Ultsch, AM De Vos and AA Kossiakoff. 1994. The X-ray structure of a growth hormone-prolactin receptor complex. *Nature* 372:478-81.

Syed RS, SW Reid, CW Li, JC Cheetham, KH Aoki, B Liu, H Zhan, TD Osslund, AJ Chirino, J Zhang, J Finer-Moore, S Elliott, K Sitney, BA Katz, DJ Matthews, JJ Wendoloski, J Egrie and RM Stroud. 1998. Efficiency of signaling via cytokine receptors depends critically on receptor orientation. *Nature* 395:511-516.

Tian S, P Lamb, AG King, SG Miller, L Kessler, JI Luengo, L Averill, RK Johnson, JG Gleason, LM Pelus, SB Dillon and J Rosen. 1998. A small, nonpeptidyl mimic of granulocyte-colony stimulating factor. *Science* 281:257-159.

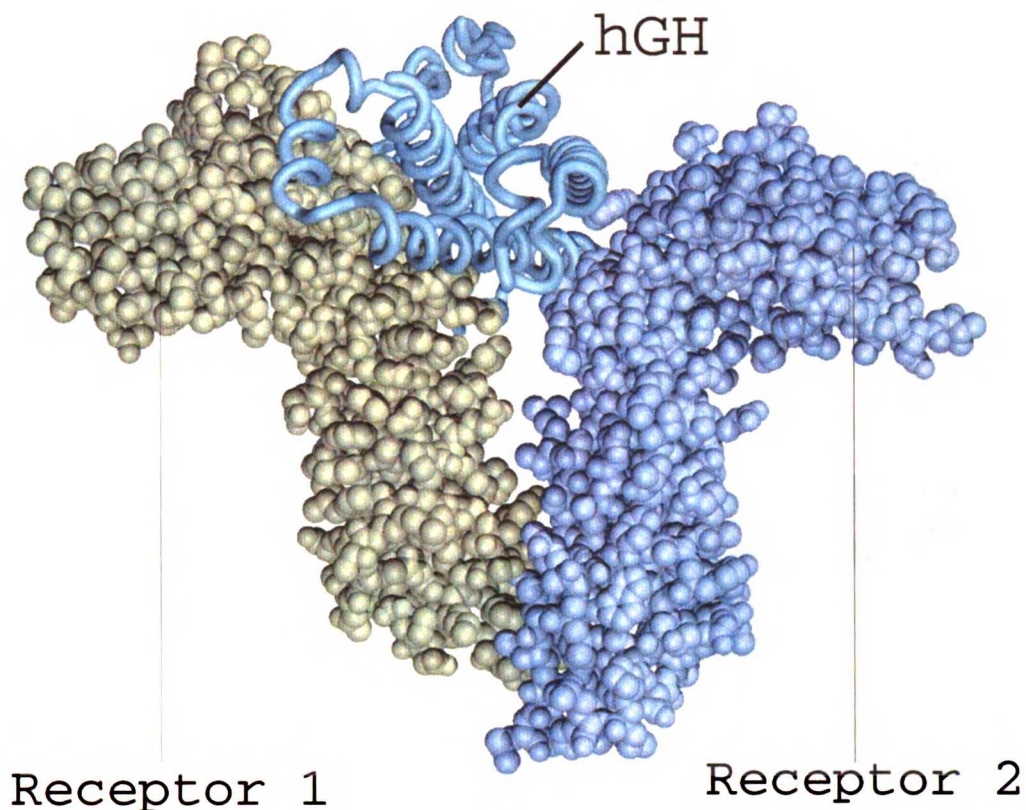
Weiner SJ, PA Kollman, DA Case, UC Singh, C Ghio, G Alagona, J S. Profeta and P Weiner. 1984. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc* 106:765.

Wells JA. 1996. Binding in the growth hormone receptor complex. *Proc Natl Acad Sci USA* 93:1-6.

Wells JA and AM de Vos. 1993. Structure and function of human growth hormone: implications for the hematopoietins. *Annu Rev Biophys Biomolecular Struct* 22:329-51.

Wrighton NC, P Balasubramanian, FP Barbone, AK Kashyap, FX Farrell, LK Jolliffe, RW Barrett and WJ Dower. 1997. Increased potency of an erythropoietin peptide mimetic through covalent dimerization. *Nature Biotech* 15:1261-5.

Wrighton NC, FX Farrell, R Chang, AK Kashyap, FP Barbone, LS Mulcahy, DL Johnson, RW Barrett, LK Jolliffe and WJ Dower. 1996. Small peptides as potent mimetics of the protein hormone erythropoietin. *Science* 273:458-64.



**Figure 1. Docking sites of human growth hormone binding to its receptor.**  
 (On this page and the following two pages.) Above is a representation of the whole hGH/hGHR complex. On the following pages, across each row is a snapshot of a molecule from the complex, with 0 as shown in the top figure, followed by sequential 90 rotations. Each row displays site points, described in table 3, as follows: first row: H1 in dark gray, H2 in magenta on hGH; second row: R1 in green and RS1 in blue on receptor 1 of the hGHR; third row: R2 in yellow and RS2 in cyan on receptor 2 of the hGHR; fourth row: H1M in red on hGH; fifth row: R1M in red on receptor 1 of the hGHR; sixth row R2M in red on receptor 2 of the hGHR. These figures and figure 2 were drawn with the conic option (Huang, et al. 1991) of MidasPlus 2.1 (Ferrin, et al. 1988), and figure 5 was drawn with MidasPlus 2.1.

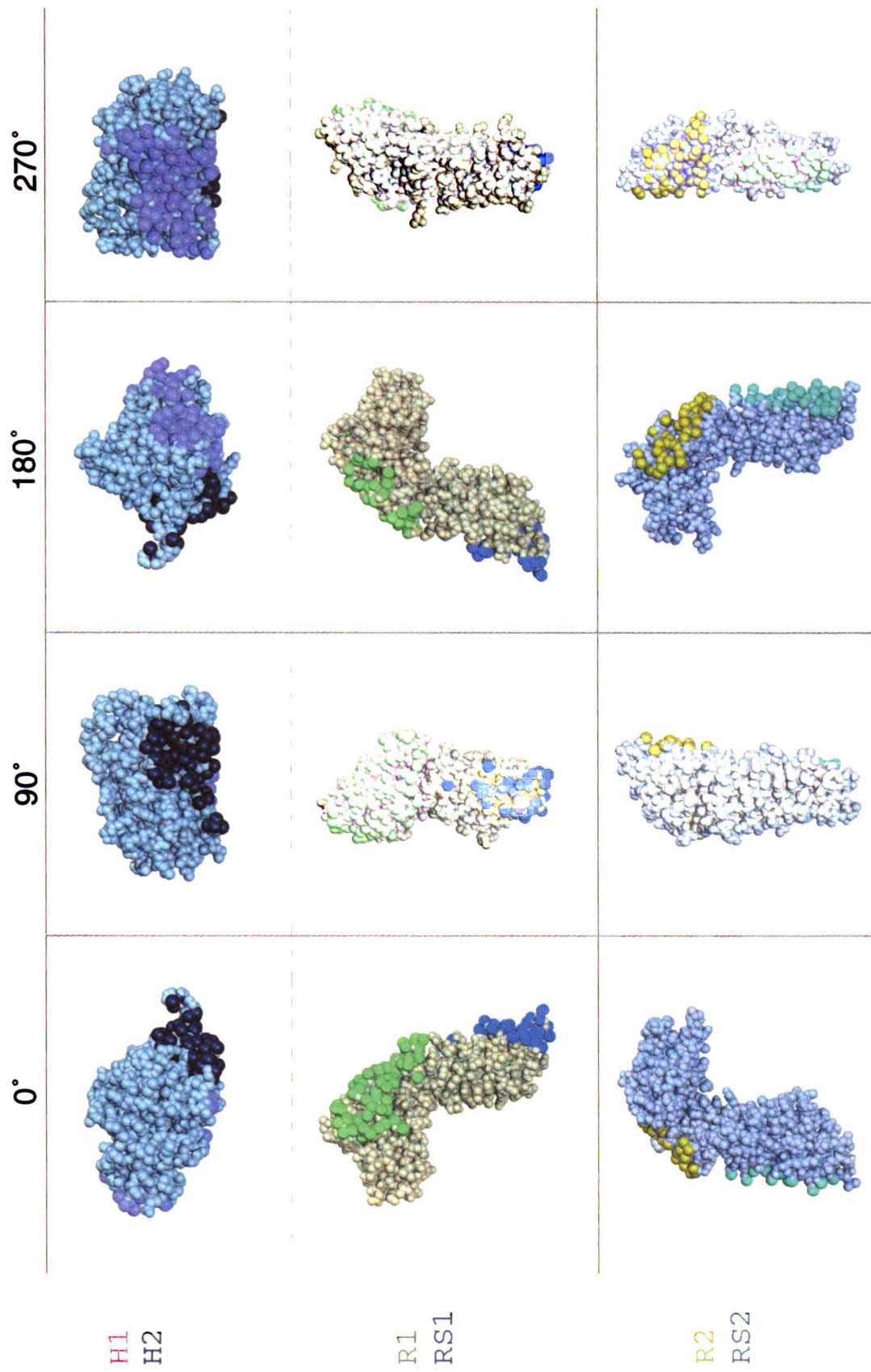


Figure 1, continued (part 2 of 3)

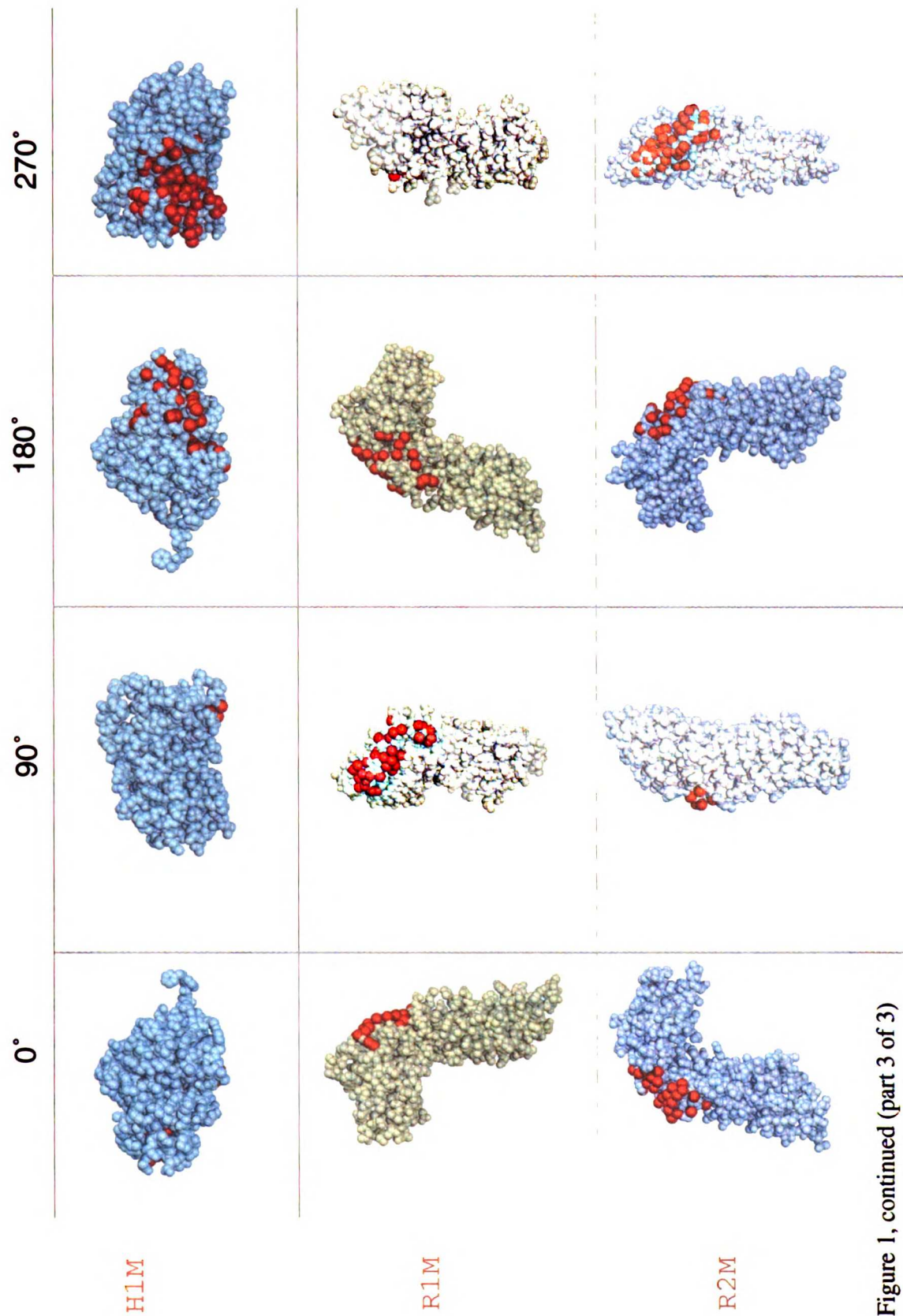
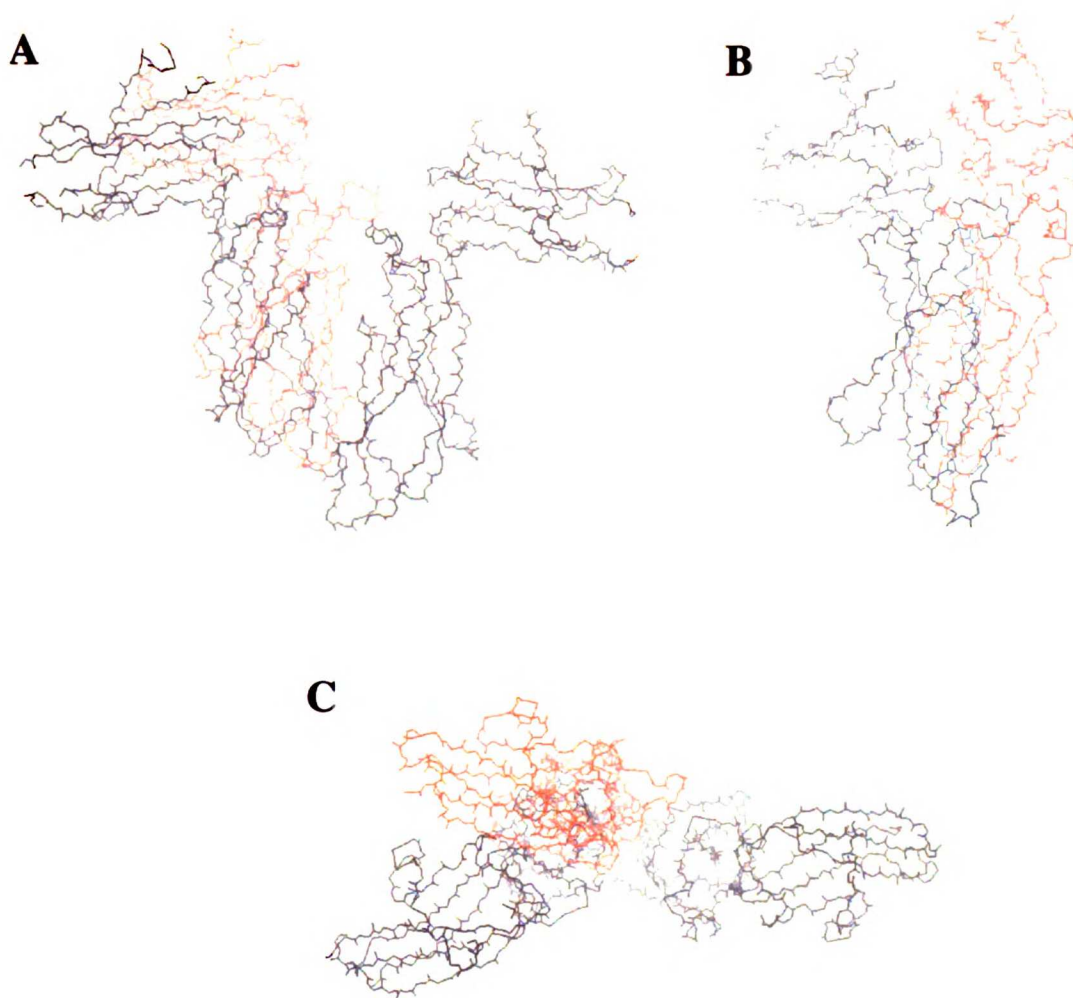
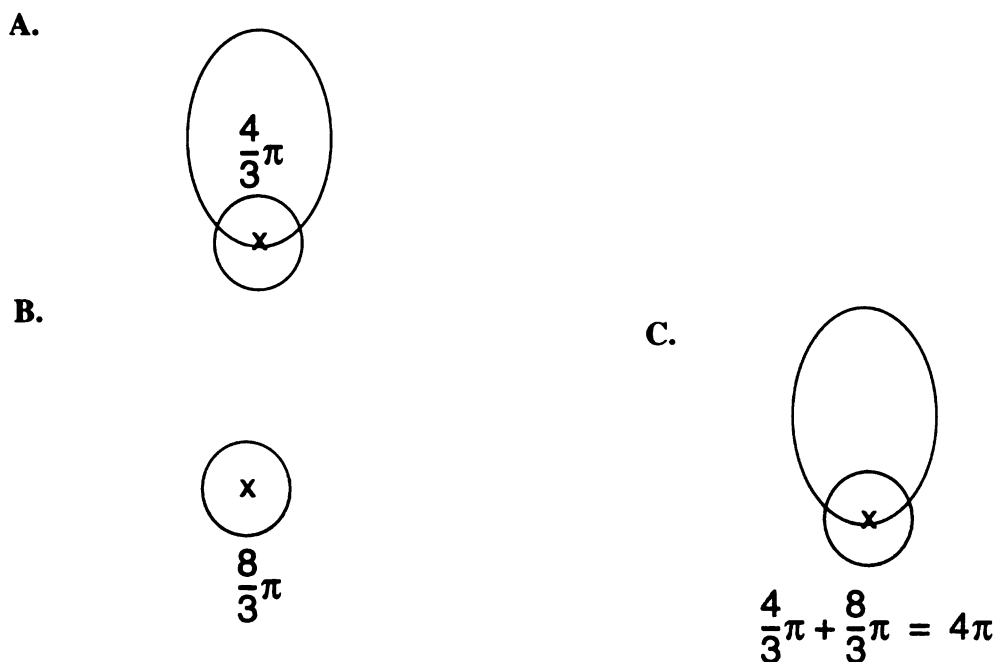


Figure 1, continued (part 3 of 3)



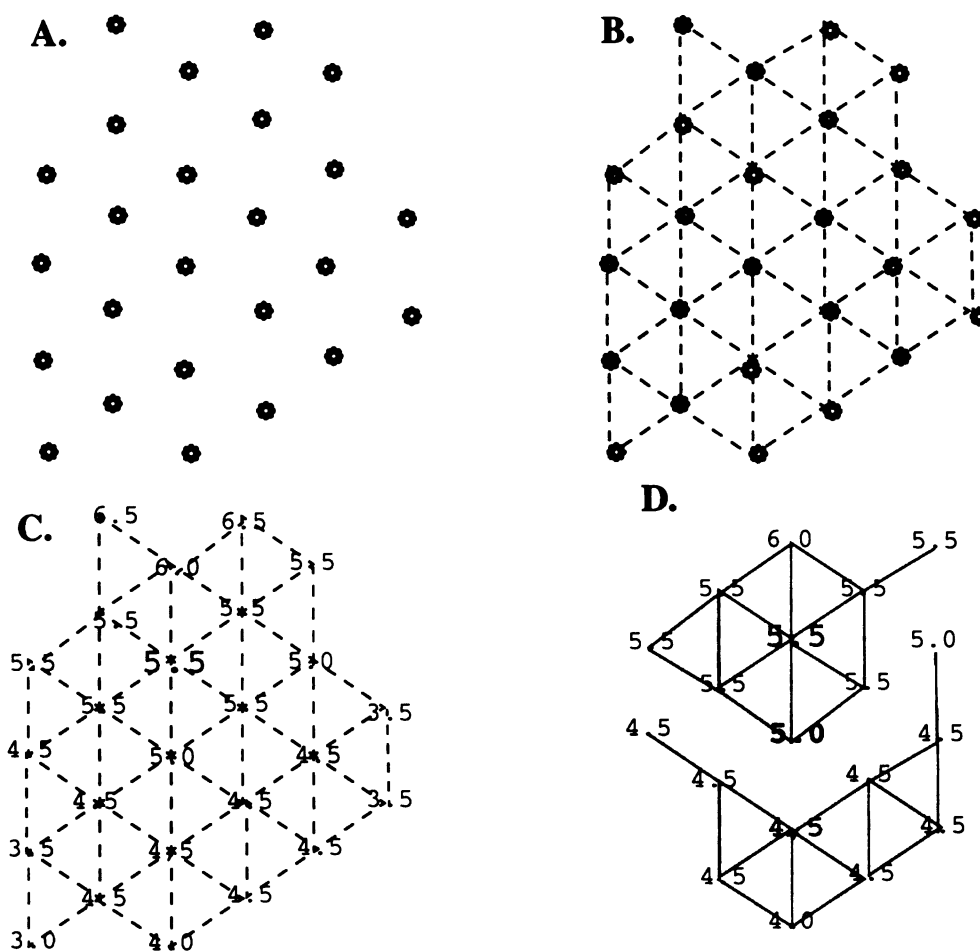
**Figure 2. Orientation of receptor monomers docked by receptor stem sites.** The orientations, shown in red, were generated by docking the receptor stem sites on receptor 1, RS1, to receptor 2 stem sites, RS2. The RMSd of the DOCK orientation to the minimized orientation of receptor 1 is 19.2Å. A. Dimer form of the hGHR. B. Viewing receptor 1, 90° from the view in A. C. The dimerized receptor molecules, viewed from above, looking down into the hormone-binding site.



**Figure 3. Description of the surface solid angle.**

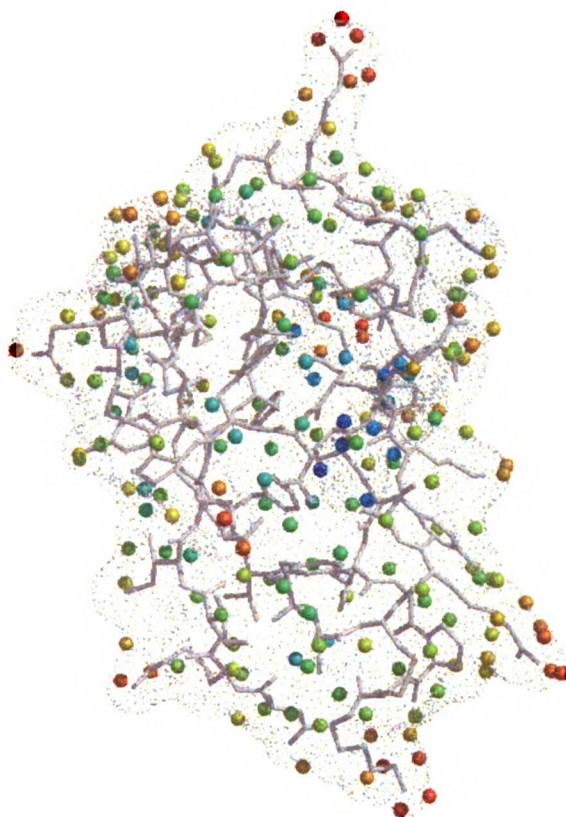
**A.** The probe sphere center is marked by an “x”, and the ellipse is a representation of a ligand, with its interior shaded. Approximately one-third of the probe sphere surface is inside the ligand, so the surface solid angle of the point at the center of the probe sphere is  $(1/3) \cdot 4\pi$ . **B.** Similar to A, except two-thirds of the sphere lies inside of the receptor surface, so the solid angle of the point at the center of the probe sphere is  $(2/3) \cdot 4\pi$ . **C.** When complexed, complementary surfaces on the ligand and receptor match, and their solid angles are complementary, summing to  $4\pi$ . Solid angle values are calculated using the Connolly algorithm (Connolly 1986a).



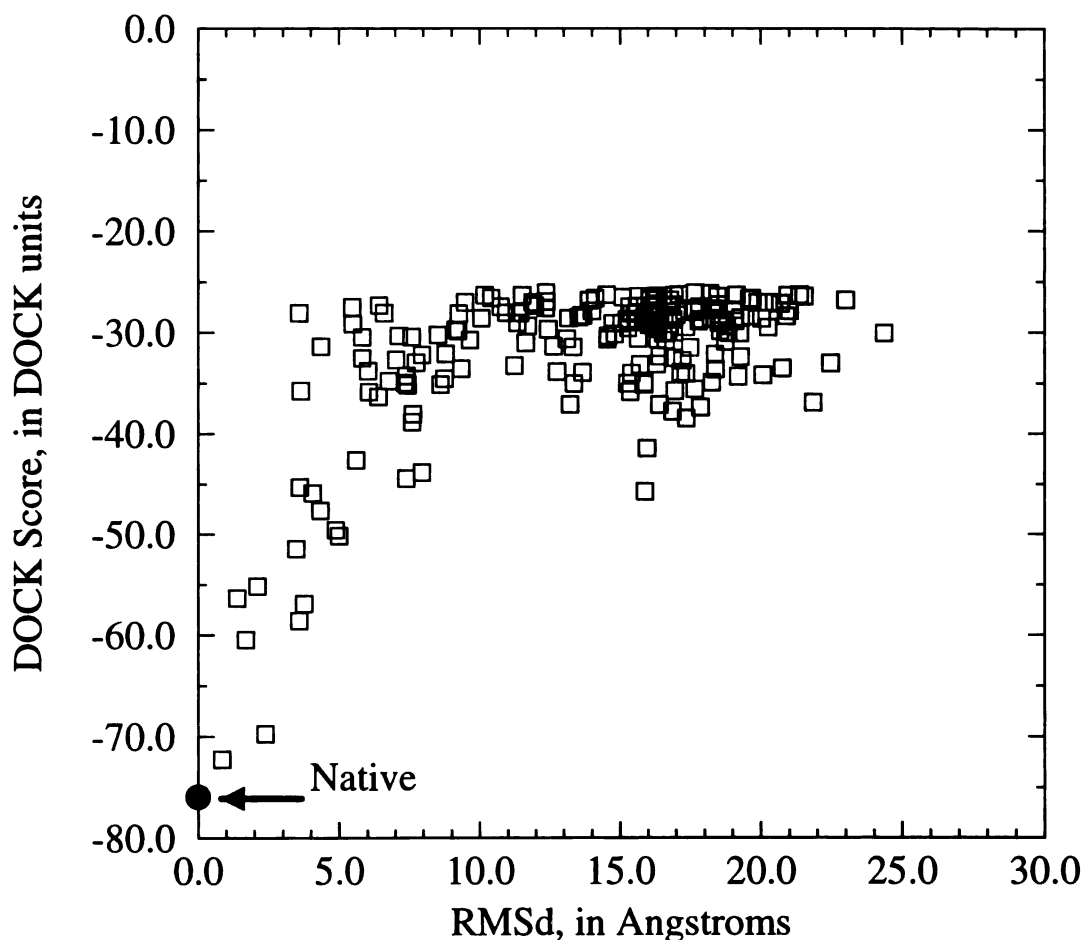


**Figure 4. Building site points from surface points.**

A. Points are placed on the macromolecular surface by Connolly's MS (Connolly 1983a) program. B. Near-neighbor relationships are drawn among surface points. C. Solid angle values of surface points are calculated for each surface point. Starting from a point selected at random, the hub, regions are formed by collecting nearest neighbors with similar solid angle values. In this example, the hub is shown in red with a solid angle value of 5.5 steradians. Surface points within the neighborhood of the hub are collected into the region if their solid angle value is within a range of values defined by the hub, which is 0.5 in this example. (The range used in this study is  $\pi/8$  (0.39) steradians.) Each region grows outward from the central hub until there are no near neighbors with solid angle values within the range. Once a point is in a region, it cannot be placed into another region. D. Regions are shown as points connected by solid lines. Two regions are shown: one in red formed with the hub of 5.5 steradians, and another in blue formed with the hub of 4.5 steradians. The point in bold face between the two hubs, with a value of 5.0, is in the red region; however, because it is a near neighbor of both hubs and its solid angle value is within 0.5 steradians of the solid angle both hubs, then it would have been included in the blue region if it had been built, first.



**Figure 5. Site points on trypsin inhibitor, colored by relative value.** Solid angle values vary from  $0.16\pi$  steradians (deep red) at the distal portions of the molecule to  $3.7\pi$  steradians (dark blue) on surface which surrounds buried waters in the interior of the molecule. These site points were generated with a probe sphere with a radius of  $5.0\text{\AA}$ .



**Figure 6. DOCK score vs. RMS deviation of DOCK orientations for chymotrypsin and turkey ovomucoid third domain.**

Orientations are generated with shape-based site points and version 4 of UCSF DOCK. The DOCK score refers to the AMBER-based force field score. As the RMSd increases and the orientations are further from the crystal structure orientation, the DOCK score is less favorable. The minimized native structure has a score of -75.9 DOCK units, and the best DOCK orientation, with an RMSd of 0.85Å (comparing all non-hydrogen atoms) had a score of -72.2 DOCK units.

## Chapter 4



### **DISCOVERY OF INHIBITORS OF HIV-1 INTEGRASE BY DOCKING TO AN UNCHARACTERIZED SITE AT THE DIMER INTERFACE**



*This work was done with Malin Young, Luke Hoffman and Keith Burdick in the Kuntz laboratory, and in collaboration with the Andrew Leavitt laboratory in the Department of Laboratory Medicine at UCSF (George Robles, Charles Yoh and Ann Tang), the George Kenyon laboratory, formerly in the Department of Pharmaceutical Chemistry at UCSF and now at the University of Michigan (Karl Maurer) and the Robert Stroud laboratory in the Department of Biochemistry and Biophysics at UCSF (Julian Chen and Yolanta Krucinski).*

## **Abstract**

We have identified a class of HIV-1 integrase (IN) inhibitors using the DOCK program. With biochemical and enzymatic data suggesting that IN acts as a dimer or higher-order oligomer, we selected a site near the interface of two monomers in the crystal structure of the catalytic core domain (residues 55-199) of Avian Sarcoma Virus (ASV) IN. The site is located 15Å from the active sites. Using the DOCK program, we searched 170,000 molecules in the Available Chemicals Database (ACD) 95.1 for small molecules that could bind to this site. Molecules were assayed for their effect on 3'-processing and strand transfer, two initial functions of HIV-1 IN. From the search, 30 molecules were selected for screening, revealed one molecule with an IC<sub>50</sub> of 4 μM. From this inhibitor, we used similarity searching to find 5 additional molecules that inhibit IN activity with IC<sub>50</sub> values less than 5 μM. These molecules have been tested for their toxicity on a mammalian cell line and have been found to have TC<sub>50</sub> values greater than 50-fold in excess of their IC<sub>50</sub> values. Structural analysis is being performed at this time, using X-ray crystallography, and the effects of this class of inhibitors on the oligomerization of IN is being examined using analytical ultracentrifugation.

## **Introduction**

While there has been great success with combination therapies for HIV-1 targeting the viral protease and reverse transcriptase, resistant strains are beginning to compromise this success. Hope lies in finding other targets and other therapies. One of these additional targets is the viral integrase (IN) enzyme.

IN plays a key role in the viral replicative cycle. After transcription of viral RNA to DNA by the viral reverse transcriptase, IN performs two reactions on the newly-formed viral DNA: 3'-processing in cytoplasm and strand transfer, or integration, in the nucleus. 3'-processing is the cleaving of a two bases from the 3'-ends of the newly transcribed viral DNA. Integration involves cleaving the host DNA and covalently linking the 3' ends of the viral DNA into the host genome.

Historically, IN has been a difficult protein to characterize, in part due to its lack of solubility. It has been determined that the active state of the molecule is at least a dimer, and perhaps a higher-order oligomer (Engelman, et al. 1993). Structural information has been difficult to obtain on the whole molecule, but three domains of the protein have been identified and, in the case of HIV-1 IN, each domain has been characterized by either NMR or x-ray crystallography. The N-terminal domain is often referred to as the "zinc-finger" domain because of the presence of a characteristic HHCC motif. It is comprised of residues 1-49 in HIV-1 IN. An NMR structure of the HIV-1 IN N-terminal domain has been determined (Zheng, et al. 1996). The C-terminal domain is referred to as the DNA-binding domain. It is comprised of residues 213-288 in HIV-1 IN, and is involved in multimerization and non-specific binding of DNA. Both an x-ray crystal structure (Eijkelen-

boom, et al. 1995) and an NMR structure (Lodi, et al. 1995) have been determined for the C-terminal domain.

The catalytic core domain, residues 50-212 of HIV-1 IN, includes the key catalytic residues, Asp 64, Asp 116 and Glu 152. In addition to the HIV-1, the avian sarcoma virus (ASV) integrase catalytic core structure has been determined. The ASV and HIV-1 IN structures have very similar folds, but their sequences are only 25% identical in the catalytic core domain. (See figures 1 and 2.) Because of the lack of information about the enzyme, we hoped that the identification of small molecules that bound to IN would assist in structure and function determination. Since our work began, several additional structures of HIV-1 IN and ASV IN (Bujacz, et al. 1996b)(Bujacz, et al. 1996a)(Jenkins, et al. 1995)(Goldgur, et al. 1998)(Maignan, et al. 1998) have become available, including one ASV IN structure with an inhibitor (Lubkowski, et al. 1998).

## **Methods**

### ***Computational methods***

Our approach is outlined in figure 3. We began our search for inhibitors by first selecting an IN structure, then selecting possible sites. At the onset of this project (winter, 1994), two X-ray crystal structures of the IN catalytic core were available: HIV-1 IN and ASV IN. Because the active site HIV-1 IN was disordered, as characterized by the absence of a key loop and high B-factors, and contained the mutation F185K which resulted in an inactive form of the enzyme, we selected the ASV IN structure.

The structure of the catalytic core domain of ASV integrase was resolved to 2.3 Å. The active site of ASV IN was well-ordered, and the key active site residues, D64, D121

and E154 were coordinated by a manganese ion. This structure was a monomer in the asymmetric unit, but the dimer was built by mathematically transforming the coordinates of the monomer position the second monomer. The transform was provided by the Wlodawer group. (When examining coordinates in the Protein Data Bank, the dimer is built by using the UNCRYST module of MidasPlus 2.1 to build the crystal form of the protein, then selecting two monomers in the proper orientation from the resulting crystal.)

We ran the site-descriptor program SPHGEN (Kuntz, et al. 1982) on the entire surface of the ASV IN catalytic core structure in its dimer form. In addition to the active site, we searched for cavities created by the formation of the dimer. We found one cavity 15Å from the active sites. This cavity was highly basic, surrounded by lysine, arginine and histidine residues. The homologous site on the HIV-1 IN did not include the F185K mutation. (See figures 4 and 5.) With the knowledge that the active form of IN is at least a dimer, we targeted this site for docking.

Spheres were selected for the active site and dimer site by visual inspection. We selected spheres so that they were no closer than 1Å apart and preferred those near the molecular surface.

We docked to both the dimer site and the active site using DOCK3.5 and standard settings (Meng, et al. 1993). We optimized our runs to generate an average of 5,000 matches per molecule. DOCK generates orientations from these matches, and scores molecules by both a contact score and a force-field score. The contact score is a count of the number of atoms the ligand contacts in the docking site on the enzyme in the proposed DOCK orientation. The force-field score is determined by calculating the electrostatic and van der Waals forces between the ligand and enzyme, with the ligand in its docked orien-



tation. A distance-dependent dielectric function is used in determining the electrostatic forces, with the dielectric constant,  $\epsilon$ , equal to  $4r$ .

We searched the Available Chemical Database (ACD), version 95.1 (Molecular Design Limited 1995, 1997). This version of the ACD has 170,000 molecules. The database is divided by charge and then further partitioned into sections of 10,000 molecules. The 200 top-scoring molecules from each section were visually examined in their docked orientation. 40 molecules from the active site and 30 molecules from the dimer site were selected from these top-scoring molecules.

After a first screen of the DOCK molecules, we used two methods to find molecules similar to our DOCK hits. One method, the ISIS program (Molecular Design Limited 1995, 1997), uses 2-dimensional substructure keys to find molecules with similar substructure. The second method, from Daylight (Daylight 1996), uses the 2-dimensional connectivity paths within the target molecule to search the database. The two lists from these search methods were concatenated and a subset was selected based upon availability and 2-dimensional similarity to the target molecule, judged visually.

### ***Assays***

The selected compounds were purchased from vendors assayed for their effect on HIV-1 IN activity, on both 3'-processing and strand transfer in the Andrew Leavitt laboratory at UCSF. (I refer to the small molecule model as "molecule" and the purchased salt form of a molecule as "compound".) Briefly, for 3'-processing, IN is incubated with radiolabeled substrate DNA. 3'-processing cleaves two bases, GT, from the 3' end. The products are examined by gel electrophoresis and measured by autoradiography, showing one band for processed DNA and a second band for unprocessed DNA. For integration, IN is

incubated with the radiolabeled target DNA and the products, longer DNA oligonucleotides, are again measured by electrophoresis and autoradiography (Leavitt, et al. 1993).

Because many of these molecules were commercial dyes, it was found that purification and de-salting significantly affected activity. All IC<sub>50</sub> and TC<sub>50</sub> data presented in the Results was derived from compounds acquired from vendors, then de-salted and purified in the Kenyon laboratory.

We found one lead molecule from our search of the dimer site, screening at 100 μM. We expanded our list of leads by searching the ACD for similar molecules. Using MDL and Daylight's similarity searching algorithms, we selected an additional 30 molecules for testing. These compounds were screened at 100 μM, and the IC<sub>50</sub> values of compounds which reduced integrase activity at 100 μM was determined. This similarity search revealed an additional 10 molecules with IC<sub>50</sub> values less than 20 μM.

Additionally, while selecting molecules from the ASV IN DOCK results, Geoff Skillman noticed that several molecules on the IN "hit" list were identical to molecules known to inhibit HIV-1 Reverse Transcriptase (RT), as shown by an ongoing effort in the Kuntz and Kenyon labs. Because molecules known to inhibit RT have been shown to inhibit IN, the compounds synthesized for their possible inhibition of RT from the Kuntz/Kenyon collaboration were also tested for their inhibition of integrase. Additional molecules based related to those found in the ACD were synthesized in the Kenyon laboratory (Skillman, et al. 1999) and assayed in the Leavitt laboratory.

The lead compounds with IC<sub>50</sub> values less than 15 μM were assayed for toxicity and viral infectivity. The toxicity assay was developed in the Leavitt laboratory by Ann Tang. CEM-SS cells, a human T-cell line, were grown in RPMI media supplemented with

10% fetal bovine serum, Penn-Strep and L-glutamine. Assays were performed in 100 microliter volumes with cells at 50,000/ml and inhibitor at final concentrations of 10, 100, 200, 500, 1000  $\mu$ M. Cells were incubated in the presence of the compounds for 6 days. 100 microliter of Calcein AM (from Molecular Probes, Inc.) at 4  $\mu$ M in D-Phosphate-buffered saline (D-PBS) was added to each well. (Calcein AM enters cells with intact membranes and is converted by intracellular esterase from a virtually non-fluorescent compound to an intensely fluorescent one.) The plate was incubated at room temperature for 30-45 minutes and read with PerSeptive Biosystems' Fluorescence Multi-well Plate reader at 485 nm for excitation and 530 nm for emission, using the CytoFluor II software. Each compound was assayed twice in quadruplicate. (Leavitt and Tang 1999)

The viral infectivity of the compounds with  $IC_{50}$  values less than 15  $\mu$ M was measured. HOS cells (human osteosarcoma cell line) were grown in 6-well plates to 20% confluence on the day of infection. Compounds were incubated with the cells one hour prior to infection. The virus, which carries a hygromycin (an antibiotic) selectable marker, was diluted 1:10 in complete media with polybrene (8  $\mu$ g/ $\mu$ l) and compound, and incubated at room temperature for 30 minutes. The virus was then serially diluted in media with polybrene and compound, and each well was infected with 1 ml. At 2.5 hours post-infection, the wells were washed three times with 3 ml D-PBS, then 2 ml of media with compound were added to each well. After 20 hours of growth in the presence of compound, the cells were grown in the absence of compound and in the presence of hygromycin at 200  $\mu$ g/ml. (Infected cells grow in the presence of hygromycin, while non-infected cells die.) Media was exchanged every two days, and the cells were stained with crystal violet at 9 days

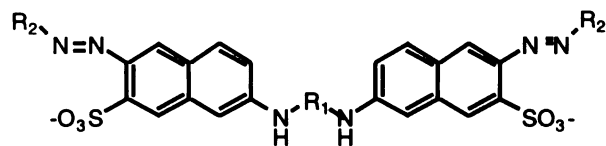
post-infection. Each experiment was performed twice in duplicate, except for MFCD0036441.

## **Results**

The results from the ongoing collaboration between the Kenyon and Leavitt labs on molecules originally synthesized as possible inhibitors of RT are reported. Also reported are results from the collaboration between the Kuntz, Leavitt, Kenyon and Stroud labs to find and to characterize new inhibitors of HIV-1 IN using DOCK.

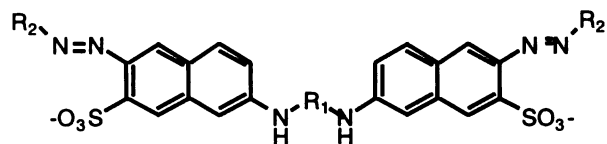
### ***RT compounds***

The IC<sub>50</sub> and TC<sub>50</sub> values for these molecules are displayed in table 1. Synthesis, purification and characterization of many of these molecules are reported in a recent article.(Skillman, et al. 1999)



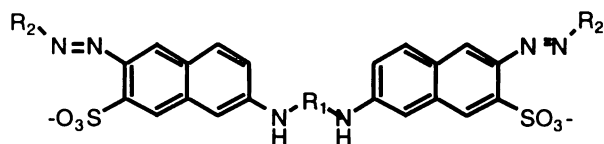
**Table 1: Inhibition of HIV IN activity.**

ID	R1	R2	HIV-1 IN IC <sub>50</sub> (μM) 3'- processing	HIV-1 IN IC <sub>50</sub> (μM) strand transfer	HIV-1 IN TC <sub>50</sub> (μM)	RT IC <sub>50</sub> (μM) or (% activity)
1 <sup>a</sup>			0.3 0.4	0.5 0.7	90/75	0.39
2			2.7 0.9	3.2 1.0	150/150	0.09
3			0.4 0.5	0.7 2.0*	80/100	0.15
4			0.6 0.4	0.8 0.6	80/60	1.1
5			0.5 0.2	0.4 0.6	140/135	0.80
10			2 >20	2 >20	low solubility	2.5 <i>e</i>
11			2.7	1.3	low solubility	2.5 <i>e</i>
12			4 >20	0.6 >20	low solubility	10.0 <i>e</i>



**Table 1: Inhibition of HIV IN activity.**

ID	R1	R2	HIV-1 IN IC <sub>50</sub> (μM) 3'- processing	HIV-1 IN IC <sub>50</sub> (μM) strand transfer	HIV-1 IN TC <sub>50</sub> (μM)	RT IC <sub>50</sub> (μM) or (% activity)
13			0.3 0.09	0.4 0.25	170/160	0.40
14			0.5 0.4	1.0 0.4	375/230	1.0
17			0.3 0.3	0.7 0.6	low solubil- ity	0.5
18			1	1	260/410	0.22
6			5 <i>b</i>	2.3 <i>b</i>	50/70	0.25
7			3 <i>c</i>	1 <i>c</i>	<i>d</i>	(0.8%) <i>g</i>
15			1 0.5	2.5 0.4	85 <i>d</i>	(1.8%) <i>g</i>
16			0.8 0.5	1.0 0.1	340/300	0.72



**Table 1: Inhibition of HIV IN activity.**

ID	R1	R2	HIV-1 IN IC <sub>50</sub> (μM) 3'- processing	HIV-1 IN IC <sub>50</sub> (μM) strand transfer	HIV-1 IN TC <sub>50</sub> (μM)	RT IC <sub>50</sub> (μM) or (% activity)
8			2.5 2.5	3.5 1-5*	820/790	(52%) <b>g</b>
9			2.5 1.0	0.5 2.0	450/630	(8.6%) <b>g</b>

\* -- Uninterpretable results.

**a** -- Molecules are numbered so they are compatible with the Leavitt Laboratory's nomenclature; however, they are ordered in this table for ease of display, by linker.

**b** -- No material available for a second full screening, but active at 1μM screen.

**c** -- Not active on 1μM screen.

**d** -- No material available for toxicity test.

**e** -- Assayed at National Cancer Institute

**f** -- Not assayed on RT

**g** -- These data for RT assays are from initial screens and represent percent-activity compared to a control at 1uM. IC50 values were not determined for these compounds.

### **DOCK compounds**

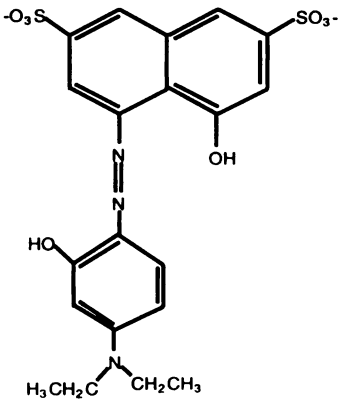
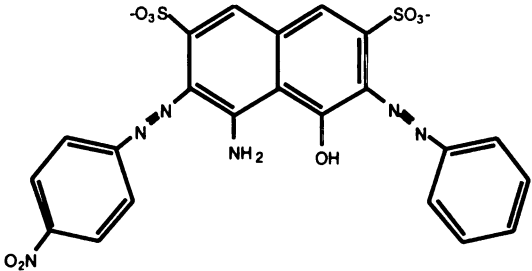
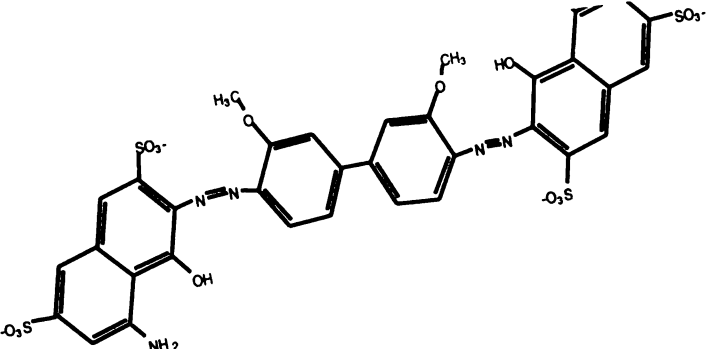
From the 170,000 molecules, 40 top-scoring molecules from both the active site run and 30 top-scoring molecules from the dimer site were selected. One molecule with an IC<sub>50</sub> value of less than 100 μM inhibitor was revealed from the active site search. We

chose not to take this molecule into further development. One molecule MFCD0070629, selected using the dimer sites a target had an  $IC_{50}$  value of 2-3  $\mu M$ .

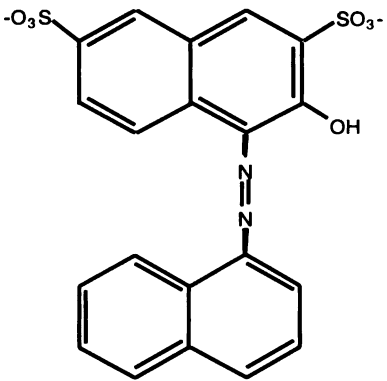
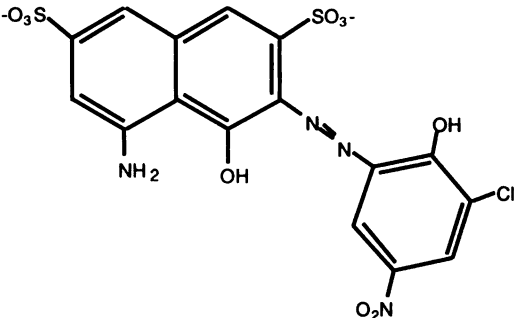
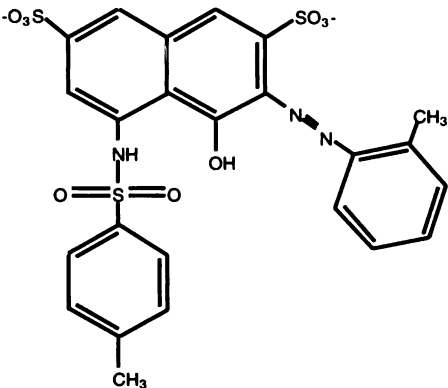
Using MFCD0070629, we searched the ACD 95.2 for similar compounds using both the Daylight similarity search and MDL's ISIS similarity search. 21 additional molecules were selected, purchased and assayed. Of these, 6 revealed consistent  $IC_{50}$  values below 15  $\mu M$ . The two-dimensional structures, the  $IC_{50}$  and  $TC_{50}$  values for these molecules and for MFCD0070629 are shown in table 2. One molecule, MFCD0030706, yielded a promising  $IC_{50}$  of 4-6  $\mu M$  (data not shown), but later was proven to have an incorrect structure in the ACD, and the data were not reproducible. We were particularly interested in this molecule because it had only one sulfonic acid group.



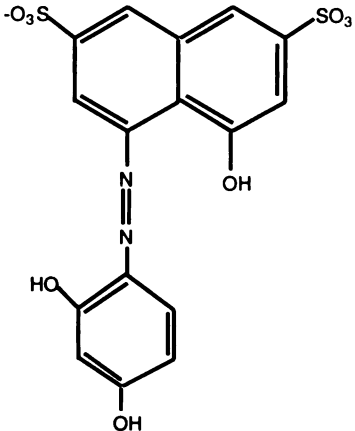
**Table 2: Inhibition and toxicity data, IN DOCK compounds**

MFCD and 2-D structure	IC <sub>50</sub> 3'- processing; Strand transfer ( $\mu$ M)	TC <sub>50</sub> * ( $\mu$ M)	Thera- peutic index*: TC <sub>50</sub> / IC <sub>50</sub>
<p>70629</p> 	<p>2.2; 2.6</p>	<p>240/240</p>	<p>109</p>
<p>4017</p> 	<p>4.0; 6.4</p>	<p>250/290</p>	<p>63</p>
<p>36441</p> 	<p>0.3; 0.6</p>	<p>80/180/ 150</p>	<p>267</p>

**Table 2: Inhibition and toxicity data, IN DOCK compounds**

MFC-D and 2-D structure	IC <sub>50</sub> 3'- processing; Strand transfer ( $\mu$ M)	TC <sub>50</sub> * ( $\mu$ M)	Thera- peutic index*: TC <sub>50</sub> / IC <sub>50</sub>
<p>3887</p> 	<p>4.3; 2.5</p>	<p>400/320</p>	<p>128</p>
<p>21475</p> 	<p>6.0; 4.3</p>	<p>370/420</p>	<p>86</p>
<p>143975</p> 	<p>15.6; 20.0</p>	<p>100/130</p>	<p>6.4</p>

**Table 2: Inhibition and toxicity data, IN DOCK compounds**

MFCD and 2-D structure	IC <sub>50</sub> 3'- processing; Strand transfer ( $\mu$ M)	TC <sub>50</sub> * ( $\mu$ M)	Thera- peutic index*: TC <sub>50</sub> / IC <sub>50</sub>
59024 	9.7; 14.0	500/540	55

\*The therapeutic index was calculated by dividing the lowest-value TC<sub>50</sub> by the lowest value IC<sub>50</sub> (3'-processing or strand transfer).

The viral infectivity assays shows that MFCD0070629, MFCD0004017 and MFCD0036441 strongly suppress viral infectivity in culture at 100  $\mu$ M, but MFCD0003887, MFCD0143975 and MFCD0059024 do not. (Some seem to actually increase viral infectivity.) MFCD0021475 was not tested. (Leavitt and Tang 1999)

## Discussion

We have identified a class of inhibitors of HIV-1 IN by docking to an uncharacterized site in the catalytic core of the homologous protein, ASV IN. These compounds have IC<sub>50</sub> values ranging from 0.3-20  $\mu$ M, and TC<sub>50</sub> values ranging from 80-540  $\mu$ M. (See

table 2.) We also have found a set of inhibitors from a library of compounds originally prepared as RT inhibitors. These molecules have IC<sub>50</sub> values in the nanomolar range.

While micromolar inhibitors are not always considered exciting leads, IN has been difficult to characterize, and relatively few leads have been reported. Other researchers have identified micromolar inhibitors (Artico, et al. 1998, Cushman, et al. 1995, Eich, et al. 1996, Farnet, et al. 1998, Fesen, et al. 1993, Hong, et al. 1997, Mazumder, et al. 1996, McDougall, et al. 1998, Mekouar, et al. 1998, Neamati, et al. 1997a, Neamati, et al. 1998a, Neamati, et al. 1997b, Neamati, et al. 1998b, Neamati, et al. 1997c, Nicklaus, et al. 1997, Ojwang, et al. 1995, Puras Lutzke, et al. 1995, Raghavan, et al. 1995, Robinson, et al. 1996a, Robinson, et al. 1996b, Zhao, et al. 1997a, Zhao, et al. 1997b, Zhao, et al. 1997c), but few small molecule sub-micromolar inhibitors have been discovered (Zhao, et al. 1997b)(Artico, et al. 1998)(McDougall, et al. 1998)(Neamati, et al. 1998a)(Mekouar, et al. 1998). Of the many leads in the literature, there is one published example describing the binding mode of the molecules (Hong, et al. 1997)(Lubkowski, et al. 1998), and one example of the lead compound inhibiting viral replication (Robinson, et al. 1996b).

Our class of inhibitors is not entirely novel. Other researchers have found leads that are also naphthelene-disulfonic acids (Hong, et al. 1997). Those leads were discovered by a three-point pharmacophore search of inhibitors previously found by random screening. It was not determined in their studies whether the leads affected viral replication, nor was the mode of inhibition explored.

### ***Possible mechanism of inhibition***

While the Wlodawer structure suggests that a similar molecule binds near the active site of ASV IN, we found our inhibitor by docking to the dimer site, 15Å from the

active site. This site is highly basic, and our lead compound is negatively charged, with two sulfonic acid groups. The active site has the three key acidic residues, D64, D116 and E152 in HIV-1 IN. In the Wlodawer structure, the disulfonic acid is bound near the active site, but on an opposite face from the active site residues. Their site is highly solvent-exposed in a native monomer or dimer form of the enzyme, but when the crystallographic unit cell is built back, the inhibitor molecule is stacked against another inhibitor molecule and is shielded from solvent by its neighboring monomer. In contrast, preliminary structural data of MFCD0070629 with HIV-1 IN catalytic core domain from Julian Chen in the Robert Stroud lab shows poorly resolved density in the dimer site which may be the inhibitor.

If our inhibitors are binding to the dimer site, there are several possible mechanisms of inhibition. First, the molecule may change the dimer form of the molecule, shifting the relative positions of the monomers. (Preliminary evidence from Chen's structure suggests that this is not the case.) Another possible mechanism is that the presence of the inhibitor changes the association constant of the monomer to itself. The molecule may require large structural changes to bind to DNA which might involve re-organization of the dimer (or higher order oligomer) interface. The inhibitor could strengthen or weaken the association of the monomers and thus prevent this reorganization. Finally, while this site has not been identified as a DNA-binding site by cross-linking studies, it may be possible that the site binds to DNA and the inhibitor disrupts the binding of DNA. (We have not determined whether or not the inhibitor is competitive with DNA.)

Regardless of the mechanism, the presence of two sulfonic acid groups is a major hurdle for further development of these molecules as drug candidates, regardless of the

toxicity and viral infectivity data. However, it may be possible to use this lead to find other lead molecules, or to change the sulfonic acid groups to carboxylic acids. And while many only consider the integration function of IN, which occurs in the nucleus, it is key to remember that 3'-processing occurs in the cytoplasm. Thus, a drug may only need to cross one membrane, not two, to reach a point of action.

### ***Ongoing work and future directions***

Among our collaborators, the current efforts are characterizing the mechanism of interaction and binding of our inhibitors to HIV-1 IN. In the Leavitt laboratory, analytical ultracentrifugation experiments are in progress to measure the association constant of the catalytic core of HIV-1 IN in the absence and presence of the inhibitor. Yolanta Krucinski and Julian Chen in the Stroud laboratory have crystallized many forms of HIV-1 IN. Many of these crystals were co-crystallized with inhibitors, and the color of the inhibitor is visible in the crystal. (See table 3 for a list of crystals and inhibitors examined to date.) Chen is currently solving a structure of HIV-1 IN crystallized in the presence of an inhibitor. Experiments have been proposed with the Steve Hughes laboratory at the NIH, to characterize DNA binding to HIV-1 IN in the absence and presence of inhibitor using surface plasmon resonance. Additionally, work has been proposed with the George Kenyon lab, now at the University of Michigan, to synthesize new inhibitors based upon this class of molecules.

**Table 3: Crystals of HIV-1 IN catalytic core domain**

Inhibitor	Growth	Conditions	Diffraction (Å)	X-ray Source	Space Group
4017	co-crystal	Cd <sup>2+</sup> , SO <sub>4</sub> <sup>-</sup>	2.0	ALS	P3 <sub>2</sub>
4017	soak	Cd <sup>2+</sup> , SO <sub>4</sub> <sup>-</sup>	2.5	UCSF	P3 <sub>2</sub>
(apo)	--	Cd <sup>2+</sup> , SO <sub>4</sub> <sup>-</sup>	1.5	ALS	P3 <sub>2</sub>
36441	soak	Cd <sup>2+</sup> , SO <sub>4</sub> <sup>-</sup>	2.2	UCSF	P3 <sub>2</sub>
70629	soak	Cd <sup>2+</sup> , SO <sub>4</sub> <sup>-</sup>	2.2	UCSF	P3 <sub>2</sub>
4017	co-crystal	PEG	4.0	SSRL	P3 <sub>2</sub> <sup>a</sup>
4017	co-crystal	(NH <sub>4</sub> ) <sub>2</sub> SO <sub>4</sub>	2.5	ALS	C222 <sub>1</sub> <sup>b</sup>

Unit cell dimensions for all crystals with Cd (first 5 rows of table) are  $a=49\text{Å}$ ,  $b=49\text{Å}$ ,  $c=103\text{Å}$ ,  $\alpha=90^\circ$ ,  $\beta=90^\circ$ ,  $\gamma=120^\circ$ .

<sup>a</sup>Unit cell dimensions are  $a=44\text{Å}$ ,  $b=44\text{Å}$ ,  $c=123\text{Å}$ ,  $\alpha=90^\circ$ ,  $\beta=90^\circ$ ,  $\gamma=120^\circ$ .

<sup>b</sup>Unit cell dimensions are  $a=45\text{Å}$ ,  $b=70\text{Å}$ ,  $c=168\text{Å}$ ,  $\alpha=90^\circ$ ,  $\beta=90^\circ$ ,  $\gamma=90^\circ$ .

I recommend that future computational work include docking to the site on HIV-1 IN, using the structure from the Stroud lab, the latest version of the ACD, and version 4 of DOCK. The site on HIV-1 IN is quite different from ASV IN. It is more solvent-exposed and may require some expertise to avoid the pitfalls of solvent-exposed sites. We may also consider examining the two smaller domains for pockets, since they are required for full activity of IN. (The catalytic core cannot perform integration, but can only perform the reverse reaction.) Finally, I suggest revisiting the inhibitor MFCD0030706 to determine if the original measured IC<sub>50</sub> value of 4-6  $\mu\text{M}$  was anomalous.

## **Acknowledgments**

We greatly appreciate receiving coordinates from the Wlodawer lab and the Davies lab prior to their availability from the PDB. Thanks to Ann Tang and Andy Leavitt for the descriptions of their assays, Yolanta Krucinski and Julian Chen for the crystallographic data, and Karl Maurer for many synthetic chemistry lessons.

This work was a collaborative effort within the Kuntz Laboratory from its inception. Luke Hoffman provided guidance for the initial docking to ASV IN. Docking and similarity searching throughout the project was shared with Malin Young. Selections were made from the first ACD docking with votes from Keith Burdick as well as Malin, Luke and myself. Shinichi Katakura also made suggestions for synthesis of new compounds, and built a model of our lead compound at the time, 70629. Finally, Geoff Skillman provided an awesome amount of input throughout the project, as well as sharing his ChemDraw template for table 1.

Finally, this work was supported by the AIDS Program Project Grant, NIH GM56531



## References

- Artico M, R Di Santo, R Costi, E Novellino, G Greco, S Massa, E Tramontano, ME Marongiu, A De Montis and P La Colla. 1998. Geometrically and conformationally restrained cinnamoyl compounds as inhibitors of HIV-1 integrase: synthesis, biological evaluation, and molecular modeling. *Journal of Medicinal Chemistry* 41:3948-60.
- Bairoch A and R Apweiler. 1997. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J Mol Med* 75:312-316.
- Bujacz G, J Alexandratos, ZL Qing, C Clement-Mella and A Wlodawer. 1996a. The catalytic domain of human immunodeficiency virus integrase: ordered active site in the F185H mutant. *Febs Letters* 398:175-8.
- Bujacz G, M Jaskolski, J Alexandratos, A Wlodawer, G Merkel, RA Katz and AM Skalka. 1996b. The catalytic domain of avian sarcoma virus integrase: conformation of the active-site residues in the presence of divalent cations. *Structure* 4:89-96.
- Couch GS, EF Petterson, CC Huang and TE Ferrin. 1995. Annotating PDB files with scene information. *Journal of Molecular Graphics* 13:153-158.
- Cushman M, WM Golebiewski, Y Pommier, A Mazumder, D Reymen, E De Clercq, L Graham and WG Rice. 1995. Cosalane analogues with enhanced potencies as inhibitors of HIV-1 protease and integrase. *Journal of Medicinal Chemistry* 38:443-52.
- Daylight, Thor, and Merlin Toolkits, v4.42. 1996. Daylight Chemical Information Systems, Santa Fe, NM.
- Eich E, H Pertz, M Kaloga, J Schulz, MR Fesen, A Mazumder and Y Pommier. 1996. (-)-Arctigenin as a lead structure for inhibitors of human immunodeficiency virus type-1 integrase. *Journal of Medicinal Chemistry* 39:86-95.

Eijkelenboom AP, RA Lutzke, R Boelens, RH Plasterk, R Kaptein and K Hard. 1995. The DNA-binding domain of HIV-1 integrase has an SH3-like fold. *Nature Structural Biology* 2:807-10.

Engelman A, FD Bushman and R Craigie. 1993. Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. *EMBO Journal* 12:3269-3275.

Farnet CM, B Wang, M Hansen, JR Lipford, L Zalkow, WE Robinson, Jr., J Siegel and F Bushman. 1998. Human immunodeficiency virus type 1 cDNA integration: new aromatic hydroxylated inhibitors and studies of the inhibition mechanism. *Antimicrobial Agents and Chemotherapy* 42:2245-53.

Ferrin TE, CC Huang, LE Jarvis and R Langridge. 1988. The MIDAS display system. *Journal of Molecular Graphics* 6:13-27.

Fesen MR, KW Kohn, F Leteurtre and Y Pommier. 1993. Inhibitors of human immunodeficiency virus integrase. *Proceedings of the National Academy of Sciences of the United States of America* 90:2399-403.

GCG, v8.1. 1996. Genetics Computer Group, Madison, WI.

Goldgur Y, F Dyda, AB Hickman, TM Jenkins, R Craigie and DR Davies. 1998. Three new structures of the core domain of HIV-1 integrase: an active site that binds magnesium. *Proceedings of the National Academy of Sciences of the United States of America* 95:9150-4.

Hong H, N Neamati, S Wang, MC Nicklaus, A Mazumder, H Zhao, TR Burke, Jr., Y Pommier and GW Milne. 1997. Discovery of HIV-1 integrase inhibitors by pharmacophore searching. *Journal of Medicinal Chemistry* 40:930-6.

Huang CC, EF Pettersen, TE Klein, TE Ferrin and R Langridge. 1991. Conic: A fast renderer for space-filling molecules with shadows. *Journal of Molecular Graphics* 9:230-236.

Jenkins TM, AB Hickman, F Dyda, R Ghirlando, DR Davies and R Craigie. 1995. Catalytic domain of human immunodeficiency virus type 1 integrase: identification of a soluble mutant by systematic replacement of hydrophobic residues. *Proceedings of the National Academy of Sciences of the United States of America* 92:6057-61.

Kuntz ID, JM Blaney, SJ Oatley, R Langridge and TE Ferrin. 1982. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology* 161:269-88.

Leavitt A and A Tang. 1999. Personal Communication.

Leavitt AD, L Shiue and HE Varmus. 1993. Site-directed mutagenesis of HIV-1 integrase demonstrates differential effects on integrase functions in vitro. *Journal of Biological Chemistry* 268:2113-9.

Lodi PJ, JA Ernst, J Kuszewski, AB Hickman, A Engelman, R Craigie, GM Clore and AM Gronenborn. 1995. Solution structure of the DNA binding domain of HIV-1 integrase. *Biochemistry* 34:9826-33.

Lubkowski J, F Yang, J Alexandratos, A Wlodawer, H Zhao, TR Burke, Jr., N Neamati, Y Pommier, G Merkel and AM Skalka. 1998. Structure of the catalytic domain of avian sarcoma virus integrase with a bound HIV-1 integrase-targeted inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* 95:4831-6.

Maignan S, JP Guilloteau, Q Zhou-Liu, C Clement-Mella and V Mikol. 1998. Crystal structures of the catalytic domain of HIV-1 integrase free and complexed with its metal co-

factor: high level of similarity of the active site with other viral integrases. *Journal of Molecular Biology* 282:359-68.

Mazumder A, S Wang, N Neamati, M Nicklaus, S Sunder, J Chen, GW Milne, WG Rice, TR Burke, Jr. and Y Pommier. 1996. Antiretroviral agents as inhibitors of both human immunodeficiency virus type 1 integrase and protease. *Journal of Medicinal Chemistry* 39:2472-81.

McDougall B, PJ King, BW Wu, Z Hostomsky, MG Reinecke and WE Robinson, Jr. 1998. Dicafeoylquinic and dicafeoyltartaric acids are selective inhibitors of human immunodeficiency virus type 1 integrase. *Antimicrobial Agents and Chemotherapy* 42:140-6.

Mekouar K, JF Mouscadet, D Desmaele, F Subra, H Leh, D Savoure, C Auclair and J d'Angelo. 1998. Styrylquinoline derivatives: a new class of potent HIV-1 integrase inhibitors that block HIV-1 replication in CEM cells. *Journal of Medicinal Chemistry* 41:2846-57.

Meng EC, DA Gschwend, JM Blaney and ID Kuntz. 1993. Orientational sampling and rigid-body minimization in molecular docking. *Proteins: Structure, Function & Genetics* 17:266-278.

MDL Available Chemicals Database, v95.2 and v97.1, and ISIS, v2.1. 1997. Molecular Design Limited, Incorporated, San Leandro, CA

Neamati N, H Hong, A Mazumder, S Wang, S Sunder, MC Nicklaus, GW Milne, B Proksa and Y Pommier. 1997a. Depsides and depsidones as inhibitors of HIV-1 integrase: discovery of novel inhibitors through 3D database searching. *Journal of Medicinal Chemistry* 40:942-51.

Neamati N, H Hong, JM Owen, S Sunder, HE Winslow, JL Christensen, H Zhao, TR Burke, Jr., GW Milne and Y Pommier. 1998a. Salicylhydrazine-containing inhibitors of HIV-1 integrase: implication for a selective chelation in the integrase active site. *Journal of Medicinal Chemistry* 41:3202-9.

Neamati N, H Hong, S Sunder, GW Milne and Y Pommier. 1997b. Potent inhibitors of human immunodeficiency virus type 1 integrase: identification of a novel four-point pharmacophore and tetracyclines as novel inhibitors. *Molecular Pharmacology* 52:1041-55.

Neamati N, A Mazumder, S Sunder, JM Owen, M Tandon, JW Lown and Y Pommier. 1998b. Highly potent synthetic polyamides, bisdistamycins, and lexitropsins as inhibitors of human immunodeficiency virus type 1 integrase. *Molecular Pharmacology* 54:280-90.

Neamati N, A Mazumder, H Zhao, S Sunder, TR Burke, Jr., RJ Schultz and Y Pommier. 1997c. Diarylsulfones, a novel class of human immunodeficiency virus type 1 integrase inhibitors. *Antimicrobial Agents and Chemotherapy* 41:385-93.

Nicklaus MC, N Neamati, H Hong, A Mazumder, S Sunder, J Chen, GW Milne and Y Pommier. 1997. HIV-1 integrase pharmacophore: discovery of inhibitors through three-dimensional database searching. *Journal of Medicinal Chemistry* 40:920-9.

Ojwang JO, RW Buckheit, Y Pommier, A Mazumder, K De Vreese, JA Este, D Reymen, LA Pallansch, C Lackman-Smith, TL Wallace and et al. 1995. T30177, an oligonucleotide stabilized by an intramolecular guanosine octet, is a potent inhibitor of laboratory strains and clinical isolates of human immunodeficiency virus type 1. *Antimicrobial Agents and Chemotherapy* 39:2426-35.

Puras Lutzke RA, NA Eppens, PA Weber, RA Houghten and RH Plasterk. 1995. Identification of a hexapeptide inhibitor of the human immunodeficiency virus integrase protein

by using a combinatorial chemical library. *Proceedings of the National Academy of Sciences of the United States of America* 92:11456-60.

Raghavan K, JK Buolamwini, MR Fesen, Y Pommier, KW Kohn and JN Weinstein. 1995. Three-dimensional quantitative structure-activity relationship (QSAR) of HIV integrase inhibitors: a comparative molecular field analysis (CoMFA) study. *Journal of Medicinal Chemistry* 38:890-7.

Robinson WE, Jr., M Cordeiro, S Abdel-Malek, Q Jia, SA Chow, MG Reinecke and WM Mitchell. 1996a. Dicafeoylquinic acid inhibitors of human immunodeficiency virus integrase: inhibition of the core catalytic domain of human immunodeficiency virus integrase. *Molecular Pharmacology* 50:846-55.

Robinson WE, Jr., MG Reinecke, S Abdel-Malek, Q Jia and SA Chow. 1996b. Inhibitors of HIV-1 replication [corrected; erratum to be published] that inhibit HIV integrase. *Proceedings of the National Academy of Sciences of the United States of America* 93:6326-31.

Skillman AG, KW Maurer, DC Roe, MJ Stauber, D Eargle, TJA Ewing, A Muscate, M.V. Madaglia, R Fisher, E Arnold, H Gao, RB II, PL Boyer, S.H. Hughes, ID Kuntz and GL Kenyon. 1999. A Novel mechanism for inhibition of HIV-reverse transcriptase. *In preparation*.

Zhao H, N Neamati, H Hong, A Mazumder, S Wang, S Sunder, GW Milne, Y Pommier and TR Burke, Jr. 1997a. Coumarin-based inhibitors of HIV integrase. *Journal of Medicinal Chemistry* 40:242-9.

Zhao H, N Neamati, A Mazumder, S Sunder, Y Pommier and TR Burke, Jr. 1997b. Arylamide inhibitors of HIV-1 integrase. *Journal of Medicinal Chemistry* 40:1186-94.

Zhao H, N Neamati, S Sunder, H Hong, S Wang, GW Milne, Y Pommier and TR Burke, Jr. 1997c. Hydrazide-containing inhibitors of HIV-1 integrase. *Journal of Medicinal Chemistry* 40:937-41.

Zheng R, TM Jenkins and R Craigie. 1996. Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization, and enhances catalytic activity. *Proceedings of the National Academy of Sciences of the United States of America* 93:13659-64.

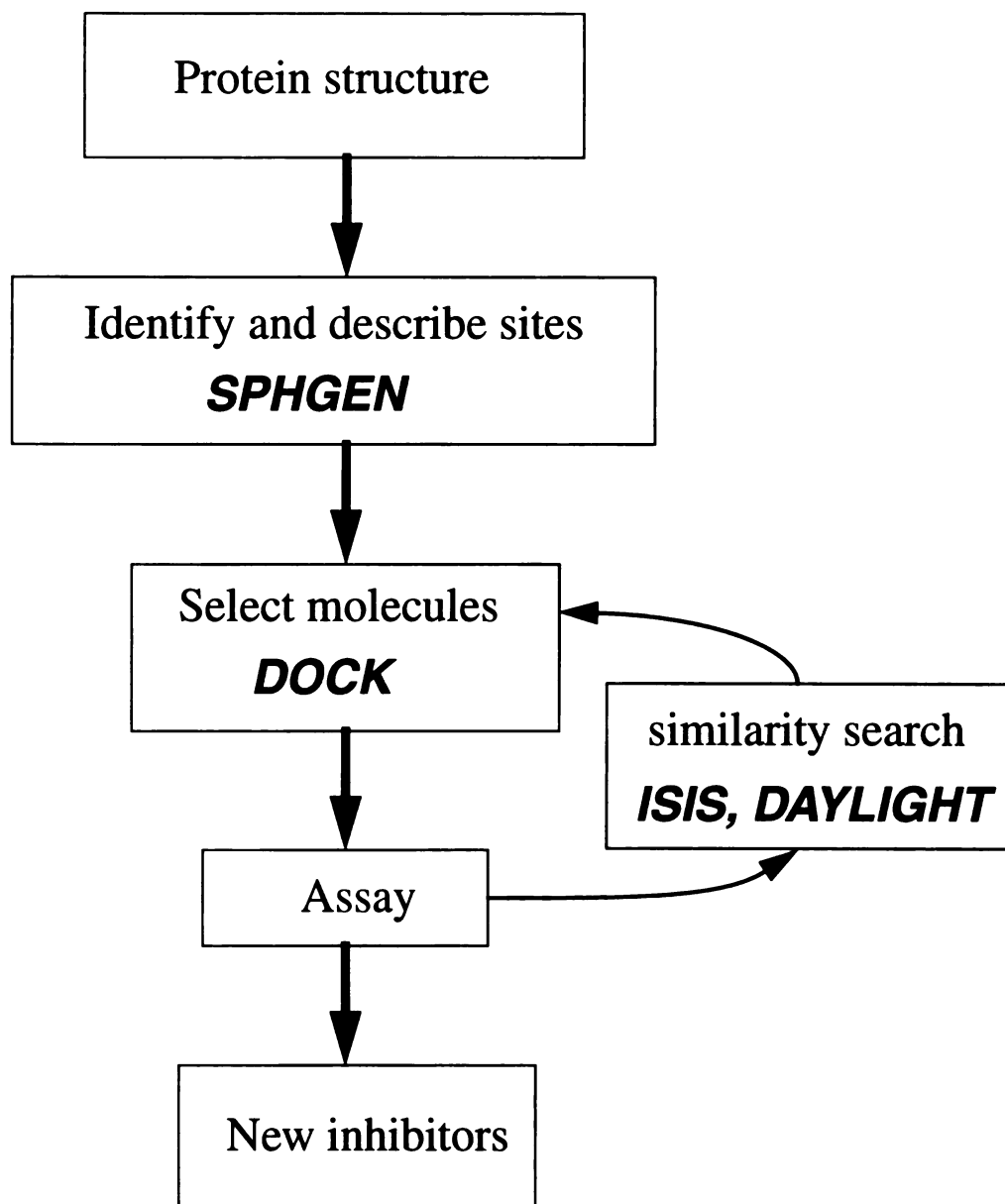


**Figure 1. Structural alignment of the catalytic core domains of HIV-1 IN and ASV IN.**

The HIV-1 IN structure (2itg) is shown in green and the ASV structure (1asu) is shown in magenta. The molecules were aligned by using the “match” feature in MidasPlus 2.1 (Ferrin, et al. 1988)(Huang, et al. 1991)(Couch, et al. 1995), aligning the alpha carbons of ASV IN R74 with HIV-1 IN K71 and ASV IN H103 with HIV-1 IN K103.







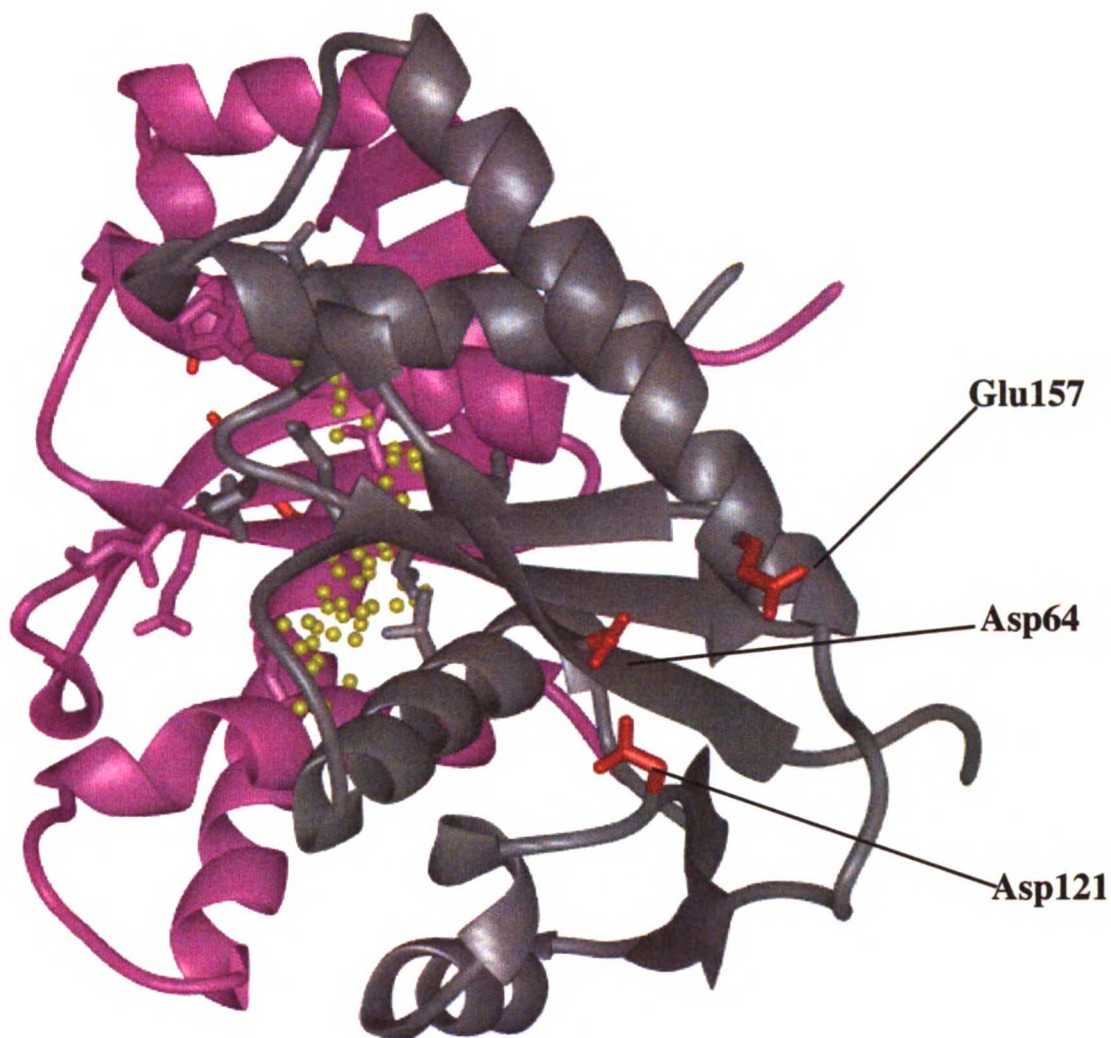
**Figure 3. Computational approach.**

This flow chart outlines our computational approach and the computer programs used to select molecules, including SPHGEN, DOCK, DAYLIGHT and MDL ISIS.

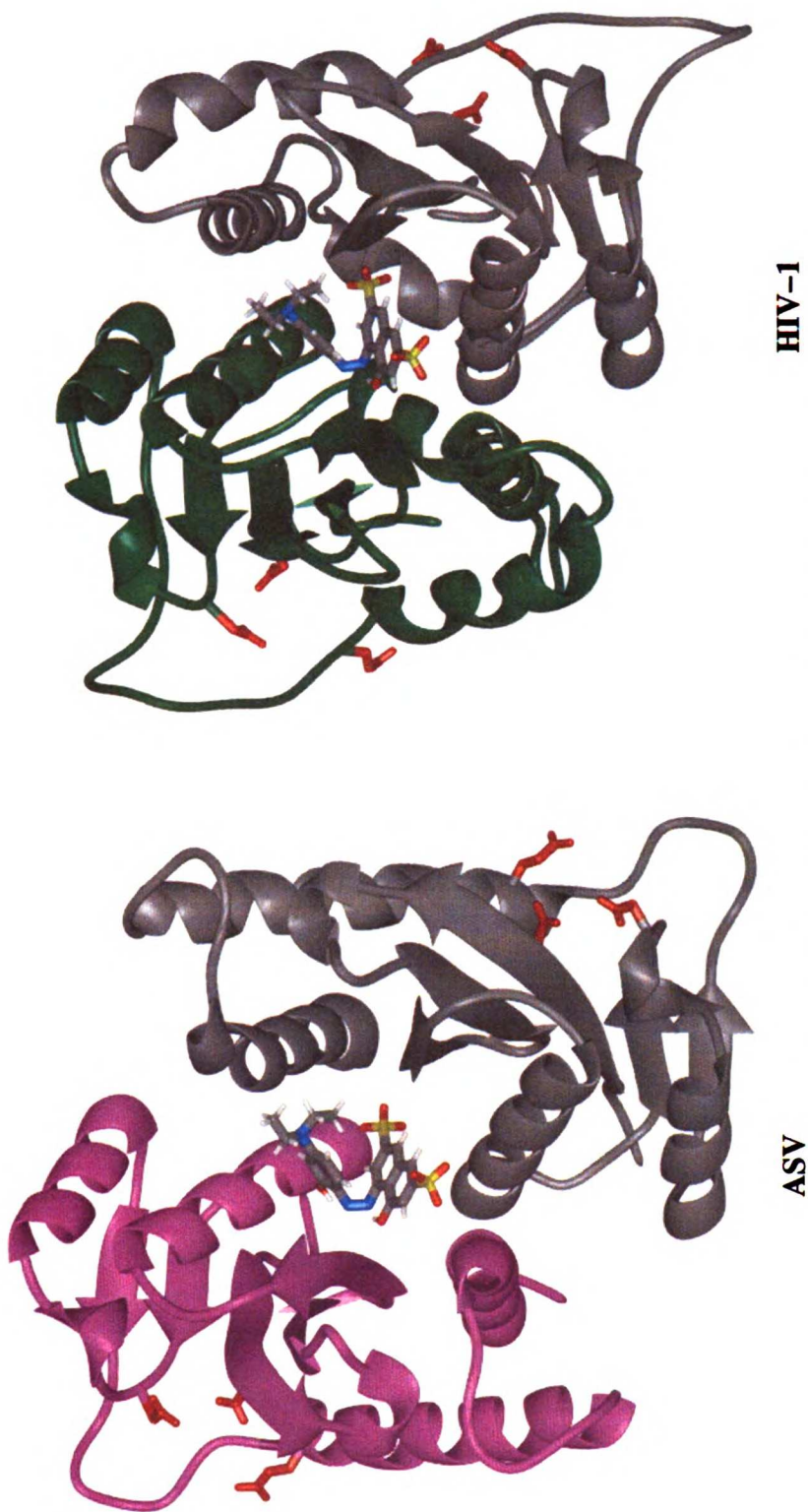


**Figure 4. Docking sites on ASV IN catalytic core domain.**

One monomer of ASV IN is shown in magenta, and the other is in gray. The sphere centers from SPHGEN are shown as small yellow spheres, and the active site residues, D64, D121, and E157 are shown in red.

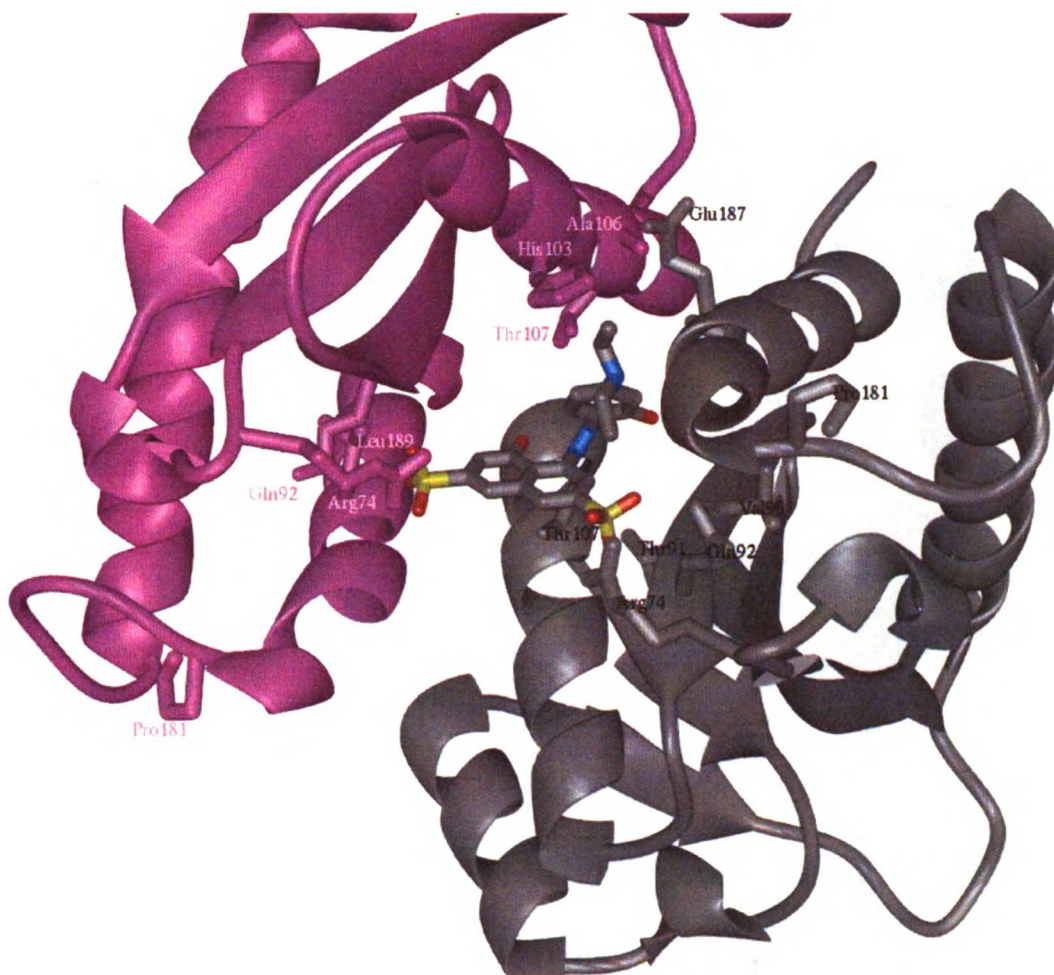


**Figure 5. Docking sites on ASV IN catalytic core domain, side view.**  
The molecule and spheres are shown as in Figure 4, but turned 90° about an axis through the dimer site.



**Figure 6. ASV IN and HIV-1 IN with MFC0070629 docked into the dimer site.**

ASV IN is on the left, with one monomer in magenta and the other in gray, and HIV-1 IN is on the right, with one monomer in green and the other in gray.



**Figure 7. MFCD0070629 in HIV-1 IN dimer site.**

The small molecule was docked to 2itg. H171 has been identified as a key residue near the undetermined density in the Chen structure. Other residues are shown because they lie near the ligand in this orientation.



**Figure 8. MFCD0070629 in HIV-1 IN dimer site.** The small molecule was docked to 2itg. H171 has been identified as a key residue near the undetermined density in the Chen structure. Other residues are shown because they lie near the ligand in this orientation.

## Chapter 5



### COMPUTATIONAL APPROACHES TO SAR BY NMR BY DOCK



*This work was done with Kaiqi Chen in the Kuntz laboratory and in collaboration with Philip Hajduk from the Steve Fesik laboratory at Abbott Laboratories.*



## **Abstract**

The experimental method, SAR by NMR, from the Fesik laboratory at Abbott screens up to 1,000 small molecules each day (Shuker, et al. 1996). Analogously, DOCK computationally screens small molecules. To compare the methods, DOCK was used to screen a database of molecules to a target that was previously screened using the SAR by NMR method and compared results. In the first run, molecules were selected from the Available Chemicals Directory (Molecular Design Limited 1995, 1997) and assayed them using SAR by NMR. In the second run, a database of molecules which had been previously assayed was docked. Two sets of site points were compared: one from SPHGEN and the other from SURFSPH. It was found that, with an early release of version 4 of DOCK, greater sampling was required than had been executed in the runs, and that adjusting the score to account for DOCK's preference of larger molecules improved our results.

## **Introduction**

In recent years, new strategies have been advanced to speed the discovery of bio-active compounds. These include high-throughput screening, combinatorial chemistry and structure-based drug design. One protocol combines structural information and computation in a process called “database mining”. By searching databases of molecules quickly, UCSF DOCK can computationally screen compounds in a few days, while it might take months screening assay in the laboratory. In addition to screening, DOCK gives a hypothetical binding mode which can assist with lead optimization development.

Analogous to docking, a new structure-based screening method has been developed in the laboratory of Stephen Fesik at Abbott Laboratories (Hajduk, et al. 1997b). Called SAR by NMR, for “structure-activity relationships by nuclear magnetic resonance”, this technique uses heteronuclear NMR to find small molecules that bind to a target protein at millimolar affinity. Libraries of small molecules are scanned by NMR, searching for two or more fragments with low affinity that can be linked to produce a larger, better binding molecule. Samples are screened quickly, pooling 10 small molecules at a time with the target protein. The experiment takes about 10 minutes for each sample. It is expected that 1,000 molecules can be sampled each day. Upon finding one low affinity ligand, the library is again screened in the presence of the first ligand and target protein to find a second (non-competitive) ligand. Using NMR, this technique gives not only binding affinity but, using NOE data collected separately, it also gives binding locations of the fragments.

To test this method, the Fesik group used the target of FK506 binding protein (FKBP). They found a ligand with a binding affinity of 2  $\mu\text{M}$ , pipercolinic acid, in their library of 1000 molecules, then scanned for a second fragment that bind to FKBP in a saturated concentration of pipercolinic acid. The second ligand, a benzanilide derivative, bound at 110  $\mu\text{M}$  affinity. The two molecules were covalently linked with several different linkers, then assayed, producing new, low-affinity ligands to FK506, binding at 19-228 nM (Shuker, et al. 1996).

Additional lead compounds have been discovered using this method, including a 15 nM inhibitor of stromelysin (Hajduk, et al. 1997c), a metalloprotease, and 10  $\mu\text{M}$  inhibitor of the DNA binding domain of human papillomavirus E protein (Hajduk, et al. 1997a). Both target proteins, as well as FKBP, are small, with 83 residues for E protein, 176 residues for the catalytic domain of stromelysin, and 107 residues for FKBP. A limitation of the SAR by NMR method is that the protein targets must be highly soluble and less than 40 kilodaltons.

I wished to evaluate the performance of DOCK in comparison to the findings of SAR by NMR, and in particular to evaluate the scoring methods available within the DOCK suite of programs. I also hoped to carry forth a replication of the NMR experiment computationally, selecting a single, well-binding compound from the ACD and then scanning for a second compound with the first ligand docked into the site on the target molecule.

DOCK not only suggests that a molecule might bind a target, but also presents a likely binding mode. Thus, I hoped to test DOCK's ability to select ligands, and also to compare and analyze the binding modes of the ligands to those found with SAR by NMR

Our first step in this process was to perform the initial screen of the database, searching for small molecules that bound to our target protein.

This first step was implemented with two related methods. One method was to examine the Available Chemicals Directory (ACD) (Molecular Design Limited 1995, 1997) of 220,000 molecules using a variety of different docking and scoring schemes, and select a subset of up to 100 molecules for testing. The second method was to dock the list of molecules previously assayed by Abbott and allow DOCK to rank them. A list of 4,000 molecules previously assayed by Abbott on FKBP and their binding affinities was obtained. These 4,000 molecules were then docked and scored, examining the effects of sampling, scoring method and site point selection.

For a test case, our collaborators, the Fesik laboratory, chose FKBP, because significant testing had been completed at Abbott and several structures were available via the Protein Data Bank, and Abbott generously released the coordinates of their NMR structure of FKBP bound to ascomycin to us. FKBP is a small, highly soluble protein that is ubiquitous in mammalian cells. It functions as a peptidylprolyl cis-trans isomerase, but when bound to FK506, the complex binds to calcineurin, at a site 10Å from calcineurin's active site, thus non-competitively inhibiting the phosphatase. The inhibition of calcineurin blocks T cell activation, and FK506 has long been used as an immunosuppressant.

## **Methods**

Two rounds of docking FKBP to small-molecule libraries were executed. The first round used DOCK 4.0beta to the ACD 95.2 (Molecular Design Limited 1995, 1997), and

the second round used DOCK 4.0 to a database of 7,000 compounds made available to us by Abbott Laboratories.

### ***Part I: Searching the ACD***

Using the average structure of ascomycin bound to FKBP, site points were selected by running the program SPHGEN (Kuntz, et al. 1982), and then choosing spheres that lay near the heavy atoms of ascomycin. 42 spheres were selected, shown in figure 1.

Molecules were scored using the AMBER-based force field score. Scoring grids were generated using the CHEMGRID program (Meng, et al. 1992) from the DOCK 3.5 suite of programs (Meng, et al. 1993). Orientations were generated using DOCK 4.0Beta, and the uniform-sampling option, which generates a user-selected number of orientations for each molecule by adjusting the matching parameters until the selected number of orientations is found. Each molecule was sampled until 100 orientations were generated. A portion of the ACD 95.2 was docked that included molecules with no formal charge and 5 to 35 non-hydrogen atoms.

Scoring grids were also generated from an alignment of 10 crystal and NMR structures of FKBP available from the Protein Data Bank and Abbott, using the ensemble grid program from Ron Knegt (Knegtel, et al. 1997). This program creates a “geometry-weighted grid”, and is heavily dependent upon the alignment of the structures. Where the structures are poorly aligned, the grid has no repulsive terms. Upon docking a subset of molecules from the ACD, it was determined that this method was not optimal for our purposes, as many of the molecules had docked positions that were deeply buried in the protein.

After docking the database, the 1,000 highest-scoring compounds were selected and clustered using Daylight fingerprints (Daylight 1996) and hierarchical clustering (Ewing 1996). While the dock run included molecules with up to 35 heavy atoms, molecules with 20 or fewer non-hydrogen atoms and a maximum molecular weight of 350 daltons (which excluded molecules with several chlorine, bromine, iodine or phosphorus atoms) were chosen. All molecules with the same score as the 1,000th best-scoring molecule were included, resulting in a total of 1,010 molecules.

Daylight fingerprints are based on atom-atom connectivity, and thus are two-dimensional. A similarity index of 0.4 was selected after considering values from 0.2-0.9. Working from the largest to the smallest clusters, I selected one molecule from each cluster to be put forth for assaying. By pre-clustering, a diverse set was assured, as similar molecules cluster together. The selection within each cluster was based upon availability, molecular weight, and the octanol/water partition coefficient as calculated by ClogP (Daylight 1996). Molecules from vendors within the United States were favored, as it is easier and more affordable to purchase from domestic vendors. Smaller molecules were preferred to larger molecules, since the search was for fragments that would later be joined to other fragments. Finally, ClogP was used to calculate the partition coefficient by examining fragments within the molecule. Each calculation from ClogP is given an associated error message, which evaluates the accuracy of the calculation. ClogP is not accurate for charged molecule. Entire clusters were excluded if the ClogP program could not accurately assess any of the molecules within the cluster, or if there were no molecules with ClogP values less than 3.0. Within a cluster, if two molecules were of similar ClogP value

and from U.S. distributors, we chose the better-scoring molecule. 32 molecules were selected.

Assays were performed in the Fesik laboratory at Abbott Laboratories, using the SAR by NMR method. Samples were diluted to 1mM, sometimes in the presence of DMSO, and pooled in groups of 10, then scanned in the presence of the target protein, FKBP, for their effect on the chemical shift in the heteronuclear single-quantum coherence (HSQC) spectra of the backbone amide hydrogens. The molecules were given a score of 0 to 5: 0 represented no binding, or no change in the observed spectra; 5 representing a binding with a dissociation constant better than 1 mM.

### ***Part II: Ranking a database***

In part I, 32 molecules were selected, but only 24 molecules were assayed. This was due to difficulties obtaining molecules from companies outside of the United States, and occasionally from some of the companies within the United States. It was decided, with our collaborators in the Fesik lab, to change the experiment and examine a database of 7,000 molecules that were available to the Fesik laboratory, all of which were included in the ACD.

Of the 7,997 molecules on the list from Abbott, several were duplicates, and several were not available in the ACD 95.2 nor the ACD 97.1. Removing the unavailable and duplicate molecules resulted in a database of 7,434 molecules, that ranged in size from 32.0 daltons to 588.3 daltons with a mean molecular weight of 199.2 and standard deviation of 62.2. Similarly, the average size as measured by the number of non-hydrogen atoms was 13.5 with a standard deviation of 4.3. Of the 7,434 molecules in the new data-

base, 4,112 had been assayed for binding to FKBP, and of these molecules, 3,974 were available in our databases.

As a comparison for site point selection, spheres were generated using the program SURFSPH (Oshiro and Kuntz 1998). These sites are shown in figure 2. The SURFSPH program was developed to place site points close to the molecular surface, and thus can be used on shallow sites like the FKBP site. On its initial run, SURFSPH generates many more spheres than could be used for docking, but the program automatically ranks the spheres based upon the relative energy of a probe atom at the sphere site, and then selects sites of lower energy. Using the SURFSPH software, I generated an initial set of 2,200 spheres which were reduced to 93, then chose 47 spheres by visual inspection.

The database was docked with version 4.0 of DOCK. These molecules were docked using both uniform sampling, with a target of 500 orientations per molecule, and standard sampling, with a distance tolerance of 0.25Å and a minimum distance of 2.0Å, and with the SURFSPH spheres and the SPHGEN spheres described in part I. Scoring grids were generated using the GRID program available with version 4.0 of DOCK and the averaged NMR structure of FKBP with ascomycin from Abbott Laboratories. Orientations were scored using the force field score, the van der Waals component of the force field score, contact score and chemical scoring, all available with version 4.0 of DOCK.



## Results

Results are reported in the same manner as Methods, with part I reviewing the search of the ACD and part II discussing the analysis of the database of molecules from Abbott Laboratories.

### *Part I: Searching the ACD*

Beginning with the top 1,010 molecules with fewer than non-hydrogen atoms and molecular weight less than 350 daltons, hierarchical clustering was used to define groups of similar molecules, thus ensuring a more diverse selection. After examining the clusters generated with similarity values from 0.2 to 0.9, the results were clustered with a similarity of 0.4, judging from the size of the clusters, the number of singletons (clusters of one molecule) and visual inspection of the similarity of the molecules. Table 1 shows the results of clusters formed with similarity values from 0.4 to 0.7. Other values were not considered because the clusters were obviously too large or too numerous.

**Table 1: Clustering of 1,010 highest-scoring molecules**

Similarity index	Number of clusters	Number of singletons	Size of largest cluster
0.4	153	11	54
0.5	305	88	37
0.6	475	246	26
0.7	639	450	14

Of the 23 largest clusters, no selections were made from one cluster because the ClogP values from this cluster could not be accurately assessed. Of the next 55 largest clusters, several clusters were excluded because the molecules were not available from

domestic distributors, the molecules contained one or more nitro-groups, or all members of the cluster had ClogP values that exceeded 3.0. Molecules with nitro-groups were excluded because DOCK's scoring disproportionately favors molecules with large group dipole moments. They are also not handled well in the ClogP calculation. Finally, 32 molecules were selected from 32 clusters.

Of the 32 molecules, 24 were purchased and assayed in the Fesik laboratory at Abbott Laboratories, using the SAR by NMR method (Shuker, et al. 1996). (As mentioned earlier, there were difficulties obtaining molecules from companies outside of the United States, so the experiment and assay moved forward with the 24 molecules that could be obtained.) The results of these assays are shown in Table 2. One molecule bound to FKBP with a dissociation constant of 110  $\mu\text{M}$ . This molecule is shown in its DOCK orientation in figure 3.

**Table 2: Results of docking the ACD to FK506 binding protein**

MFCD number	Score at 0.25mM	Score at 0.5mM	Score at 1.0mM	K <sub>D</sub> (mM)
5020	0	0	0	>5
30828	0	0	--	>5
89405	0	1	--	>5
5694	3	4	5	0.11
91321	0	0	--	>5
96021	0	0	--	>5
20671	0	0	--	>5
142731	0	1	1	>5
2183	0	2	1	>5
88985	0	0	--	>5
88669	1	1	--	>5
57074	0	1	1	>5
158981	0	0	--	>5
91212	0	1	--	>5
5503	1	0	--	>5
5261	0	0	--	>5
42801	0	1	1	>5
24203	2	1	0	>5
94561	0	0	--	>5
160177	0	1	1	>5
160514	2	2	2	>5
186424	0	0	--	>5
213333	0	0	--	>5
5004	0	0	--	>5

Scores are a qualitative indication of the change in chemical shift of the HSQC spectrum of FK506 binding protein in the presence of compound, at each concentration (0.25, 0.5 and 1.0 mM). Scores of 0-1 indicate little binding to the protein, and scores of 3 or higher indicate binding. Affinities are estimated as exceeding 5mM for all except 5694. The dissociation constant (K<sub>D</sub>)

was measured by fitting the shift in the peaks of the HSQC spectrum as a function of ligand added. This data was provided by Philip Hajduk from Abbott Laboratories.

Reviewing and re-scoring of these molecules revealed that a bug in DOCK 4.0Beta did not properly calculate scores when using DOCK 3.5 grids, and that our level of uniform sampling, 100 orientations, was insufficient for this site and these molecules. This was determined by re-docking the 1,010 top-scoring molecules from the original run using DOCK 4.0. Upon re-docking, the distribution of scores that had seen in the original run could not be obtained (figure 4). Additionally, it was determined that the appropriate number of orientations for uniform sampling of this site would be much higher than 100 (figure 5). For a quick determination of the appropriate amount of sampling, the distribution of scores for the 1,010 top-scoring compounds was examined for a strong correlation. Sampling for 100, 200, 500 and 1,000 orientations, the distribution changed little between 500 and 1,000 orientations. Thus, for this site, the appropriate sampling would be greater than 200 orientation and fewer than 500 orientations. This result was used when docking the second database of available molecules from Abott.

## ***Part II: Ranking a database***

### ***Selecting site points and examining sampling methods***

The database of 3,974 assayed molecules from Abbott Laboratories was docked using the sphere sets from SURFSPH and using the previously-generated site points from SPHGEN. The database was also docked with uniform sampling, set to 500 orientations, and standard sampling with a distance tolerance of  $0.25\text{\AA}$  and a minimum distance of  $2.0\text{\AA}$ . The results of the SPHGEN dockings to the SURFSPH dockings are compared in figures 6-9.

Some molecules could not be docked with standard sampling. These molecules probably had matches, but the orientations generated by the matches did not have favorable scores. The same phenomenon occurred with uniform sampling, but less frequently, since DOCK was able to change distance tolerances until it found 500 orientations, some of which scored favorably.

The correlation of scores from SPHGEN and SURFSPH site points is best with energy scoring and uniform sampling (figure 9), and SURFSPH sites generally give better-scoring orientations than SPHGEN sites. With contact scoring and either sampling, SURFSPH sites produce higher-scoring orientations, although the correlation is poor (figures 6 and 8). The more favorable results with SURFSPH sites is likely to be a result of the close proximity to the surface when compared to SPHGEN sites, thus increasing favorable van der Waals interactions.

With standard sampling and energy scoring, regression analysis between scores from SURFSPH and SPHGEN site points suggests that the SPHGEN sites give slightly better scores than SURFSPH spheres (figure 7). The correlation ( $R=0.865$ ) is adequate and if one examines the number of points with scores more favorable than -25, there are more favorably-scoring orientations from SURFSPH sites than from SPHGEN sites. This suggests that the overall results from SURFSPH sites may be more favorable when using standard sampling. For the same amount of sampling, some very high-scoring orientations were obtained with SURFSPH sites that were not obtained with SPHGEN sites.

Standard sampling was used rather than uniform sampling because while uniform sampling found orientations for all molecules, producing the user-set number of orientations for each molecule regardless of its size, default sampling searches exhaustively for

all possible orientations within the user-defined distance tolerance and minimum distance. Thus, while it may only generate a few orientations for a large molecule, it may produce 5,000 for a smaller molecule, and find a better-scoring orientation in that exhaustive search. The database was re-docked using the SURFSPH site points, increasing the distance tolerance slightly to 0.35Å to examine the performance of DOCK in finding possible ligands.

### ***Scoring***

Of the 3,974 assayed molecules available in the database, dissociation constants ( $K_D$ ) were determined for 132 molecules of which 27 had a  $K_D$  of 1mM or less. For the remaining molecules, binding was measured in pools of 10 with the target protein, and the binding data is reported as greater than 4mM or greater than 10mM.

To examine DOCK's performance, the DOCK results were ranked from most favorable score to least favorable score. The question was then posed: was DOCK able to separate the wheat from the chaff, and did we find the hits among the non-hits in the database? This was answered by examining the enrichment curve, a plot of the fraction of the database searched on the x-axis versus the fraction of hits recovered on the y-axis (figures 10-15). The ideal enrichment would be to score  $M$  experimental hits as the  $M$  top-scoring molecules in a database of  $N$ . For a random enrichment,  $M$  molecules evenly distributed over the database of  $N$  would be found. The ideal and random curves for these data of 27 "hits" in a database of 4,000 are shown in figure 10.

When comparing three scoring methods: contact scoring, energy scoring, and vdw scoring (the van der Waals component of the energy score) none of these scoring methods are ideal (figure 11). Overall, contact scoring finds all of the hits before either other

method, and all methods do better than random. However, if this were a “real” dock run, only a small number of compounds would be assayed, perhaps 100. 100 compounds comprises 2.5% of the database of 4,000. When the first 25% of the database searched is examined, again the contact scoring finds the most hits, as well as at 2.5% (figure 12).

Because the DOCK scoring methods do not account for entropic changes, including solvation effects, it incorrectly favors larger molecules. To attempt to account for this, the vdw portion of the force-field score was used, which performed slightly better than the energy score, was more quantitative than the contact score, and was found to correlate best in another system (Sun, et al.). The score was adjusted by adding a fraction of the molecular weight of the molecule to the score. A range of factors was examined, from 0.1 to 0.8. Over the entire database, adjusting the scores resulted in poorer performance, in that nearly the entire database was searched before all hits were found (figure 13). Upon examining the initial slope of the enrichment curve, by adjusting the vdw score, 20% of the hits were found after searching only 3% of the database, or 120 molecules (figure 14).

A similar examination of enrichment was done on scores normalized by the molecular weight of the molecule, simply dividing by the molecular weight. These enrichment curves performed worse than random searches.

## **Conclusions**

In this study, a new 100 $\mu$ M-binding ligand to FK506 binding protein was discovered upon searching a large database and testing only 24 molecules. The use of DOCK was explored to rank databases of molecules. DOCK’s scoring methods were found to favor larger molecules, but some of the bias can be corrected.

Two issues made this project inherently difficult: the target, FKBP, and the size of the molecules in this target. First, the target, FKBP, was selected as our test system because it has been well-characterized, structurally and chemically. Unfortunately, the binding site of FK506 to FKBP was very solvent-exposed, and the DOCK scoring methods do not account for solvent effects. Also, more orientations were generated for “open” sites, because there are more ways to place the molecule in the site than there would be if it were a pocket. The second issue was that this was a search for **small, millimolar-binding** molecules, while DOCK successes have generally been larger, micromolar-binding molecules (for example (Hoffman, et al. 1997)(Gschwend, et al. 1997) (Somoza, et al. 1998)). These issues highlighted two prominent considerations within DOCK: scoring and sampling.

These studies have revealed that overall, DOCK’s vdw, energy and contact scoring can select well-binding molecules out of a database, but the best-binding molecules are not selected as the highest-scoring molecules. Overall, it does perform better than a random search. The scoring can be slightly altered to correct for DOCK’s favoring larger molecules, but adjusting the score by adding a factor based on the molecular weight improves enrichment over the first 10% of the database search and then diminishes enrichment over the remaining 90% of the database search. Because the top-scoring molecules are selected for testing, these adjusted scores are an advancement.

Numerous other methods come to mind to explore “fixes” to DOCK scoring, particularly of these kinds of sites, which are quite common in non-enzymatic systems, particularly systems that bind other large molecules, like DNA-binding proteins. The use of solvation correction methods being developed within the Kuntz laboratory (Zou and Sun



1998) could be explored, particularly using it over the database of 4,000 molecules. Simpler solvation correction methods based upon the exposed surface area of the ligand could also be explored, as developed by Dan Gschwend et al. (Gschwend 1995), or changing the sampling of orientations slightly by placing dummy atoms in the solvent as has been done successfully in a previous study (Somoza, et al. 1998), which would force the molecules into the shallow pocket and thus spend more “time” sampling closer to the surface. This might improve the overall performance as measured by CPU time, finding better-scoring orientations faster, but not by score selection, as it exhaustively searched all orientations that met the distance criterion. Finally, other scoring methods could be examined, for example, empirical methods such as DOCK’s chemical scoring.

It can be difficult to determine the sufficient amount of sampling. A standard method has been to select a sample of the database and sample it at different distance tolerances and minimum distances, then examine the results and judge, by examination, whether the sampling was sufficient. Our initial approach assumed that this question would be answered by uniform sampling, which would allow us to generate a set number of orientations for each molecule. Upon reflection, however, this resulted in incomplete sampling, in that the results did not correlate between runs as determined by rank correlation (data not shown) and the distribution of raw score. In the end, standard sampling methods and visual examination were used, since standard sampling methods resulted in better enrichment (data not shown), even with a large number of orientations, than uniform sampling. It is likely that this is due to the fact that 500 orientations is an insufficient amount of sampling for some smaller molecules, but exceeds the necessary sampling for

larger molecules which would not fit in the site. Another issue is that sufficient sampling is a site-dependent and database-dependent phenomenon.

Another question is how does one best measure performance? Here, enrichment was measured, focusing on finding the hits in DOCK's top-ranked molecules. Overall, DOCK was able to locate some hits in the top 100 compounds of the 4,000 molecules in the database with either contact, van der Waals, or energy scoring, and was able to improve enrichment when adjusting the van der Waals for favoring larger molecules. Statistical methods may reveal a better measurement of performance among series of DOCK runs on the same database.

The future of this project could lie in the exploration of other scoring methods, however, because of the site-dependence of the model, it may be better to examine other targets. The Fesik laboratory has now published two additional searches, one to stromelysin E, a metallo-protease, and another to the human papillomavirus E2 protein, which binds to DNA. The metalloprotease site is not solvent-exposed, and may be a better target for the existing DOCK scoring methods.

Despite that the site was difficult to model using existing DOCK methodology, one hit out of the uncharged portion of the ACD was found, and several of the 27 hits out of the pre-screened database were located using different scoring and site characterization methods. It is still the case that DOCK provides many false positives and many false negatives, but still provides a good number of true positives. Perhaps additional scoring considerations will improve results, particularly for solvent-exposed sites like FKBP.

## Acknowledgments

This work would not have been possible without data generously provided by Philip Hajduk in Steve Fesik's laboratory at Abbott Laboratories. I also thank my office mates, Connie Oshiro and Geoff Skillman, for many helpful discussions, and Todd Ewing for assistance with the new software, version 4 of DOCK.

## References

Daylight, Thor, and Merlin Toolkits, v4.42. 1996. Daylight Chemical Information Systems, Santa Fe, NM.

Ewing T. 1996. Personal communication.

Gschwend DA. 1995. Molecular docking towards drug discovery: improving interaction specificity. Pharmaceutical Chemistry. San Francisco, UCSF.

Gschwend DA, W Sirawaraporn, DV Santi and ID Kuntz. 1997. Specificity in structure-based drug design: identification of a novel, selective inhibitor of *Pneumocystis carinii* dihydrofolate reductase. *Proteins* 29:59-67.

Hajduk PJ, J Dinges, GF Miknis, M Merlock, T Middleton, DJ Kempf, DA Egan, KA Walter, TS Robins, SB Shuker, TF Holzman and SW Fesik. 1997a. NMR-based discovery of lead inhibitors that block DNA binding of the human papillomavirus E2 protein. *Journal of Medicinal Chemistry* 40:3144-50.

Hajduk PJ, RP Meadows and SW Fesik. 1997b. Discovering high-affinity ligands for proteins. *Science* 278:497,499.

Hajduk PJ, G Sheppard, DG Nettlesheim, ET Olejniczak, SB Shuker, RP Meadows, DH Steinman, GM Carrera, PA Marcotte, J Severin, K Walter, H Smith, E Gubbins, R Simmer, TF Holzman, DW Morgan, SK Davidsen, JB Summers and SW Fesik. 1997c. Discovery

of potent nonpeptide inhibitors of stromelysin using SAR by NMR. *Journal of the American Chemical Society* 119:5818-5827.

Hoffman LR, ID Kuntz and JM White. 1997. Structure-based identification of an inducer of the low-pH conformational change in the influenza virus hemagglutinin: irreversible inhibition of infectivity. *Journal of Virology* 71:8808-20.

Knegtel RM, ID Kuntz and CM Oshiro. 1997. Molecular docking to ensembles of protein structures. *Journal of Molecular Biology* 266:424-440.

Kuntz ID, JM Blaney, SJ Oatley, R Langridge and TE Ferrin. 1982. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology* 161:269-88.

Meng EC, DA Gschwend, JM Blaney and ID Kuntz. 1993. Orientational sampling and rigid-body minimization in molecular docking. *Proteins* 17:266-278.

Meng EC, BK Shoichet and ID Kuntz. 1992. Automated Docking with Grid-Based Energy Evaluation. *Journal of Computational Chemistry* 13:505-524.

Available Chemicals Database, v95.2 and v97.1, and ISIS, v2.1. 1997. Molecular Design Limited, Incorporated, San Leandro, CA

Nicholls A, K Sharp and B Honig. 1991. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* 11:281-206.

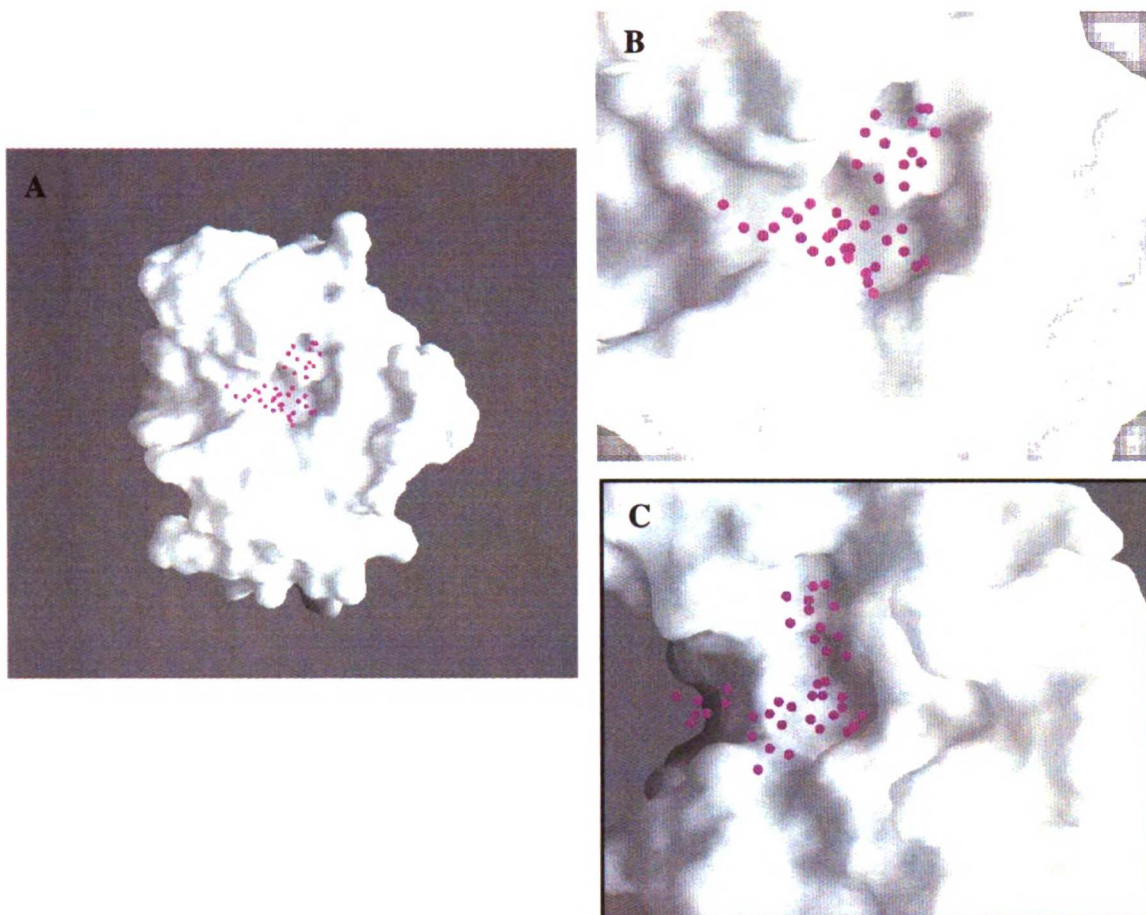
Oshiro CM and ID Kuntz. 1998. Characterization of receptors with a new negative image: use in molecular docking and lead optimization. *Proteins* 30:321-36.

Shuker SB, PJ Hajduk, RP Meadows and SW Fesik. 1996. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 274:1531-4.

Somoza JR, AG Skillman, NR Munagala, CM Oshiro, RM Knegtel, S Mpoke, RJ Fletterick, ID Kuntz and CC Wang. 1998. Rational design of novel antimicrobials: blocking purine salvage in a parasitic protozoan. *Biochemistry* 37:5344-5348.

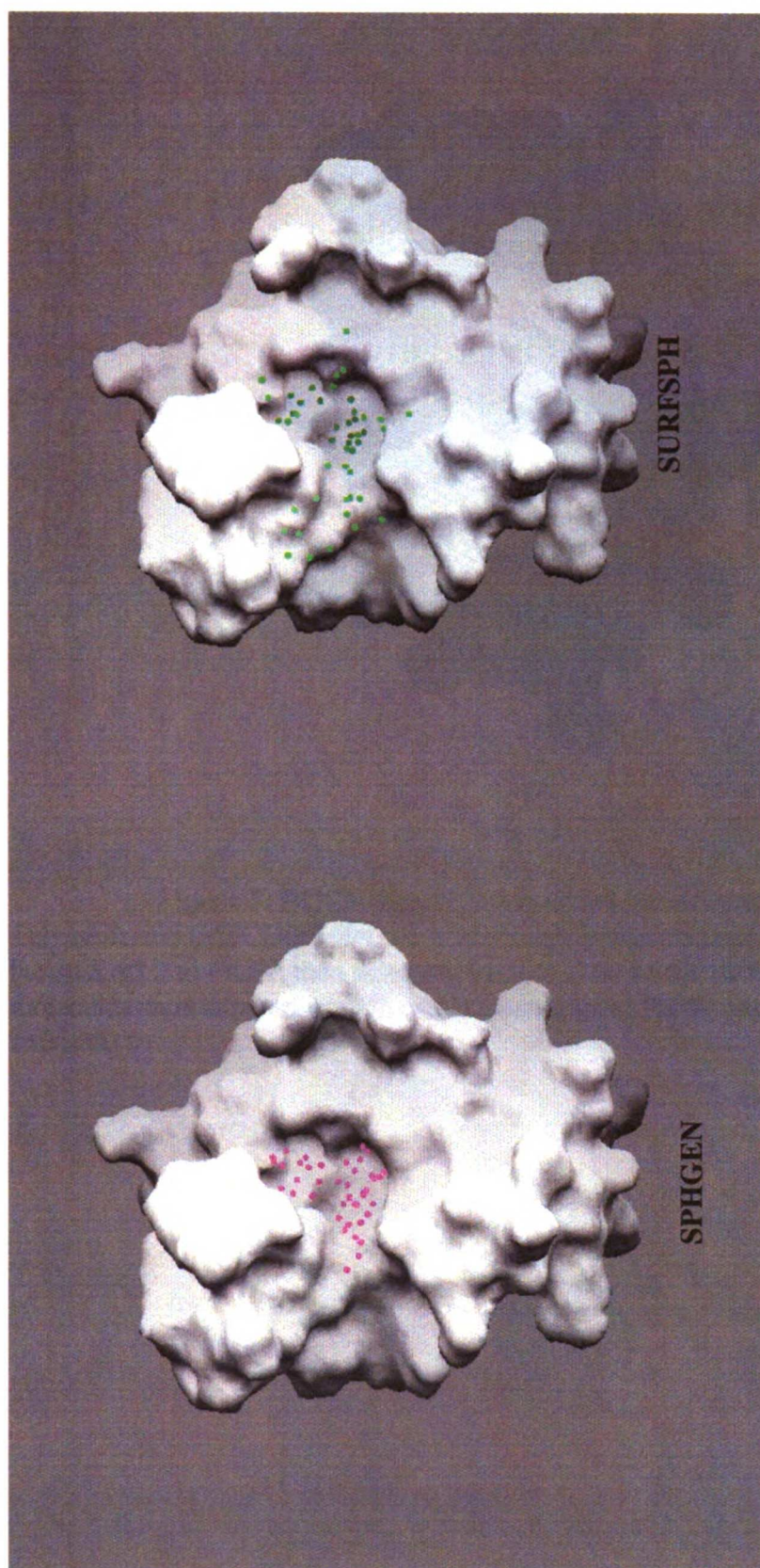
Sun Y, TJA Ewing, AG Skillman and ID Kuntz. 1998. CombiDOCK: structure-based combinatorial docking and library design. *Journal of Computer-aided Molecular Design* 12:597-604.

Zou X, Y Sun and ID Kuntz. 1999. Inclusion of solvation in ligand binding energy calculations using generalized-born model. Submitted to *Journal of the American Chemical Society*.



**Figure 1. FK506 binding protein.**

Spheres were generated with SPHGEN and 42 spheres were selected that were near the heavy atom sites of ascomycin in the Abbott NMR structure. A. The FK506 binding protein is shown as a molecular surface, and the sphere centers are shown in magenta. B. A closer view of the pocket shown in A. C. The pocket, viewed from the right of the point of view seen in A. Note that the pocket is very shallow and that some sphere are far from the molecular surface. Figures 1 and 2 were generated with GRASP (Nicholls, et al.).



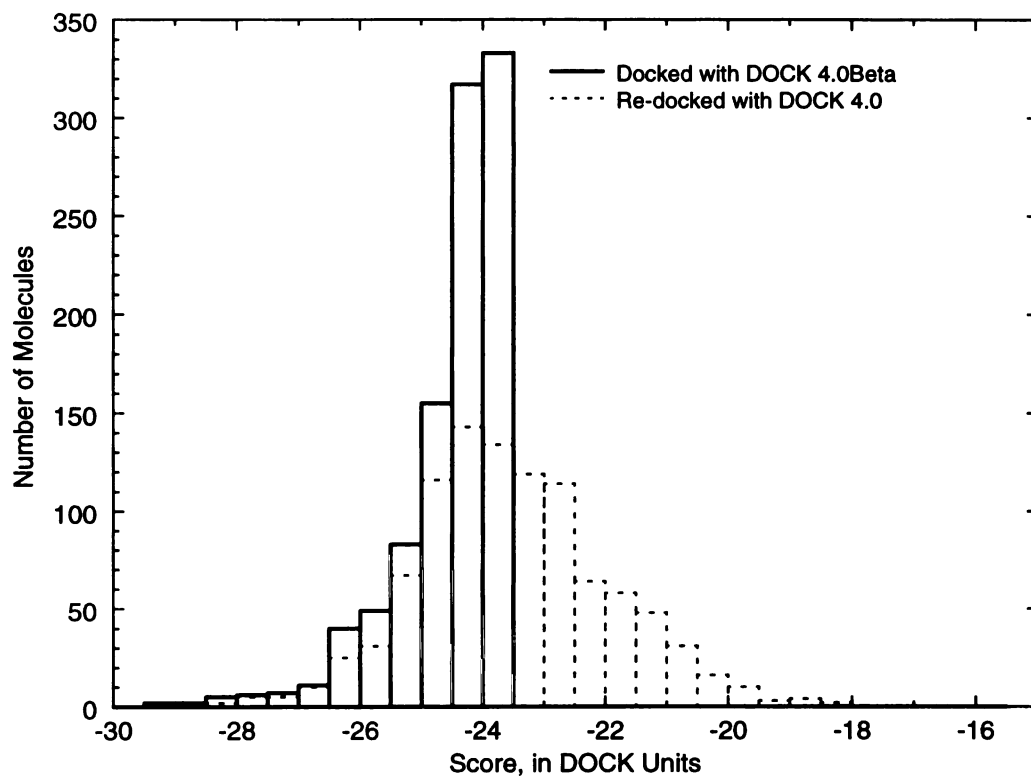
**Figure 2. SPHGEN spheres and SURFSPH spheres in FK506 binding site of FKBP.** On the left is FK506 binding protein with spheres centers from SPHGEN (shown in magenta). On the right is FK506 binding protein with spheres from SURFSPH, in green. Note that the SURFSPH spheres cover a larger portion of the pocket, because they can place site points on non-concave surfaces.



**Figure 3. DOCK ligand in predicted orientation in FKBP site.**

This molecule (MFCD005694, N1-(6-indazolyl)sulfanilamide) was selected by docking the ACD95.2 to FKBP, then selecting from clusters of the 1000 top-scoring molecules. 24 molecules were assayed, and this molecule bound to FKBP with a dissociation constant of 110  $\mu\text{M}$ .





**Figure 4. DOCK scores of top 1,000-scoring molecules from docking ACD95.2 to FKBP.**

A histogram of the original scores is shown by a solid line. A histogram of the same molecules re-docked with the 9/97 release of DOCK4.0 is shown by a dashed line.

**Figure 5. Determining the appropriate amount of sampling.**

(On the following page.) The top 1000 molecules from docking ACD95.2 to FKBP were re-docked with the uniform\_sampling parameter set to 100, 200, 500 and 1000 orientation. The upper plot shows the rank correlation, determined by taking the best score from all dockings for each molecule, then ranking the molecules by that score. The rank correlation is then and has the range (-1,1). This plot shows that the optimal amount of sampling

$$\frac{\sum_{i=1}^N (x_i - \bar{x}) - (y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

is between 200 and 500. The top scores for all molecules (a rank correlation value of 1) was not obtained with any amount of sampling. The lower plot is a histogram representation of the data shown in the upper plot. The x-axis is the DOCK energy score and the y-axis is the number of molecules. Note that the profile is the same for 500 orientations and 1000 orientations, suggesting, as seen in the upper plot, that the optimal sampling is between 200 and 500 orientations.

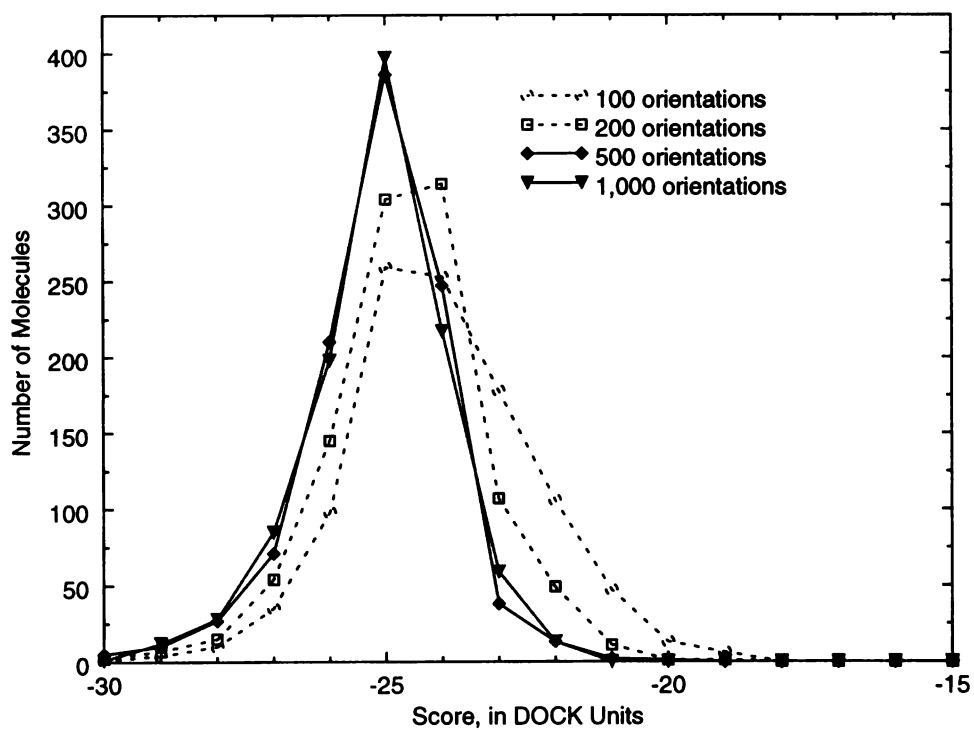
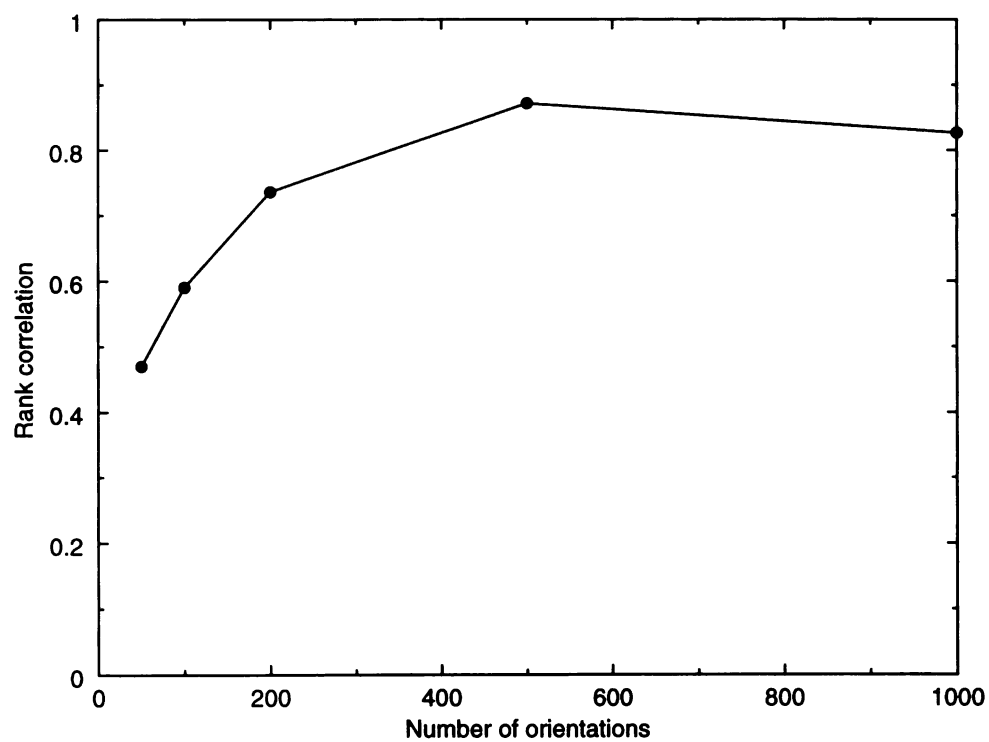
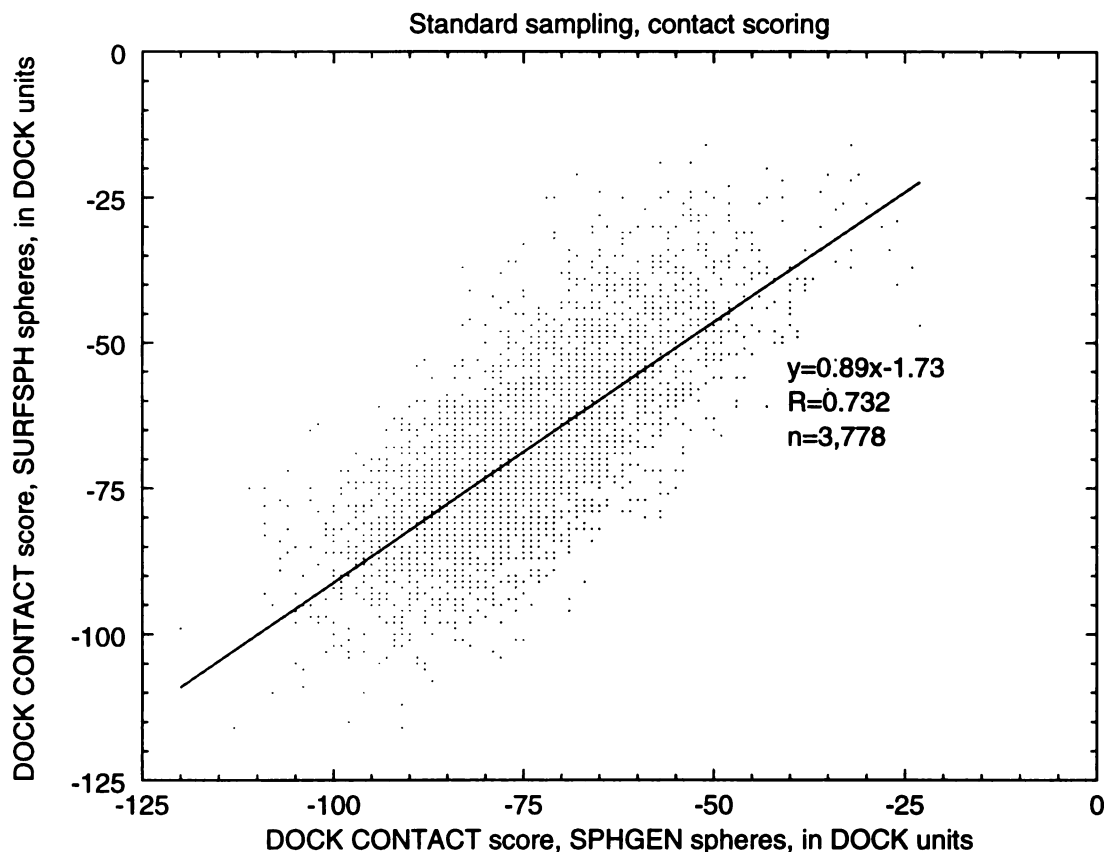


Figure 5

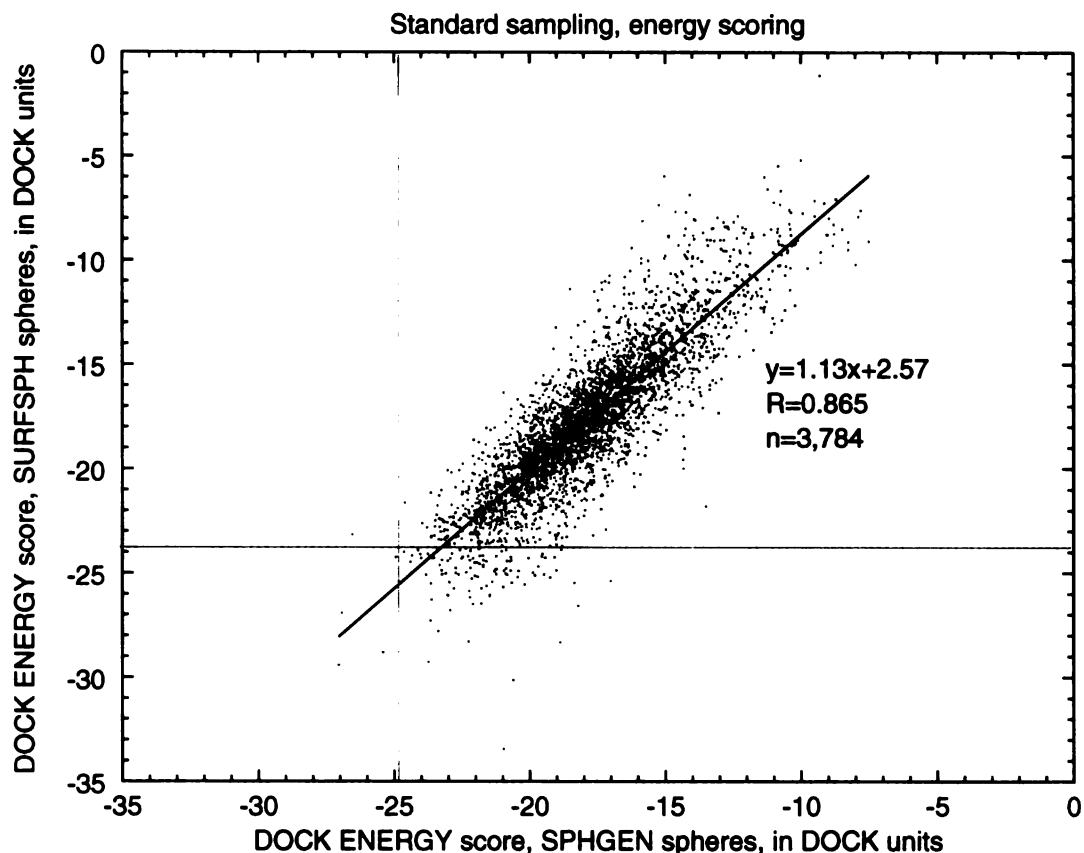
## SURFSPH vs. SPHGEN spheres



**Figure 6. Results from docking with SURFSPH spheres compared to results from SPHGEN spheres with contact score.**

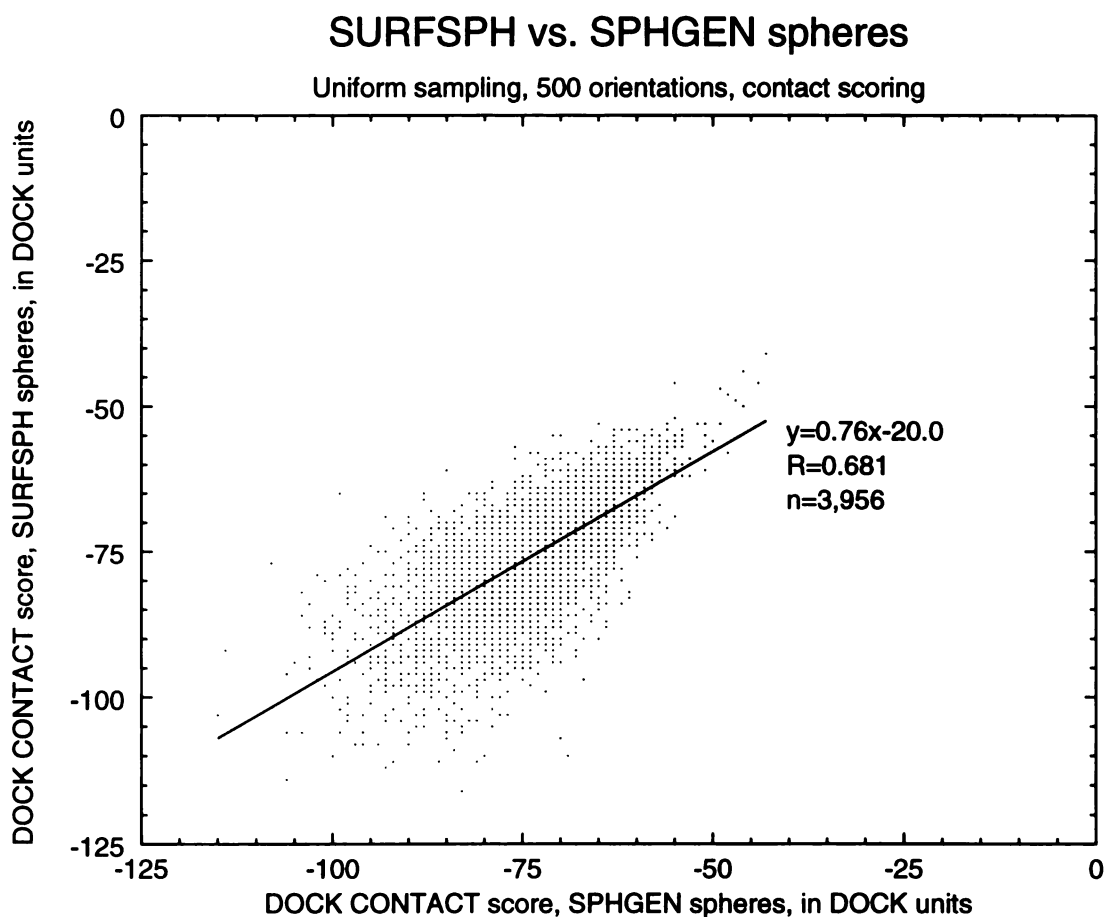
Each molecule was docked to the site using the default sampling parameters (distance\_minimum of 2.0Å and distance\_tolerance of 0.25Å). The y-axis is the best score obtained with SURFSPH spheres, and the x-axis is the best score obtained with SPHGEN spheres. While there is not a strong correlation between scores (with an  $R^2$  of 0.54), the SURFSPH spheres perform slightly better than the SPHGEN spheres finding orientations that have a more favorable contact score. Of the 3,974 molecules in the database, orientations with favorable contact scores were obtained with both site point methods for all but 196 of the molecules.

## SURFSPH vs. SPHGEN spheres



**Figure 7. Results from docking with SURFSPH spheres compared to results from SPHGEN spheres with energy score.**

Each molecule was docked to the site using the default sampling parameters (distance\_minimum of 2.0Å and distance\_tolerance of 0.25Å). The y-axis is the best score obtained with SURFSPH spheres, and the x-axis is the best score obtained with SPHGEN spheres. There is a stronger correlation between scores for energy score than for contact score (with an  $R^2$  of 0.75 versus 0.54 for contact score). The regression analysis suggests that SPHGEN spheres perform slightly better than SURFSPH spheres in finding orientations with a more favorable energy score. However, if we examine the number of orientations with scores more favorable than -25, we see more orientations from the SURFSPH sites (below the red line) than from the SPHGEN sites (to the left of the green line). Of the 3,974 molecules in the database, we were able to obtain orientations with favorable scores with both site point methods for all but 190 of the molecules.

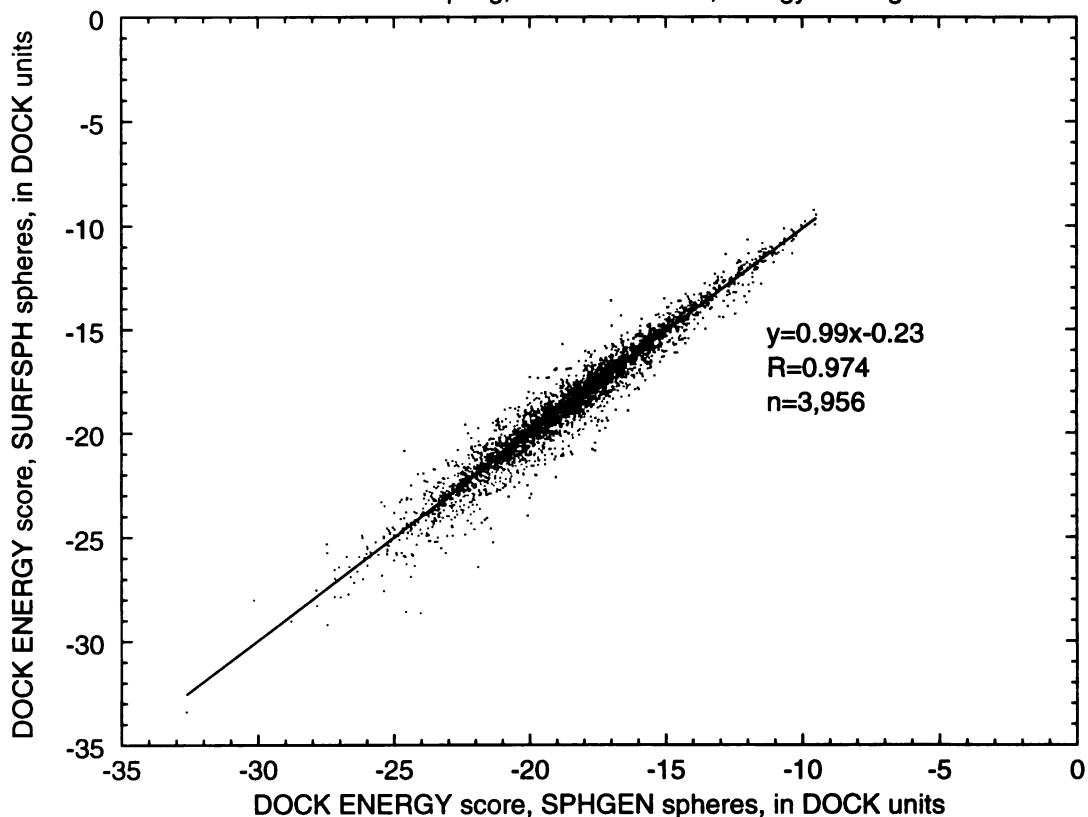


**Figure 8. Results from docking with SURFSPH spheres compared to results from SPHGEN spheres with contact score.**

Each molecule was docked to the site using the uniform sampling for 500 orientations. The y-axis is the best score obtained with SURFSPH spheres, and the x-axis is the best score obtained with SPHGEN spheres. The correlation between scores is poor, with an  $R^2$  of 0.55. SPHGEN spheres do not perform as well as SURFSPH spheres in finding orientations with a more favorable contact score. Of the 3,974 molecules in the database, orientations with favorable scores were obtained with both site point methods for all but 18 of the molecules.

## SURFSPH vs. SPHGEN spheres

Uniform sampling, 500 orientations, energy scoring



**Figure 9. Results from docking with SURFSPH spheres compared to results from SPHGEN spheres with energy score.**

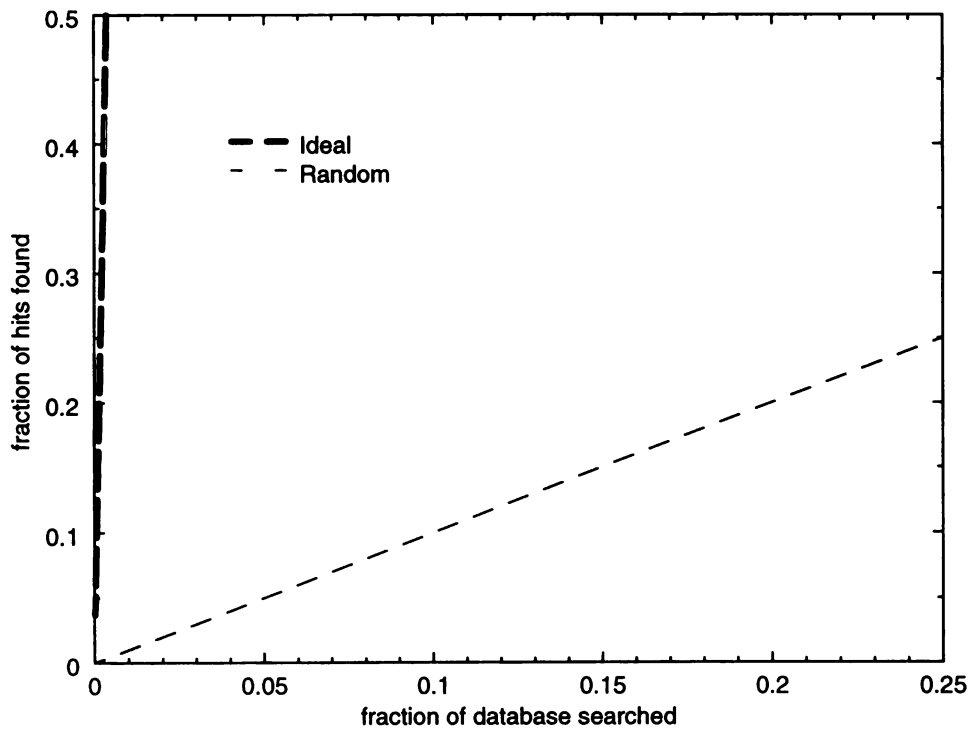
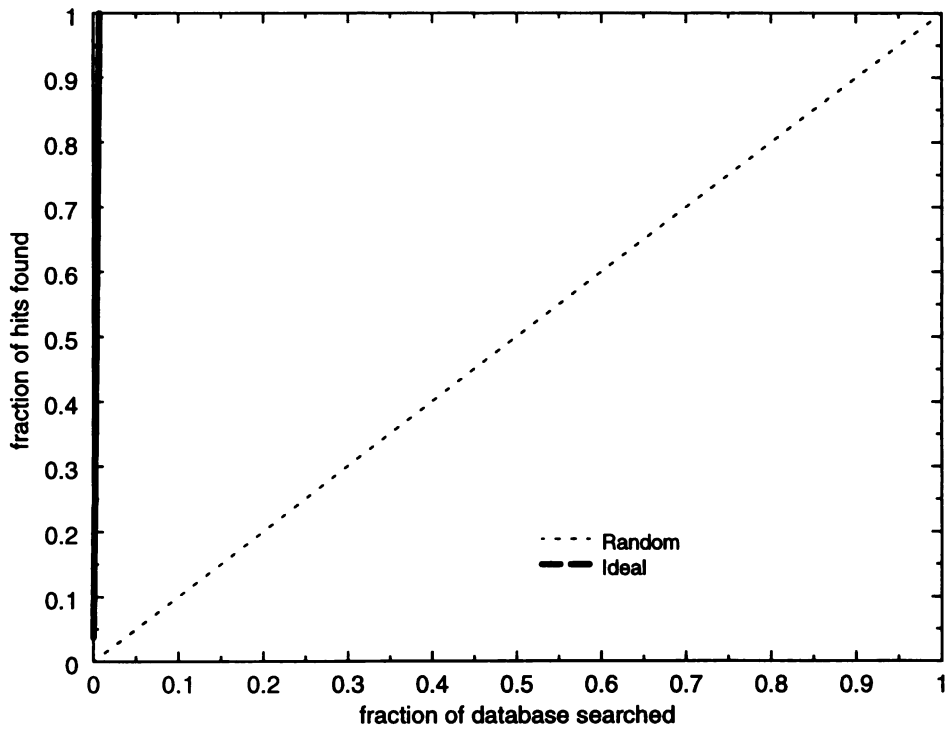
Each molecule was docked to the site using the uniform sampling for 500 orientations. The y-axis is the best score obtained with SURFSPH spheres, and the x-axis is the best score obtained with SPHGEN spheres. The scores are highly correlated, with an  $R^2$  of 0.95 and a slope of nearly 0.99. The resulting orientations from SPHGEN spheres and SURFSPH spheres are nearly identical in energy score. Of the 3,974 molecules in the database, orientations with favorable scores were obtained with both site point methods for all but 18 of the molecules.

**Figure 10. Ideal and random enrichment curves.**

(On following page.) The enrichment curve is a representation of enrichment given by DOCK when compared to random sampling of a database. Given a database of 1,000 molecules and 10 “hits”, in the ideal case (upper plot), DOCK will rank the 10 hits as its top-scoring molecules. On the x-axis is a ranking of the molecules, normalized from 0 to 1. On the y-axis is the number of hits found, normalized from 0 to 1. In our ideal case, the first 10 molecules are found in the first 0.1% of the database. On the lower plot is a representation of random enrichment, where hits will be ranked anywhere from best to worst.

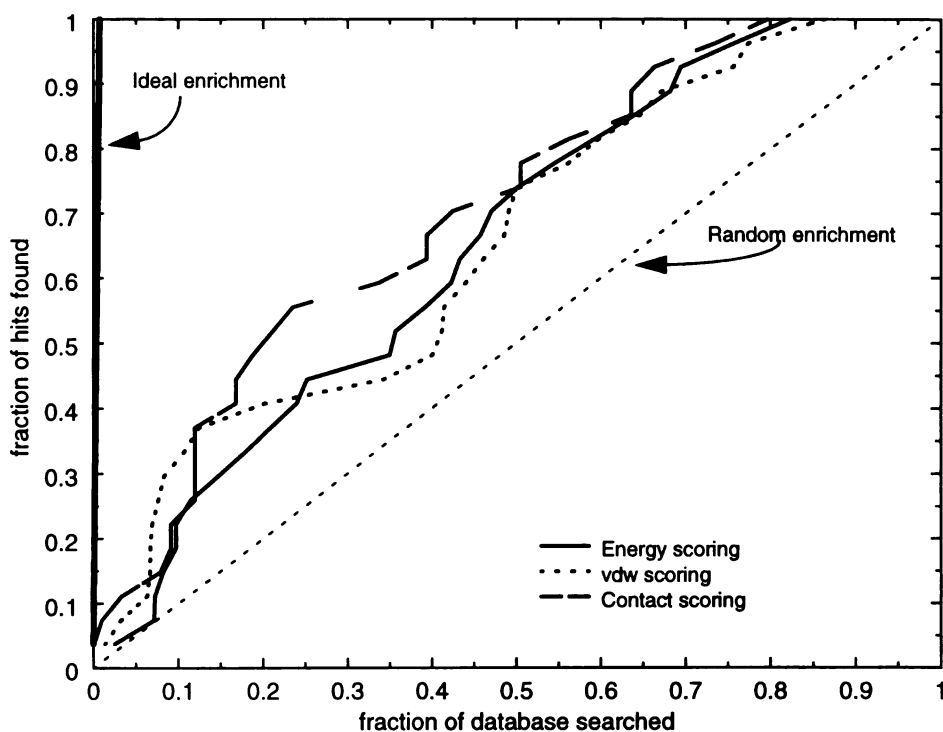


### Ideal vs. Random Enrichment



**Figure 10**

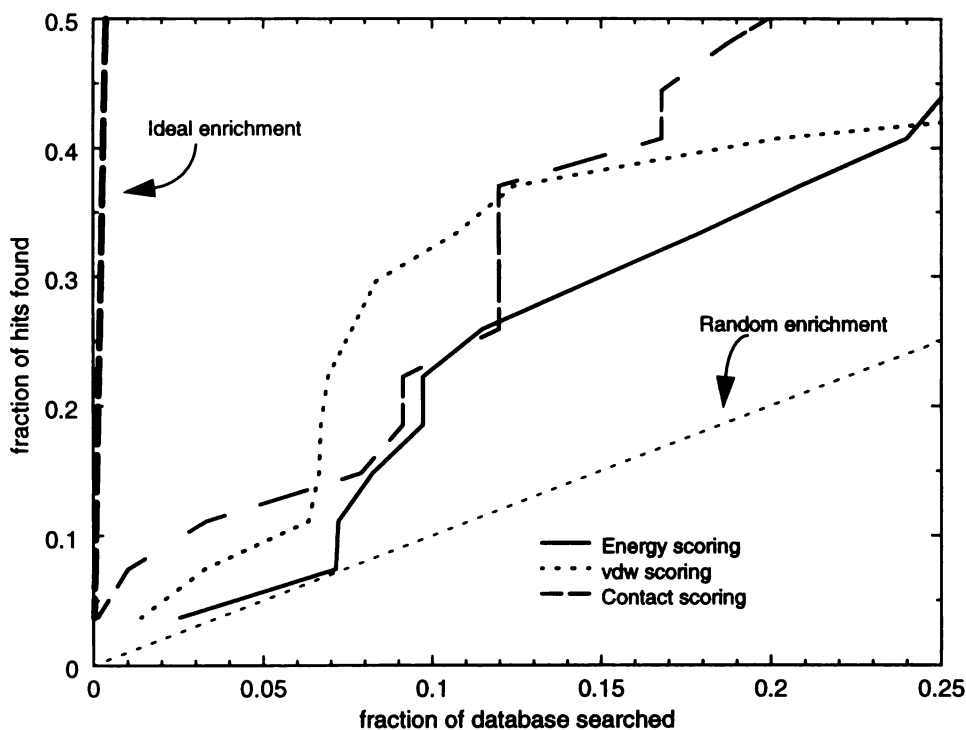
## Enrichment curve



**Figure 11. Enrichment curve.**

Three scoring methods are shown: energy scoring with a solid line, van der Waals portion of the energy score with a dotted line, and contact scoring with a dashed line. The dotted line from (0, 0) to (1, 1) represents random enrichment, and the bold, dashed line on the left side of the plot is ideal enrichment. The database includes 3,974 molecules. 27 of the molecules are hits, with dissociation constants less than 1mM. All scoring methods do better than random, but significantly poorer than ideal.

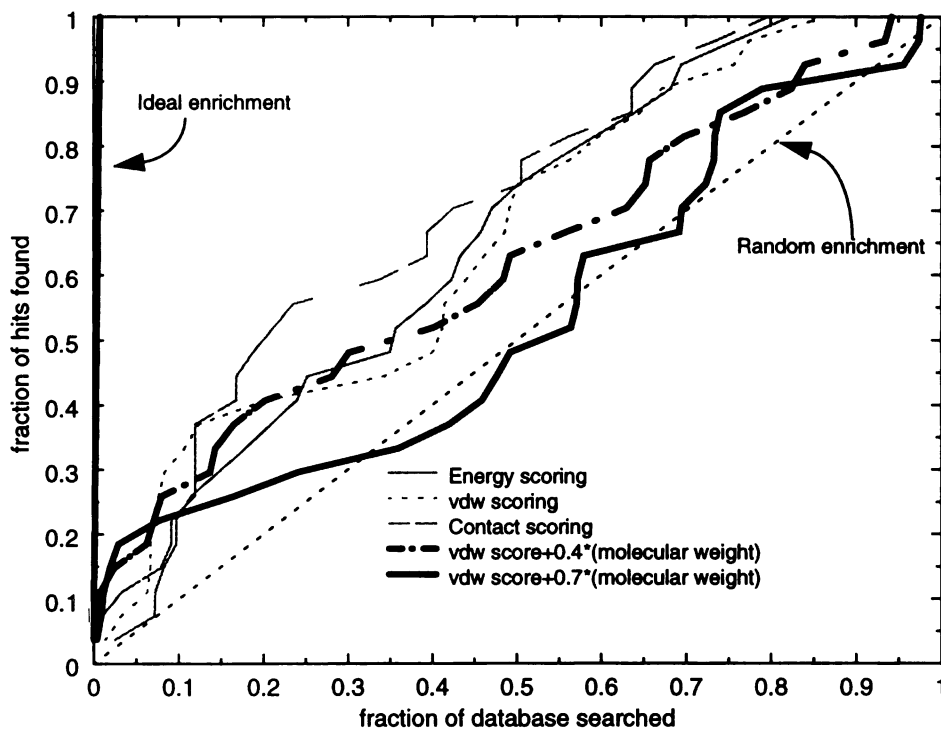
## Enrichment curve



**Figure 12. Enrichment curve, first 25% of the database.**

Shown is the first 25% of the database, or 994 molecules, of the database shown in figure 11. In a database search, these molecules might be selected for screening because they are ranked highest. The three scoring methods are shown as in figure 11. A dotted line represents random enrichment, and a bold, dashed line on the left side of the plot is ideal enrichment. Of the three scoring methods, contact scoring has the best results, ranking 16 of the 27 hits in the top 1,000. Energy scoring and vdw scoring do poorer, finding 12 and 11 of the hits, respectively. The database includes 3,974 molecules. 27 of the molecules are hits, with dissociation constants less than 1mM.

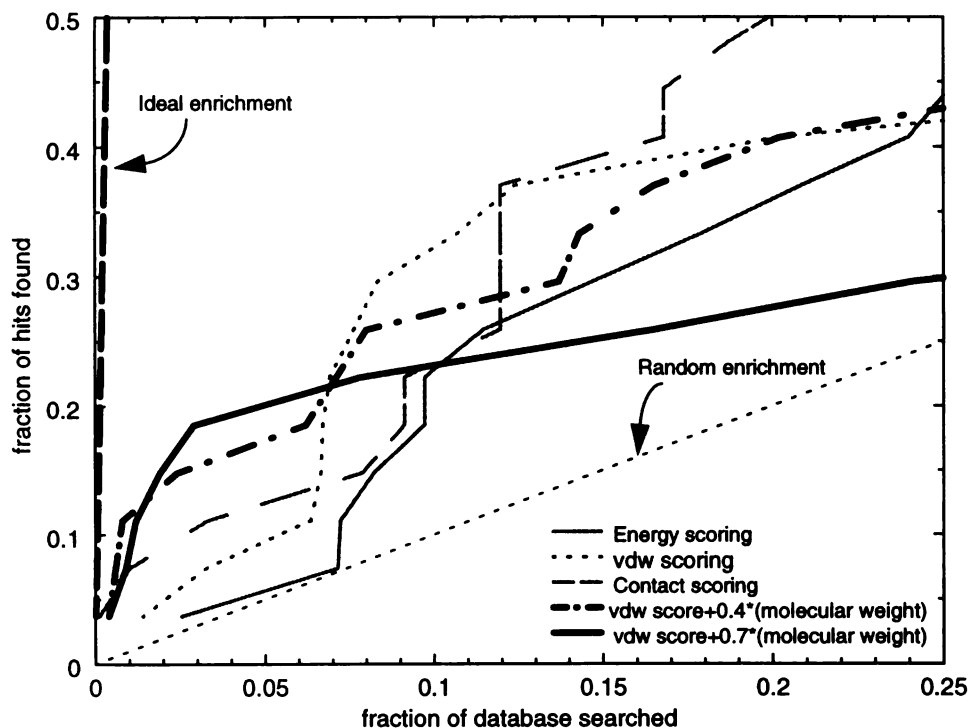
### Enrichment curve



**Figure 13. Weighting factors to correct bias.**

A weighting factor was added to the van der Waals component of the energy score to correct for DOCK's bias of larger molecules. Shown are two factors, 0.4 multiplied by the molecular weight and 0.7 multiplied by the molecular weight. Shown in gray are the enrichment curves for energy score, vdw score and contact score, as shown in figures 11 and 12. Both correction factors find more of the hits in the top-ranked molecules than the non-weighted scores. The van der Waals score corrected by 0.7 times the molecular weight appears to find the most hits in the first 10% of the database, but finds fewer hits later in the search, occasionally doing more poorly than a random screen.

## Enrichment curve



**Figure 14. Weighting factors to correct bias, first 25% of the database.**

Scoring methods and scores with weighting factors are represented as in figure 13. Both weighting factors improve results in the top 5% of the database, finding 4 hits (with a weighting factor of 0.4) and 5 hits (with 0.7) in the first 199 molecules. However, if the top 10% of the database is assayed, neither correction factor provides improvement over vdw scoring, alone.

## **Chapter 6**



# **EXPLORATION OF MACROMOLECULAR INTERACTIONS: DEVELOPMENT AND IMPLEMENTATION OF DESCRIPTORS FOR MACROMOLECULAR DOCKING**

## **EPILOGUE**



Receptors are compelling targets for docking. They are inherently difficult to model because they tend to be solvent-exposed, particularly cell surface receptors. Frequently receptors bind to macromolecules, so binding sites are large and have regions that are convex or flat, as well as regions that are concave. In addition to the receptor binding sites, receptor molecules bound to their target are possible targets for docking.

This dissertation is the development of techniques for macromolecular docking, the use of those techniques to explore the interactions of a complex macromolecular system, and the application of existing docking methods to model two sites -- a site at a macromolecular interface and a site with characteristics common to macromolecular interfaces. These methods have been developed for the purpose of docking to receptor molecules and exploring the interactions of receptor molecules.

In chapters 2 and 3, I demonstrated that shape-based site points can be used to describe the surface shape of macromolecules, and to dock macromolecules. The advantage offered by these site descriptors is a reduction in the CPU time required for docking large systems. I also used these site points to explore a three-body system, human growth hormone and its receptor. I found that I could recapitulate experimental findings, in that I could dock the hormone to the first receptor molecule in the absence of the second receptor molecule, but I could not dock the hormone easily to the second receptor molecule without the contributions of the first receptor. Additionally, I found that it is possible to dock the human growth hormone complex with a subset of sites, called the functional epitope. Mutational analysis has identified the functional epitope as the source of most of the binding energy between hormone and receptor. Finally, I found that the receptor mole-

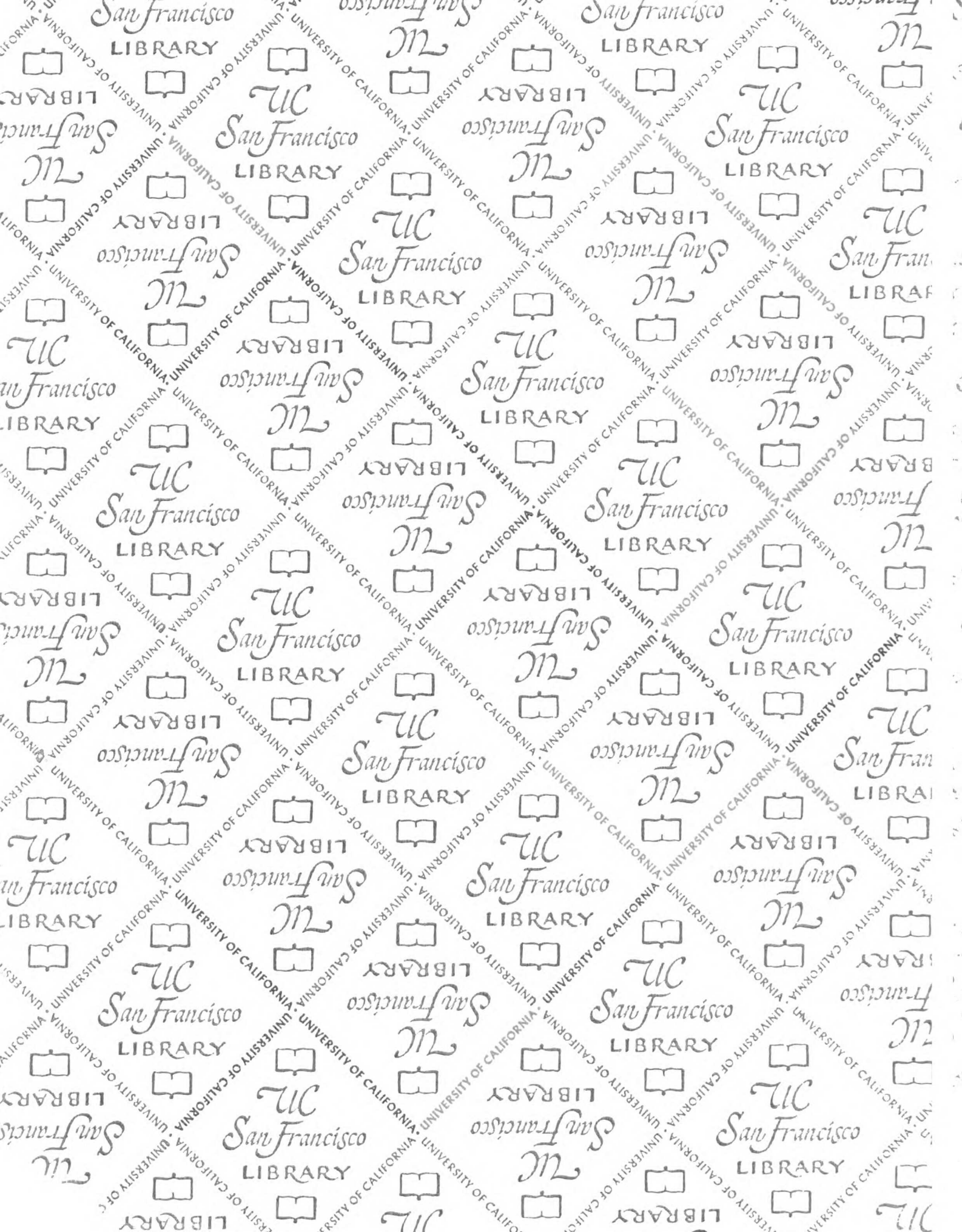
cules prefer an alternate orientation in the absence of hormone. This orientation was 19Å RMSd from the native complex, and was similar to one found in a recent structure of erythropoietin receptor bound to an antagonist, and may represent an “off” state of the receptor.

In chapter 4, I used DOCK to find small-molecule inhibitors of HIV-1 integrase. I selected a site that was created by the formation of the dimer. The role of the site was not known, but the enzyme was a dimer or higher-order oligomer in its active form, and a molecule bound to this site may disrupt or hinder the formation of oligomers. A class of small-molecule inhibitors with inhibition constants ( $IC_{50}$ ) in the low micromolar range was discovered, including one molecule with an  $IC_{50}$  value of 300 nanomolar. I also reported on molecules that were synthesized as possible HIV-1 reverse transcriptase inhibitors, but were assayed to HIV-1 integrase because of the similarity to molecules discovered when docking to HIV-1 integrase. Several of these molecules had  $IC_{50}$  values in the nanomolar range.

In chapter 5, I explored the use of DOCK to screen and to rank a database of small molecules to the FK506 binding protein (FKBP). The binding site for FK506 on FKBP had many of the characteristics of a macromolecular binding site, in that it was a shallow bowl and solvent-exposed in the absence of FK506. I discovered a new millimolar-binding molecule from a database of over 150,000 molecules, from which only 24 were assayed. I also explored the use of adjusting the DOCK score to counteract the inherent bias of DOCK’s scoring function toward larger molecules, and found that these adjustments improved DOCK’s ability to select well-binding molecules.



I have demonstrated that it is possible to dock to receptor surfaces, both with small- and macro-molecules. Also, I have discovered orientations of complexes of macro-molecules that were not found in crystal structures, but may represent alternative states of the complex, as shown with human growth hormone receptor. I have also shown that it is possible to discover an inhibitor of an enzyme by docking to a site that may affect oligomerization of the macromolecule. Finally, I have shown that it is possible to offset the preferences of DOCK's scoring functions, which are particularly apparent in a solvent-exposed site, and improve results.



# For reference

Not to be taken  
from the room.

7065272



3 1378 00706 5272

