Title
Development and application of de-novo structure based design algorithms

Author
De Jesus Haresco, Jose Teodorico

Publication Date
2002

Peer reviewed|Thesis/dissertation

Development and Application of De-Novo Structure Based Design Algorithms

by

Jose Teodorico De Jesus Haresco III

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

For Gina, Julian, and Eaden

Acknowledgements

As young graduate students, we might imagine ourselves on the fringes of science, skillfully meandering the changing landscape of discovery, innovation, and invention. We might further envision that at the end of our short time in academia we will know all there is to know in our fields, confident that all but the most jagged of intellectual precipices will crumble beneath our vision.

Nothing could be further from the truth.

After four years at one of the top-ranked research institutions, having worked alongside the most experienced, brightest, most promising, and most gifted scientists in the world I am perhaps more in awe of scientific discovery, its application, and those who dedicate their lives to uncovering the unknown. This thesis is the culmination of my evolution from a mere spectator of science to one who contributes to its future. While it is a singular drop in our knowledge, I hope that it proffers some insight into the development and application of computational chemistry algorithms.

I have had the privilege to spend the last several years being mentored by the most perceptive, intellectually astute, caring, and concerned mentor I could ever have hoped for. Working with Tack Kuntz has been one the highlights of my life for a number of reasons. He never failed to push my intellectual limits, pique my curiosity, guide my efforts, or provide general counsel when I needed it most. He gave me the freedom and the license to pursue my more non-academic interests in business, and supported me every step of the way. He took interest in, and showed incredible amounts of concern for, my family. He saw me as I am, a person with many facets and interests, and did not attempt to change that. Rather, he simply urged me to bin my interests appropriately.

Simple words on this page cannot communicate how grateful I am for his tutelage and friendship.

Similarly, I have to thank R. Kip Guy and Fred Cohen for serving on my thesis committee and making my research much more difficult, but inherently much more interesting. This thesis is also dedicated in part to the memory of Peter Kollman, whose passing only one month after he chaired my oral examination consistently reminds me that every moment in life should be spent learning, contributing, and enjoying the company of those closest you.

My cohorts in graduate school deserve a very special thank-you. In particular, Sandy Waugh, Rey Banatao, and Sean Mooney, who collectively formed the other members of the Quicksilver Genomics team and helped spur my imagination and dreams to greater heights. In addition, Scott Pegg, Demetri Moustakas, Natasja Brooijmans, Geoff Skillman, Jim Arnold, Liping Zhang, and Chern Singh Goh have all played crucial roles in my graduate school education. Without a doubt, the ability to commiserate is absolutely critical to the completion of a graduate degree.

I have to additionally thank the odd collection of lawyers, scientists, CEO's, and venture capitalists that took me under their collective wings and gave me an outlet for all my interests and creativity, wrote absurdly glowing letters of recommendations for business school, opened their rolodexes, and not to mention made it easier to support a family of four on a graduate student stipend. Rich Vandenbroek (Cooper Hill Partners), Camille Samuels-Pearson (Versant Ventures), Buzz Burlock (Origin Capital), Jeanne Cunicelli (Bay City Capital), J. Leighton Reed (Aviron and Alloy Ventures), James Sabry (Cytokinetics), Joel Kirschbaum (UCSF OTM), Chris Scott, Eric Greenberg (Acumen

Simple words on this page cannot communicate how grateful I am for his tutelage and friendship.

Similarly, I have to thank R. Kip Guy and Fred Cohen for serving on my thesis committee and making my research much more difficult, but inherently much more interesting. This thesis is also dedicated in part to the memory of Peter Kollman, whose passing only one month after he chaired my oral examination consistently reminds me that every moment in life should be spent learning, contributing, and enjoying the company of those closest you.

My cohorts in graduate school deserve a very special thank-you. In particular, Sandy Waugh, Rey Banatao, and Sean Mooney, who collectively formed the other members of the Quicksilver Genomics team and helped spur my imagination and dreams to greater heights. In addition, Scott Pegg, Demetri Moustakas, Natasja Brooijmans, Geoff Skillman, Jim Arnold, Liping Zhang, and Chern Singh Goh have all played crucial roles in my graduate school education. Without a doubt, the ability to commiserate is absolutely critical to the completion of a graduate degree.

I have to additionally thank the odd collection of lawyers, scientists, CEO's, and venture capitalists that took me under their collective wings and gave me an outlet for all my interests and creativity, wrote absurdly glowing letters of recommendations for business school, opened their rolodexes, and not to mention made it easier to support a family of four on a graduate student stipend. Rich Vandenbroek (Cooper Hill Partners), Camille Samuels-Pearson (Versant Ventures), Buzz Burlock (Origin Capital), Jeanne Cunicelli (Bay City Capital), J. Leighton Reed (Aviron and Alloy Ventures), James Sabry (Cytokinetics), Joel Kirschbaum (UCSF OTM), Chris Scott, Eric Greenberg (Acumen

Sciences), and John Hefti (Signature Biosciences and Prometheus): You all validated, welcomed, and fueled my desire to walk the razor sharp line between science and its application.

I would also like to thank my parents and in-laws, who despite admitting that they had no idea what I was doing here at UCSF consistently supported our family's ongoing journey through graduate school in every conceivable way Mom, I wouldn't be here if you hadn't worked so hard to bring me to this country. Dad, none of this would have been possible without your constant support. We only wish you were geographically closer.

Julian and Eaden. You two boys make every hard day, every worry, and stressful thought disappear in the face of Thomas the Tank Engine, drool, and incessantly random outburst of song. I've learned a great deal from seeing the world through your eyes. Your smiles, hugs, and kisses are worth the world's weight in gold-pressed latinum. Nothing material that life can bring will ever compare to the joy you bring me.

Lastly, my most sincere, heartfelt, and loving thanks go to my wife, friend, and partner, Gina. You put up with all the work, financial uncertainty, an often stressed out husband, and heaven knows what else. I cannot have done this without you. Despite all the science that has entered my head over the last four years, the most valuable lesson I have learned has come from you. You have taught me that every moment with those you love must be spent in loving them, and that there must never be any doubt in their minds that your love for them comes before anything else in this life. Life, in all its wonders, is far too short. Spend it in love. We did it, honey. *We finally did it.*

Abstract

Computational chemistry is a mainstay of modern drug discovery techniques. Based largely on the use of 3-dimensional databases, *in-silico* screening and optimization exist alongside traditional medicinal and organic chemistry methods. This thesis describes the development and application of *de-novo* computational methods for use in computational chemistry.

It begins in chapter I by addressing a number of common limitations in database mediated screening methods by describing the development and validation of a *de-novo* structure based design algorithm known as ADAPT. ADAPT attempts to circumvent problems of chemical diversity and library size by using rigid chemical fragments to build populations of whole molecules that evolve towards high binding compounds. We showed that ADAPT was capable of regenerating known chemical scaffolds and generating novel scaffolds which could be further explored. In chapter II, I further develop the idea of de-novo design through the structure-based design, optimization, and in-vitro evaluation of the first small molecule mimetic of a β-peptide inhibitor of PDZ domains. The scope of the project, spanning early design and conception to in-vitro evaluation, served to validate the computational chemistry techniques and methods developed and used in our laboratory. Lastly, chapter III details early work focused on improving the current paradigms for protein-protein docking. The algorithm and results described therein illustrate how a reduction in the dimensionality of the search problem by docking in two dimensions instead of three enables us to achieve results attained by traditional 3-dimensional approaches in significantly shorter amounts of time.

Table of Contents

List of Figures

Chapter I

Chapter II

Chapter III

Introduction

In the two decades that structure-based drug design has existed, it has evolved from arcane academic theory into a mainstay of pharmaceutical drug development. Researchers rely on computational chemistry to routinely answer questions relevant to a number of fields including chemistry, biology, bioengineering, and nanotechnology. Nevertheless there remains a need for the continued research and development of new algorithms and tools that enhance our understanding and deepen our analysis of both simple and complex molecular systems. Constant innovation and discovery in the field enables us to apply conventional approaches beyond the confines of their original design objectives. The body of this work builds on the existing work in by Kuntz and co-workers by extending and hypothesizing upon the functionality of the DOCK algorithm beyond the common paradigm of database screening.

Molecular docking, as first implemented by Kuntz and co-workers in the early 1980's, led to our ability screen 3-dimensional databases for putative small molecule inhibitors of a given target [1-7]. The approach was fundamentally based on measuring the geometric and energetic components of a two-body interaction. That simple approach has since been extended to enable compound optimization, computational toxicology, and a number of other applications. Chapter I describes another extension of the DOCK paradigm into a scoring function that drives *de-novo* molecular design in a program known as ADAPT. Our goal was to develop a method capable of developing new ligands using information from the three dimensional (3D) structure of a protein target without the prior knowledge of other ligands. Current computational *de-novo* methods use the 3D structural information of the target to narrow the search for interesting ligands. One

strategy is to join compound fragments, which fit favorably into the active site[8,9] either by constructing linking atoms or by searching a database of possible linking structures.[10-15] This strategy often leads to many novel and potentially useful compounds that are very dependent on the placement and selection of the fragments. [16] Another strategy is to construct a molecule within a binding site by 'growing' it from a seed atom or fragment, using the site's steric and/or energetic properties to guide the growth of the compound. [17-21] Molecules built in this manner are often sensitive to the position and choice of the seed atom used and are difficult to synthesize. [16,22] Scott C-H Pegg and I designed the ADAPT program to circumvent the limitations of conventional 3D database searching by allowing the program to build populations of small molecules from rigid chemical fragments and using a genetic algorithm to evolve these populations towards high-scoring compounds. DOCK forms the basis of a genetic fitness function that weeds out low scoring members of the population. The ADAPT approach does not rely on pre-existing knowledge of known compounds against a target. Instead, it uses a well-validated method (molecular docking) to guide the design process. The program proved itself capable of generating strong, though not optimal, chemical scaffolds against three known targets.

Once an initial scaffold is identified, molecular docking can continue to play a central role in optimization of that scaffold. Structure based design played an important role in the design of inhibitors against glyceraldehyde-3-phosphate dehydrogenase[23], HIV-1 protease, kinesin[24], and HIV-1 fusion protein[25]. This design-optimization paradigm is illustrated in the second chapter, which describes the *de-novo* structure based design, optimization, and in-vitro validation of the only known small molecule β-strand mimetic inhibitor of PDZ domains. PDZ domains were originally described as conserved

structural elements in the 95 kDa post-synaptic density protein, PSD-95; the Drosophila tumor suppressor, Discs-large Dlg; and the tight junction protein, zonula occludens-1 ZO-1. The domains are small, usually of 80 amino acids, and often present multiply in adaptor proteins and with other protein interaction domains. They have emerged as important modular protein interaction domains in a wide variety of eukaryotic organisms. In general, PDZ domains mediate subcellular organizing of protein complexes bringing proteins with various functions to one locus. They have been found to be associated with ion channels, transmembrane receptors, structural proteins, and regulatory enzymes.[28-31] The central role of PDZ domains in mediating cell signaling and cell-cell interaction is emerging in diverse fields from neuroscience to protein trafficking to cancer. Because of the structural similarity between several reported structures of PDZ domains, both with and without bound ligand, we considered it feasible to design novel ligands to MAGI-3 PDZ2 domain by working with the structure of PSD-95 PDZ3 domain bound to KQTSV.[32] An ideal small molecule mimetic of the peptide ligand must be able adopt a conformation similar to that of the native ligand and present critical side chains to the domain in similar orientations. Molecules designed using structure as a starting point for the rational design process have an advantage in that they are inherently directed at the biologically relevant forms of their target, and are capable of successfully binding to the desired targets. Bolin and co-workers used a similar approach in the rational design of inhibitors of inhibitors of antigen presentation by HLA-DR class II MHC molecules.[31] Furet et al. used extensive crystallographic data to design a high-affinity antagonist of the Grb2-SH2 domain.[32] A collaborative project between the Kuntz and Guy laboratories, this project spans the range of activities from early scaffold design to the in-vitro

measurement of the compound's activity against the desired target, and confirms the utility of structure-based design protocols.

Scientists have also extended the same principles used in docking small molecules to target receptors to docking proteins against other proteins. This problem poses both a theoretically and computationally more difficult problem because the scale of the objects involved, the increase in the degrees of freedom in the computation, and the dynamic complexity of the protein-protein binding phenomenon. Some protein-protein docking methods evaluate the complementarity of molecular surfaces to determine how well two proteins fit together. Other approaches focus on matching surfaces[33-37], and some enhance the search for geometric fit by matching the positions of surface normals and spheres[38-42]. Some shape algorithms model the hydrophobic effect during association from the change in the solvent-accessible surface area of the molecules[43], while others employ a simplified scheme to estimate electrostatic, hydrophobic, or desolvation contributions to dimer formation[44]. Most of these methods have focused on rigid body docking, and in general achieve respectable success (typically 1-3 A RMS). Several groups have additionally focused on finding the correct orientation among a large number of false positives[44]. Recently, the use of soft-body docking and side-chain enumeration has yielded greater sensitivity to the protein-docking results[44]. It should be noted however, that the use of non-complexed models instead of co-crystallized complexes to dock partners to each other does not necessarily generate similar results in terms of time and accuracy. One last class of approaches transforms the problem from three dimensions into two. Sternberg and coworkers extended their Fourier transform-based approach from small molecule docking to protein-protein docking[45]. Several groups have extended this

method to show that a reduction in dimension yields high sensitivity (< 2 A RMS in some cases) while increasing the speed of the analysis [46]. In chapter III this thesis expands upon this area by investigating the use of a 2-dimensional method for protein-protein docking. The method requires the use of a projection algorithm that transfers 3D data such as the solvent accessible molecular surface, the surface electrostatic and van der waals potentials, and any other relevant chemical information into a 2-dimensional map. The process is reminiscent of projection algorithms that result in topographical maps. The central hypothesis underlying this particular project is that the reduction in the computational degrees of freedom enables a much faster analyses of putative protein-protein interactions, without a reduction in sensitivity or specificity set by the most current 3-dimensional docking methods. The approach is evaluated in one well studied system as a proof-of-concept pilot study.

Science advances as function of the amount and complexity of our knowledge of the world around us. I have always believed, however, that the practical application of our knowledge leads to additional advances for humanity at large. To the layman, our ability to design better therapeutics, treat diseases, and ease pain are more tangible manifestations of our advancements. This thesis is built upon that belief by focusing on the extension and application of molecular docking.

References

1. Walters, P. W., Stahl, M., Murcko, M. A. *Drug Des Today*, 3 **(1998)** 160.
2. Goodford, P. *J. Med. Chem.*, 28 **(1985)** 849.
3. Miranker, A., Karplus, M. *Proteins*, 11 **(1991)** 29.
4. Eisen, M. B., Wiley, D. C., Karplus, M., Hubbard, R. E. *Proteins*, 19 **(1994)** 199.
5. Miranker, A., Karplus, M. *Proteins*, 23 **(1995)** 472.
6. Pearlman, D. A., Murcko, M. A. *J. Me. Chem*, 39 **(1996)** 1651.
7. Bohm, H.-J. J. *Comput.-Aided Mol. Design*, 6 **(1992)** 61.
8. Lauri, G., Bartlett, P. J. *Comput.-Aided Mol. Design*, 8 **(1994)** 51.
9. Roe, D. C., Kuntz, I. D. J. *Comput.-Aided Mol. Design*, 9 **(1995)** 269.
10. Lewis, R. A., Leach, A. R. J. *Comput.-Aided Mol. Design*, 8 **(1994)** 467.
11. DeWitte, R., Shakhnovich, E. *J. Am Chem Soc*, 118 **(1996)** 11733.
12. Bohacek, R. S., McMartin, C. *J. Am Chem Soc*, 116 **(1994)** 5560.
13. Rotstein, S. H., Murcko, M. A. J. *Comput.-Aided Mol. Design*, 7 **(1993)** 23.
14. Rotstein, S. H., Murcko, M. A. *J. Med Chem*, 36 **(1993)** 1700.
15. Moon, J. B., Howe, W. J. *Proteins*, 11 **(1991)** 314.
16. Joseph-McCarthy, D. *Pharma Therapeutics.*, 84 **(1999)** 179.
17. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J Mol Biol* (161) **1982**, 269-288.
18. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D.; Blaney, J. M.; Jorgensen, E. C.; Connolly, M. L.; Ferrin, T. E.; Langridge, R.; Oatley, S. J.; Burridge, J. M.; Blake, C. C. *J Comput Aided Mol Des* **2001**, *15*, 411-428.

19. Zou, X.; Sun, Y.; Kuntz, I. D. *Journal of the American Chemical Society* **1999**, *121*, 8033-8043.

20. Connolly, M. L.; Skillman, A. G.; Kuntz, I. D.; Blaney, J. M.; Jorgensen, E. C.; Ferrin, T. E.; Langridge, R.; Oatley, S. J.; Burridge, J. M.; Blake, C. C. *Science* **1983**, *221*, 709-713.

21. Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. *J. Mol. Graphics* **1988**, *6*, 13-27, 36-17.

22. DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R.; Blaney, J. M.; Jorgensen, E. C.; Connolly, M. L.; Ferrin, T. E.; Langridge, R.; Oatley, S. J.; Burridge, J. M.; Blake, C. C. *J Med Chem* **1988**, *31*, 722-729.

23. Ewing, T. J. A., Kuntz, I. D. *J. Comp. Chem.* **1997**, 18 1175.

24. Aronov, A. M.; Munagala, N. R.; Kuntz, I. D.; Wang, C. C. *Antimicrobial Agents and Chemotherapy* **2001**, *45*, 2571-2576.

25. Hopkins, S. C.; Vale, R. D.; Kuntz, I. D. *Biochemistry* **2000**, *39*, 2805-2814.

26. Bewley, C.A., Lousi, J.M., Ghirlando, R., Clore, G.M., *J Biol Chem.* **2002** Apr 19;277(16):14238-45.

27. Hendrix, D.K., Klein, T.E., Kuntz, I.D. *Protein Sci.* **1999** May;8(5):1010-22.

28. Itoh, M.; Sasaki, H.; Furuse, M.; Ozaki, H.; Kita, T.; Tsukita, S. *J Cell Biol* **2001**, *154*, 491-497.

29. Kuwahara, H.; Araki, N.; Makino, K.; Masuko, N.; Honda, S.; Kaibuchi, K.; Fukunaga, K.; Miyamoto, E.; Ogawa, M.; Saya, H. *J Biol Chem* **1999**, *274*, 32204-32214.

30. Liu, T. F.; Kandala, G.; Setaluri, V. *J Biol Chem* **2001**, *276*, 35768-35777.

31. Bolin, D. R.; Swain, A. L.; Sarabu, R.; Berthel, S. J.; Gillespie, P.; Huby, N. J.; Makofske, R.; Orzechowski, L.; Perrotta, A.; Toth, K.; Cooper, J. P.; Jiang, N.; Falcioni, F.; Campbell, R.; Cox, D.; Gaizband, D.; Belunis, C. J.; Vidovic, D.; Ito, K.; Crowther, R.; Kammlott, U.; Zhang, X.; Palermo, R.; Weber, D.; Guenot, J.; Nagy, Z.; Olson, G. L. *J Med Chem* **2000**, *43*, 2135-2148.

32. Furet, P.; Garcia-Echeverria, C.; Gay, B.; Schoepfer, J.; Zeller, M.; Rahuel, J.; Campbell, R.; Cox, D.; Gaizband, D.; Belunis, C. J.; Vidovic, D.; Ito, K.; Crowther, R.; Kammlott, U.; Zhang, X.; Palermo, R.; Weber, D.; Guenot, J.; Nagy, Z.; Olson, G. L. *J Med Chem* **1999**, *42*, 2358-2363.

33. Gabb, H.A., Jackson, R.M., Sternberg, M. J.E. *J Mol Bio.* 1997. **272**, 106-120.

34. Chen, R., Weng, Z. *Proteins: Struc Func Gen.* 2002. **47**, 281 – 294.

35. Norel, R., Petrey, D., Wolfson, H.J., Nussinov, R. *Proteins: Struc Func Gen.* 1999. **36**, 307-317

36. Camacho, C.J., Vajda, S. *Proc Nat Acad Sci.* 2001. **98,** 10636-10641.

37. Jackson, R.M., Gabb, H.A., Sternberg, M.J.E. *J Mol Bio.*1998. **276**, 265-285.

38. Hendrix, D.K., Klein T.E., Kuntz, I.D. *Prot Sci.* 1999. **8**, 1010-1022.

39. Jones, S., Thornton, J.M. *J Mol Bio.* 1997. **272**, 121-132.

40. Zhang, C., Chen, J., DeLisi, C. *Proteins: Struc Func Gen.* 1999. **34**,255-267.

41. Janin, J. *Prog Bio-phys Mol Bio.* 1995. **64**,145-166.

42. Shoichet, B.K., Kuntz, I.D. *Chem Biol.* 1996. **3**,151-156.

43. Jiang, F., Kim, S. *J Mol Bio.* 1991. **219**, 79-102.

44. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesen, A.A., Aflalo, C. Wodak, S.J. 1992. *Proc Natl Acad Sci.* 1992. **89**, 2195-2199.

45. Lorder, D.M., Udo, M.K., Shoichet, B.K. *Protein Sci.* **2002** Jun;11(6):1393-408.

46. Smith, G.R., Sternberg, M.J.. *Curr Opin Struct Biol.* **2002** Feb;12(1):28-35.

Gloss to Chapter I

Chapter I is based entirely on the development and validation of ADAPT, a genetic algorithm for searching chemical space. The interest in this approach stemmed from a desire to sidestep the limitations of a 3-dimensional chemical database. Most structure based design methods require the use of such in-silico 3-dimensional databases to generate promising lead compounds. The commonly cited drawbacks of the dependence on 3-dimensional databases are the lack of chemical diversity, the inability to simultaneously screen for multiple properties, and the inability to incorporate chemical optimization into a rapid screen.

The approach described in this chapter overcomes many of these traditional limitations by implementing *de-novo* chemical design, a methodology that will be further illustrated in chapter II. The principles behind *de-novo* design dictate an independence from whole molecule databases, and instead focus on the use of atomic or chemical fragment databases to access combinatorially larger numbers of chemical entities. However, our ability intelligently search through this combinatorial space is dependent on our screening function (in this case, DOCK), and our ability to optimize this function in the context of chemical space.

The approach uses a genetic algorithm to traverse the chemical space defined by the accessed by a library of chemical fragments that are combined to build whole molecules. Genetic algorithms are optimization schemes that model Darwinian evolution. They have the advantage of being able to optimize across many different functions simultaneously, producing a user-defined number of results. Genetic algorithms are not known to be the greatest optimizing functions because of their inability to converge to a

single consistent solution. Rather, they often return larger numbers of good, though not optimal, solutions. Genetic algorithms are a consistently used in computational biology for a variety of applications including protein folding, de novo drug design, and bioinformatics.

In the context of drug-design, genetic algorithms enable us to screen large numbers of compounds using multiple criteria and identify a user-defined number compounds that have strongly meet those criteria. We began the development of the ADAPT program with high hopes for a finely-tuned *de-novo* design program. However, it instead proved itself to be more useful for generating general solutions to a design problem, rather than something that could 'pick a needle out of a haystack.' This lesson turned out to be quite useful in a later project (discussed in chapter 2) that relied on a generalized approach to the design of a small molecule peptide mimetic. ADAPT returned very strong results, but never identified the optimum solution. In some ways, this typical GA behavior validated our implementation of the algorithm. Many of the previous adaptations of a genetic algorithm in computational biology resulted in similar results.

This chapter is presented in the form of a journal article published in the *Journal of Computer-Aided Molecular Design.* The project actually grew out of my rotation project with Tack during my first year in graduate school – a 'toy 'genetic algorithm that showed the evolution between a cube and a sphere. While inherently impractical, it was a good means of developing a foundation for understanding the practical implementation of genetic algorithms. This project was also my first lengthy programming project outside of the classroom. Working with Scott Pegg was beneficial because the interaction created

numerous opportunities to share theoretical insights, as well as learn programming methods from an experienced programmer. In the beginning, Scott and I shared full responsibility for developing the code and the various algorithms that would eventually become the program. Midway, we decided that the code would seem more consistent if one person rewrote the code into a more efficient implementation of our ideas. Subsequent to the completion of the code, Scott and I split up the validation experiments. I focused primarily on showing proof-of-concept on the lengthier HIV-1 RT exercise. Incidentally, the first book Tom Ferrin had me read before entering graduate school was 'An Introduction to Genetic Algorithms' by Goldberg.

CHAPTER I

~•~

A Genetic Algorithm for Structure-Based De Novo Design

~•~

Scott C.-H., Pegg, Jose J. Haresco, Irwin D. Kuntz

Accepted for Publication in

Journal of Computer-Aided Molecular Design, 15: 911-933,2001

Abstract

Genetic algorithms have properties that make them attractive in de novo drug design. Like other de novo design programs, genetic algorithms require a method to reduce the enormous search space of possible compounds. Most often this is done using information from known ligands. We have the developed the ADAPT program, a genetic algorithm which uses molecular interactions evaluated with docking calculations as a fitness function to reduce the search space. ADAPT does not require information about known ligands. The program takes an initial set of compounds and iteratively builds new compounds based on the fitness scores of the previous set of compounds. We describe the particulars of the ADAPT algorithm and its application to three well-studied target systems. We also show that the strategies of enhanced local sampling and re-introducing diversity to the compound population during the design cycle provide better results than conventional genetic algorithm protocols.

Introduction

The ultimate goal of structure-based de novo drug design is the ability to develop new ligands using information from the three dimensional (3D) structure of a protein target without the prior knowledge of other ligands. The search space of all possible chemical compounds is far too large to be enumerated [1], while the locations of bioactive drugs within this space tend to be spare, non-contiguous, and very difficult to predict a priori. Current computational methods use the 3D structural information of the target to narrow this search space. One strategy is to join compound fragments, which fit favorably into the active site [2,3] either by constructing linking atoms or by searching a database of possible linking structures. [4-9] While this strategy leads to many novel and potentially useful structures, the resulting compounds are very dependent on the placement and selection of the fragments.[10] Another strategy is to construct a molecule within a binding site by 'growing' it from a seed atom or fragment, using the site's steric and/or energetic properties to guide the growth of the compound. [11-15] This strategy can also result in useful structures, but the molecules built are sensitive to the position and choice of the seed atom used and often leads to molecules that are difficult to synthesize. [10,16]

Genetic algorithms have recently been applied to the problem of de novo drug design. Based on the concepts of natural selection, genetic algorithms operate by creating a set of proposed solutions to a problem of interest (a 'population'), evaluating them ('fitness pressure'), and then using the best solutions to develop a new set of proposed solutions ('breeding'). Genetic algorithms have several properties which make them attractive in de novo ligand design, the most important being the ability to perform well

when the search space is large and incompletely understood, the fitness functions are not exact, and a global minimum is not required. Unlike incremental growing schemes, genetic algorithms can use properties of whole compounds to guide the generation of new compounds. Properties such as absorption, distribution, metabolism, and excretion (ADME) cannot easily be calculated from molecular fragments or partially constructed compounds, and they are of increasing importance in today's drug development climate.[17] Another potentially attractive feature of genetic algorithms is the production of an ensemble of solutions. A diverse set of proposed compounds can help to avoid patent issues and provide 'back up' compounds of different chemical classes, which can be used if the initial compounds fail during drug development.

As with other computational de novo design strategies, genetic algorithms require a method of reducing the allowable search space. Most of the currently published genetic algorithm strategies use fitness functions that rely upon knowledge of a template ligand or pharmacophore to restrict the space of allowable compounds. Such fitness functions include similarity to a specific known ligand [18], fitness to a pharmacophore template derived from putative or known ligands [19-21], and the presence of properties previously determined by structure-activity relationships.[22] These strategies use implicit information about the target, and are not solely structure-based strategies. Exceptions include the method of Glen and Payne in which a set of pharmacophoric constraints is built directly from a protein active site[20] and a genetic algorithm implementation which used experimentally measured binding affinities as a fitness function.[23] It has been suggested that current molecular docking methods, in which the steric and energetic fit of a ligand to a target site is estimated, could be used as a fitness function.[24] Such a

function would restrict the search space using information from the 3D structure of the target without using knowledge of existing ligands. Recently, Sheridan et. al. published a further development of their genetic algorithm for designing combinatorial chemistry compound libraries, which incorporates docking calculations into the fitness function.[25,26] While similar to our approach, their algorithm has been demonstrated on compounds built only via linear combinations of fragments, and uses a docking calculation in which only a small number of rigid conformations of each compound are scored.

We present here the ADAPT program, a genetic algorithm for structure-based de novo drug design which builds compounds as acyclic graphs of fragments and incorporates the use of flexible docking calculations into the fitness function. We first give the details of the genetic algorithm implementation, and then show its application to ligand design problems in which we demonstrate the searching ability of the genetic algorithm, the effects of forcing local sampling, and the performance enhancement of adding extra diversity to the compound population while the genetic algorithm is running. We also demonstrate application of the program to three protein target systems in which we can recover fundamental properties of ligands known to bind to these targets.

Methods

Genetic Algorithms

Genetic algorithms are a type of stochastic search algorithm originally developed by Holland [27,28] based on the concepts of Darwinian evolution.[29] All genetic algorithms start with a population of proposed solutions to a given problem. In the de novo drug design case, the population is a set of small molecules and the problem at hand is that of building bioactive compounds with desirable properties. The population is subjected to a fitness pressure in which the individuals are evaluated in terms of how well they solve the given problem. The individuals are then bred to create a new population, with the more 'fit' individuals being chosen to breed more often. This breeding is usually done using an adaptation of the crossover and mutation events observed in DNA replication. The genetic algorithm cycle continues though generations of evaluation and breeding until sufficiently good solutions have been found. Figure 1 shows a genetic algorithm cycle and its application to the de novo design problem. The following are the specifics of our particular adaptation of genetic algorithms to structure-based de novo ligand design.

Compound representation

In the ADAPT program, compounds are represented by an acyclic graph of at most 16 fragments, with the user defining an upper and lower bound on the number of fragments per molecule. The fragments themselves are defined by the user via an adapted subset of the SMILES line notation [30], and each fragment can have, at most, 8 user-defined points to which other fragments may be attached. The fragments reside in a linear order in a file such that chemically similar fragments are located next to each other in the

fragment file (this order can also be hand-edited). This order is determined using a combination of binning and "gray-scaling" based on Daylight fingerprints[31]. Fragments are first binned by the number of 1's in their fingerprint and the bins are sorted according to the number of 1's per fingerprint. The fragments are then sorted using a walk, which starts with a randomly chosen fragment from the bin holding fragments with the fewest number of 1's. The next fragment in the walk is chosen from the same bin as the fragment with the largest number of 1's in common with current fragment. When the current bin is empty, the next fragment is chosen from the next bin. The user may designate one fragment to be a 'scaffold', in which case each compound must have one (and only one) occurrence of that particular fragment.

Starting generation

The starting generation can either be a diverse set of compounds, or it can be based on a user-defined compound. A diverse population of compounds is generated by choosing fragments at random and assembling them at random attachment points one at a time. Iteratively adding a node to the graph via a single new edge creates a connected graph with N nodes and N-1 edges, which must be acyclic. When the starting generation is seeded with a user-defined compound, the initial generation is built by randomly adding, subtracting, or swapping at most two fragments from the original structure. The user must define the original compound as an acyclic graph in terms of fragments (from the fragment file) and their connectivities. A copy of the original compound is included unchanged in the first generation.

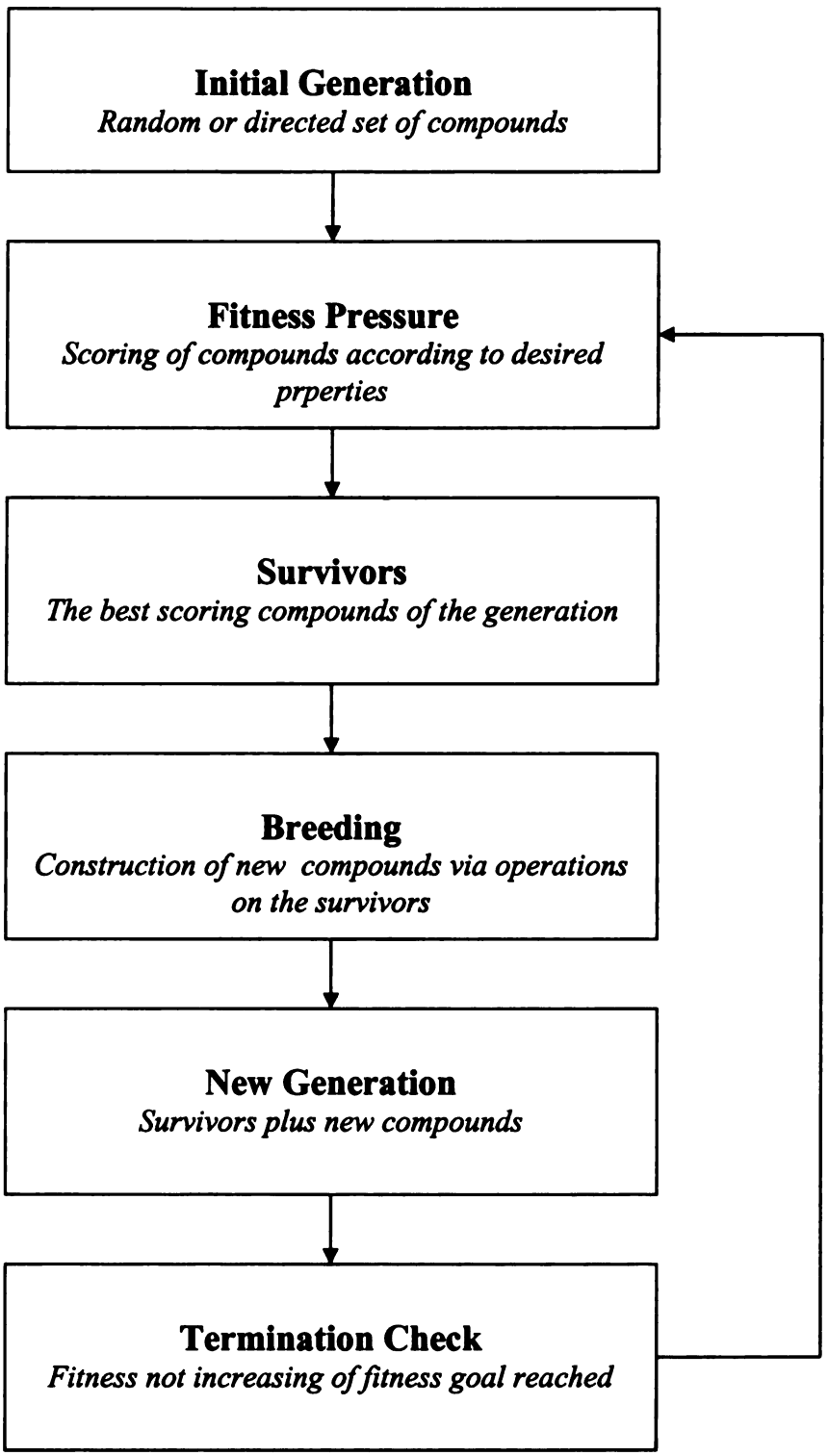
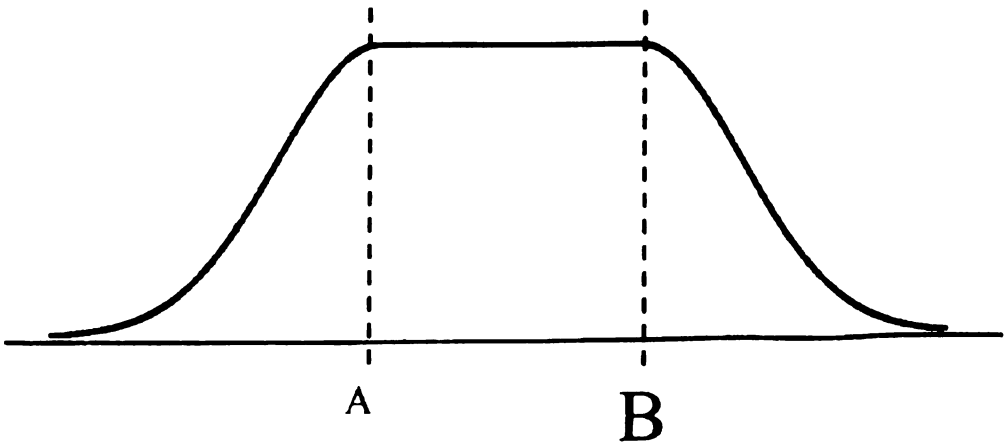


Figure 1. The genetic algorithm cycle and its application to drug design. An initial set of compounds is scored according to a set of desired properties, such as estimated binding affinity. The best scoring compounds of the generation are considered 'survivors' and are used in the construction of new compounds. The next generation, consisting of survivors of these new compounds, is then subjected to the fitness pressure. The cycle typically continues either for a set number of generations, or until the fitness scores no longer improve

Fitness evaluation

The major fitness evaluation of the ADAPT program is performed by the DOCK4.0 program[32] which gives an estimate of the steric and energetic fit of a ligand to a receptor site, taking into account ligand flexibility. Compounds are first converted to SMILES strings by our genetic algorithm, and then to a single three dimensional conformer using the CONCORD[33] module of the SYBYL program.[34] The DOCK score is computed as a pairwise potential between atoms of the ligand and atoms of the active site. The scoring function contains two terms, a standard 6-12 Van derWalls term and a 1/r electrostatic term.[32]

Also implemented are functions to measure simple physical properties of the assembled molecules such as CLogP (using Daylight's "clogp" program), molecular weight, number of rotatable bonds, and number of hydrogen bond donors/acceptors. These functions are scored according to a plateaued normal distribution (see figure 2) in which values between the user-defined plateau range return 1.0 and otherwise return values between 0 and 1 according to a user-defined normal distribution that makes up the edges of the plateau. Each fitness measure has a user-defined coefficient, and the overall



$$S(x) = \begin{cases} 1 & \text{if } A < x < B \\ N(\mu, \sigma) & \text{if } x < A \text{ or } x > B \end{cases}$$

Figure 2. The plateau fitness function. Values of properties within the defined limits A and B return a score of 1. For values outside this range, the scoring function returns the value of the normal distribution (with mean and standard deviation defined by the user) that has been split in half and moved to the edges of the plateau.

fitness of a compound is simply the linear sum of each measure multiplied by its coefficient,

$$fitness = C_{dock} S_{dock} + C_{mw} S_{mw} + C_{rot} S_{rot} + C_{hbond} S_{hbond} + C_{clogp} S_{clogp}$$

where C_n is the coefficient for score S_n, which (with the exception of S_{dock}) is defined by

$$S_n(x) = \begin{cases} 1 & if \;\; A_n \le x \le B_n \\ N(\mu,\sigma) & if \;\; A_n > x \;\; or \;\; B_n < x \end{cases}$$

where $N(\mu,\sigma)$ is a normal distribution with mean μ and standard deviation σ defined by the user, and A_n and B_n are boundaries determined by the user for each of the fitness functions (see figure 2). Once each compound has an overall fitness score, the top scoring fraction (as defined by the user) of compounds is used as the parents for the next generation. The user may also choose whether the parents are kept unchanged in the next generation, or whether only their children continue.

Breeding

Breeding of a set of compounds to produce the next generation is done via adaptations of crossover and mutation to the graph representation of the compounds. Crossover occurs by swapping a subset of nodes between two graphs. The number of nodes is chosen at random according to a decaying exponential distribution (biasing the exchange towards a smaller number of nodes), and the nodes themselves are chosen by choosing a random node and then performing a random walk through the graph to build a connected set of nodes. The sets of nodes swapped are either of equal size, where the crossover process is isomorphic to a recoloring of nodes according to type, or are of

unequal sizes, in which case the crossover process is equivalent to a recoloring of nodes followed by the addition or removal of nodes with one edge per node being added or removed. Because the sets of nodes swapped are themselves acyclic graphs, the cases of equal and unequal set sizes both lead to acyclic graphs. The parents for crossover are chosen via an 'elitist' scheme, in which the probability of being chosen is proportional to a compound's relative fitness within a generation, the more fit being chosen more often.

Mutation can occur to either the identity of the nodes or the connectivity of the graph. The connectivity is mutated by choosing an edge at random and reconnecting it to a node with an open attachment point at random as long as it doesn't induce a cycle in the graph. Because there are N-1 edges for N nodes in the graph, checking for a cycle is equivalent to looking for a disconnected node, an algorithm that has complexity O(N). Node identity is mutated by moving up or down in the fragment definition file. The probability of a mutation is determined by the user, while the distance moved is determined by a decaying exponential also defined by the user, such that short distance moves are more likely. When the fragments are ordered in terms of chemical similarity, this scheme causes the genetic algorithm to preferentially sample local chemical space.

Stopping

The ADAPT cycle will stop when the fitness of the generations ceases to improve, the program has found a generation with fitness better than a defined goal, or the program has gone a set number of generations. The user can set a threshold for the required improvement of the fitness, as well as a window over which these values can be averaged. When using a fitness goal, the algorithm may get stuck in a local minimum

where the fitness of the generations is no longer improving, but the fitness value is not near the goal. In this case, the ADAPT program adds 'diversity' to the current generation by randomly adding, subtracting, or swapping at most two fragments from each compound. This allows the program to 'jump' to another random (but not too distant) point in chemical space from which it may have a better chance of reaching the fitness goal. The user defines the number of generations the ADAPT program may progress before this diversity is added.

Results

We chose three well studied protein targets to test the ADAPT program on: cathepsin D, dihydrofolate reductase, and HIV-1 reverse transcriptase. The cathepsin D target was chosen because we have experimental binding data from a previous combinatorial chemistry project[35] to which we could compare the results of ADAPT applied to the same combinatorial scheme. Dihydrofolate reductase was chosen because it would provide a larger search space than a simple combinatorial chemistry scheme, but one in which there exists a well-studied ligand (methotrexate) that would allow us to test the effects of seeding the initial population with a known ligand. HIV-1 reverse transcriptase was chosen because the non-nucleoside binding site accepts a wide variety of known ligands, representing a large search space, but with a specific structural theme that we could attempt to rediscover with the ADAPT program.

Cathepsin D

In a structure-based design collaboration, a peptidomimetic scaffold with three substituent positions was combined with a 10 x 10 x 10 set of side chain groups in an effort to develop non-peptide inhibitors to Cathepsin D.[35] All 1,000 resulting compounds were synthesized and tested experimentally for inhibition. Figure 3 shows the set of fragments and the scaffold. Our interest here was to see if we could determine, using the ADAPT program, which of the side chain groups, and at which positions, would produce the best inhibitors.

With current computational molecular docking methods, 1,000 compounds is a small number to evaluate, and represents a small chemical search space. In order to create a larger, yet still practical search space, we let any of the 25 fragments be at any of the 3 scaffold attachment sites, giving us 15,625 unique compounds in our search space. We performed ten runs of the ADAPT program, each going through 50 generations, using DOCK as the only measure of compound fitness. The values of other ADAPT parameters are given in table 1. By the 50th generation, the average fitness of the compounds within a generation was no longer improving rapidly (see figure 4). Also, by the 50th generation, the number of unique compounds was much lower, indicating that we were converging on specific structures in our chemical search space (see figure 5).

Figure 6 shows a histogram of the fragments occurring at each position in the final generations of the ten runs of ADAPT. At positions R1, R2, and R3, there were 8, 7, and 7 fragments, respectively, which occurred more than once in the final generations. A combinatorial library constructed by using the fragments and positions identified by ADAPT would result in a 8 x 7 x 7 = 392 compound library. This library would contain

four of the seven inhibitors identified experimentally at 100nM in the synthesized 1,000 compound library, including the single compound with the lowest inhibition constant. There could, in fact, be more inhibitors in the ADAPT based library, but experimental data exists for only 24 of the 392 compounds. Evaluated on the basis of just DOCK scores, the ADAPT based library has a better score distribution than the synthesized library (see figure 7).

None of the 23 inhibitors identified in the synthesized library at 330nM were found as individuals in the final generations of the ADAPT runs, however. This is not surprising, since the fitness space defined by the DOCK score does not mirror precisely the binding landscape of the cathepsin D active site, and the chemical space searched by ADAPT was roughly 15 times larger than the synthesized library. In fact, the known inhibitors, while still mostly in the top 15% of the landscape have an overall worse DOCK score distribution than the ADAPT produced compounds, which are all within the top 1% (see figure 8). Nevertheless, the DOCK score provided enough information to allow ADAPT to identify the fragments present in the majority of the best inhibitors.

To investigate how well our genetic algorithm is sampling the space of possible compounds, we scored the fitness of all 15, 625 possible compounds using DOCK. This allowed us to see the overall ranking of any compound generated by the genetic algorithm. Figure 9 compares the genetic algorithm with a random search in terms of the fractional ranking expected at random divided by the fractional ranking of the worst scoring survivor of a given population. This gives us a measure of 'enrichment' over a random search. We also ran our genetic algorithm ten times under the same conditions, but with a fragment set that was randomly ordered instead of grayscale. While both the

grayscaled and randomized fragment sets result in a positive enrichment over a random search, the grayscaling of the fragment set to force local sampling produced much better enrichment.

Overall, the ADAPT program was able to select the fragments (and their positions) that were consistently present in the best inhibitors determined experimentally. The program's inability to directly produce the known inhibitors in the final generations was likely due to their ranking in the binding landscape being slightly different than their ranking in the landscape defined by DOCK scores. Our experiments with the cathepsin D system also show that forcing local sampling by allowing fragments to mutate only to other fragments that are chemically similar produces a more efficient search for high scoring compounds.

Parameter	Value
Population size	12
Fragments_per_compound_max	3
Fragments_per_compound_min	3
Use_scaffold	TRUE
Dock_fitness	TRUE
Num_dock_runs	3
Dock_fitness_coefficient	1.0
Clogp_fitness	FALSE
Weight_fitness	FALSE
Rotatable_bond_fitness	FALSE
H_bond_fitness	FALSE
Breeding_fraction	0.5
Keep_parents	TRUE
Breeding_types	Crossover, mutation
Mutation_types	Identity
Mutation_probability	0.5
Mutation_lambda	0,5

Table 1. ADAPT parameter settings used in the cathepsin D experiments.

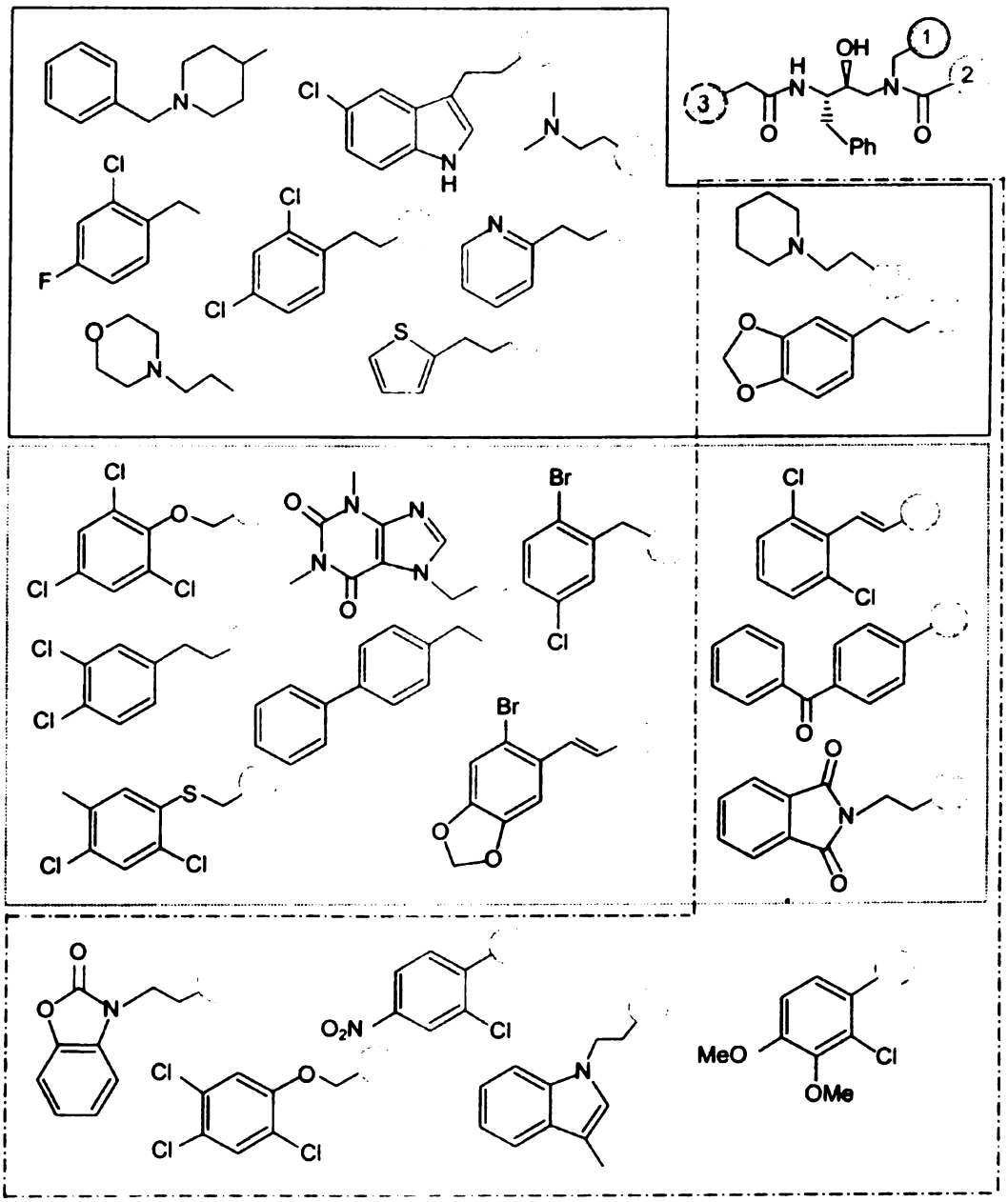


Figure 3. The fragments used in the cathepsin D experiments. At upper right is the scaffold with three R groups labeled. The fragments bounded by the solid lines were tested experimentally at R1, those bounded by the dashed lines were tested experimentally at R2, and those bounded b by the dashed and dotted line were tested experimentally at R3. Fragments in multiple regions were tested at two(or three) positions. In the ADAPT experiments, all 25 fragments were allowed to be at any of the three scaffold positions

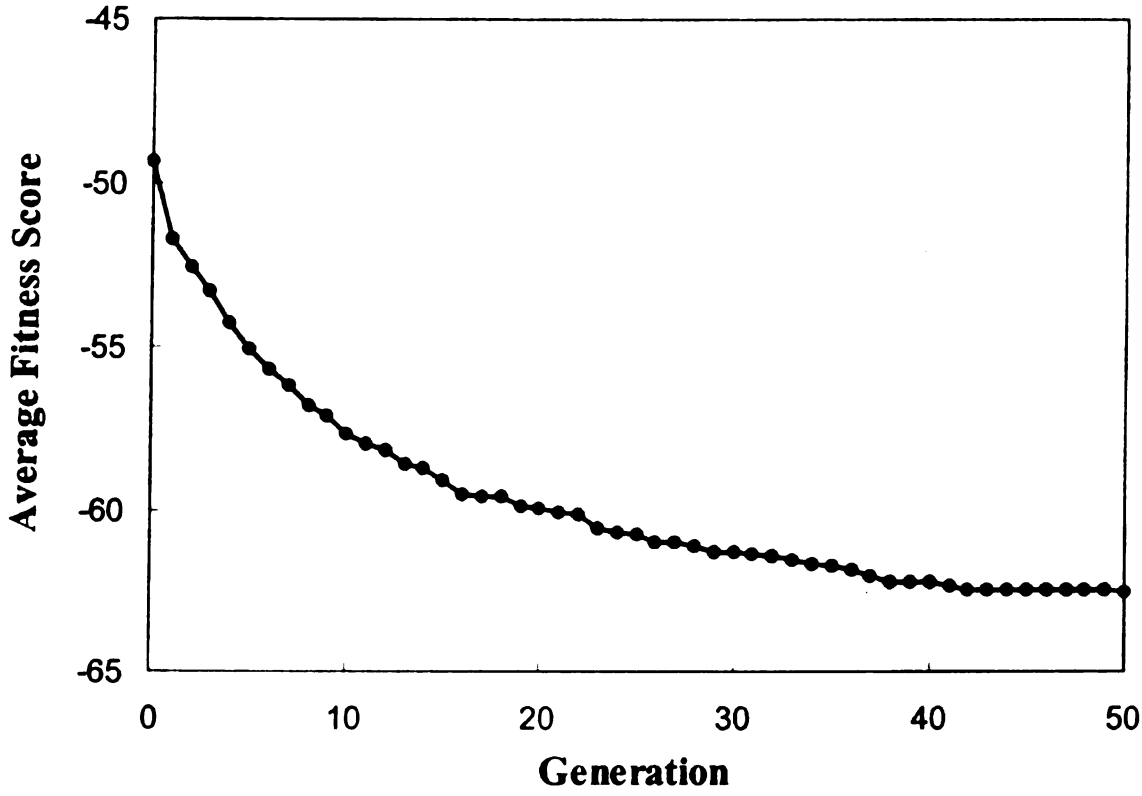


Figure 4. Improvement of the fitness score in the cathepsin D experiments. The values are the fitness scores of the surviving compounds, averaged over all ten runs of the ADAPT program

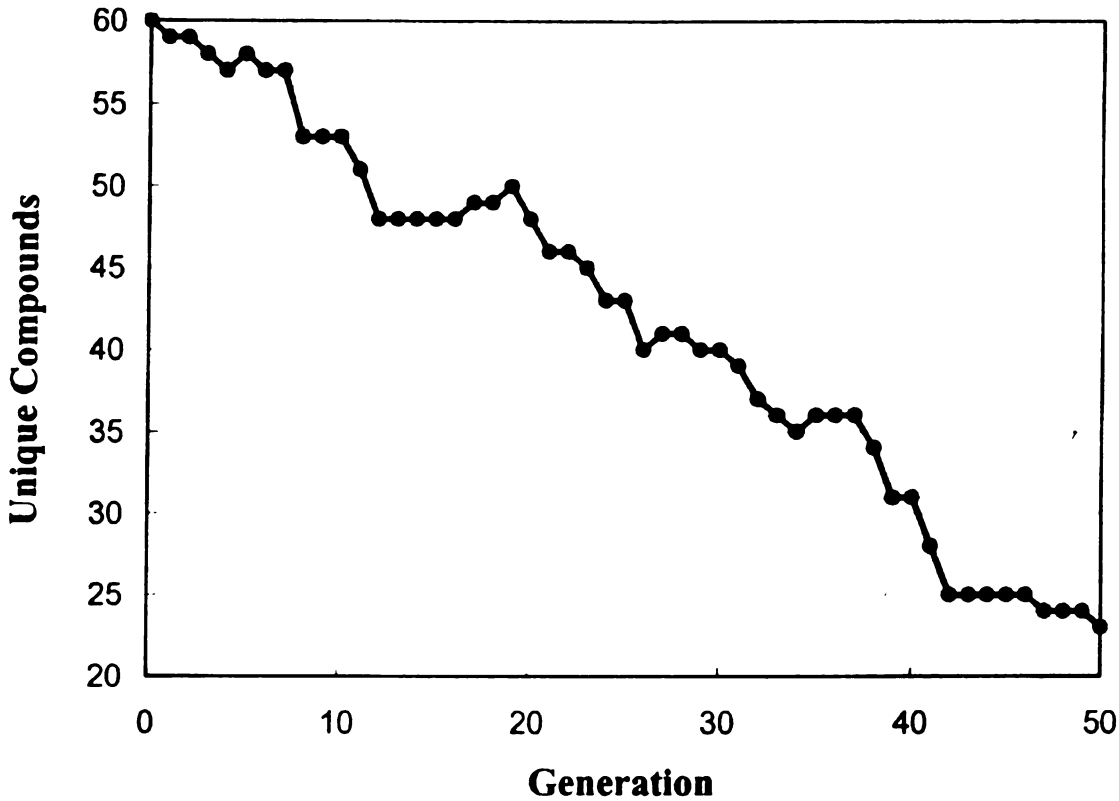


Figure 5. The number of unique compounds per generation in the cathepsin D experiments. The values represent the number of unique survivors across all ten runs of the ADAPT program.

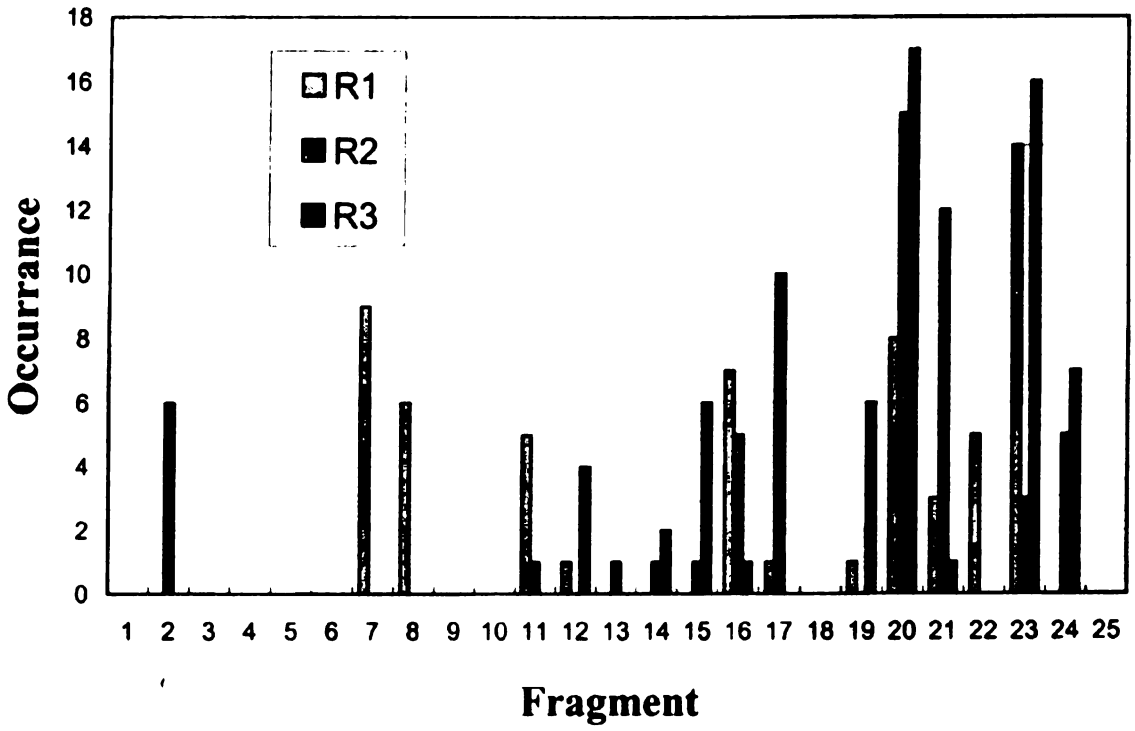


Figure 6. Fragments observed in the final generations of the ten ADAPT runs in the cathepsin D experiments

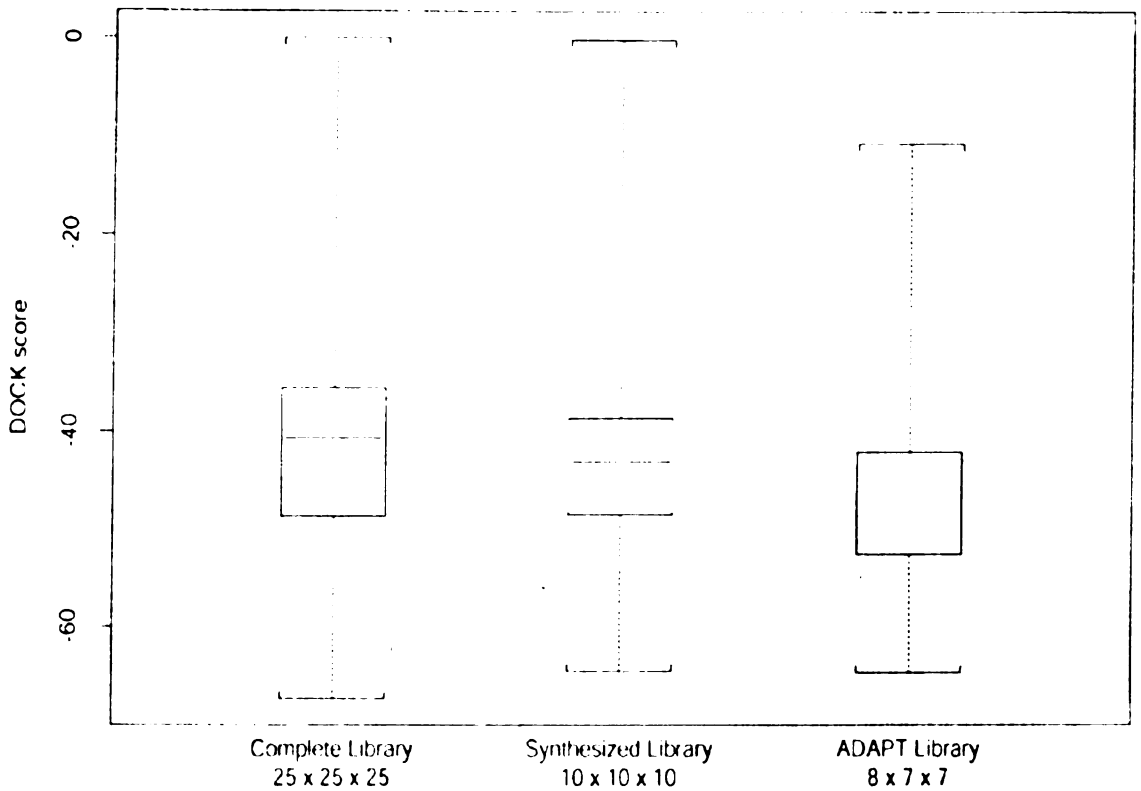


Figure 7. The distributions of rank (defined by DOCK score) of the complete set of possible compounds(25 x 25 x 25), the synthesized compound library (10 x 10 x 10) and the library based on fragment identified by ADPAT (8 x 7 x 7).

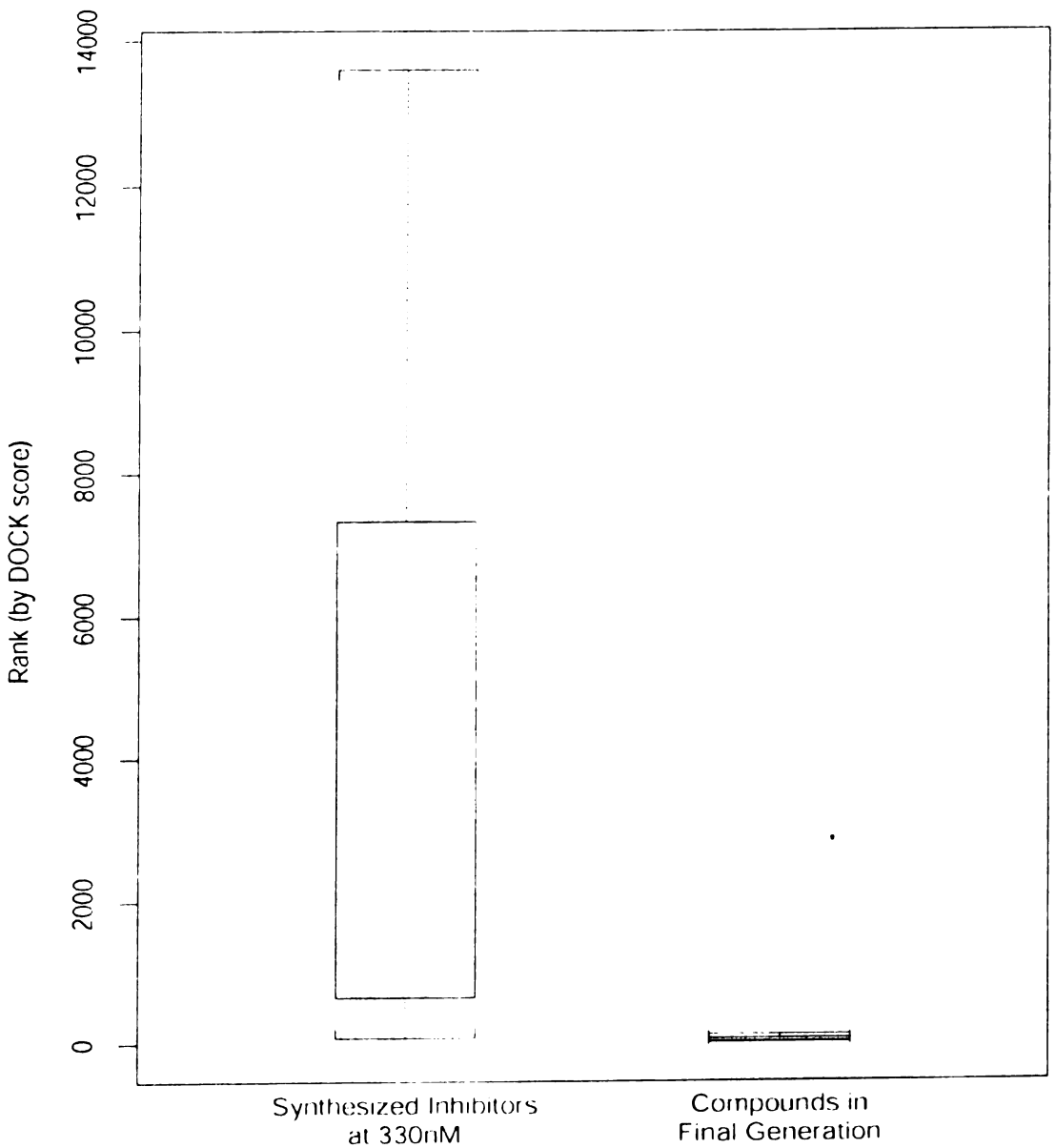


Figure 8. The distributions of rank (defined by DOCK score)of the 23 experimentally determined inhibitors at 330 nM, and the 23 compounds from the final generations of the ten ADAPT runs.

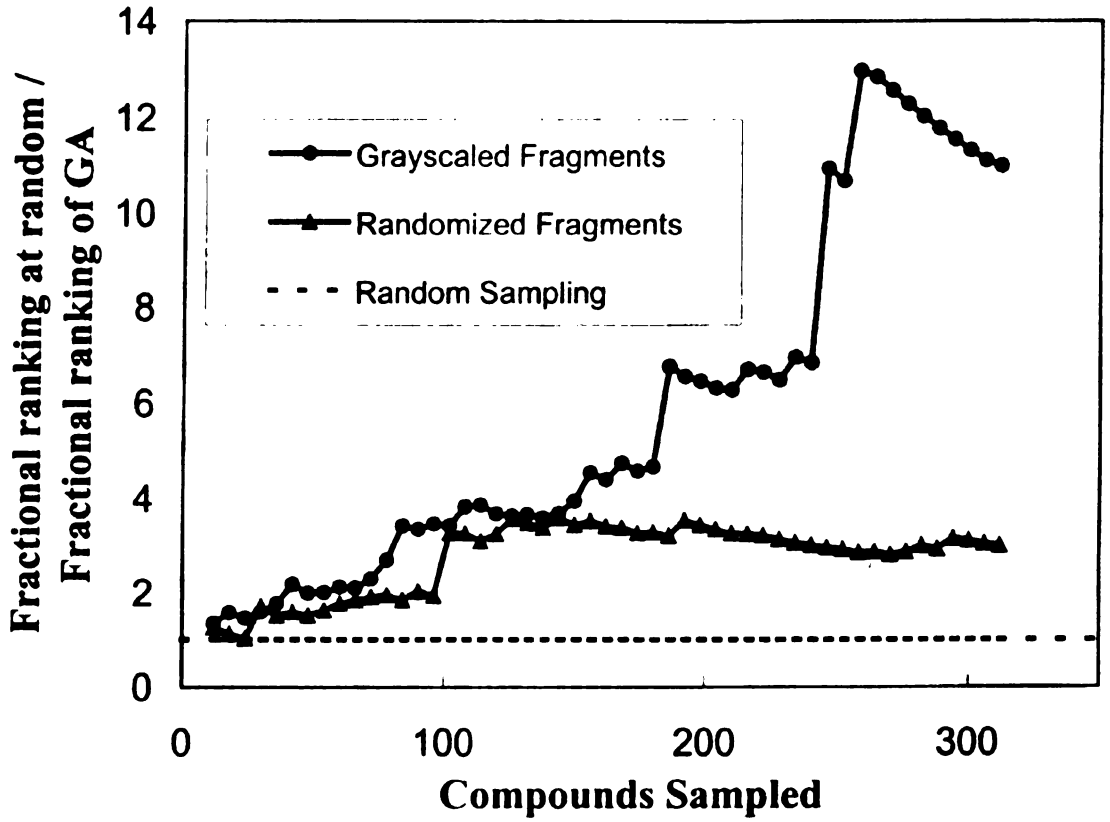


Figure 9. Enrichment of sampling using the ADAPT program over sampling at random. The enrichment is calculated by dividing the fractional ranking expected via random sampling (in this case sampling six at a time without replacement and taking the rank of the worst scoring individual) by the fractional ranking of the worst scoring survivor of each generation, averaged over ten runs.

Dihydrofolate Reductase

The enzyme dihydrofolate reductase (DHFR) with the small molecule inhibitor methotrexate represents a well-studied system in structure based drug design.[36-38] We applied the ADAPT program to this system to investigate three topics: how well DOCK would guide the construction of compounds in a larger search space, the effects of seeding the initial population with a known ligand, and the effects of adding diversity to the population during longer runs of the program. The fragment set consisted of 13 fragments that make up methotrexate, logical modifications to those fragments (e.g. replacement of a nitrogen in the pteridine ring with a carbon atom), and some basic "building blocks", fragments such as single carbon and nitrogen atoms. Figure 10 shows methotrexate and some of the fragments from the fragment set. Included in the set are building blocks general enough to allow virtually any acyclic hydrocarbon to be constructed. With 32 fragments to choose from and compounds allowed to have between 3 and 13 fragments, a lower bound for the size of the chemical space is approximately 3.5×10^8 unique compounds.

We performed three sets of ten runs each with ADAPT, each set consisting of identical parameters except for the random number seed, with DOCK as the only fitness function. For the first set, the initial generation was based on methotrexate and the ADAPT program was forced to run a minimum of 30 generations before stopping based on a lack of improvement in the fitness score. For the second and third sets, the initial generation was built randomly but the second set was forced to run a minimum of 30 generations while the third set was forced to run exactly 100 generations. Each run used a

population size of 20, of which half were kept as survivors in each generation, and used both crossover and mutation in breeding.

Figure 11 shows the distributions of the DOCK scores for the compounds in the final generations of each of the three sets of runs, plus a distribution of the DOCK scores of (unoptimized) randomly assembled fragments. All three ADAPT runs create compounds that score much better than random, with the runs seeded with methotrexate having both the narrowest and best scoring distribution. None compounds in the final generations of the unseeded runs to 30 generations had DOCK scores better than methotrexate while 94% of the compounds in the final generations of the seeded runs had scores better than methotrexate. However, 28% of the compounds in the final generations of the unseeded runs that were run to 100 generations had DOCK scores better than methotrexate.

Despite generating a wide variety of structures, our genetic algorithm identified a general motif present in known DHFR inhibitors. There were 98 unique structures among the survivors in the final generations of the seeded runs, a selection of which is shown in figure 12a, along with their DOCK scores. All of the structures had an essentially chain-like structure with a fused ring fragment at one end. Of the 98 compounds, all but two retained the pteridine ring fragment at one end, while the chain extending from it varied widely. In the second set of runs, in which the initial generations were not based on methotrexate, there was still a large bias towards a fused ring structure at one end and a flexible chain coming off of it, but the pteridine ring fragment was not seen in a high fraction of the compounds in the final generations (21/100) Figure 12b shows a selection of good scoring structures generated by the second set of runs. The DOCK scores of these

Parameter	Value
Population size	20
Fragments_per_compound_max	13
Fragments_per_compound_min	3
Use_scaffold	FALSE
Dock_fitness	TRUE
Num_dock_runs	3
Dock_fitness_coefficient	1.0
Clogp_fitness	FALSE
Weight_fitness	TRUE
Weight_range	410 to 450
Weight_standard_deviation	60
Weight_fitness_coefficient	-10.0
Rotatable_bond_fitness	FALSE
H_bond_fitness	FALSE
Breeding_fraction	0.5
Keep_parents	TRUE
Breeding_types	Crossover, mutation
Mutation_types	Connectivity, Identity
Mutation_probability	0.5
Mutation_lambda	1.0

Table 2. ADAPT parameter settings used in the dihydrofolate reductase experiments

compounds were all poorer than that of methotrexate and the compounds generated by the first run. These results are in accordance with work by Verkhivker et al.[39] who demonstrated that the DHFR/methotrexate binding landscape would accept mutations to the end of the chain away from the pteridine fragment, and would accept virtually no changes in the pteridine structure. In the set of longer unseeded runs, compounds were generated with DOCK scores better than methotrexate, along with a higher occurrence of the pteridine fragment (56/100). Figure 12c shows a selection of compounds generated from the third set of runs.

To investigate the effects of adding diversity to the population during a long run of the program, we ran the ADAPT program for 1,000 generations, adding diversity back to the population every 200 generations. We compared this to doing 5 separate runs of 200 generations each. Figure 13 shows the fitness score distribution of the 5 final generations of the individual runs and the fitness score distribution of the 5 generations just before diversity was added to the population. The distribution of compounds at the generations just before diversity was added has a better average fitness score and best score.

The results of our experiments using on the DHFR/methotrexate system show that the ADAPT program can evolve compounds with the basic motif of known ligands. However, in a search for compounds with better fitness than a known ligand, using the ligand to seed the initial population results in the faster production of better scoring, yet still diverse compounds. Compounds with fitness scores better than the known ligand can still be produced by ADAPT, but it requires runs with significantly more generations.

Our results also show that when performing long runs in search of good scoring compounds, it can

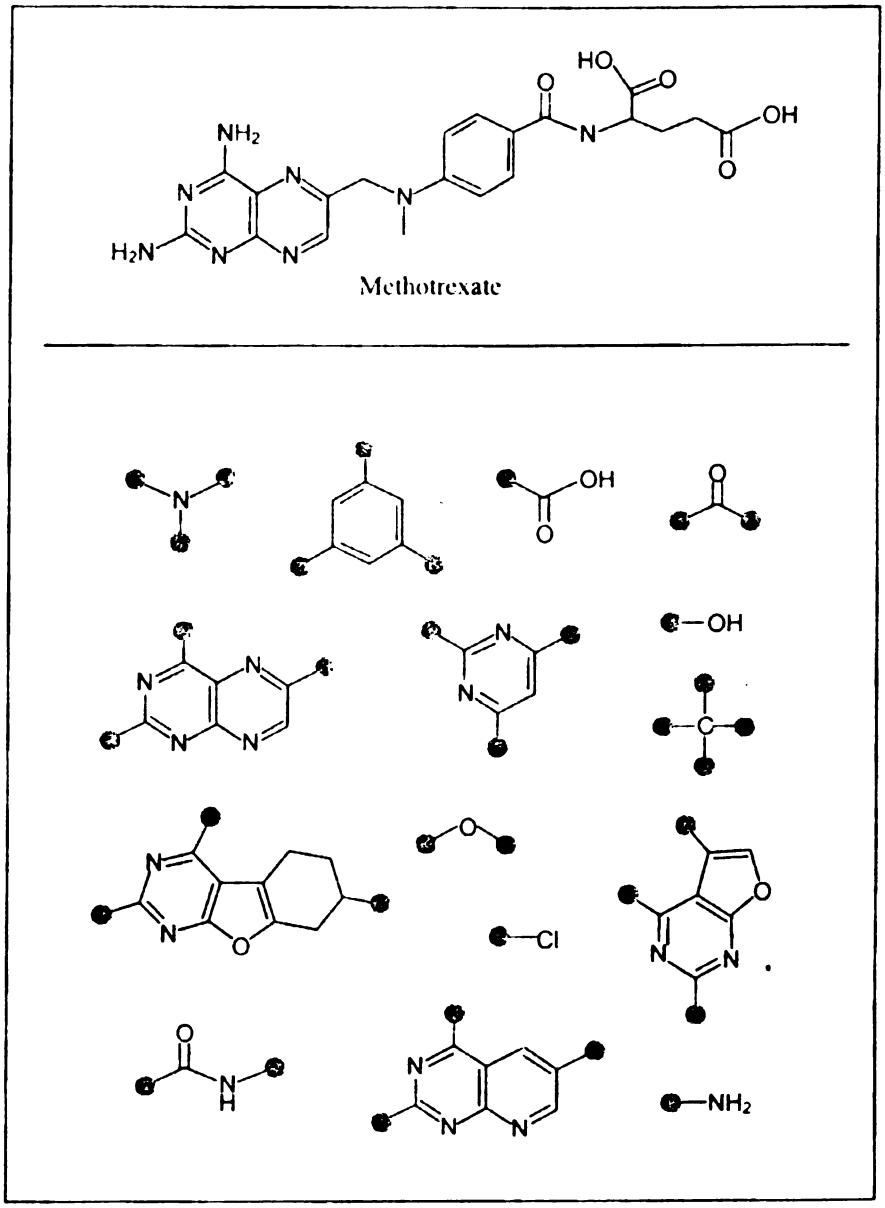


Figure 10. The known DHFR inhibitor methorexate and fragments from ADAPT runs. Fragments shown represent enough fragments to reconstruct methotrexate, as well as a diverse selection of general fragments that were added to the set. The gray circles represent attachment points.

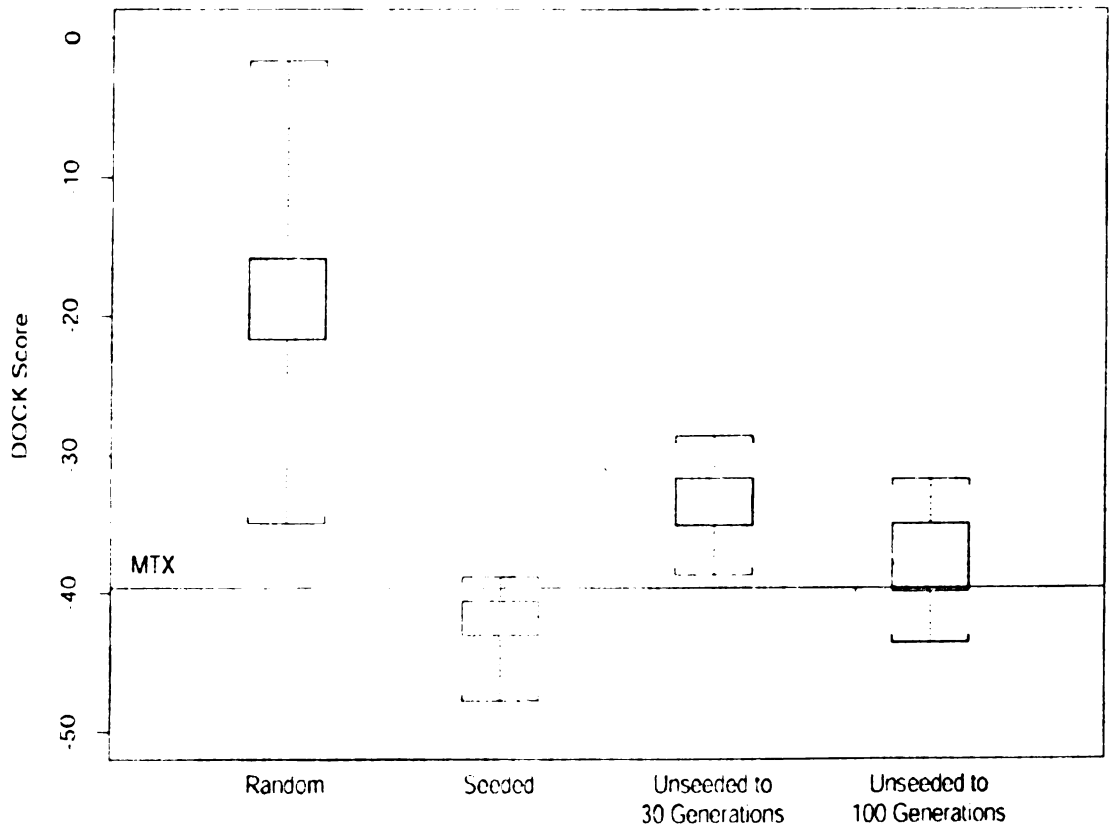


Figure 11. The distributions of DOCK scores for the compounds in the final generation of the ADAPT runs and of randomly assembled compounds. The distance between the top and bottom of the shaded box represents the inter-quartile range, the dark line represents the median, and the brackets represent the extreme values of the distribution. The random distribution represents 1,000 randomly assembled compounds, while the other three represent the approximately 100 compounds (10 from each of 10 runs). The horizontal line shows the DOCK score of methotrexate (-39.61).

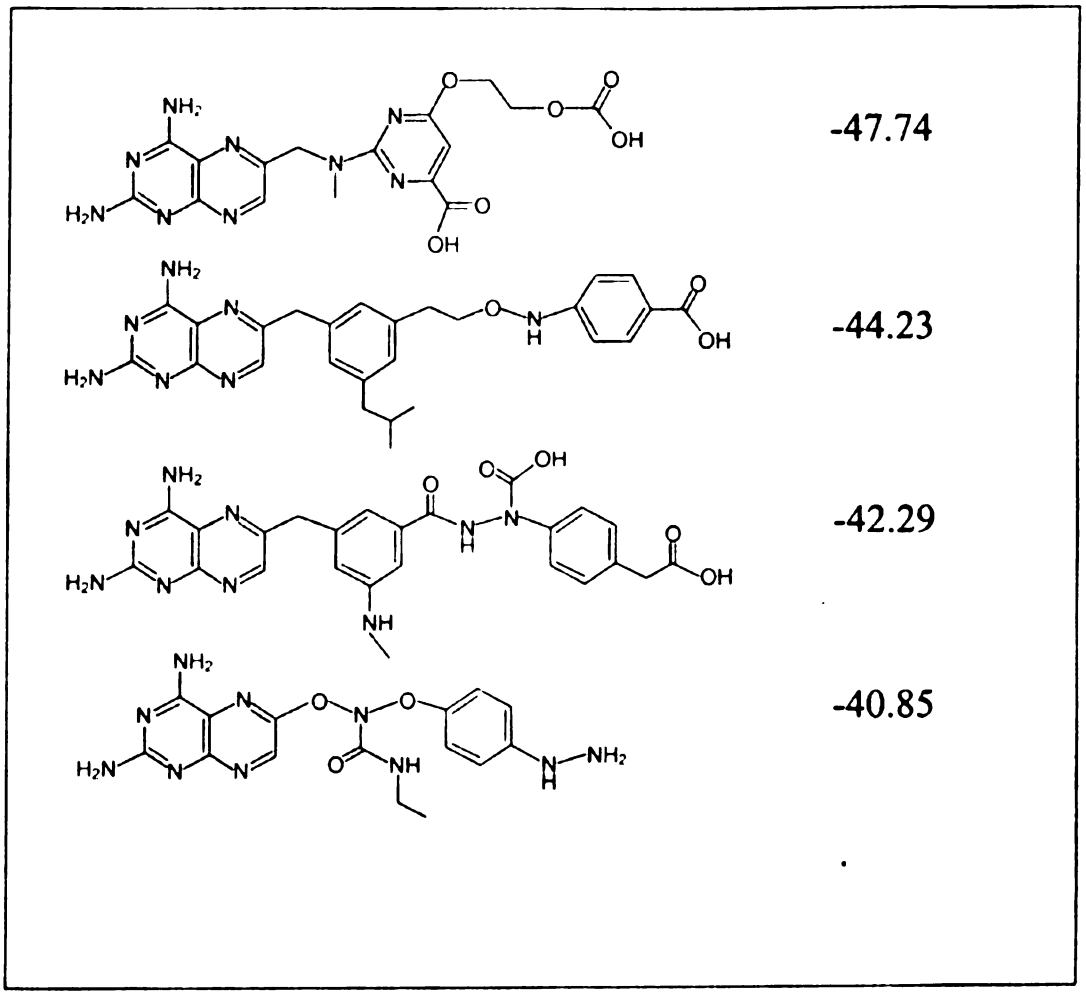


Figure 12A. Representative structures and their DOCK scores from the final generation of ADAPT runs seeded with methotrexate.

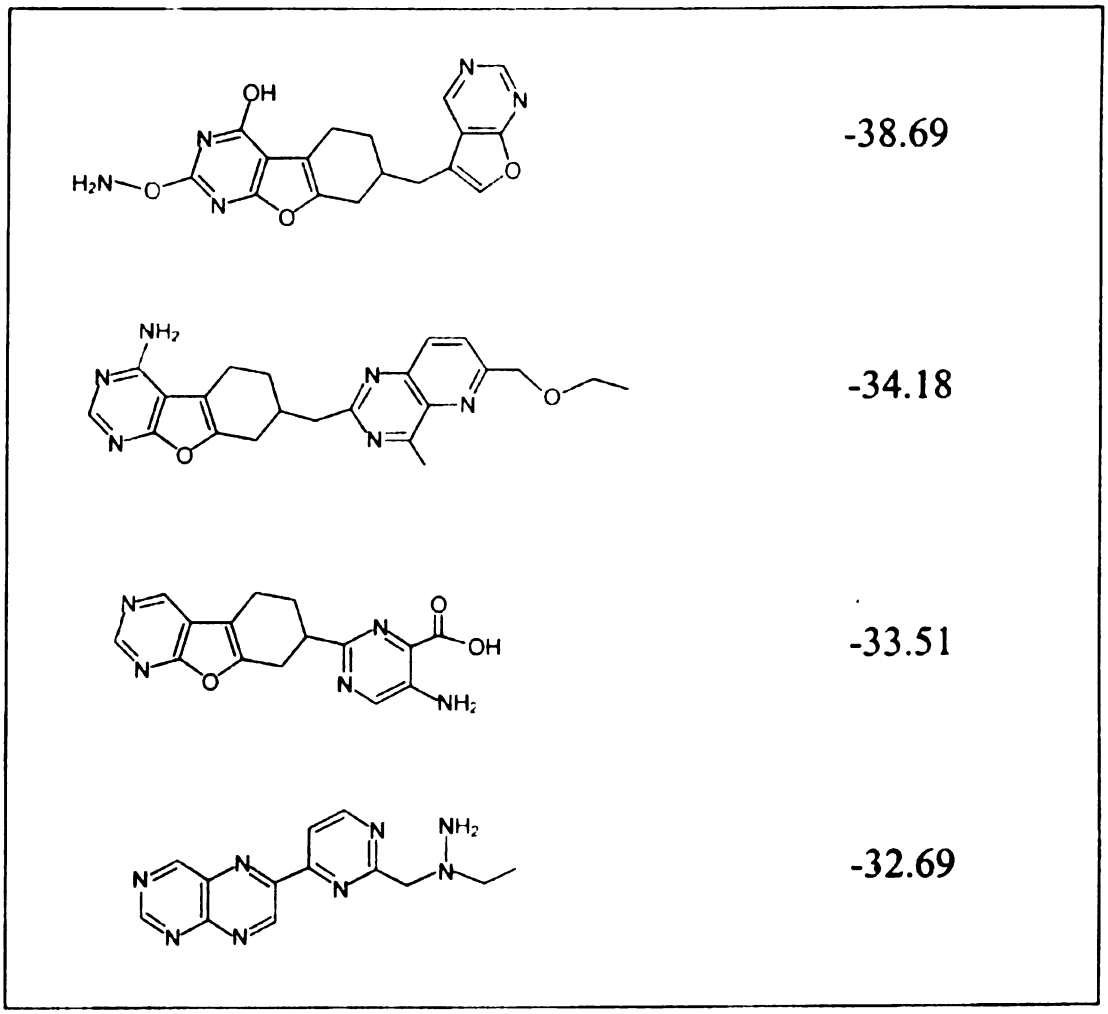


Figure 12B. Representative structures and their DOCK scores from the final generation of ADAPT runs which were unseeded and run to 30 generations.

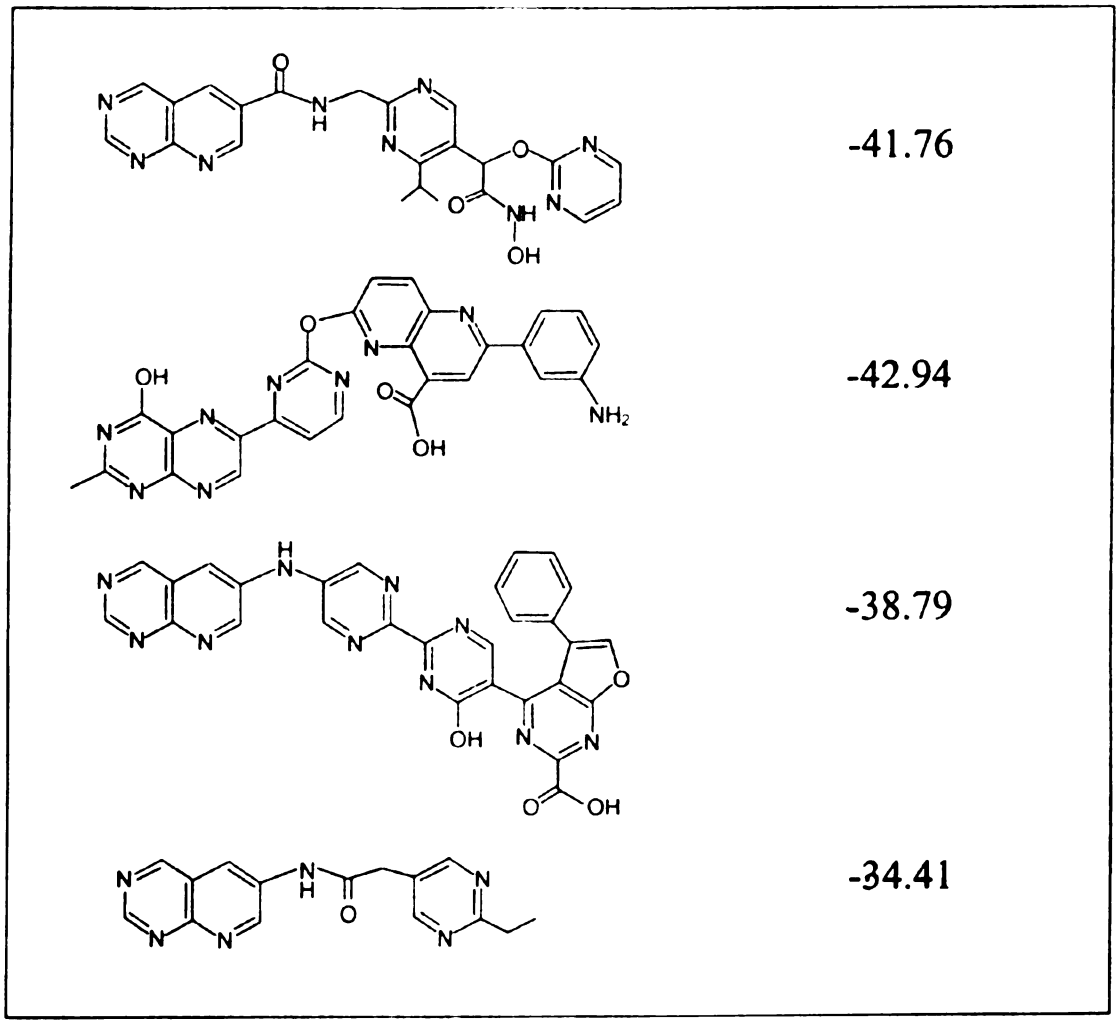


Figure 12C. Representative structures and their DOCK scores from the final generation of ADAPT runs which were unseeded and run to 10 generations.

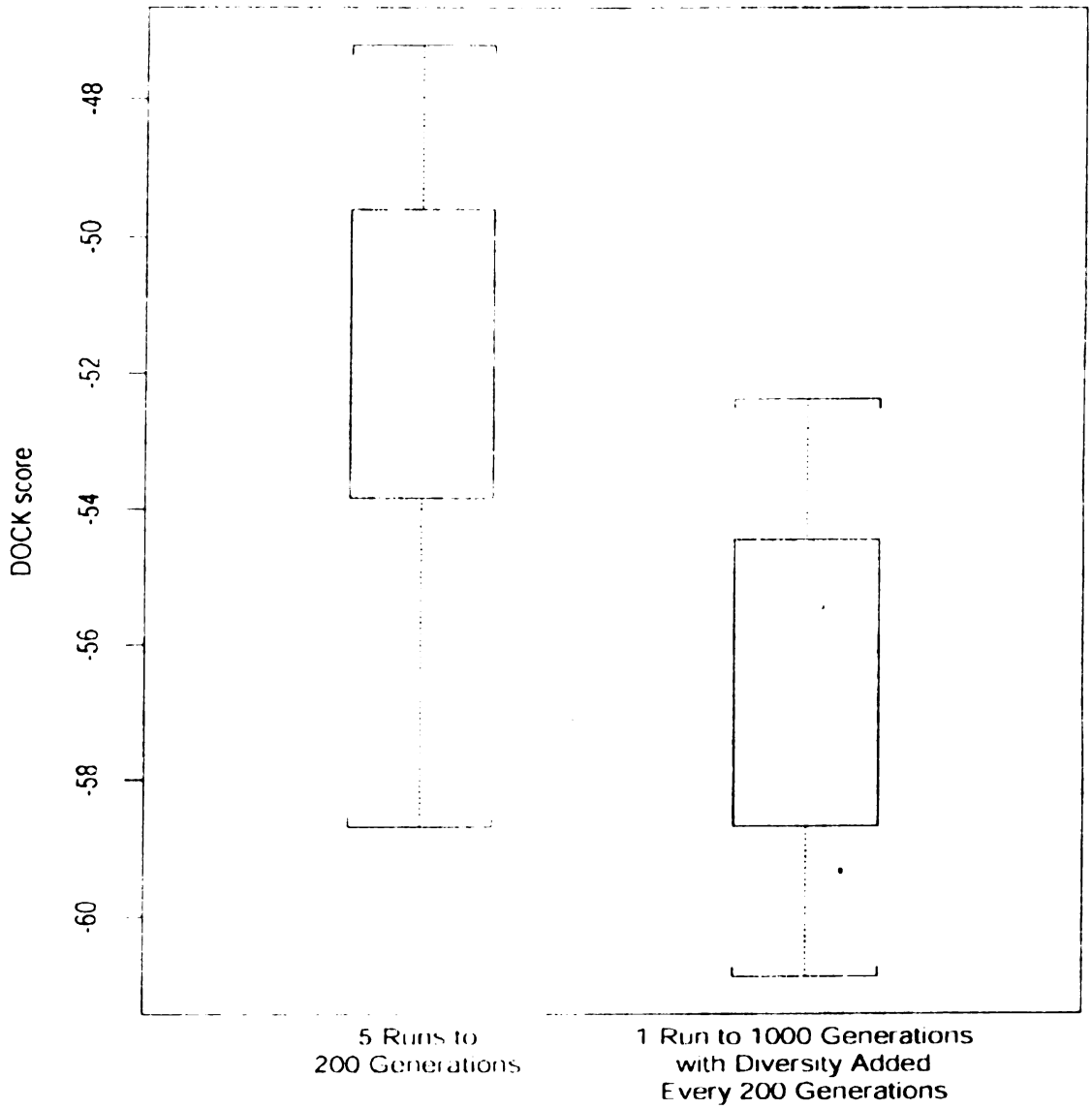


Figure 13. The distributions of DOCK scores for the compounds from the final generations of 5 runs to 200 generations and the compounds from the generations just before diversity was re-introduced in a run to 1,000 generations, adding diversity every 200 generations. Both distributions represent 100 compounds. The distance between the top and bottom of the shaded box represents the inter-quartile range, the dark line represents the median, and the brackets represent the extreme values of the distribution.

HIV-1 Reverse Transcriptase

The non-nucleoside inhibitors (NNI) of HIV Reverse Transcriptase (HIV RT) are a chemically diverse class of compounds that non-competitively inhibit enzymatic activity by binding to a specific site on the enzyme. Previous work indicates that a vast majority of these inhibitors adopt a 'butterfly-like' geometry when bound in the NNI pocket.[40,41] The axis of the butterfly's body is aligned almost parallel to the β9- β10 hairpin. Wing I is defined as the portion of the NNI binding site that lies closest to the polymerase active site, while the section that lies distant from the active site is wing II.[41] We applied the ADAPT program to the HIV-RT system in order to evaluate its ability to reproduce scaffolds with 'butterfly-like' shapes in their docked geometries, as well as its ability to rediscover known scaffolds.

We chose a model of HIV-1 RT complexed with 1-[(2-hyrdoxyethoxy)methyl]-6-(phyenylthio)thymine (HEPT) (pdb identifier 1RTI) as the target for our application. In this model HEPT adopts butterfly geometry in its bound state. The NNI binding site was characterized using *sphgen* and GRID as previously described.[32] Twenty-six chemically diverse non-nucleoside inhibitors were fragmented into rigid segments in order to form a library of 65 chemical fragments which ADAPT used to build compounds. With 65 fragments and 4 to 12 fragments per compound, a lower bound for the number of compounds that can be constructed using this library is approximately 5×10^{12}. Table 3 lists all the NNI's used to create the fragment library. Figure 14 shows how Thiocarboxyanilide UC781 was fragmented into its rigid fragments. We ran the ADAPT program ten times, using a population size of 30 and using a combination of DOCK score and a molecular weight score as the fitness function Other key parameters included the

use of crossover, mutation probability of 0.5, weight plateau boundaries of 214 to 415 (determined as the range of weights of known NNIs), and a weight plateau distribution with standard deviation 100. After ten runs a total of 5,250 compounds were generated.

In order to characterize the 'butterfly' geometry criterion, five other HIV-RT models complexed with different non-nucleoside inhibitors that adopt the butterfly shape were superimposed onto the 1RTI NNI binding site containing HEPT. The MidasPlus[42] program was used to superimpose 1BQM (HIV-1 RT complexed with Quinoxaline), 1DTT (HIV-1 RT complexed with a phenylethyl thioureathiazole (PETT) derivative), 1HNI (HIV-1 RT complexed with an α-APA derivative), 1REV (HIV-1 RT complexed with 8-Choloro TIBO), and 1VRT (HIV-1 RT complexed with Nevirapine) into 1RTI's reference frame. Thirty-one residues (G93a-P97a, L100a-L101a, L103a, S105a-T107a, V179a-Y183a, Y188a-S191a, P225a-W229a, G131a-D237a, Y318b) immediately surrounding the NNI binding site served as the primary reference for orienting all the models into the correct reference frame. The root-mean-squared-deviations for the 31 residues in 1BQM, 1DTT, 1HNI, 1REV , and 1VRT to 1RTI are 2.86, 1.85, 2.33, 1.56, and 1.78 angstroms, respectively.

Once superimposed, each atom in all six inhibitors was classified as belonging to either wing I or wing II of the butterfly shape. We performed the classification by dividing the butterfly along a plane containing N11 of Nevirapine, S' of HEPT, and the Cδ of L100a, and partitioning the heavy atoms of the ligands into each wing. Figure 15 is a 2D projection of the atoms onto a plane perpendicular to that used to partition the heavy atoms. The origin represents an end-on view of a vector connecting the N11 of Nevirapine and the S' atom of HEPT. The angle between best fit lines through all the

TNK-651	Thiadozyl Dialkylcarbamate RD 4-2024
α APA	Arylpyridothiodiazepine MEN 10979
8-Chloro TIBO (Trivirapine)	DABO 12e
BHAP U90152 (Delavirdine)	Hept Pyridinone Hybrid 18a
Pyridinone L-697,661	Benzyloxymethylpyridinone
Thiocarboxyanilide UC781	Alkyl(aryltho)uracil
Indole carboxamide L 737,126	Diarylsulfone NPPS
Benzathiodiazine-1-oxide NSA287474	Pyrryl Arylsulfone
Quinaxolinone 13a	Highly Substituted Pyrrole
Benzoxazinone DMP 266 (Sustiva)	MKC-442
Pyrrolobenzoxazepinone 4	Imidazodipyridodiazepine UK129,485
Imidazopyridazine 33	Phenylthiazolylthiourea (PETT)

Table 3. The 26 non-nucleoside inhibitors used to create the HIV-RT fragment library.

Parameter	Value
Population size	30
Fragments_per_compound_max	12
Fragments_per_compound_min	3
Use_scaffold	FALSE
Dock_fitness	TRUE
Num_dock_runs	1
Dock_fitness_coefficient	1.0
Clogp_fitness	FALSE
Weight_fitness	TRUE
Weight_range	214 to 415
Weight_standard_deviation	100
Weight_fitness_coefficient	-10.0
Rotatable_bond_fitness	FALSE
H_bond_fitness	FALSE
Breeding_fraction	0.5
Keep_parents	TRUE
Breeding_types	Crossover, mutation
Mutation_types	Connectivity,Identity
Mutation_probability	0.5
Mutation_lambda	1.0

Table 3. ADAPT parameter settings used in the HIV-1 Reverse Transcriptase experiments

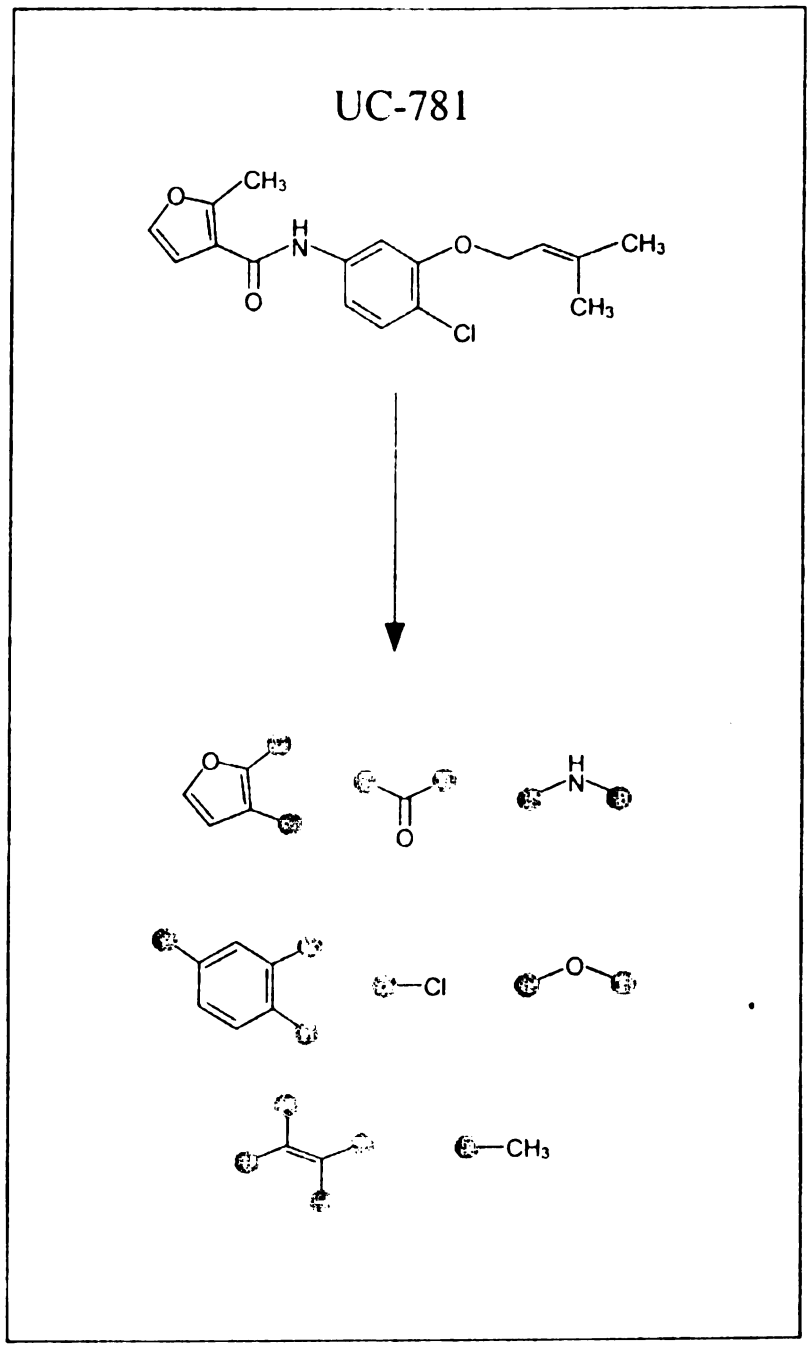


Figure 14. Decomposition of UC-871 into rigid fragments. A gray circle indicates an attachment point.

wings in each atom is 138°. In order to calculate the angle between the wings of each ligand, the positions of all atoms in each wing are averaged to yield a 'wing-point' description of the wing's size position. The angle between the wings is represented by the angle between the line connecting the wing-point in wing I to the origin, and the line connecting the origin to the wing-point in wing II. The angles for the six known non-nucleoside inhibitors range from 105°-167° (HEPT: 141°, Quinoxaline: 106°, PETT: 149°, α-APA: 157°, 8-Chloro TIBO: 167°, Nevirapine: 133°). In addition to facilitating angle calculations, the distance of each wing-point from the origin can be considered a measure of the relative size of each wing.

We analyzed the compounds generated by the ADAPT program in order to determine how well the compounds fulfill the butterfly geometry. To count as a butterfly compound, 50% of a ligand's atoms must fill both wings; this was done to eliminate those compounds that fill only one of the wings of the NNI binding site. 3,363 of the 5,250 (64.1%) compounds fulfilled this criteria. The atoms in the six known NNI's examined are all found within 3.5 angstroms of the plane used to divide the NNI binding site. Further, these same atoms can all be found within 2 angstroms of a best-fit plane through the atoms in each wing. We used these distance criteria to assign the atoms in each of the 3,363 compounds as belonging to either wing I or wing II of the NNI binding site. These atoms were used to calculate the wing-point representing the atoms of each wing in each compound. These points were projected onto the 2D plane perpendicular to the partition between the wings. The angle between the lines connecting each wing-point to the origin represents the ensemble average angle between the atoms in each wing of the compound.

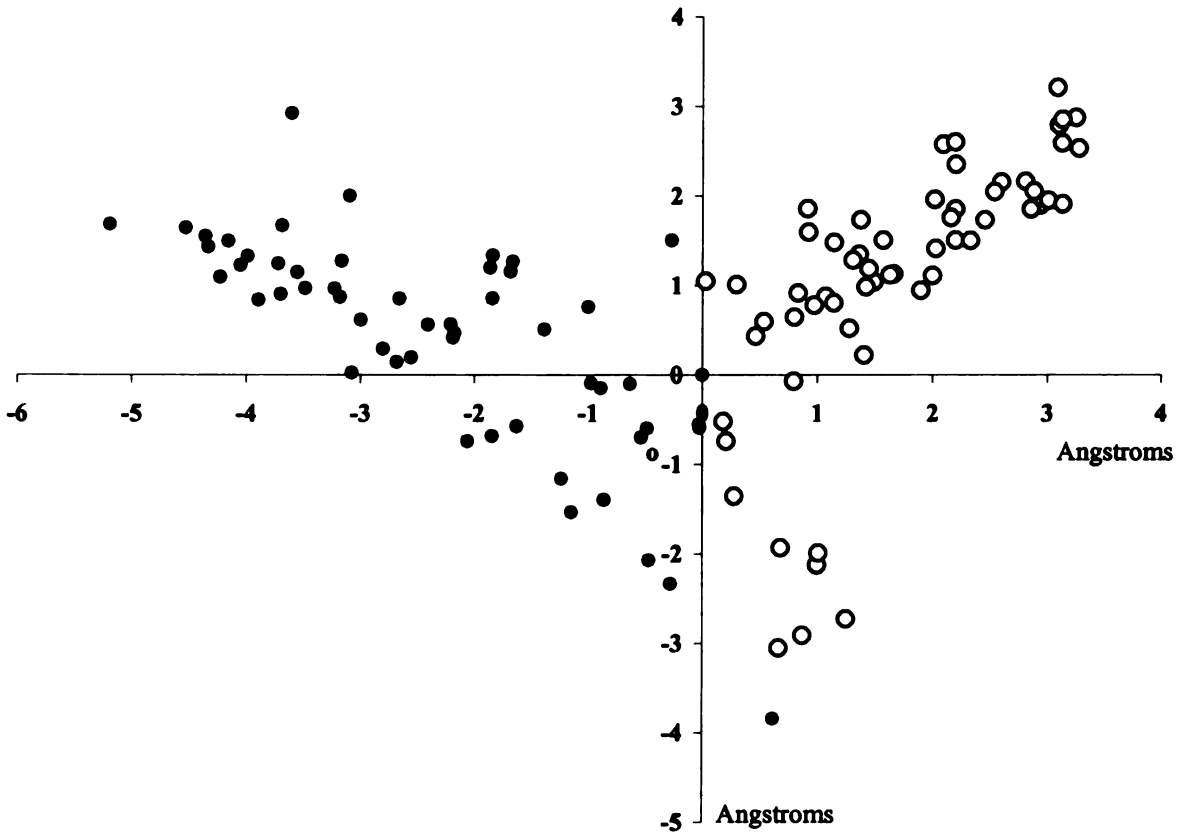


Figure 15. 2D projection of heavy atoms fro HEPT, Quinoxaline, PETT, α-APA, Nevirapine, and 8-Chloro TIBO bound into the NNI binding site. Filled in circles are atoms that occupy the wing I section, and empty circles occupy wing II. The origin represents a vector between N11 of Nevirapine and S1 of 8-Chloro TIBO. The angle between the linear best-fit lines through all the atoms in each wing is 138°.

Figure 16 is a 2D projection of all the wing-points; as before, the origin represents an end-on view of the vector connecting the N11 of Nevirapine and the S' of HEPT. Each point represents the average coordinate of a ligand's wing. Lines drawn through each wing intersect at an angle of 120°. This value agrees with the angle measured previously by Ding et al.[41] Figure 17 is a distribution of the angles in figure 16. The distribution has a mean of 142° and a standard deviation of 21°. Figure 18 is a plot of the angles between the wings as viewed against the distances of the wing points from the origin. The distance of a wing-point from the origin is a rough measure of a wing's size. The wings of a non-nucleoside inhibitor can adopt a range of angles, however the data shows that the angle between the wings decreases as the average wing size increase. This is to be expected since the geometric constraints of the binding site on the non-nucleoside inhibitors are more pronounced when the compounds occupy more space within the site.

The ADAPT program was able to reproduce four known NNI scaffolds. Effavirenz (Sustiva ™), Pyrrolobenzodiazepinone, PETT, Dyarryl Sulfone –like scaffolds were found among the butterfly like compounds (figure 19). Of particular note is the PETT-like compound generated by the algorithm. Its real-world counterpart, MSC-127, was the result of an extensive synthetic chemistry program that explored over 750 variations of the lead PETT scaffold.[43] Overall, the ADAPT program was able to reproduce a geometric constraint, the 'butterfly' motif of known NNI's from the use of a molecular docking fitness function. We also had some modest success in reproducing known NNI scaffolds.

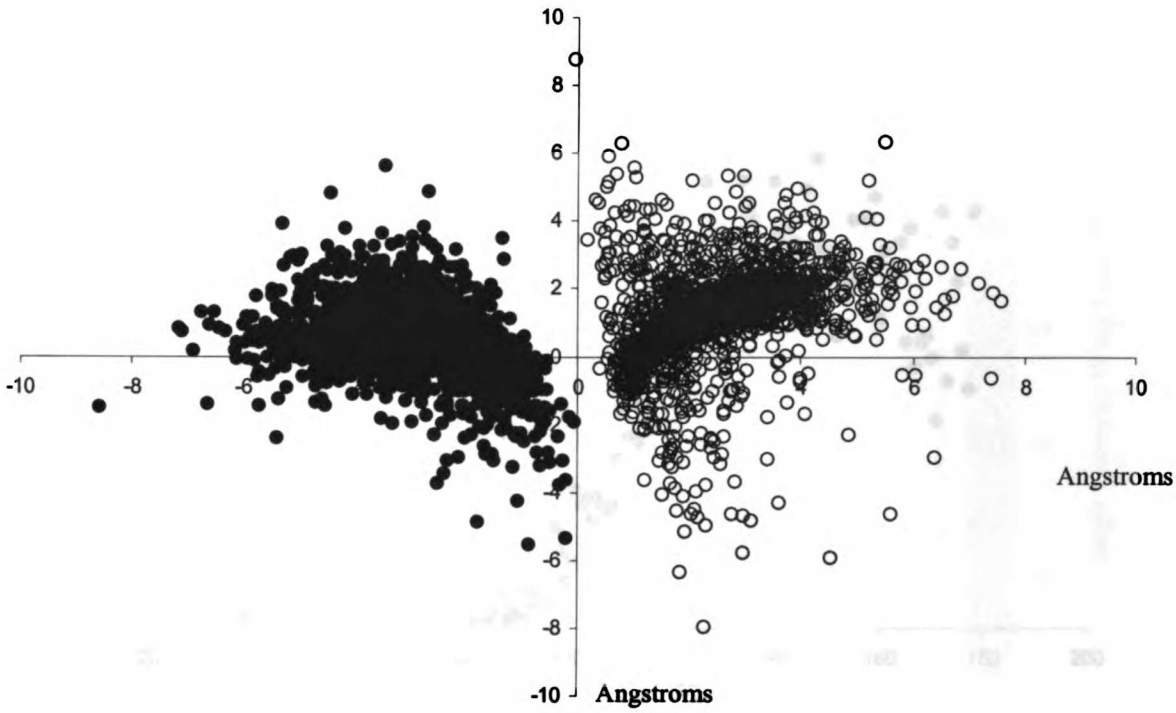


Figure 16. 2D projection of the wings of all ligands that fulfill the criteria of having a butterfly geometry. Filled circles are wing-points that occupy the wing I section, and empty circles occupy wing II. The angle between the best-fit lines through each wing is 120°.

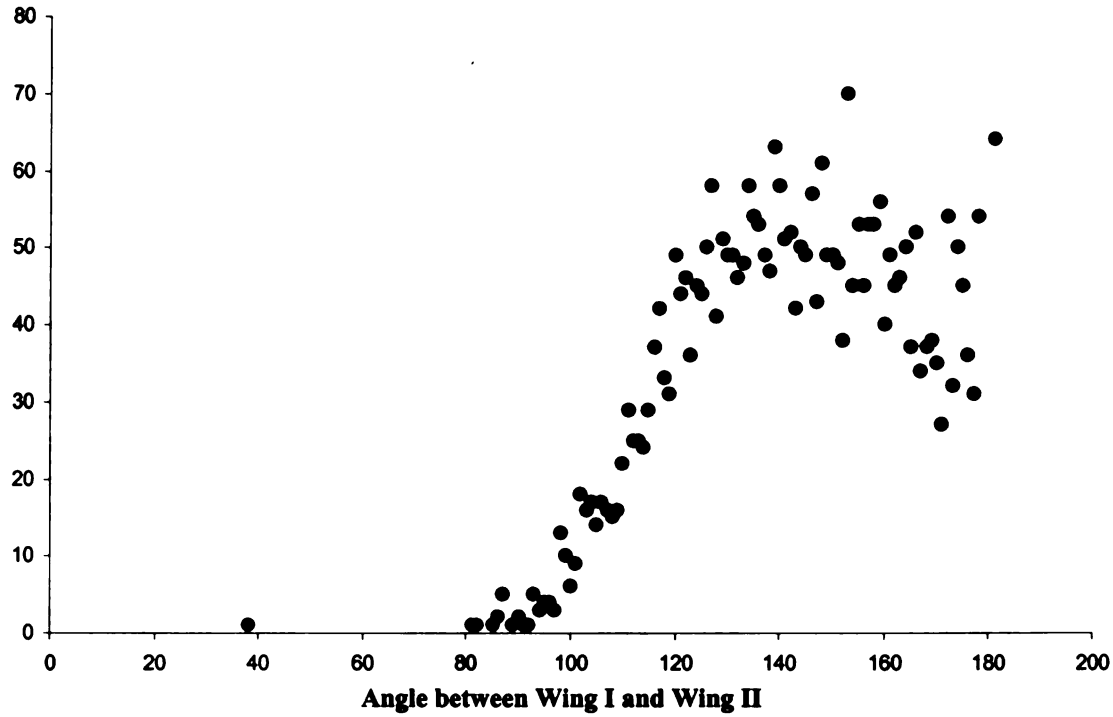


Figure 17. The distribution of angles between the wings of 3,363 compounds generated by the genetic algorithm that fulfill the butterfly geometry. The distribution has a mean of 142° and a standard deviation of 21°.

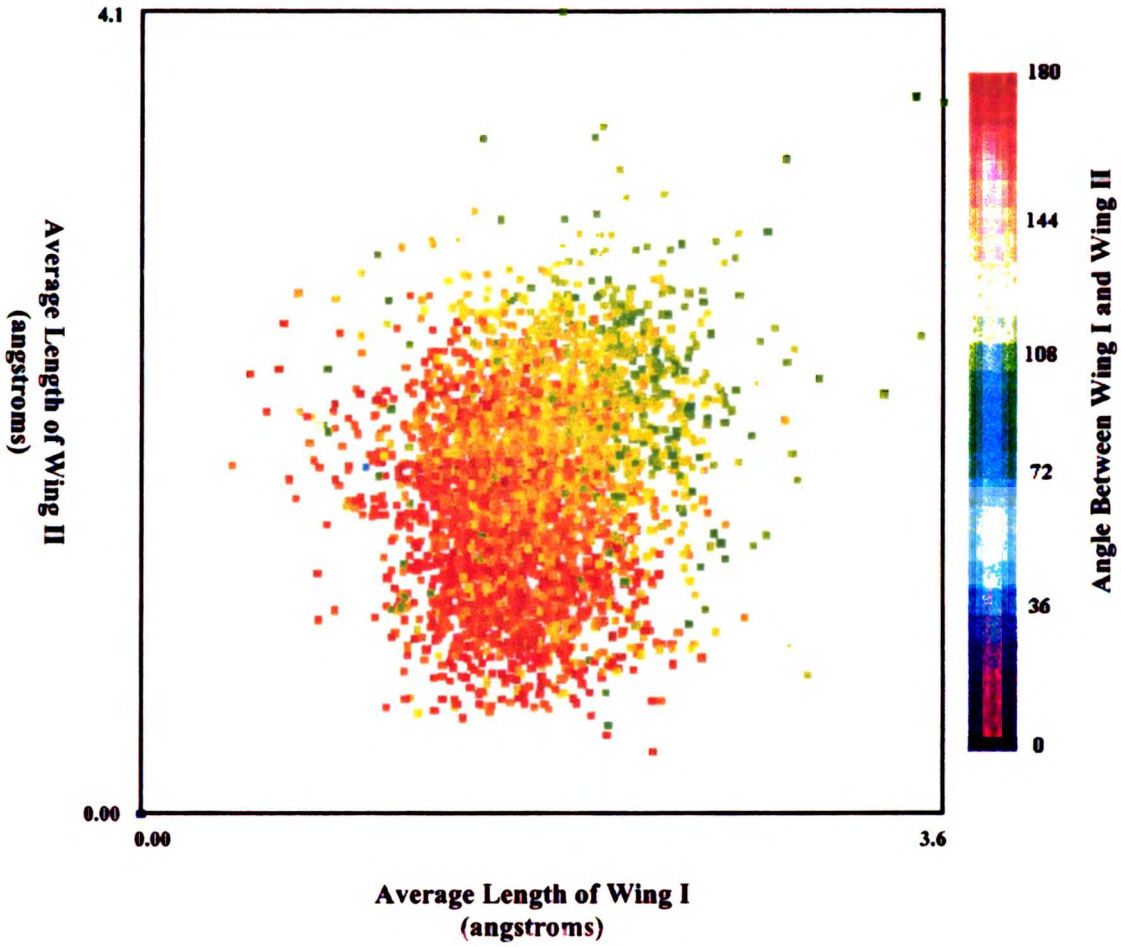


Figure 18. Angle between wing I and wing II as viewed against the distance of each wing-point from the origin. Distancvesd from the origin provide a rough measure of a wing's size. As the size of the wings increase, they are forced to adopt smaller angles.

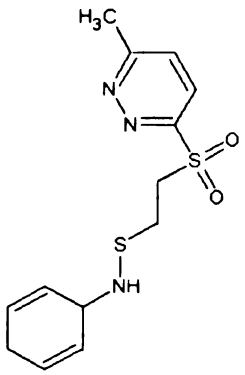
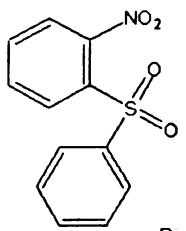
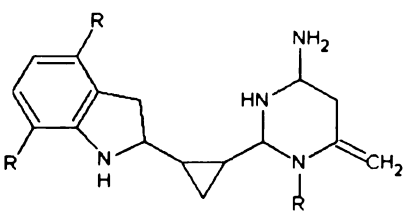
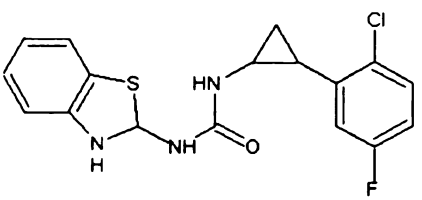
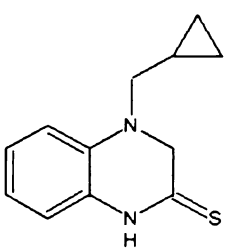
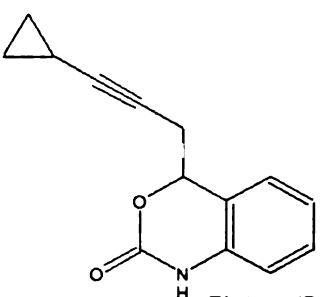
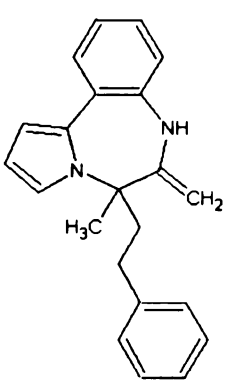
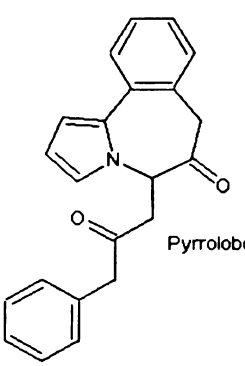
Algorithm Generated Compounds	Actual Non-nucleoside Inhibitors
	 Diarryl Sulfone
	 Optimized PETT (MSC- 127)
	 Efavirenz (Benzoxazinone, DMP 266) Sustiva
	 Pyrrolobenzodiazepinone

Figure 19. The algorithm recovered three of the known NNI inhibitors. Of particular interest was the generation of a compound similar to MSC-127 (second row). This compound was the result of an extensive synthetic optimization process that screened over 750 variations of the basic PETT scaffold.

Discussion

We have developed the ADAPT program, a genetic algorithm for structure-based de novo design which uses molecular docking as a fitness function. We discuss here the utility of design algorithms that use a binding estimate as a fitness function compared to the prior practice of a fitness function based on analysis of known ligands. We then consider the benefits and hazards of forced local sampling with the ADAPT program, and the difficulty in measuring the success of de novo design algorithms.

The ADAPT program shares the advantages of other genetic algorithms for de novo design in that it can optimize under many constraints simultaneously and provide a rich set of diverse solutions, however the use of molecular docking as a fitness function introduces both particular advantages and disadvantages. The major advantage is that information about known ligands to a particular active site is not required. This is an increasingly important advantage as genomic sequencing projects and post-genomic analysis produce a growing number of new protein targets for which no binding information is known. Another feature of avoiding the use of known ligands is an increased ability to discover more diverse compounds, a clear advantage in situations where there may be competing patent issues, or where it is desirable to explore backup candidates of different chemical classes. Information about known ligands can, of course, be useful. The flexibility of ADAPT allows us to explore the value of this additional information, as in our experiments with DHFR, where compounds based on methotrexate were readily generated.

The major disadvantages to using molecular docking fitness functions are that the calculations can be fairly slow, especially when they include ligand flexibility, and that

any errors in the scoring function are propagated into the compound generation process. The computational overhead forces the population sizes of the ADAPT program to be relatively small and thus may cause us to loose some sampling ability. The genetic algorithm of Sheridan et. al. opts for a larger population and fewer generations. The optimal settings of parameters such as population size remains an area of open research in genetic algorithm theory, with some arguing for high population sizes[44] and others for small populations, especially when the cost of fitness evaluation is high.[45,46] As with most genetic algorithms, ADAPT spends the vast majority of its time evaluating the fitness function. Even when docking with relatively low orientational sampling, it took approximately 5 hours to complete 100 generations with a population of size 20 on an Silicon Graphics R10000 processor (the housekeeping of the genetic algorithm, including crossover and mutation, took less than 10 seconds per generation.) Scoring errors from the docking calculations may neglect certain important molecular interactions (e.g. hydrogen bonding, desolvation[47,48]) and produce compounds that "score" well, but do not necessarily have good experimental binding affinities. Genetic algorithms based on pharmacophore and template models suffer from essentially the same problem, however. Fortunately, the easy of programming genetic algorithms means that as new scoring functions become available, they can be readily implemented and tested.

Fitness functions using molecular docking calculations can be designed to take advantage of information not readily available in fitness functions based on pharmacophore or molecular templates. One can, for example, build a fitness function that advocates specificity by rewarding compounds that dock well into the active site of one target while penalizing compounds that dock well into the active site of another

target. One could also search for compounds that dock well to entire families of targets by favoring only compounds that bind to multiple active sites.

Another advantage of the ADAPT program is the ability to let the user to specify not only the specific fragment set used to build the compounds, but also the fragment attachment points. This feature allows the user a fair amount of control over the possible chemistry and can help to avoid the generation of synthetically difficult compounds. Combinatorial chemistry schemes, such as that used in our experiments with cathepsin D, are a clear example of this utility.

Our experiments with the cathepsin D system also show the advantage of forcing a genetic algorithm to sample locally in chemical space. This makes sense, because it leads to focused sampling near optimal solutions. We enforced local sampling by only allowing mutations of fragment types to fragments similar in chemistry. Our implementation of this is somewhat crude, however, and consideration of alternate strategies such as binning fragments such that fragments can only be mutated or swapped with fragments within the same bin is part of future work. Sheridan and Kearsley demonstrated a performance improvement in their genetic algorithm through the use of a type of binning scheme.[25] It may also prove fruitful to combine the genetic algorithm search strategy with a more localized branch-and-bound style search.

Forcing mutations to be local in nature not unexpectedly increases the chance of the optimization process being trapped in local minima. We have found that adding diversity back to the population during a run of the ADAPT program can help alleviate this by moving the algorithm to a random, but not too distant place in the search space. Repeatedly optimizing from these semi-diverse points can lead to generating good

scoring compounds faster than starting repeatedly from a completely diverse population. Clearly there may be an optimal amount of diversity that should be added to provide a good restarting point—too little diversity puts the algorithm back where it started and too much diversity forces the algorithm to weed through many poor scoring structures. This optimal amount of diversity may be system invariant and finding it is a consideration in future work.

Measuring the success of any de novo design program is difficult, primarily because the number of unique compounds that can be generated is much larger than the number of compounds for which we have experimental binding data. One possible measure of success is to see how easily an algorithm can reproduce a known ligand that has been shown to bind experimentally. In this case the ADAPT program poorly in the cathepsin D system, and has only limited success with HIV-1 RT and DHFR where the chemical search space was deliberately much less restricted. In the DHFR experiments, however, ADAPT produced compounds with fitness better than the known inhibitor, indicating that our failure to reproduce the known inhibitor was at least partly the result of the fitness function not having the known inhibitor as a minimum. Another possible measure of success for a de novo design program is whether it can reproduce the trends seen in known ligands. In this case the ADAPT program does fairly well in identifying the 'fused-ring plus chain' motif seen in DHFR ligands and the 'butterfly' motif seen non-nucleoside HIV-1 RT inhibitors, and does very well in identifying the side chain fragments most likely to produce good inhibitors in the cathepsin D combinatorial chemistry system. The difference in ability between finding known ligands and finding

trends appears to be mostly the result of the DOCK score being an inexact measure of true binding affinity.

Ideally we would like to be able to measure the success of a de novo design algorithm by being able to accurately (and rapidly) compute the binding affinities of compounds produced by the algorithm and simply analyzing their distribution. Binding affinity is often not the only quality criterion of the user, however, and success is likely to be measured differently under different conditions—there is no definitive magic formula of properties for good drug candidates. Nevertheless, genetic algorithms are quite attractive in their abilities to construct diverse solutions to problems with large search spaces and many constraints. As ligand affinity scoring methods (and those which consider other important properties in the drug development process) improve, genetic algorithms (or hybrid variants) for structure based de novo design will become more valuable and subject to greater scrutiny as strictly optimization methods in comparison to other de novo design algorithms.

Acknowledgments

We thank Geoff Skillman, Connie Oshiro, Jim Arnold, and Ken Foreman for helpful discussions. This work was funded by NIH Biotechnology Training Grant GM08388 (SP) and NIH General Medical Sciences Grant GM31497 (IDK and JJH).

References

1. Walters, P. W., Stahl, M., Murcko, M. A. Drug Des. Today, 3 **(1998)** 160.
2. Goodford, P. *J. Med Ch*em, 28 **(1985)** 849.
3. Miranker, A., Karplus, M. *Proteins*, 11 **(1991)** 29.
4. Eisen, M. B., Wiley, D. C., Karplus, M., Hubbard, R. E. *Proteins*, 19 **(1994)** 199.
5. Miranker, A., Karplus, M. *Proteins*, 23 (b) 472.
6. Pearlman, D. A., Murcko, M. A. *J. Med Chem*, 39 (b) 1651.
7. Bohm, H.-J. *J. Comput.-Aided Mol Design*, 6 (b) 61.
8. Lauri, G., Bartlett, P. *J. Comput.-Aided Mol Design*, 8 **(1994)** 51.
9. Roe, D. C., Kuntz, I. D. *J. Comput.-Aided Mol Design*, 9 **(1995)** 269.
10. Lewis, R. A., Leach, A. R. *J. Comput.-Aided Mol Design*, 8 **(1994)** 467.
11. DeWitte, R., Shakhnovich, E. *J. Am Chem Soc*, 118 **(1996)** 11733.
12. Bohacek, R. S., McMartin, C. *J. Am Chem Soc*, 116 **(1994)** 5560.
13. Rotstein, S. H., Murcko, M. A. *J. Comput.-Aided Mol Design*, 7 **(1993)** 23.
14. Rotstein, S. H., Murcko, M. A. *J. Med Chem.*, 36 **(1993)** 1700.
15. Moon, J. B., Howe, W. J. *Proteins*, 11 **(1991)** 314.
16. Joseph-McCarthy, D. *Pharmacol Therapeutics*, 84 **(1999)** 179.
17. Lipinski, C. A., Lombardo, F., Dominy, B. W., Feeny, P. *J. Adv Drug De. Rev*, 23 **(1997)** 3.
18. Schneider, G., Lee, M.-L., Stahl, M., Schneider, P. *J. Comput.-Aided Mol Design*, 14 **(2000)** 487.
19. Westhead, D. R., Clark, D. E., Frenkel, D., Li, J., Murray, C. W., Robson, B., Waszkowycz, B. *J. Comput.-Aided Mol Design*, 9 **(1995)** 139.

20. Glen, R. C., Payne, A. W. R. *J. Comput.-Aided Mol Design*, 9 **(1995)** 181.

21. Douguet, D., Thoreau, E., Grassy, G. *J. Comput.-Aided Mol Design*, 14 **(2000)** 449.

22. Devillers, J. *J. Chem. Inf Compu. Sci*, 36 **(1996)** 1061.

23. Weber, L., Wallbaum, S., Broger, C., Gubernator, K. Angew. Chem. Int. Ed. Engl., 34 **(1995)** 2280.

24. Blaney, J. M., Dixon, J. S., Weininger, D., Molecular Graphics Society Meeting on Binding Sites: Characterising and Satisfying Steric and Chemical Restraints, Weininger, D. (Ed.), 1993, York, U.K., pp.

25. Sheridan, R. P., Kearsley, S. K. *J. Chem. Inf Compu. Sci*, 35 **(1995)** 310.

26. Sheridan, R. P., SanFeliciano, S. G., Kearsley, S. K. *J. Mol Graph Model*, 18 **(2000)** 320.

27. Holland, J. H., Adaptation in Natural and Artificial Systems; University of Michigan Press, Ann Arbor, MI, **1975**.

28. Holland, J. H., Adaptation in Natural and Artificial Systems; 2nd ed.; MIT Press, Cambridge, MA, **1992**.

29. Darwin, C., The Origin of Species; Dent Gordon, London, **1973**.

30. Weininger, D. *J. Chem Inf. Comput Sci*, 30 **(1990)** 237.

31. Daylight Toolkit, v. 4.6, Daylight Chemical Information Systems, Inc., Santa Fe, NM, USA.

32. Ewing, T. J. A., Kuntz, I. D. *J. Comp Chem*, 18 **(1997)** 1175.

33. CONCORD, Tripos, Inc., St. Louis, MO.

34. SYBYL, v6.5, Tripos, Inc., St. Louis, MO.

35. Kick, E. K., Roe, D. C., Skillman, G. A., Liu, G., Ewing, T. J. A., Sun, Y., Kuntz, I. D., Ellman, *J. Am Chem & Biol*, 4 **(1997)** 297.

36. Queener, S. F., Bartlett, M. S., Jay, M. A., Durkin, M. M., Smith, J. W. *Antimicrob. Agents Chemother.*, 31 **(1987)** 1323.

37. Bertino, J. R. *J. Clin. Pharmacol*, 30 **(1990)** 291.

38. Blaney, J. M., Hansch, C., Silipo, C., Vittoria, *A. Chem. Rev.*, 84 **(1984)** 333.

39. Verkhivker, G. M., Rejto, P. A., Bouzida, D., Arthurs, S., Colson, A. B., Freer, S. T., Gehlhaar, D. K., Larson, V., Luty, B. A., Marrone, T., Rose, P. *J. Mo. Reco.*, 12 **(1999)** 371.

40. De Clerq, E. *Antiviral Res*, 38 **(1998)** 153.

41. Ding, J., Das, K., Moereels, H., Koymans, L., Andries, K., Janssen, P. A., Hughes, P. A., Arnold, E. *Nat Struct Biol*, 2 **(1995)** 407.

42. Ferrin, T. E., Huang, C. C., Jarvis, L. E., Langridge, R. *J. Mol Graph*, 6 **(1988)** 13.

43. Artico, M. *Il Farmaco*, 51 **(1996)** 305.

44. Goldberg, D. E., Third International Conference on Genetic Algorithms, Schaffer, J. D. (Ed.), 1989, Fairfax, VA, Morgan Kaufmann, June 4-7, **1989**, pp. 70.

45. Grefenstette, J. *J. IEEE Transactions on Systems, Man, and Cybernetics*, 16 **(1986)** 122.

46. Schaffer, J. D., Caruana, R. A., Eshelman, L. J., Das, R., Third International Conference on Genetic Algorithms, Schaffer, J. D. (Ed.), 1989, Fairfax, VA, Morgan Kaufmann, June 4-7, **1989**, pp. 51.

47. Caflisch, A., Karplus, M. *Pers Drug Disc Design*, 3 **(1995)** 51.

Gloss to Chapter II

This chapter describes the design, optimization, and validation of a small molecule mimetic of a beta sheet peptide. This project was collaboration between myself and Naoaki Fuji and Kathleen Pendol from R. Kip Guy's laboratory. The object was to discover a scaffold capable of mimicking the binding properties of a four amino acid long beta peptide that bound to PDZ domains.

In general, PDZ domains mediate subcellular organizing of protein complexes bringing proteins with various functions to one locus. PDZ domains were originally described as conserved structural elements in the 95 kDa post-synaptic density protein, PSD-95; the Drosophila tumor suppressor, Discs-large Dlg; and the tight junction protein, zonula occludens-1 ZO-1. The domains are small, and often occur multiply in adaptor proteins and with other protein interaction domains. They are important modular protein interaction domains in a wide variety of eukaryotic organisms. The central role of PDZ domains in mediating cell signaling and cell-cell interaction is emerging in diverse fields from neuroscience to protein trafficking to cancer.

The intended use of this chemical scaffold was a chemical genetics approach to deciphering the functions of specific PDZ domains. Chemical genetics is an approach that uses small molecules to regulate the expression of certain gene products, instead of attempting to regulate the genes themselves through mutagenesis methods. Chemicals have some advantages over genetic technologies: they are versatile research tools that can be rapidly adopted by many labs and used for precise control of protein function in cells where genetic manipulation is difficult.

This project grew out of a desire to apply some of the principles and algorithms we focus on here in the lab towards a practical problem. It was originally meant to be a side-project, but soon took on a life of its own. While the traditional approaches (database screening) were implemented in the outset, it was *de-novo* approach that led to the design of the scaffold. I used of tried and true modeling tools and basic observation to design a pteridine-based scaffold that Naoaki Fuji was extremely confident in. Working with a truly talented and insightful medicinal chemist was very fulfilling, and brought a much-needed sense of reality to the work I was doing in our laboratory. I optimized the scaffold in-silico based on chemical feasibility, as well as the scaffolds structural similarities to known inhibitors. Similarly, the scaffold progression from simple drawings to an assay validated compound with measurable activities and good SAR helped me attain a singular goal most computational graduate students have: to start a project and see our approaches and techniques validated with traditional wet-lab methods, and thus tell a 'complete' design story. It was by far one of the most fulfilling projects I have been involved in. Of course, the fact that this is really the first small molecule compound to target these domains adds a certain level of pride in this work.

The content in this chapter was written originally for submission to the *Journal of the American Chemical Society*. This version the manuscript is the result of a first-draft synthesis of everyone's methods built on an introduction, structure-based design discussion, and conclusion that I wrote. It has undergone numerous revisions since then by a number of people working on the project, and remains a work-in-progress. The breadth of techniques and general relevance prompted Kip to strive for a publication in a journal with wider readership. At the time of writing this thesis, additional cell-based

assay data was not available yet. A favorable cellular activity profile will add another layer of richness to this project, and enable us to pursue another series of publications that reach a wider audience. In late-2002, this manuscript was redacted for submission to the journal *Nature*.

CHAPTER II

~•~

De-novo Structure-Based Design, Synthesis, and Evaluation of a Novel Non-peptide Indole Scaffold as a β strand Mimetic Ligands to PDZ Domains

~•~

Naoaki Fuji, Jose J Haresco, Kathleen Pendol, Irwin D. Kuntz, R. Kip Guy

Originally written for the Journal of the American Chemical Society.
Later rewritten for submission to Nature

Abstract

PDZ domains have been found to mediate critical protein interactions that enforce localization and organization of proteins in a variety of submembranous protein complexes. The ability to modulate the functions of PDZ domains would greatly enhance our knowledge of mechanisms of cellular signaling that depend on these actions. To date, there have been no reports of non-peptide small molecules that bind specifically to PDZ domains. We describe herein the structure-based design, synthesis, and in-vitro evaluation of novel, non-peptide small molecules specifically targeted to the ligand-binding pocket of PDZ domains. Molecular scaffolds were designed from the co-crystal structure of the PSD-95 PDZ3 domain complexed with its native peptide ligand, CRIPT (KQTSV). Iterative molecular docking and modeling resulted in an optimized scaffold that formed the basis for the organic synthesis of several scaffold with consideration for the previously reported binding specificity of the MAGI-3 PDZ2 domain. Following synthesis, these compounds were evaluated for affinity by an in vitro fluorescence polarization competition assay. One of these molecules was found to bind with an IC_{50} in the low μM range. This application effectively demonstrates the utility and efficiency of structure-based design methods for use in the development of small molecule inhibitors of protein interaction. In addition, this work suggests the possibility of developing specific small molecule inhibitors for class I PDZ domains.

Introduction

PDZ domains were originally described as conserved structural elements in the 95 kDa post-synaptic density protein, PSD-95; the Drosophila tumor suppressor, Discs-large Dlg; and the tight junction protein, zonula occludens-1 ZO-1. The domains are small, usually of 80 amino acids, and often present multiply in adaptor proteins and with other protein interaction domains. They have emerged as important modular protein interaction domains in a wide variety of eukaryotic organisms. In general, PDZ domains mediate subcellular organizing of protein complexes bringing proteins with various functions to one locus. They have been found to be associated with ion channels, transmembrane receptors, structural proteins, and regulatory enzymes.[1-4] The central role of PDZ domains in mediating cell signaling and cell-cell interaction is emerging in diverse fields from neuroscience to protein trafficking to cancer.

For example, the membrane bound guanylate kinase PSD-95 localizes and clusters the N-methyl d-aspartate (NMDA) receptor and the Shaker potassium channel in the post-synaptic density of neurons.[5,6] In some cases, multiple PDZ domains mediate local interactions. For example, multiple proteins with PDZ domains localize LET-23, a receptor tyrosine kinase of Caenorhabditis elegans[7] crucial for the development of the vulva. Mutations of these PDZ domains result in a lack of vulval differentiation. Recently, NHERF (Na+/H+ exchanger regulatory factor), which contains two PDZ domains, has been shown to be involved localization and/or turnover of G-protein coupled receptors, platelet-derived growth factor receptor, and ion transporters.[8]

The MAGI's (membrane associated guanylate kinase of inverted orientation) are a small subfamily of the MAGUK family of adaptors that are widely expressed in various

human tissues and have six PDZ domains.[9] Some PDZ domain interactions of the MAGI have been identified and their roles are currently under investigation. An interesting interaction of the MAGI-2/3 adaptors is between PDZ2 and PTEN, a tumor suppressor.[10,11] *PTEN* encodes a lipid and protein phosphatase with a carboxy terminal PDZ domain recognition sequence that antagonizes the activity of the phosphotidylinositol 3-kinase/Akt signaling pathway. Mutations of PTEN causes down regulation of this pathway and has been associated with various human cancers.[12-14] The interaction of MAGI-2/3 and PTEN has been shown to greatly enhance regulation of Akt/PKB kinase activity by allowing the formation of a PTEN associated complex. Disruption of this interaction would be a unique way to investigate the role of the interaction in Akt signaling.

In general, PDZ domains recognize carboxy terminal peptide ligands through a conserved peptide-binding groove.[15-17] Upon binding, the ligand forms a new antiparallel β strand, with the peptide ligand making backbone contacts with β strand B in the PDZ domain.[18,19] Structural data show that the ligand adopts a specific conformation within the binding pocket implicit by the degree of order of the first five amino acids of the ligand. PDZ domain primary structure and ligand consensus sequences have been used to segregate PDZ domains into classes: class I domains recognize the sequence T/S-X-V/I-COOH, class II domains recognize F/Y-X-F/V-COOH, and class III PDZ domains recognize E/D-X-V-COOH. It is thought that these conserved ligand residues make contacts that are required for binding with the class of PDZ domain.

Analysis of structural data from co-crystals has revealed much of the atomic detail of the interaction between ligand and PDZ domain. The free carboxylate of the terminal

residue forms an important basis for ligand recognition. The Gly-Leu-Gly-Phe (GLGF) sequence common to class I PDZ domains forms the carboxylate binding loop and lies between the β strands A and B, shrouding a conserved arginine or lysine.[20-22] In addition, the carboxylate makes water mediated contacts with the conserved basic residue. The binding of the free carboxylate of the ligand in the carboxylate-binding loop of the PDZ domain directs the side chain of residue (0) to dip into a cavity formed by conserved amino acid residues in the PDZ domain.[21,22] For most PDZ domains, this pocket is lined with hydrophobic residues thus explaining why residue (0) in most PDZ ligands is hydrophobic. The size of the pocket may account for differences in the nature of the hydrophobic side chain with individual PDZ domains showing specificity for valine, isoleucine, phenylalanine, leucine, or alanine.[23]

PDZ domains generally display a greater selectivity for amino acids upstream of residue (0). Residue (-2) has been shown by crystallographic and mutational analysis to be particularly important in PDZ binding and helps determine the PDZ domain class.[23] Class I PDZ domains include a highly conserved histidine (H372) in the first position of α helix 2 which has been shown to hydrogen bond with residue (-2), usually dictating that this residue be a serine or threonine.

Mutational analysis has indicated that the identity of residue (-1) helps determine ligand specificity for individual PDZ domains within a class. Crystallographic studies have failed to explain these observations with available crystal structures not showing consistent contacts between the side chain of this residue and the PDZ domain. However, hydrophobic contacts have been seen in the structures of NHERF and syntrophin with peptide ligands. Therefore, it is predicted that residues on the PDZ domain β strand C

might impart specificity and add affinity through contacts with residue (-1) of the ligand. Also, modeling studies have indicated that interactions between this side chain and residues in β strands B and C might be responsible for imparting tight affinity when optimized.[24] Similarly, residue (-3) shows contacts to the PDZ domain in some structures, but the nature of the interaction has not been defined across PDZ classes. This may indicate that this residue confers added specificity for discrimination between PDZ domains. While it has been shown that minimally four residues will bind specifically to PDZ domains, residues further amino terminal have an indistinct role in binding and lending specificity.

Because of the structural similarity between several reported structures of PDZ domains, both with and without bound ligand, we considered it feasible to design novel ligands to MAGI-3 PDZ2 domain by working with the structure of PSD-95 PDZ3 domain bound to KQTSV.[18] An ideal small molecule mimetic of the peptide ligand must be able adopt a conformation similar to that of the native ligand and present critical side chains to the domain in similar orientations. Molecules designed using structure as a starting point for the rational design process have an advantage in that they are inherently directed at the biologically relevant forms of their target, and are capable of successfully binding to the desired targets. Bolin and co-workers used a similar approach in the rational design of inhibitors of inhibitors of antigen presentation by HLA-DR class II MHC molecules.[25] Furet et al. used extensive crystallographic data to design a high-affinity antagonist of the Grb2-SH2 domain.[26] We report here the structure-based design, organic synthesis, and in-vitro evaluation of a family of compounds based on a substituted 1,2,6,7 – indole scaffold designed to bind MAGI-3 PDZ2 domain.

Methods

Structure Based Design of Ligands to PSD-95 PDZ3

Molecular docking as implemented in DOCK explores the degrees of freedom in orientation and conformation of a small molecule in the context of a target receptor.[39,40] The process of molecular docking begins with a characterization of the solvent accessible surface of the receptor, followed by the generation of a negative image of the active site. Subsequently electrostatic, Van der Waals, and atomic contact information about the target site is generated and stored for future reference. DOCK then explores numerous orientations of the molecule and calculates the receptor-molecule binding energy for each placement. Recently, Zou and coworkers developed a version of DOCK (sDOCK) that uses the generalized Born (GB/SA) model of solvation to calculate receptor-ligand binding energies.[41] The GB/SA approach allows for the estimation of electrostatic, Van der Waals, and, additionally, hydrophobic contributions to the free energy of binding. All DOCK experiments were carried out using an SGI R10000.

The crystal structure of the third PDZ domain of PSD-95 in complex with CRIPT (protein data bank file 1BE9)[18] was used for the design of a chemical scaffold based on a 1,2,6,7 – substituted indole. To prepare the PDZ domain for docking, most crystallographic waters were removed from the crystal structure. The water molecule coordinated with R318 and the ligand carboxylate was retained in the crystal structure.[18] The solvent accessible molecular surface of the ligand binding groove, as defined by Connolly,[42] was characterized using the *dms*[43] program included with MidasPlus (MidasPlus molecular modeling software, UCSF Computer Graphics Laboratory). The program *sphgen*[39,44] was used to generate a negative image of the binding groove

consisting of spheres of various size. The heavy atoms of the β strand were then added to the sphere set, and the total sphere set was reduced by removing any spheres with a distance greater than 2.5 angstroms from the center of mass of the native structure. This resulted in a final set of 127 spheres. Gasteiger charges[45] and hydrogens were added to the receptor using the Amber 95 force field, as implemented in SYBYL 6.5,[46] and electrostatic, Van der Waals, and contact grids were generated using the *grid* program. During the scaffold optimization phase, large numbers of scaffold derivatives needed to be built. An in-house Python script was used to generate SMILES[47] strings. *CONCORD* (as implemented in Sybyl6.5) was subsequently used to convert the 2D descriptors to 3D molecules and add hydrogens and Gasteiger charges.

Medicinal Chemical Analysis and Optimization of Ligands to PSD-95 PDZ3 and MAGI-3 PDZ2

The scaffold was optimized by evaluating other ring systems and making substitutions at selected sections of the indole ring (Figure 1). Figure 2 shows the various substitutions that were evaluated. Decisions regarding the types of substitutions were based on synthetic feasibility and previously reported binding interactions of the MAGI-3 PDZ2 domain. Each position on the indole was optimized independently to ensure the synthetic feasibility of the molecules, and combinations of fragments were evaluated for synthetic feasibility. Lastly, sDOCK was used to correct for solvation effects.

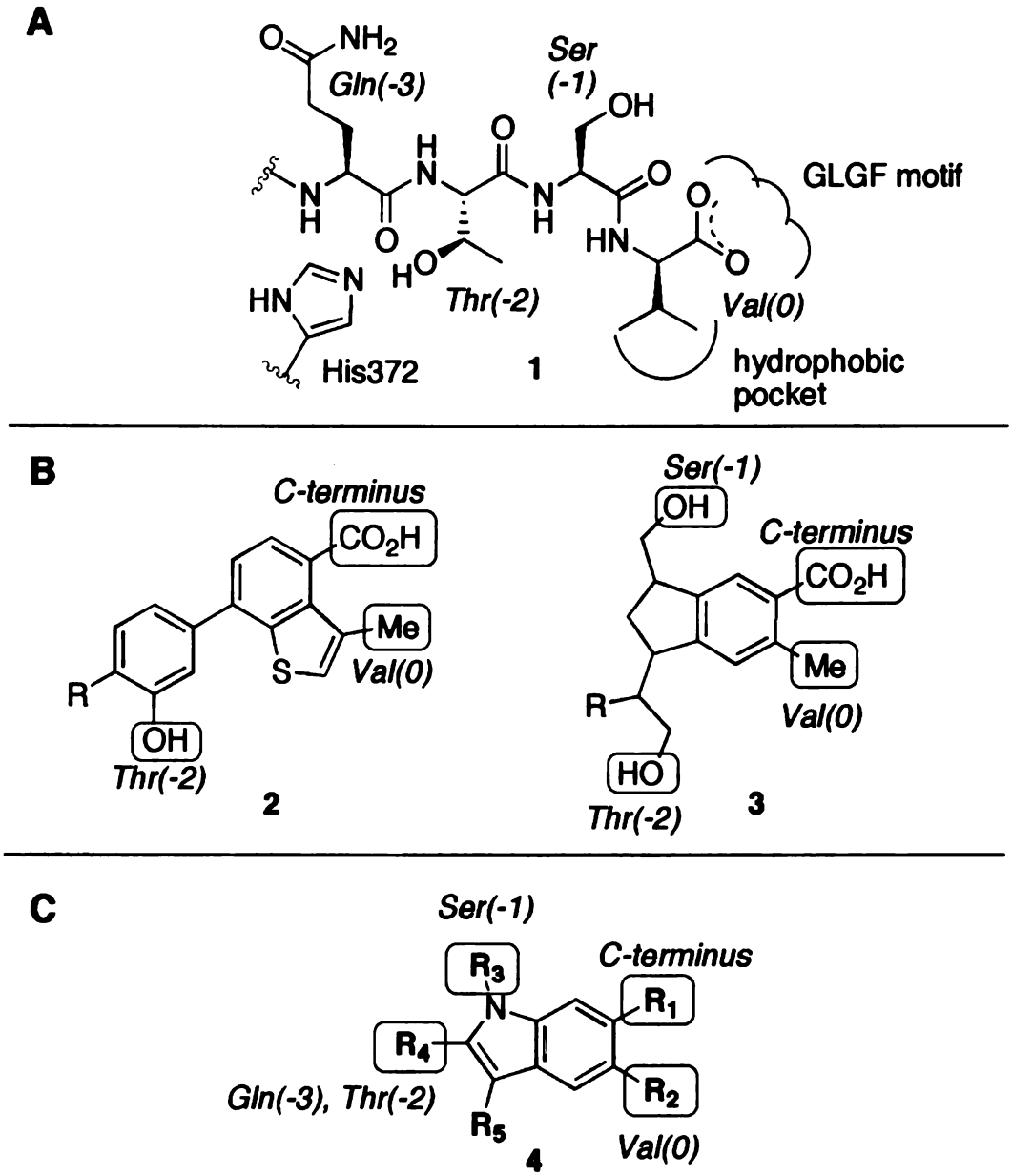
Biochemical Evaluations

This section summarizes Kathleen Pendol's biochemistry work on the project. A fluorescence polarization competition assay was used to detect binding of the compounds to MAGI-3 PDZ2. The PDZ domain was expressed as a GST construct using the Pharmacia Biotech pGEX4.1 vector. Plasmid containing BL21 DE3 cells were grown in 1x LB media and induced with 1M IPTG at $O.D._{600} = 1$ and harvested after growing for 4 hours at 37°C shaking at 250rpm. The protein was purified through a glutathione sepharose bead slurry, eluted with 50mM glutathione, and concentrated and dialyzed through 10000 MW Waters centricon filter. Protein was then quantified by Bradford assay (Coomassie Plus-200 Protein Assay Reagent by Pierce) and BCA assay (bicinchoninic acid reagents by Pierce). Protein identity and purity were checked by SDS PAGE and western blot against anti GST antibody.

A flourescein labeled carboxy terminal sequence of PTEN, OregonGreen™-PFDEDQHTQITKV-COOH, was used as a probe. For a positive control, we chose PFDEDQHTQITWV-COOH, the highest affinity peptide sequence for MAGI-3 PDZ2 known.[24] To synthesize the labeled peptide we used standard Fmoc conditions on Wang resin to build a 13 residue peptide. A typical coupling cycle includes deprotecting the terminal amino acid with 20% piperidine in dry DMF, washing the resin 2-3 times with DMF then methylene chloride and both again, and determining the existence of free amine by ninhydrin kaiser test. A slurry containing coupling reagent 2.4 equivalents of HBTU, the next N terminal amino acid to be added to the growing peptide 2.5 equivalents of Fmoc protected amino acid, 5 equivalents of DIEA in dry DMF. The amino acid was coupled over 2-3 hours and the kaiser test was used to determine

completeness. Coupling steps were repeated if a positive kaiser test resulted. This method was used for each residue of the peptide. The finished peptide was cleaved from the resin with 95% TFA with a cocktail of scavengers including thioanisole and phenol. The peptide was precipitated with ether and lyophilyzed. Peptides were purified using HPLC and identified with MALDI mass spectrometry.

To detect binding of our compounds to the PDZ2 domain of the MAGI-3 protein, a competition polarization assay was employed. The buffer included 35mM HEPES at pH 7.4, triton X-100 0.01%, KCl 10mM, NaCl 10mM, $MgCl_2$ 50μM. The protein was added to a final concentration of 100nM and the probe, 10nM. The competitor was then added to final concentration range of 10ρM to 300μM. To dissolve these hydrophobic indole scaffold compounds 0.12% DMSO was needed. This amount was also added to the control peptides for the assay. Triplicates of the samples in a total volume of 20μL were transferred to 384 well Corning opaque plate for analysis. Fluorescence polarization was measured at equilibrium by LJL Biosystems plate reader. Competition data was fit to a one-site competition expression and the IC_{50} values of the nonpeptide compounds were compared.



4a: R_1= CO_2H, R_2= CH_3, R_3= $(CH_2)_2Ph$, R_4= $CHOH(CH_2)_3CH_3$, R_5= H
4b: R_1= CO_2H, R_2= CH_3, R_3=($CH_2)_2OH$, R_4= $CHOH(CH_2)_3CH_3$, R_5= H
4c: R_1= CO_2H, R_2= CH_3, R_3= $(CH_2)_2Ph$, R_4= $(CH_2)_4CH_3$, R_5= H
4d: R_1= CO_2Me, R_2= CH_3, R_3=$(CH_2)_2Ph$, R_4= $CHOH(CH_2)_3CH_3$, R_5= H

Figure 1. Design of novel scaffolds to mimic the side chain presentation of the four carboxy terminal residues of the PDZ domain ligand. Panel A shows the four terminal residues of a class I PDZ domain ligand, indicating important contacts with the receptor. The specific side chain to side chain hydrogen bond is highlighted. Panel B shows two early designed scaffolds indicating the enforced tending toward lower molecular weight and decreased conformational flexibility in the ligand. Panel C shows the final designed scaffold and targeted variations.

Results

Structure Based Evaluation of Designed Molecules

Analysis of the co-crystal structure of PSD-95 PDZ3 with the CRIPT ligand gave the geometry of the bound ligand in its β strand conformation. Inspection of the β strand revealed that its conformation was uncharacteristic of other β strands. The phi/psi angles of an antiparallel β strand have been shown to be around 113°/180°. These angles for the 0, -1, and -2 residues are 117.7°/110.4°, -106.6°/149.3°, and -101.9°/178.4°, respectively. Presumably, this uncharacteristic conformation allows the Val(0) to dip towards the hydrophobic cavity in the PDZ domain binding groove. In order to preserve the orientations of the carboxylate and the Val(0) side chain, we measured the distances and angles using Weblab[27] between the α carbons of residues 0 and -1 and manually matched them to those of simple rigid cores. Several rigid cores, including various forms of indoles, naphthalenes, and indenes were evaluated. Only an indole shape adequately preserved these distance and angle restraints. The distance between the α carbons of residue 0 and -1 is 3.81Å, with the angle between the corresponding side chains being 180°. The distance between the indole N1 and C7 is 3.85Å. Derivatives at these positions of an indole scaffold would have the same relative orientation in the binding groove as their native side chain counterparts. Figure 1 shows the native peptide ligand in its β strand conformation, **1**, and the structures observed to mimic the orientation and position of the ligand carboxylate and the Val(0) side chain in the PDZ ligand binding pocket, including an indole scaffold, **4**.

Scaffold design and optimization ensued from efforts to maintain functional groups that account for required contacts to the PDZ domain and include positions to

extend the structure to impart additional affinity and possibly specificity. Our structurally most promising computational hit, **2** (Figure 1), did not contain synthetically convenient places to affix groups intended to mimic the -1 and -3 side chains of the ligand. Reengineering **2** to a more useful nucleus, **3**, facilitated the addition of groups mimicking the -1 position yet lacked a substitution to mimic the -3 side chain, and this compound was viewed as synthetically difficult due to the presence of two stereocenters. We developed scaffold **4** to accommodate all positions for functionalities deemed necessary for a specific inhibitor of PDZ domain interactions. Scaffold **4** included positions R1-R5 which were explored in projecting various functionalities (Figure 2) to optimize the ligand's mimicry of peptide ligand side chains into the PDZ domain binding pocket.

The initial DOCK experiments with the indole scaffold indicated that it could indeed place critical functional groups in proximity to the orientation of the β strand conformation of bound ligand in the crystal structure. Figure 3 shows molecules **4a** and **4b** in their docked orientation compared to the native β strand, **1**. A critical requirement for PDZ domains is a carboxylate to anchor the ligand into the binding pocket making crucial contacts with a highly conserved region of the PDZ domain. In structure **4a** the carboxylic acid positioned at R1 has the same orientation as its native counterpart. The favorable geometry is supported by DOCK's energy scoring: -26.31 for **4** while -20.00 for the native β strand **1**. Although a terminal amide has been experimentally shown to destroy affinity between the ligand and the PDZ domain,[24,28] we determined computationally that a carboxylic acid or amide group was the most appropriate substituent to use at this position. Both docked close enough to the conserved GLGF

residues common to class I PDZ domains. The tetrazole made similar contacts, but did not score as well. Two other substituents shown in Figure 2 were too large, and often resulted in the molecule docking with an incorrect orientation.

A group at the R2 position would be analogous to the Val (0) sidechain which fills the hydrophobic pocket on the surface of the PDZ domain. We evaluated a methyl, ethyl, and isopropyl group at this position (Figure 4). The methyl and ethyl groups performed similarly in their contribution to binding energy, solvation, and contact scores as determined by molecular docking (Figure 4A and 4B). Despite having a similar score, the isopropyl group proved to be too large for the hydrophobic pocket, causing inversion of the scaffold (Figure 4C). Because DOCK scores were comparable and the methyl substituted scaffold starting material was commercially available we chose a methyl group for R2.

The R3 position of the indole scaffold has an optimal orientation for presenting the hydroxyl groups to mimic the -1 side chain or other functional groups may be added to achieve specificity for PDZ domains. PSD-95 PDZ3 whose structure was used for these docking experiments shows a preference for hydroxyl groups at the -1 position in the peptide ligands although the structure indicates that this side chain is solvent exposed and does not make contacts with the PDZ domain. Basing our analysis on this structure resulted similarly in the hydroxyalkyl groups evaluated at R3 position not making contacts with the PDZ domain, and these R3 groups did not seem to contribute greatly to the overall scores of the various compounds. Nor did they vary greatly in their relative contributions to overall scores. Designing a ligand for the MAGI-3 PDZ2 domain

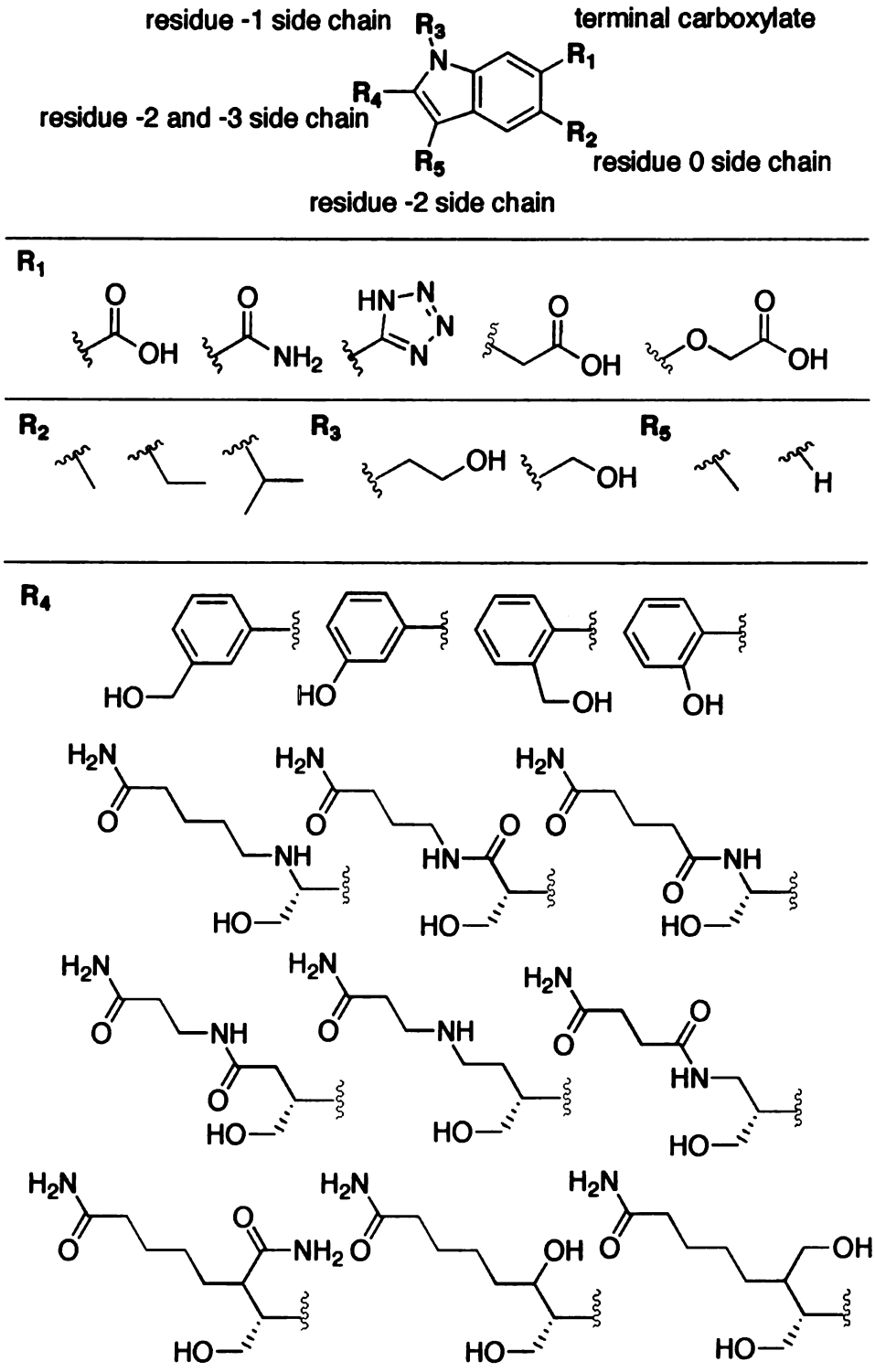


Figure 2 Composition of library used for *in silico* screen used to evaluate potential novel library scaffolds to mimic the PDZ ligand geometry. The chemset used to generate the library by showing all substituents at each position. Initially this library was screened to evaluate the scaffold. Compounds were picked from amoung those scoring best with DOCK.

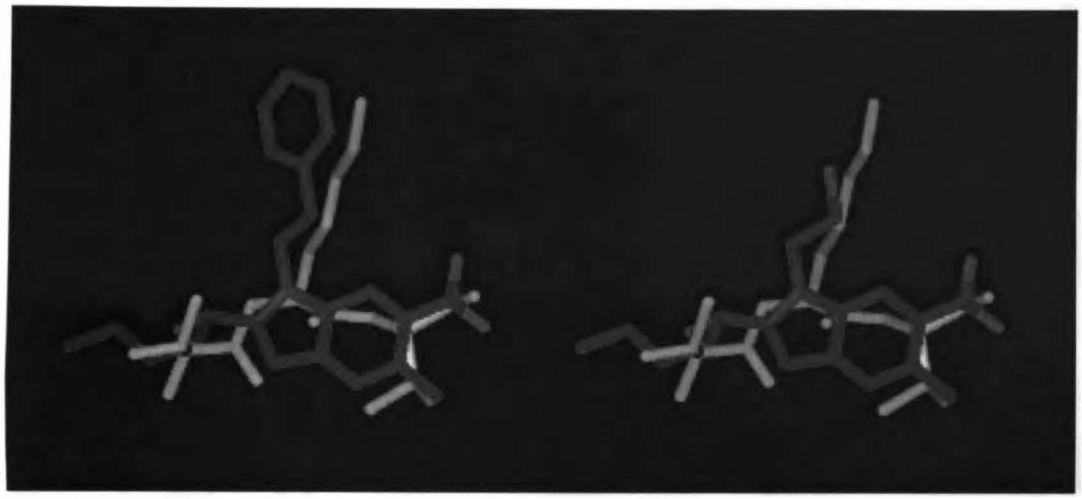
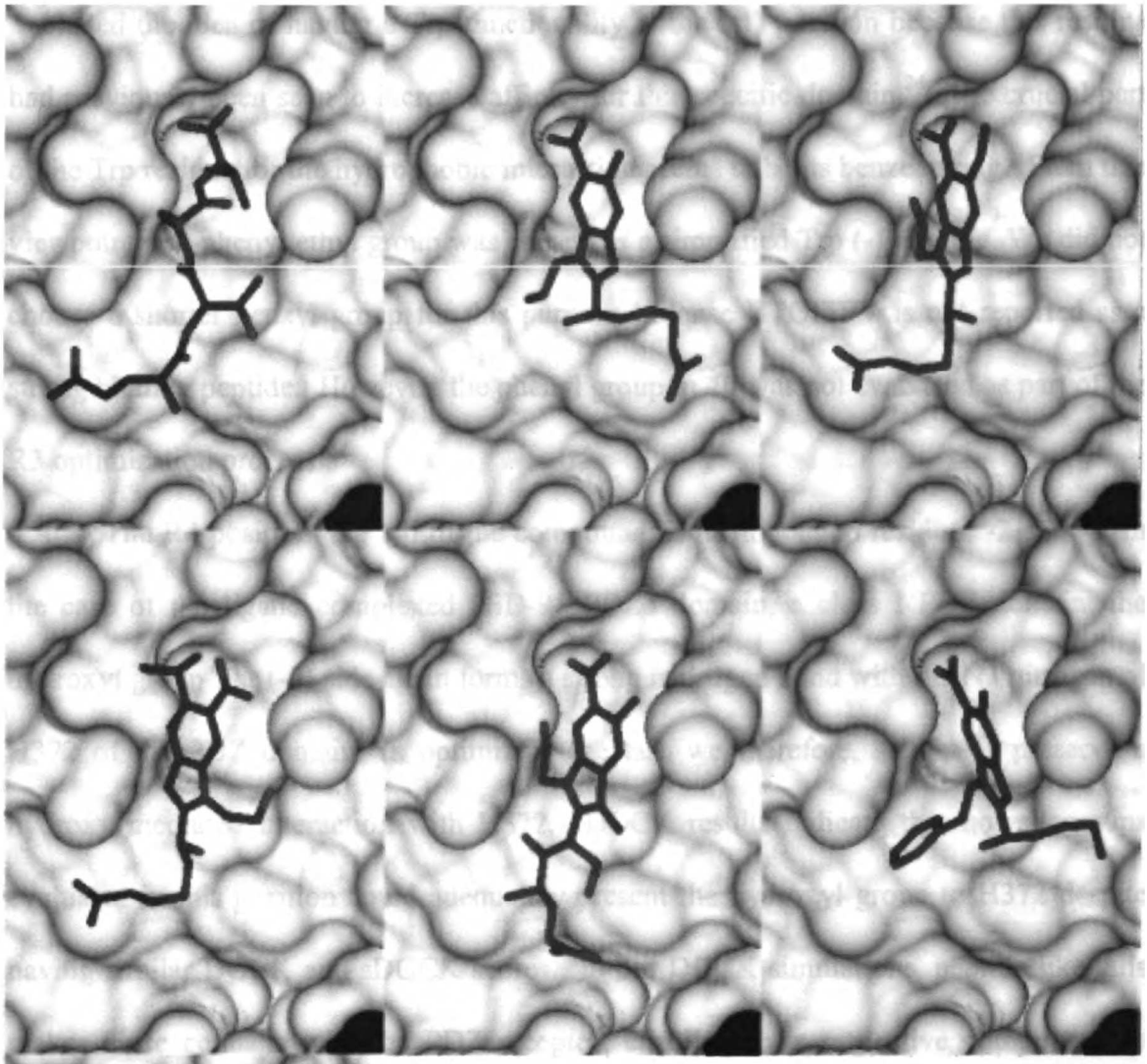


Figure 3. FJ-1 (**3a**, in color) and FJ-2 (**3b**, in color) overlayed against the PTEN peptide ligand (TKV) (in white). The carboxyl, methyl, and phenyl groups reproduce the same orientations as the c-terminus carboxylic acid, the Val side chain, and the trajectory of the Lys side chain, respectively.



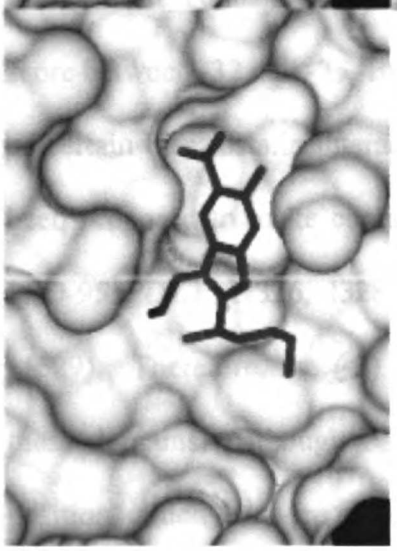


Figure 4. 4a shows the native CRIPT peptide (DOCK score: –20). Evaluation of methyl **(4b)**, ethyl **(4c)**, and isopropyl **(4d)** groups at the R2 position. The three variants have similar dock scores (**4a:** -30.54, **b** -31.1, **4c:** -30.45), but the isopropyl group causes inversion of the scaffold, with the carboxylic acid groups oriented towards the surface. This renders the isopropyl inappropriate as a choice for substitution at this position. **4e** (-34.22) shows a methylated variant of the indole scaffold. While there is sufficient room to accommodate the methyl group, it causes slight displacement of the entire molecule. **4f** and **4g** show synthesized molecules 4a (-26.31) and 4b(-27.74), respectively.

included the idea of having a Trp functionality at the (-1) position because this residue had previously been seen to increase affinity for this specific domain.[24] The crucial part of the Trp residue for the hydrophobic interaction seems to be its benzene ring. From this viewpoint, a 2-phenyl ethyl group was chosen as a simplified Trp (-1) mimic. We did not choose a shorter benzyl group for this purpose because HTQITFV is not reported as a strong binding peptide. However, the phenyl group in **4b** was not evaluated as part of the R3 optimization process.

The R4 position, the tail of the compound, is analogous to residues -2 and -3. In the case of the ligand complexed PSD-95 PDZ3 domain, a class I PDZ domain, the hydroxyl group Thr (-2) side chain forms a strong hydrogen bond with the N3 nitrogen of H372 of the PDZ domain. In optimizing the tail, we therefore, sought to preserve a hydroxy mediated interaction with H372. DOCK results indicate that not all groups evaluated in this position could adequately present the hydroxyl group to H372 despite having similar DOCK and sDOCK scores. In our DOCK simulations, peptide like tails makes close contacts with the PDZ receptor, enabling their respective compounds to score between -32 and -35. Extensions lacking peptide character directly result in the loss of certain electrostatic interactions and yield lower DOCK scores. Residues that have been proposed to make contacts with the residue (-3) are distinct in character between PSD-95 PDZ3 (N326, I328, S338, F339) and MAGI-3 PDZ2 (A617, A619, K628, M629). Therefore, the (-3) position might be ideal for providing specificity between these domains. Although PSD-95 PDZ3 prefers hydrophilic side chains for the (-3) position, which has been seen from the crystal structure to hydrogen bond with N326 and S338, for the design of an inhibitor targeting MAGI-3 PDZ2, which shows a preference

for hydrophobic side chains at this position (I, V, W, C),[24] the n-butyl extension seemed to be an appropriate fit. The n-butyl is considered also as a suitable alternative for QITWV hydrophobic isoleucine sidechain and may stabilize the complex with favorable Van der Waals contacts and hydrophobic interactions with the side chains that are proposed to be in the vicinity of the binding of this side chain.

Analysis of the indole scaffold in the context of the PDZ domain revealed a hydrophobic gap between the molecular surfaces of the indole scaffold and the domain that might accommodate a methyl group. The R5 position is the only position that does not have a corollary in the native β strand. Figure 4d shows a scaffold variant with a methyl group at the R5 position which is the size that can be accommodated in this region of the PDZ domain. Introduction of a methyl group at this position generally decreases the DOCK score or inverts the binding mode to flip. In these cases the methyl group binds into the hydrophobic cavity for R2 group in correct ways (data not shown). Thus, a simple hydrogen was chosen as R5 in initial experiments.

Figure 3 shows one scaffold variant, **4b**, that successfully makes all the necessary contacts to mimic the β strand: 1) a carboxylic acid at the R1 position mimics the peptide ligand C-terminus, 2) a methyl group at the R2 position fills the hydrophobic pocket that the Val(0) side chain normally occupies, and 3) a nonpeptide extension at the R4 position places a hydroxy group within hydrogen bonding distance to H372. Variations of **4** on its R1 and R4 extension correspond to that of the Val(0) terminal carboxylate and the Thr(-2) hydroxyl group respectively in TWV peptide. If **4b** binds to PDZ domains as a ITWV mimic, such mutations would destroy its binding activity. From this viewpoint, **4c** and **4d** were designed as mimetics of negative mutant peptides for evaluation of the hypothesis

of **4b** binding as a ITWV mimic. This chemset was targeted for synthesis and evaluation as inhibitors of PDZ domain function.

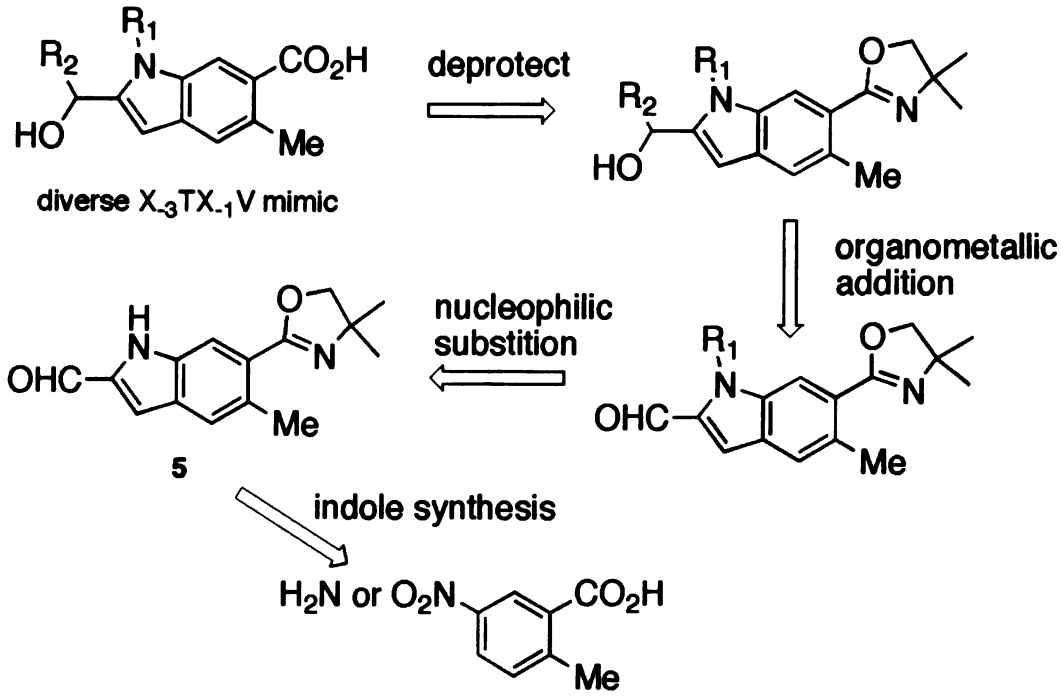
Synthetic Chemistry

This section summarizes Naoaki Fujii's medicinal chemistry work on the project. A target was the generation of an intermediate presenting essential functionality for PDZ domain binding, attachment points for generation of a diverse library, and tolerating various chemical conditions to allow for library production. Because it is essential for affinity, introduction of a protected carboxylate was planned early in the synthesis. The precursor carboxylate substituted indole scaffold would be the common intermediate to all variations of ligands. Aromatic halides are robust precursors for esters by palladium-catalyzed carbonylation,[29] but they were unsuitable for our envisioned high-throughput synthesis. We chose an oxazolin-2-yl group for the carboxylate precursor. Therefore, we designed the intermediate 5-methyl-6-(oxazolin-2-yl)indole-2-carboxaldehyde, **5**, as the foundation for our library development. The retro-synthetic analysis is shown in Figure 5. Three basic strategies for indole synthesis were studied preliminarily. For Fischer indole synthesis and its variants it is necessary to work with acidic reaction conditions with high temperature, therefore this strategy is not compatible with an acid labile oxazolin-2-yl group.[30] Reductive cyclization of 2-nitrobenzylketone[31,32] gave low yields and multiple products. The palladium-catalyzed coupling of an alkyne and 2-iodoaniline[33,34] has not been reported for the synthesis of R3 unsubstituted indoles having carboxylic acid derivatives on the 2 position. Therefore, we chose palladium-catalyzed coupling of ethylpyruvate and 2-iodoaniline[35] for synthesis of the indole-2-carboxylic acid derivatives, as shown in figure 6. Note that aldehyde **5** is useful for the syntheses of

indoles having diverse substituents on its 1 (R3)- and 2 (R4)- positions by sequential electrophilic and nucleophilic alkylation.

Compounds synthesized were assayed by fluorescence polarization competition. The positive control peptide of the native PTEN sequence competed with the labeled same sequence for the PDZ binding pocket with an IC_{50} of 4.5μM. The highest detectable IC_{50} by this assay is approximately 50μM. This is because of protein aggregation complications evident by increasing polarization values when high control peptide concentrations were added to the assay solution. Also the polarization values of the probe were affected by the high concentrations of the indole scaffold competitors. Many of the compounds produced had poor affinity for the PDZ domain and therefore characterization of their binding ability by IC_{50} is undermining the scaffold's ability to represent a ligand in the PDZ binding pocket. Compound **4b** demonstrated affinity for the MAGI-3 PDZ2 domain by displacing the peptide probe in a concentration dependent manner (Figure 7). Meanwhile, **4a** bound with less affinity and negative controls, **4c** and **4d**, in which moieties required for the PDZ ligand interaction were disrupted did not show any competition toward the PTEN sequence peptide.

Figure 5. Retrosynthetic analysis of the key intermediate **5**



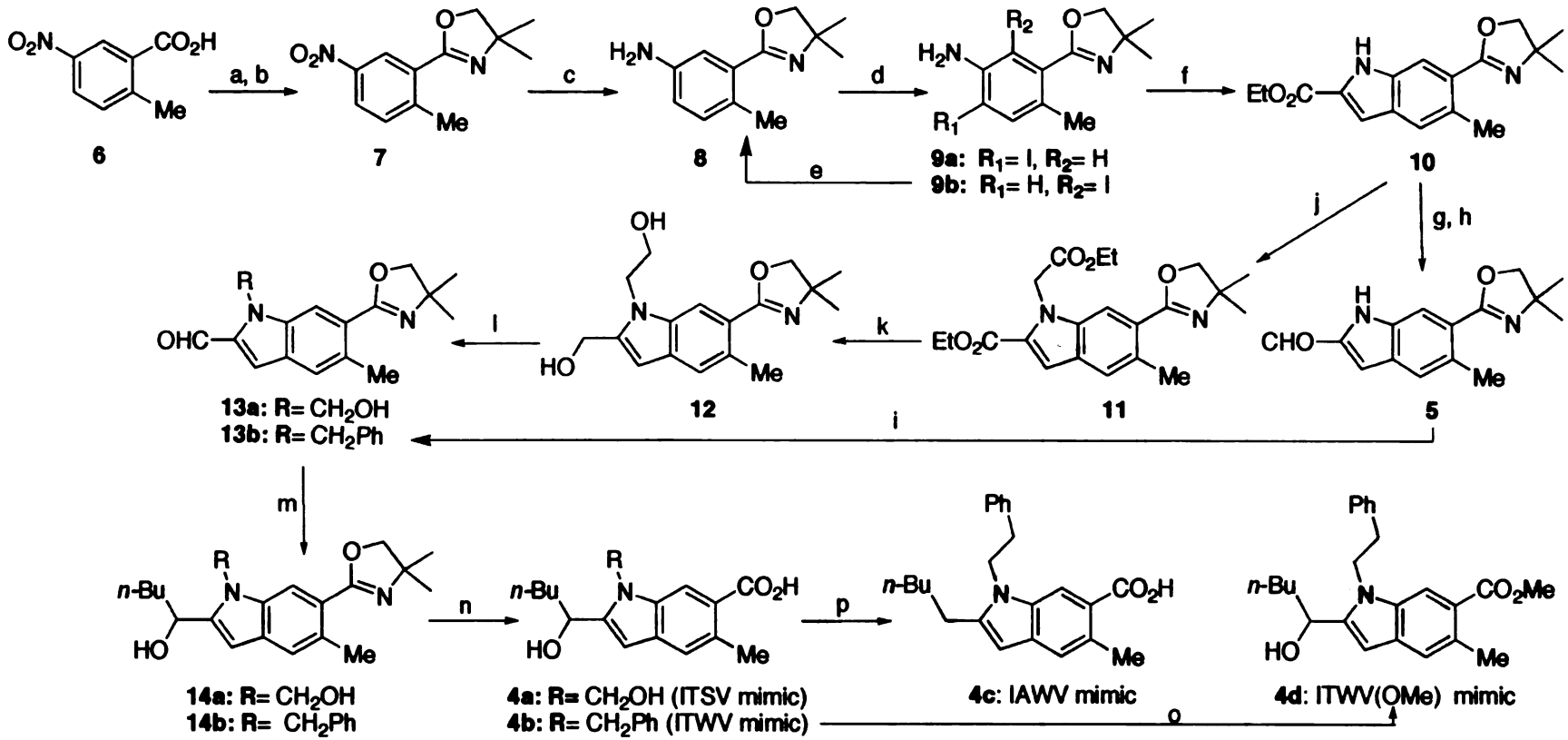
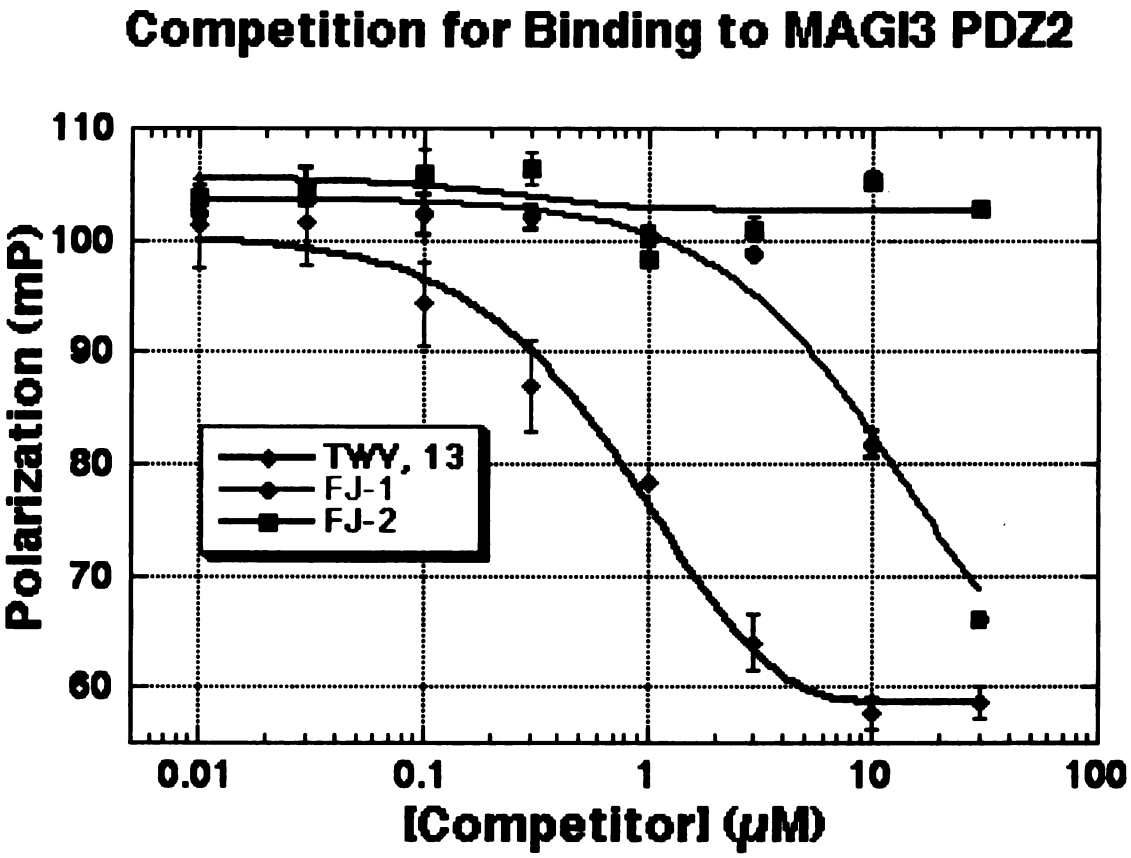
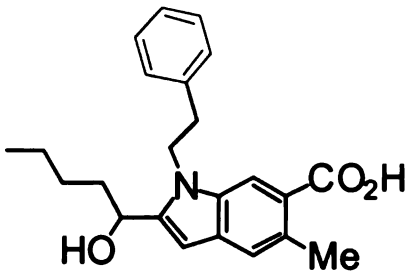
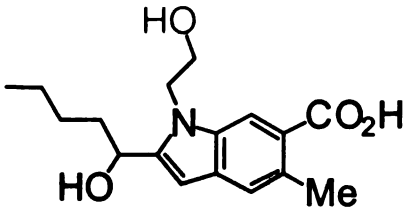


Figure 6. Synthesis of PDZ domain inhibitors: a) $HOCH_2C(CH_3)_2NH_2$ (3 eq), HBTU (1.2 eq), DIPEA (2 eq), DMF, 40 °C, 14 hr. b) $SOCl_2$ (5 eq), CH_2Cl_2, rt, 0.5 hr. c) Fe (7.8 eq), NH_4Cl (aq), EtOH, reflux, 3 hr, 92% over 3 steps. d) ICl (1.6 eq), $CaCO_3$ (30 eq), MeOH, H_2O, rt, 1 hr. **9a**: 36%, **9b**: 47%. e) H_2, Pd-C, MeOH, Et_3N, rt, 1hr, quant. f) Ethyl pyruvate (5 eq), $Pd(OAc)_2$ (0.2 eq), DABCO (5 eq), DMF, 105 °C, 50 min, 44%. g) $LiAlH_4$ (10mol), THF, reflux, 2.5 hr, 75%. h) MnO_2 (1.8eq), CH_2Cl_2, rt, 12 hr, 76%. i) $BrCH_2CH_2Ph$ (5 eq), K_2CO_3 (10 eq), DMF, 40 °C, 22 hr, 93%. j) $BrCH_2CO_2Et$ (1.05 eq), NaH (1.1 eq), DMF, 0 °C, 10 min, 74%. k) $LiAlH_4$ (20 eq), THF, reflux, 0.5 hr, 58%. l) MnO_2 (19 eq), CH_2Cl_2, rt, 12 hr, 84%. m) *n*-BuMgBr (5 eq), 0 °C, 0.5 hr, **14a**: 72%, **14b**: 87%. n) i. MeI (large excess), K_2CO_3 (3 eq), acetone, rt, 2.5 d; ii. NaOH, H_2O, MeOH, reflux, 7 hr, 2 steps overall **4a**: 65%, **4b**: 89%. o) MeI (large excess), K_2CO_3 (large excess), acetone, rt, 1.5 h, 69%. p) H_2, 10% Pd-C, HCl (20 eq), MeOH, rt, 1 hr, 25%.





FJ-1:TWV mimic



FJ-2:TSV mimic

Figure 7. FJ-1 and FJ-2 assayed against the native peptide. Competition curves show that FJ-1 effectively competes with the native peptide ligand, but at 1/10th the affinity

Discussion

DOCK has been successfully used to screen large 3-dimensional databases for small molecule inhibitors of protein and enzyme systems[37,38] and design or optimize specific chemical scaffolds. Side chains of the bound ligand in PDZ domains project along similar vectors when compared between structures. For this reason we felt a modeling approach to design ligands would be ideal. The structure based optimization assured that the PDZ domain showed a high preference for compounds that looked most similar to the native peptide while there were alternatives to the critical functional groups. The indole structure provided a useful scaffold to mimic characteristic interactions that allow for ligand binding in a PDZ domain and allow for diverse reactions to provide functional moieties capable of lending to selectivity in binding between PDZ domains.

A tentative structure activity relationship can be assigned to scaffold **4** from the results of the competition experiments. We assumed the carboxylate functionality mimicked the carboxy terminus of the peptide ligand anchoring within the pocket interacting with the conserved basic residue, a water molecule, and the backbone amide protons of the GLGF sequence of the PDZ domain. The importance of this functionality is emphasized by the competition of the negative control, **4d** in which modification of this group to a methyl ester abolishes any competition. This also suggests that the compound is binding in the orientation depicted by the DOCK evaluations. R1 position is critical for binding and gives the generic β strand sensitivity to recognize the domains, but not the specificity to distinguish between the various classes. Our structure-based evaluation suggests that R2 position is similarly important.

The contribution from residue (-1) to the stabilization of ligand binding to PSD-95 PDZ3 is unclear. Although PSD-95 PDZ3 shows a preference for hydrophilic residues in the (-1) position of proposed native ligands, (D, S, R), there have not been structural observations supporting this tendency. Groups at the R3 position do not contribute to binding with PSD-95 PDZ3 domain during the structure based optimization process 3which is consistent with the structure of PSD-95 PDZ3. The optimized peptide ligand for MAGI-3 PDZ2, HTQITWV, unambiguously shows this (-3) side chain to be useful in adding affinity.[24] Therefore, this may be a position that would allow discrimination between MAGI-3 PDZ2 and PSD-95 PDZ3. The Trp residue in HTQITWV is predicted to bind in a hydrophobic cavity between β strands B and C of MAGI-3 PDZ2 provided by residues (Met/Leu). Thus, it can be anticipated to acheive tighter affinity to MAGI-3 PDZ2 by replacement of the hydroxyethyl group of 4b with a hydrophobic moiety at R3 as in 4a. As predicted, 4a binds to the MAGI-3 PDZ2 domain although it shows a lower DOCK score than 4b. This result is consistent with the reported relative affinities of an optimized peptide, HTQITWV, and the native PTEN sequence, HTQITKV.[24] We had previously observed that modification of the Trp(-1) in HTQITWV peptide drastically reduced binding. These results indicate a large contribution to affinity by the (-1) position substituent which can be replicated by the R3 substituent. These results may develop into a discriminating effect by the R3 substituent of the compound for individual PDZ domains.

The R4 side arm positions the hydroxyl moiety that mimics the side chain hydroxyl of residue (-2). This functionality was suspected to be involved in ligand recognition by class I PDZ domains. Correspondingly, the negative mutation, **4c**,

abolishes binding to the pocket as seen with a similar mutation in native peptides. Because PDZ domains bind ligands fulfilling an interaction specific to their class at residue (-2) and this scaffold is able to reproduce this interaction, the scaffold may be appropriate to present moieties needed to bind to other classes of PDZ domains.

The hydrophobic carbon chain on R4 position seems to impart binding to the pocket. This is appropriate because while CRIPT, a native sequence binding PSD-95, includes a glutamine in the (-3) position the PTEN sequence has an isoleucine in the corresponding position. Because of this, it is understandable that the MAGI-3 PDZ2 domain might select for a hydrophobic side chain mimic on R4 position. The residues in MAGI-3 PDZ2 in this region might contribute through hydrophobic interactions. We studied several modifications of the HTQITWV peptide to understand the key interactions between the peptide and the MAGI-3 PDZ2 domain, and found that QITWV binds tighter than ITWV. According to this result, it is expected to get tighter molecules by further modification of extensions at the R4 position. It is important to note that this site has been recognized to aid the PDZ domain in the discrimination of its ligands. It can be concluded that the R4 position contributes to the binding affinity of the and therefore might be exploited to give selectivity between PDZ domains.

Conclusions

The widespread occurrence of PDZ domains as facilitators of subcellular organization makes them an important target for biological studies. Our ability to study the nature of PDZ mediated interactions, the functions of PDZ containing proteins, and the varied functions of PDZ domains themselves would be greatly enhanced by the discovery of inhibitors that bind specifically to certain PDZ domains. The scaffold described herein offers several opportunities for optimization towards specific PDZ domains. A small molecule mimetic must be able to make similar contacts with critical residues if it is to have the same specificity for PDZ domain subtypes. The similarity of binding mode between our indole scaffold and class I consensus sequence peptide (ITWV for MAGI-3 PDZ2) is clearly shown by the meaningful difference in affinity between **4a**, **4b**, **4c** and **4d**. Further variation of **4** will provide optimized binding molecules of MAGI-3 PDZ2 and tight binding molecules for other PDZ domains. PDZ domains are very widely expressed protein module and frequently exists as multiple domains of one protein, thus it is almost impossible to design selective binding molecules to every PDZ domains by structure based analysis. Our study demonstrates an efficient ligand discovery process through collaboration of structure based and medicinal chemical design. Another efficient way is construction and evaluation of structure focused chemical library. An important feature of the intermediate **5** is that it is feasible to make diverse library highly variant in the R3 and R4 position in order to discover class selective and domain selective inhibitors.

References

1. Kausalya, P. J.; Reichert, M.; Hunziker, W. *FEBS Lett* **2001**, *505*, 92-96.
2. Itoh, M.; Sasaki, H.; Furuse, M.; Ozaki, H.; Kita, T.; Tsukita, S. *J Cell Biol* **2001**, *154*, 491-497.
3. Kuwahara, H.; Araki, N.; Makino, K.; Masuko, N.; Honda, S.; Kaibuchi, K.; Fukunaga, K.; Miyamoto, E.; Ogawa, M.; Saya, H. *J Biol Chem* **1999**, *274*, 32204-32214.
4. Liu, T. F.; Kandala, G.; Setaluri, V. *J Biol Chem* **2001**, *276*, 35768-35777.
5. Kornau, H. C.; Schenker, L. T.; Kennedy, M. B.; Seeburg, P. H. *Science* **1995**, *269*, 1737-1740.
6. Niethammer, M.; Kim, E.; Sheng, M. *J Neurosci* **1996**, *16*, 2157-2163.
7. Tsunoda, S.; Sierralta, J.; Sun, Y.; Bodner, R.; Suzuki, E.; Becker, A.; Socolich, M.; Zuker, C. S. *Nature* **1997**, *388*, 243-249.
8. Voltz, J. W.; Weinman, E. J.; Shenolikar, S. *Oncogene* **2001**, *20*, 6309-6314.
9. Dobrosotskaya, I.; Guy, R. K.; James, G. L. *J. Biol. Chem.* **1997**, *272*, 31589-31597.
10. Wu, X.; Hepner, K.; Castelino-Prabhu, S.; Do, D.; Kaye, M. B.; Yuan, X.-J.; Wood, J.; Ross, C.; Sawyers, C. L.; Whang, Y. E. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 4233-4238.
11. Wu, Y.; Dowbenko, D.; Spencer, S.; Laura, R.; Lee, J.; Gu, Q.; Lasky, L. A. *J. Biol. Chem.* **2000**, *275*, 21477-21485.
12. Cantley, L. C.; Neel, B. G. *Proc Natl Acad Sci U S A* **1999**, *96*, 4240-4245.
13. Maehama, T.; Dixon, J. E. *Trends Cell Biol* **1999**, *9*, 125-128.

14. Vazquez, F.; Sellers, W. R. *Biochim Biophys Acta* **2000**, *1470*, M21-35.

15. Cowburn, D.; Voltz, J. W.; Weinman, E. J.; Shenolikar, S. *Curr Opin Struct Biol* **1997**, *7*, 835-838. Center, Durham, North Carolina, NC 27710 USA.

16. Oschkinat, H.; Cowburn, D.; Voltz, J. W.; Weinman, E. J.; Shenolikar, S. *Nat Struct Biol* **1999**, *6*, 408-410.

17. Harrison, S. C.; Cowburn, D.; Voltz, J. W.; Weinman, E. J.; Shenolikar, S. *Cell* **1996**, *86*, 341-343.

18. Doyle, D. A.; Lee, A.; Lewis, J.; Kim, E.; Sheng, M.; MacKinnon, R. *Cell* **1996**, *85*, 1067-1076.

19. Morais Cabral, J. H.; Petosa, C.; Sutcliffe, M. J.; Raza, S.; Byron, O.; Poy, F.; Marfatia, S. M.; Chishti, A. H.; Liddington, R. C. *Nature* **1996**, *382*, 649-652.

20. Tochio, H.; Zhang, Q.; Mandal, P.; Li, M.; Zhang, M. *Nat Struct Biol* **1999**, *6*, 417-421.

21. Schultz, J.; Hoffmuller, U.; Krause, G.; Ashurst, J.; Macias, M. J.; Schmieder, P.; Schneider-Mergener, J.; Oschkinat, H. *Nat Struct Biol* **1998**, *5*, 19-24.

22. Daniels, D. L.; Cohen, A. R.; Anderson, J. M.; Brunger, A. T. *Nat Struct Biol* **1998**, *5*, 317-325.

23. Sheng, M.; Sala, C. *Annu Rev Neurosci* **2001**, *24*, 1-29.

24. Fuh, G.; Pisabarro, M. T.; Li, Y.; Quan, C.; Lasky, L. A.; Sidhu, S. S. *J. Biol. Chem.* **2000**, *275*, 21486-21491.

25. Bolin, D. R.; Swain, A. L.; Sarabu, R.; Berthel, S. J.; Gillespie, P.; Huby, N. J.; Makofske, R.; Orzechowski, L.; Perrotta, A.; Toth, K.; Cooper, J. P.; Jiang, N.; Falcioni, F.; Campbell, R.; Cox, D.; Gaizband, D.; Belunis, C. J.; Vidovic, D.; Ito,

K.; Crowther, R.; Kammlott, U.; Zhang, X.; Palermo, R.; Weber, D.; Guenot, J.; Nagy, Z.; Olson, G. L. *J Med Chem* **2000**, *43*, 2135-2148.

26. Furet, P.; Garcia-Echeverria, C.; Gay, B.; Schoepfer, J.; Zeller, M.; Rahuel, J.; Campbell, R.; Cox, D.; Gaizband, D.; Belunis, C. J.; Vidovic, D.; Ito, K.; Crowther, R.; Kammlott, U.; Zhang, X.; Palermo, R.; Weber, D.; Guenot, J.; Nagy, Z.; Olson, G. L. *J Med Chem* **1999**, *42*, 2358-2363.
27. In;; Accelrys Inc.: San Diego, CA.
28. Harris, B. Z.; Hillier, B. J.; Lim, W. A. *Biochemistry* **2001**, *40*, 5921-5930.
29. Magerlein, W.; Beller, M.; Indolese, A. F. *J. Mol. Catal. A: Chem.* **2000**, *156*, 213-221.
30. Gan, T.; Liu, R.; Yu, P.; Zhao, S.; Cook, J. M. *J. Org. Chem.* **1997**, *62*, 9298-9304.
31. Shin, C.-g.; Yamada, Y.; Hayashi, K.; Yonezawa, Y.; Umemura, K.; Tanji, T.; Yoshimura, J. *Heterocycles* **1996**, *43*, 891-898.
32. Suzuki, H.; Gyoutoku, H.; Yokoo, H.; Shinba, M.; Sato, Y.; Yamada, H.; Murakami, Y. *Synlett* **2000**, 1196-1198.
33. Yasuhara, A.; Kanamori, Y.; Kaneko, M.; Numata, A.; Kondo, Y.; Sakamoto, T. *J. Chem. Soc., Perkin Trans. 1* **1999**, 529-534.
34. Zhang, H.-C.; Ye, H.; Moretto, A. F.; Brumfield, K. K.; Maryanoff, B. E. *Org. Lett.* **2000**, *2*, 89-92.
35. Chen, C.-y.; Lieberman, D. R.; Larsen, R. D.; Verhoeven, T. R.; Reider, P. J. *J. Org. Chem.* **1997**, *62*, 2676-2677.

36. Ladd, D. L.; Weinstock, J.; Wise, M.; Gessner, G. W.; Sawyer, J. L.; Flaim, K. E. *J. Med. Chem.* **1986**, *29*, 1904-1912.

37. Aronov, A. M.; Munagala, N. R.; Kuntz, I. D.; Wang, C. C. *Antimicrobial Agents and Chemotherapy* **2001**, *45*, 2571-2576.

38. Hopkins, S. C.; Vale, R. D.; Kuntz, I. D. *Biochemistry* **2000**, *39*, 2805-2814.

39. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J Mol Biol* **1982**, *161*, 269-288.

40. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D.; Blaney, J. M.; Jorgensen, E. C.; Connolly, M. L.; Ferrin, T. E.; Langridge, R.; Oatley, S. J.; Burridge, J. M.; Blake, C. C. *J Comput Aided Mol Des* **2001**, *15*, 411-428.

41. Zou, X.; Sun, Y.; Kuntz, I. D. *Journal of the American Chemical Society* **1999**, *121*, 8033-8043.

42. Connolly, M. L.; Skillman, A. G.; Kuntz, I. D.; Blaney, J. M.; Jorgensen, E. C.; Ferrin, T. E.; Langridge, R.; Oatley, S. J.; Burridge, J. M.; Blake, C. C. *Science* **1983**, *221*, 709-713.

43. Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. *J. Mol. Graphics* **1988**, *6*, 13-27, 36-17.

44. DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R.; Blaney, J. M.; Jorgensen, E. C.; Connolly, M. L.; Ferrin, T. E.; Langridge, R.; Oatley, S. J.; Burridge, J. M.; Blake, C. C. *J Med Chem* **1988**, *31*, 722-729.

45. Gasteiger, J.; Ihlenfeldt, W. D.; Roese, P.; Wanke, R. *Anal. Chim. Acta* **1990**, *235*, 65-75.

Gloss to Chapter III

As the number of proteins with unknown functions increase, it will be useful to screen the protein databank for binding partners in order to gain insight into protein function. Many protein-protein docking methods to date involve extensions to the original docking paradigm developed by Kuntz and co-workers. However, the number of calculations involved makes these methods fall short of being useful for database screening. In the most ideal of circumstances, a first pass screen of the protein databank would be verified using more thorough protein-protein docking methods. To be truly useful, a first pass screen should be very rapid, perhaps on the same timescales as rigid docking of a ligand database against a single receptor.

This chapter describes the proof-of-concept experiment behind a two-dimensional protein-protein docking method. The goal of this experiment was to show proof-of-concept of the approach, and achieve some understanding of its strengths, weaknesses, and needed improvements relative to existing methods. This project grew out of my interest to expand the role of molecular docking from its more academic and project driven applications to something as useful and ubiquitous as BLAST. The proof of concept data contained herein provides some insight into how this method might be a way to achieve that lofty goal. My only regret is not having enough time to fully explore its potential.

CHAPTER III

~•~

Protein-protein Docking using 2D Topographical Representations of Molecular Surfaces

~•~

Abstract

A next step in our understanding of entire genomes requires insight into the three-dimensional (3D) structure of proteins in the near term. Most of these protein structures will be determined by high-throughput modeling procedures. Thus, a structure-based analysis of the network of protein-protein interactions in genomes is likely to move forward with docking methodologies that are capable of dealing with the high-throughput requirements of such large-scale analyses. While numerous protein-protein docking algorithms exist, most achieve RMSD less than 2.5 Å only by spending a great deal of time establishing protein complementarity using traditional docking methods, or some variations thereof. I present here the results of a pilot project for a docking method that uses 2D topographical representations of molecular surface. The method is shown to find the correct geometry for a complex consisting of α–chymotrypsin and pancreatic human trypsin inhibitor variant 3. The results indicate that the 2D method is sensitive to gross topological features (large 'dents' or 'bulges' caused by a helix, for example), but is relatively insensitive to the effects caused by smaller structures, such as amino acid side-chains. The work presented here indicates that the method has some promise vis-à-vis existing protein-protein docking methods. That utility, however, is mitigated by a need for further development before a conclusive comparison can be made.

Introduction.

Almost all intracellular activities, including transcription and signaling cascades, are dependent on the specific interaction of macromolecules. The interaction is often a complex process involving the physical contact of two macromolecules, and the van der waals and electrostatic interactions that mediate the formation of a stable interface. From available crystal structures, we know that the units involved in macromolecular complexes are complementary in both their shape and their chemical makeup.

The desire to understand further macromolecular interaction drives the development of simulation methods aimed at modeling biophysical interactions at the atomic and subatomic levels. Protein-protein docking is one such approach. It is an extension of the same molecular docking principles that allow us to explore the interactions of small molecule inhibitors with their protein targets. Traditional docking explores the energetic components of binding by traversing the rotational and translational degrees of freedom of a small molecule in the context of a particular binding site. There are now numerous examples of how small molecule docking has successfully led to the discovery of effective inhibitors of protein and enzyme systems. Protein-protein docking methods attempt to similarly dock large proteins to their targets or partners. Some rely on traditional 3-dimensional docking methods. The simplest protein-protein docking methods evaluate the complementarity of molecular surfaces to determine how well two proteins fit together. Some approaches focus on matching surfaces[1-5], and others enhance the search for geometric fit by matching the positions of surface normals and spheres[6-10]. Some shape algorithms model the hydrophobic effect during association from the change in the solvent-accessible surface area of the

molecules[11], while others employ a simplified scheme to estimate electrostatic, hydrophobic, or desolvation contributions to dimer formation[12]. Most of these methods have focused on rigid body docking, and in general achieve respectable success (typically 1-3 Å RMS). Several groups have additionally focused on finding the correct orientation among a large number of false positives[13]. Recently, the use of soft-body docking and side-chain enumeration has yielded greater sensitivity to the protein-docking results[14]. It should be noted however, that the use of non-complexed models instead of complexes to dock partners to each other does not necessarily generate similar results in terms of time and accuracy. One last class of approaches transforms the problem from three dimensions into two. Sternberg and coworkers extended their Fourier transform-based approach from small molecule docking to protein-protein docking[13]. Several groups have extended this method to show that a reduction in dimension yields high sensitivity (< 2 Å RMS in some cases) while increasing the speed of the analysis.

The current approaches can generate a reasonable degree of accuracy. However, the lack of speed of protein-protein docking relative to small molecule docking makes the current implementations impractical for high-throughput *in-silico* screening of a macromolecule against a database of known protein structures such as the PDB. Such an analysis would greatly increase our ability understand the function of unknown proteins for which 3D models have been obtained through structural biology or structural genomics methods. While increases in computational power will certainly increase our analytical ability, fundamental advances in protein-protein docking methodologies will ultimately determine the utility of this class of approaches.

Herein I prototype a method of docking two macromolecular objects that relies on the conversion of relevant three-dimensional information about an object into two dimensions. First the details of the implementation are presented, followed by the description of several experiments conducted over the course of the development of the algorithm to establish proof-of-concept. The goals of these experiments were to show proof-of-concept of the approach, and achieve some understanding of its strengths, weaknesses, and needed improvements relative to existing methods.

Methods

3D to 2D conversion

This method was evaluated on the α-chymotrypsin / pancreatic human trypsin inhibitor variant 3 complex. Figure 1 shows the details of the method. The process begins with the construction of the solvent accessible molecular surface using the *dms* program as implemented in *MidasPlus*[22]. Then the van der waals and electrostatic potentials on the molecular surface are calculated by first using *grid* program to generate the electrostatic, attractive van der waals, and repulsive van der waals grids around the protein. The protein is prepared for grid production by using Sybyl 6.5[23] to add Amber 6 charges to the heavy atoms of the protein. No hydrogens are added. The electrostatic, attractive and repulsive potentials at each surface point are calculated by interpolating the values off of the relevant grids.

A sphere with a radius equivalent to the shortest distance between the protein's center and the molecular surface is created. The molecular surface will eventually be projected onto this sphere. The sphere's circumference and meridians are divided into 0.5 Å bins, effectively gridding the surface of the sphere. For α-chymotrypsin, this process resulted in 27,730 bins. The average perpendicular distance between a point on the molecular surface and the sphere, electrostatic potential, and van der waals potential of all the points centered on each bin is calculated. Each grid section is represented as an object that stores the average distance of the molecular surface above the grid from the sphere, the electrostatic and van der waals potential of the molecular surface at that point, and any other user-defined information (such as the curvature of the surface at that point,

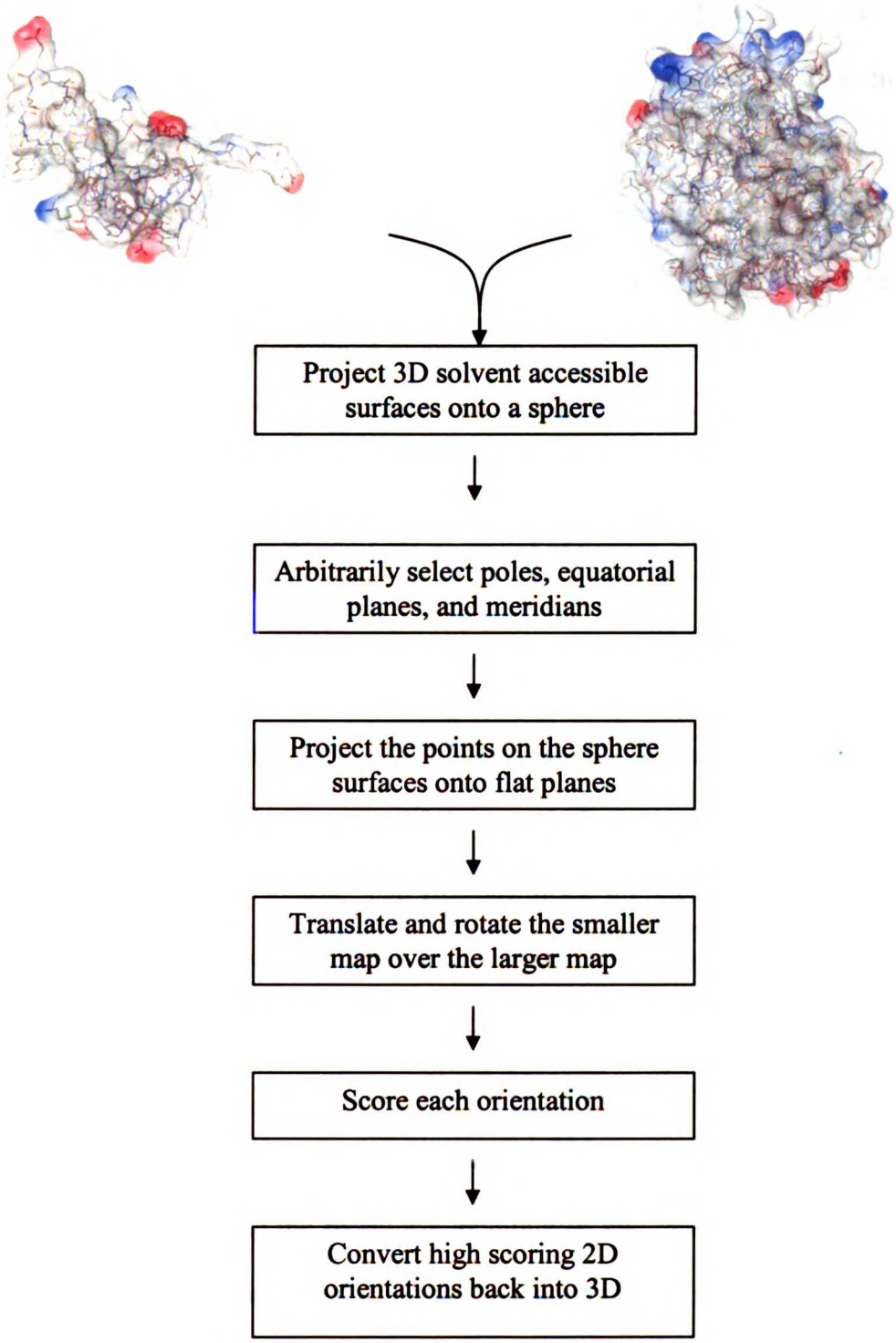


Figure 1. Algorithm for converting 3D molecular surfaces into 2D topological maps

or the direction of the surface normals, etc). Because each grid has a defined coordinate on the sphere in relation to the meridian and equator, a two-dimensional map of the sphere can be built by plotting the longitude and latitude of a grid point. This process is performed for both members of a complex, resulting in two distinct 2D maps that will be used to determine the correct orientation of the two objects. Figures 2 and 3 respectively show the 2D and 3D topographical maps of α –chymotrypsin. Figure 4 shows the 3D energy map (van der Waals + electrostatic) of α –chymotrypsin. Figures 5 and 6 respectively show the 2D and 3D topographical maps of trypsin inhibitor. Figure 7 shows the 3D energy map (van der Waals + electrostatic) of trypsin inhibitor.

Matching

The larger (receptor) of the two maps is held immobile, while the smaller is translated and rotated over the surface as part of the search for topographical complementarity. Simple translation of the smaller map (ligand) along the longitude and latitude of the larger map corresponds to moving the ligand over the surface of the receptor, with the orientation of the ligand fixed with respect to the center of receptor; i.e. the same 'face' of the ligand is presented towards the receptor surface as it moves across the surface. Of course, knowledge regarding the correct 'face' is not generally available, thus for each particular coordinate along the receptor's longitudinal and latitudinal axes, the ligand is rotated about its center in order to present all possible faces to the receptor. In two dimensions this is accomplished by building variations of the ligands map that differ in the equatorial and meridian planes used to generate them. The result is a large series of maps of the smaller ligand that collectively represent all possible orientations of the

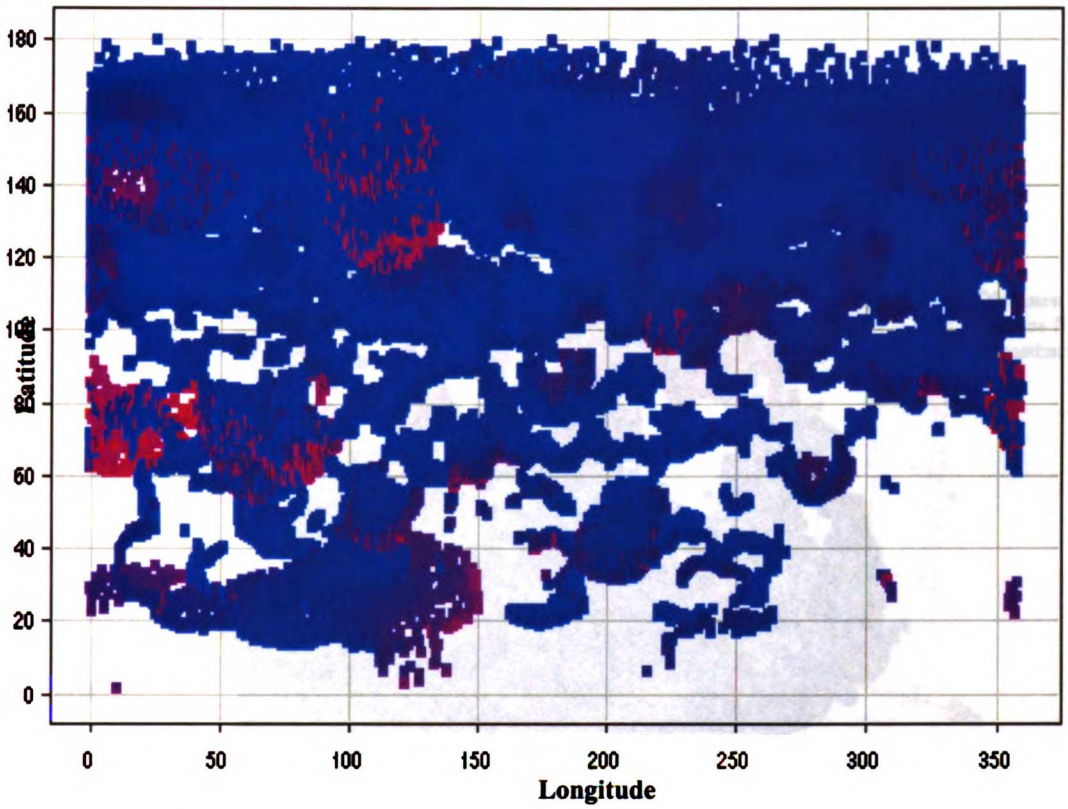


Figure 2. The result of performing the 3D to 2D projection on α-chymotrypsin's solvent accessible molecular surface. The figure shows the density of the points along the longitudinal and latitudinal axes. Variations in color (from blue to red) reflect distances of the molecular surface at a particular coordinate from the sphere surface. Light blue areas are areas closer to the surface, while red areas are farther away from the surface. As expected, features are best preserved along the equatorial regions of the map, while the extremities of the map suffer from a certain amount of distortion.

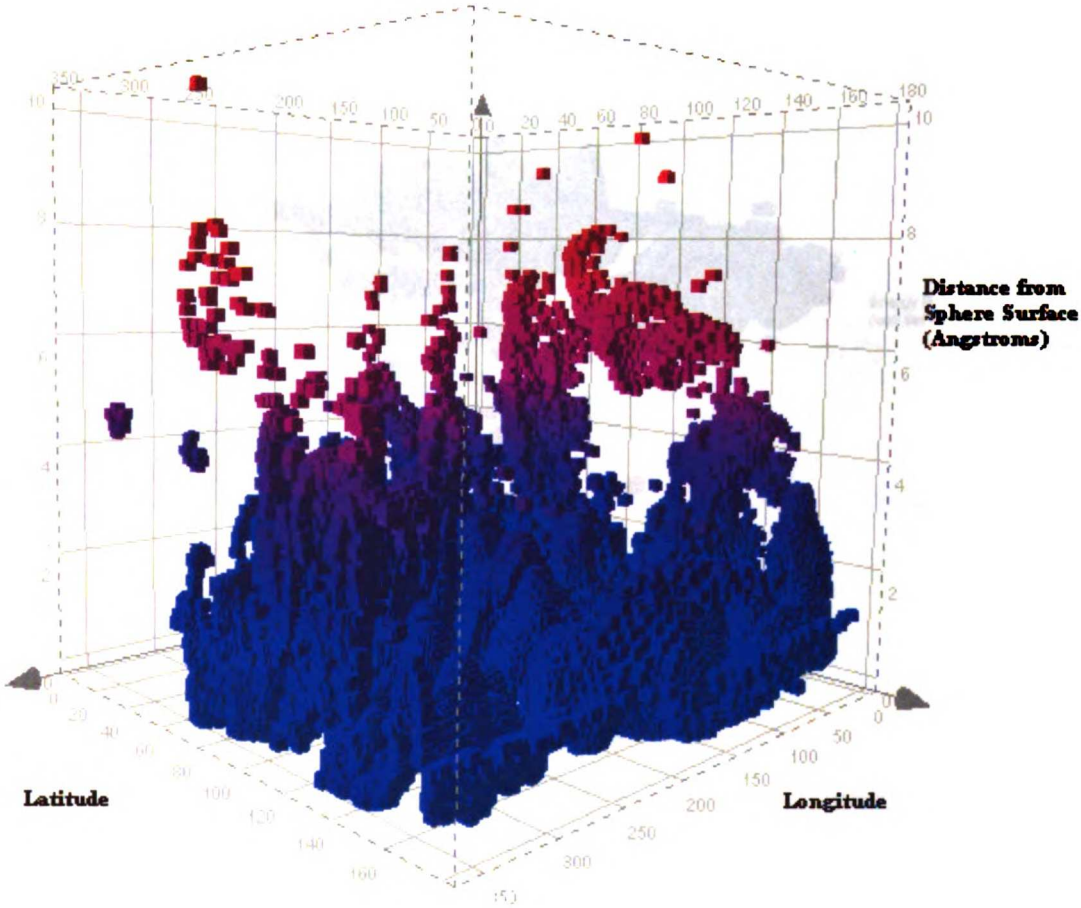


Figure 3. Projected elevations of the molecular surface of α-chymotrypsin. Distances of each from the surface of the projection sphere are plotted on the z-axis. Surfaces closer to the sphere surface are shown in blue, while surfaces farther away are shown in red. The method generally produces surfaces that preserve the local curvature of the molecular surface. The degree of precision is limited by the size of the bins used to store information of the molecular surface. This particular map generated using bins of 0.5 Å on a side; a map generated using bins of 0.1 Âs would generate 25 times the number of points.

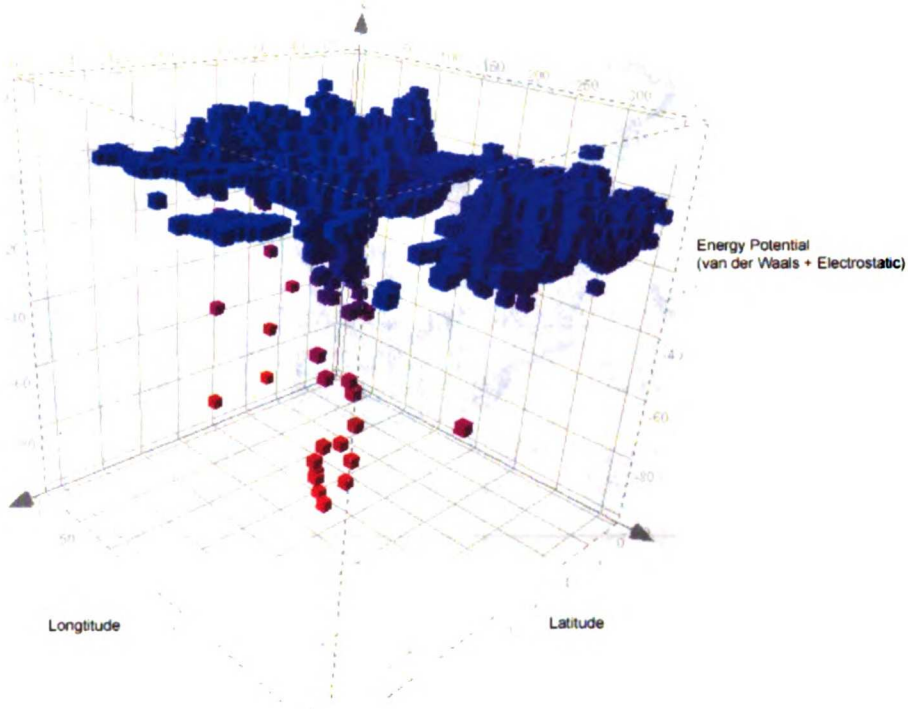


Figure 4. Projected energy surface of α-chymotrypsin. The energy potential of the sphere surface plotted on the z-axis. The color scale ranges from red (negative values) to blue (positive values). The values for the potential are derived by summing the van der waals and electrostatic potentials at a particular coordinate on the map.

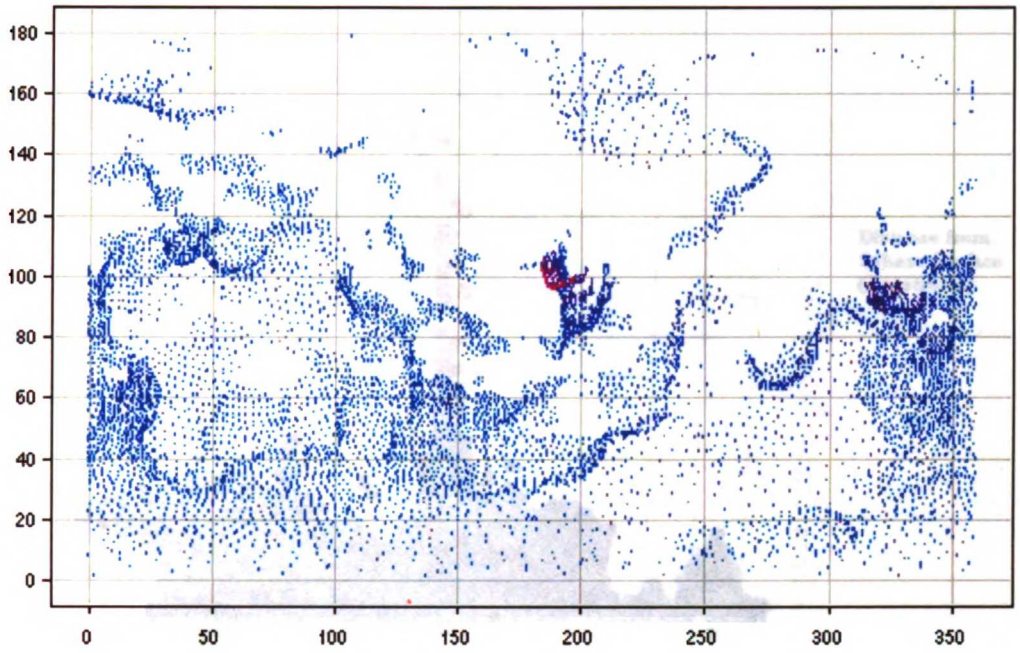


Figure 5. Similar to figure 2, this figure is the result of performing the 3D to 2D projection on trypsin inhibitor's solvent accessible molecular surface. The figure shows the density of the points along the longitudinal and latitudinal axes. Variations in color (from blue to red) reflect distances of the molecular surface at a particular coordinate from the sphere surface. Light blue areas are areas closer to the surface, while red areas are farther away from the surface. The single cluster of red dots in the center corresponds to the C-teriminus of the protein, which protrudes far above the surface of the sphere (figure 6). As expected, features are best preserved along the equatorial regions of the map, while the extremities of the map suffer from a certain amount of distortion.

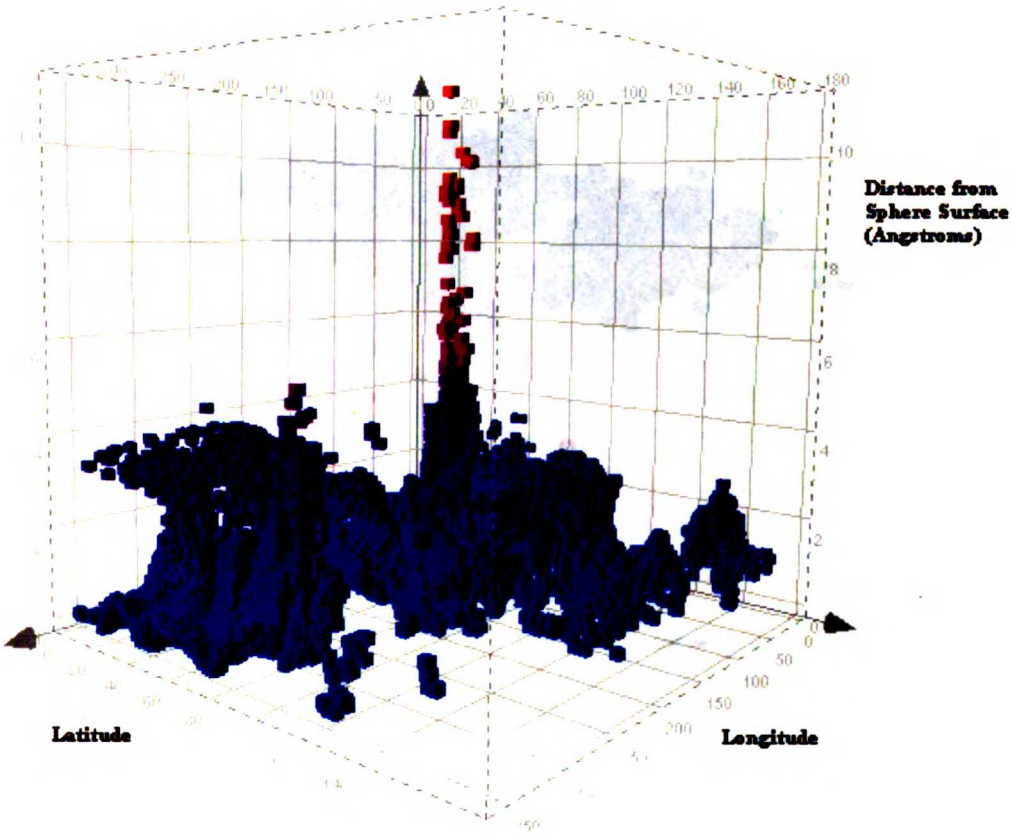


Figure 6. Projected elevations of the molecular surface of the trypsin inhibitor. Distances of each from the surface of the projection sphere are plotted on the z-axis. Surfaces closer to the sphere surface are shown in blue, while surfaces farther away are shown in red.

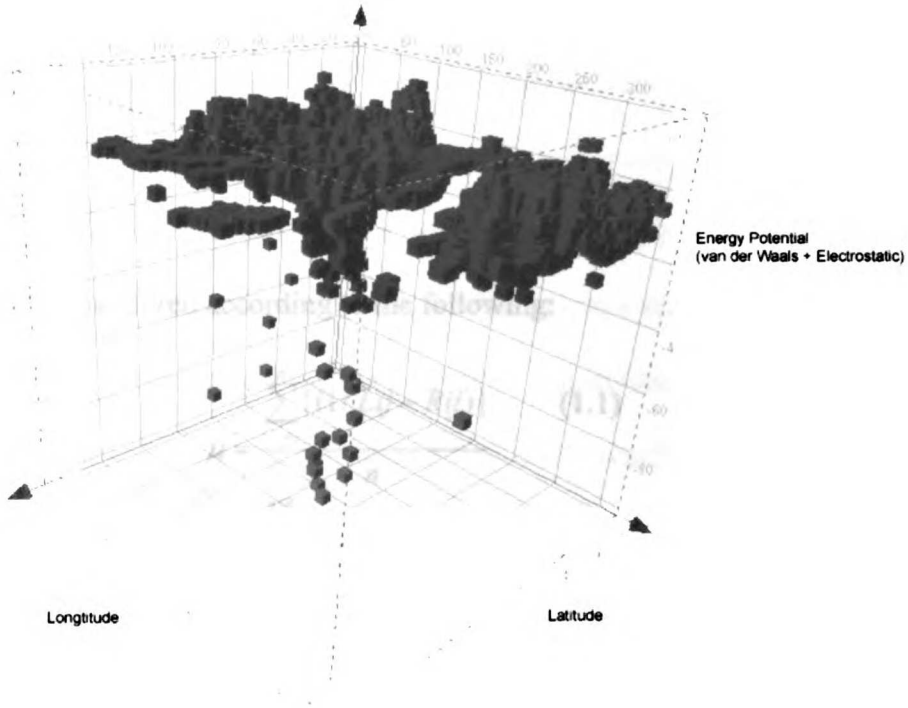


Figure 7. A 3D projection of the energy potential surface of trypsin inhibitor, with the energy potential of the sphere surface plotted on the z-axis. The color scale ranges from red (negative values) to blue (positive values). The values for the potential are derived by summing the van der waals and electrostatic potentials at a particular coordinate on the map.

ligand when its position is fixed above a particular coordinate on the receptor surface. Figure 8 shows how each 2-dimensional movement corresponds to a movement in three dimensions.

Scoring

Each orientation is scored according to the following:

$$\mu = \frac{\sum_{i,j=1}^{n} |(1/Lij - Rij)|}{n} \qquad (1.1)$$

$$s = \frac{\sum_{i,j=1}^{n/8} |1/(Lij - Rij)|}{n/8} \qquad (1.2)$$

$$Score = \sum_{x=1}^{8} \frac{s}{\mu} \qquad (1.3)$$

First, a measure of the average complementarity of the ligand map (μ) and its counterpart on the receptor is calculated (eqn 1.1). Each value element $\mathbf{L_{ij}}$ (i,j are longitude and latitude coordinates, respectively) of the ligand map is subtracted from the corresponding value on the receptor map, $\mathbf{R_{ij}}$. Complementarity exists when $\mathbf{L_{ij}}$ and $\mathbf{R_{ij}}$ are approximately equivalent in value. The inverse of the difference between $\mathbf{L_{ij}}$ and $\mathbf{R_{ij}}$ is higher for complementary values (difference equals 0), than for non-complementary values of $\mathbf{L_{ij}}$ and $\mathbf{R_{ij}}$ that result non-zero differences. The absolute value of all inverses of the differences are summed and divided by the number of points, **n**, to yield an average value for the entire map, μ. Subtracting all corresponding values $\mathbf{R_{ij}}$. from $\mathbf{L_{ij}}$ produces a matrix that is then divided into several sections. This accounts for the fact that only a small percentage of the molecular surface is ever involved in the protein-protein

interaction. In this case, the matrix is divided into eight sections, each encompassing 12.5% of the map. In the future, this value will be dynamic depending on the size of the protein. In the case of not having any *a priori* knowledge of the protein interface, several values may need to be evaluated as part of the experiment. In each section, each value element $\mathbf{L_{ij}}$ of the ligand map is subtracted from the corresponding value on the receptor map $\mathbf{R_{ij}}$. Each section is then given its own score, **s** (eqn 1.2), that results from summing the inverse of the differences of all values in that section, and dividing by the number of points in that section, **n/8**. **s** therefore, is a measure of the local complementarity. Sections with better complementarity will have higher scores than those with lower complementarity. The final score of the orientation (eqn. 1.3), is the sum over all 8 sections of **s** (the local complementarity) divided by the average complementarity of the whole map, μ. Orientations that have a high degree of complementarity will have a higher score.

This scoring scheme is limited to the evaluation of protein interactions based on topological and electrostatic features. However, it will be insufficient to measure complementarities based on van der waals interactions. The Lennard-Jones function predicts that simply taking the differences of two values in the receptor and ligand maps will lead to false positives and false negatives because equivalent values can exist at both ends of the potential function. The most straightforward way to account for van der waals interactions is to score the interaction in three dimensions, while continuing to optimize the orientations in two dimensions. The conversion of a two dimensional orientation of a single point to three dimensions is very rapid, yet will undoubtedly slow the optimization process. Nevertheless it is the most reliable method of accurately

scoring the van der waals interaction between the two proteins. A two dimensional method of calculating van der waals interactions, while desirable from the point of view of computational speed and overall algorithm efficiency, would require assumptions about point to point distances in two dimensions that are either weakly relevant or completely invalid in three dimensional terms. It is noteworthy that none of the two dimensional docking methods attempt to use van der waals or electrostatic potentials to establish complementarity of two proteins, relying instead on physical (potential of mean force, for example[25]) characteristics to derive additional information about the complex as a preprocessing step.

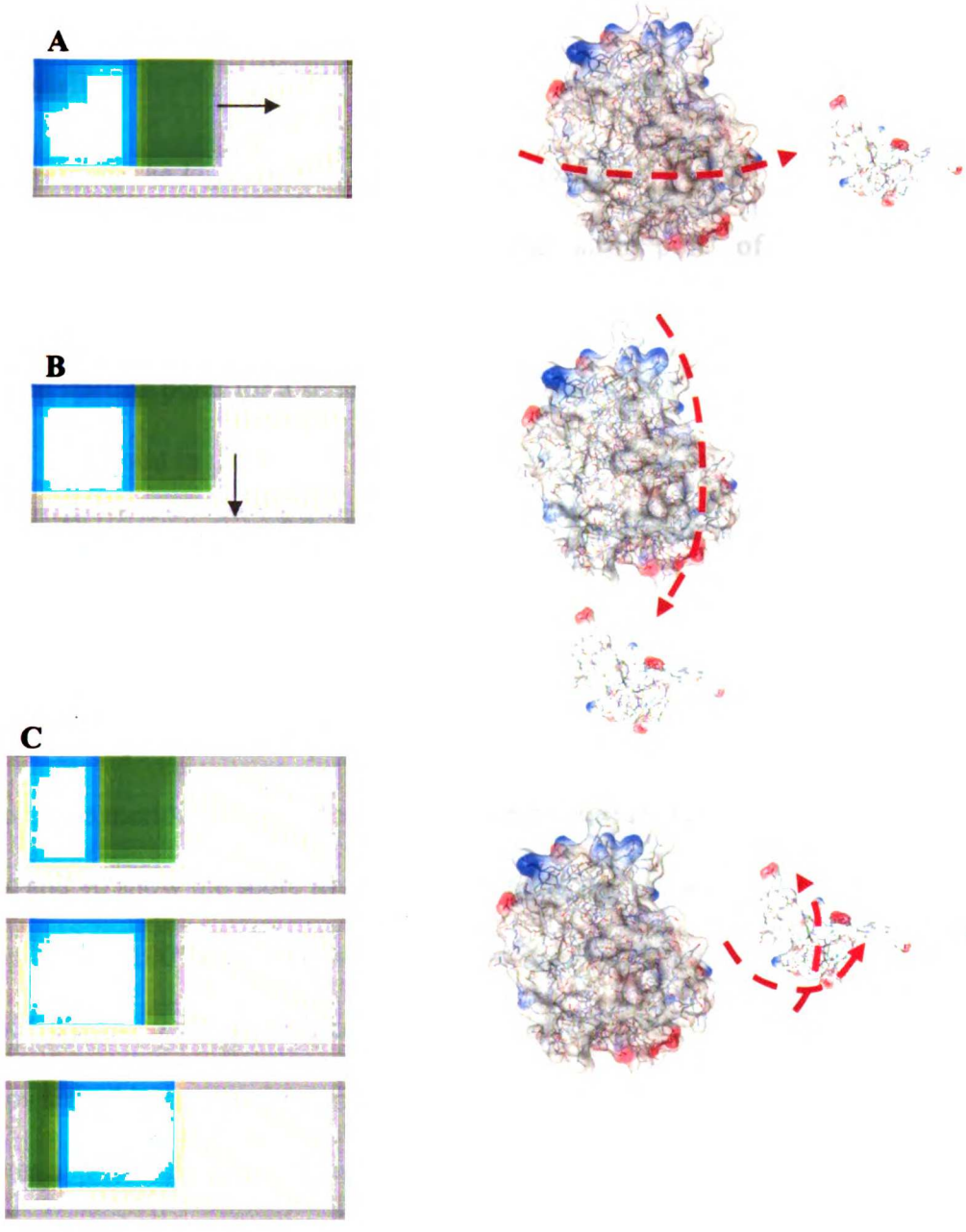


Figure 8. Correlation between movements in two dimensions with movements in three dimensions. **A and B** Translation of the ligand map (dark blue and green composite) over the receptor map (grey) corresponds to movement of ligand along the surface of the receptor. **C.** At a particular coordinate of the receptor surface, the ligand is rotated about its own center in order to evaluate complementarity. The corresponding movement in 2D is to generate and score multiple ligand maps at a single coordinate that vary in the meridians and equatorial planes used to generate the 2D representation. The lower three maps show this process. The dividing line between the blue and green sections indicate the original meridan and equatorial plane used to generate the map. Generating 2D maps using other meridians and equators along the ligand's axis result in multiple maps that vary in which face is presented to the receptor.

Results

Translating the Ligand Across the Receptor Surface

The ligand is first translated to the north 'pole' of the receptor. The poles, equators, and meridians are arbitrarily chosen in the first iteration. The north pole serves as the starting point for a series of translations along the longitudinal and latitudinal axes. At a later point in the project, we may attempt more informative definitions of the protein poles. The orientation of the ligand relative to the surface is kept constant; i.e. the ligand always presents the same 'face' to the receptor. Furthermore, the receptor to ligand distance is kept at constant, reducing the degrees of freedom that have to be explored as part of the proof-of-concept. The native orientation lies 100° 'south' of the 'north' pole along the meridian of the receptor's molecular surface (100° South, 0° East). Figures 9a, 9b, and 9c show the score as the ligand is translated in increments of ~3.2°. As the ligand approaches 100° south, the scores start to increase. Similarly, as the ligand approaches 360°, the scores start to increase. Figure 10 illustrates the translations, and figure 11 shows the highest scoring orientation encountered. A 3.2° resolution scan of the surface takes approximately 1 hour using unoptimized python code.

Rotating the Ligand at a Single Coordinate

The ligand is positioned in its native orientation relative to the receptor, and rotated around its geometric center in order to evaluate the scoring function's ability to further evaluate the complementarity of two surfaces. The results are shown in figure 12. At first glance, the data indicates that as you rotate the ligand around its center, the scores drop from a high of ~85,000 down to scores in the low teens. This is to be expected.

What are anomalous are the presence of other high scoring rotations and the lack of correlation between score and rotations in the latitudinal direction. A scoring function should be sensitive to both types of movements because neither one is completely reliable. The data indicates, however counter intuitively, that longitudinal rotations alone will eventually lead you to the correct orientation. This illustrates a lack of sensitivity on the part of the scoring function to find optima in cases where surface similarities may exist; given the spherical shape of the trypsin inhibitor used in this experiment, the scoring function seems to have found a number of false positives.

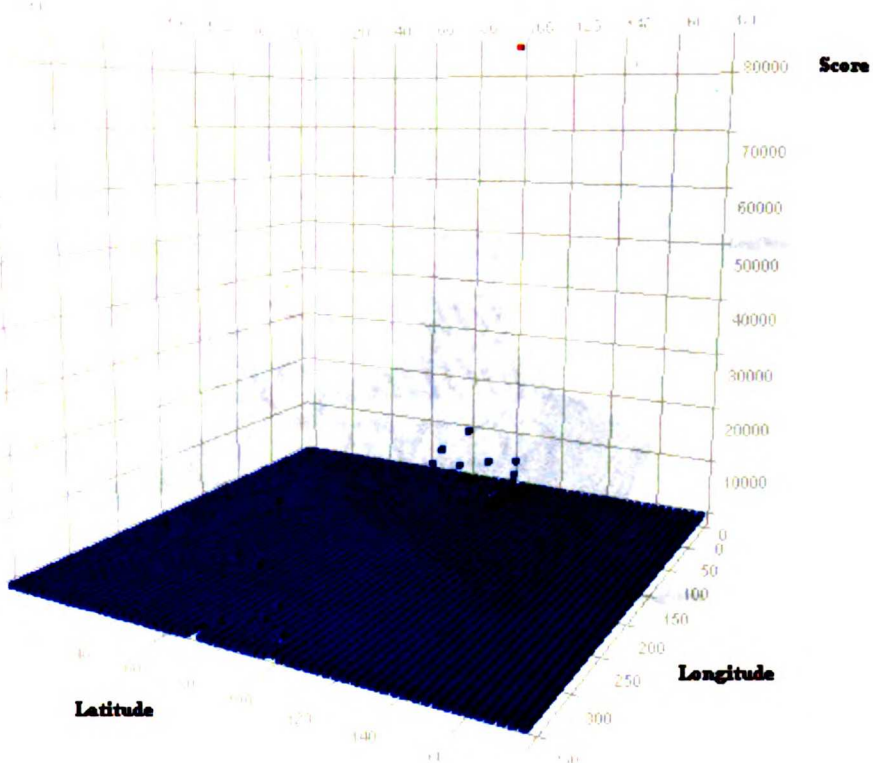


Figure 9a. The results of translating the ligand map over the receptor map, starting at the 'North' pole of the receptor, and moving in increments of ~3° along the longitude and latitude lines. The relative orientation of the ligand is kept constant as it is translated across the receptor map. The native orientation of the ligand lies at 0° longitude and 100° latitude. The scores increase as the ligand map moves closer to its native orientation.

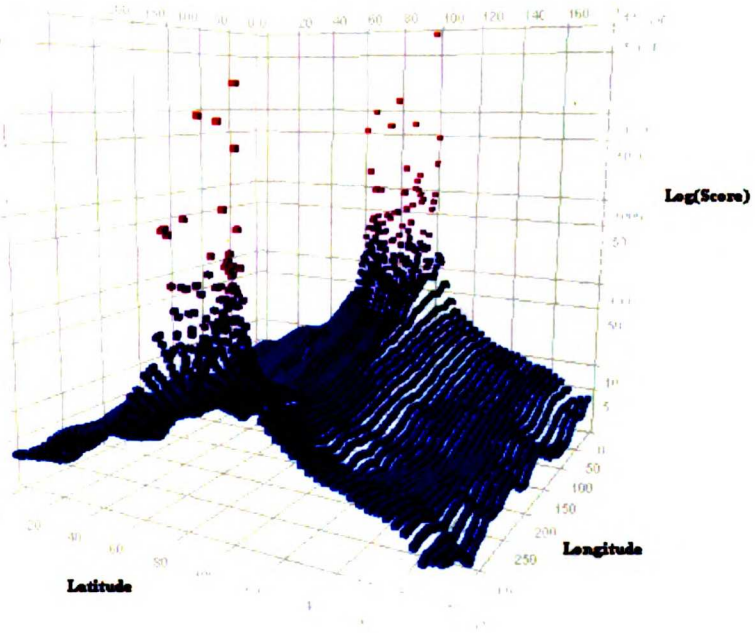


Figure 9b. The results of translating the ligand map over the receptor map, starting at the 'North' pole of the receptor, and moving in increments of ~3° along the longitude and latitude lines. Scores are plotted on a log scale. The relative orientation of the ligand is kept constant as it is translated across the receptor map. The native orientation of the ligand lies at 0° longitude and 100° latitude. The scores increase as the ligand map moves closer to its native orientation.

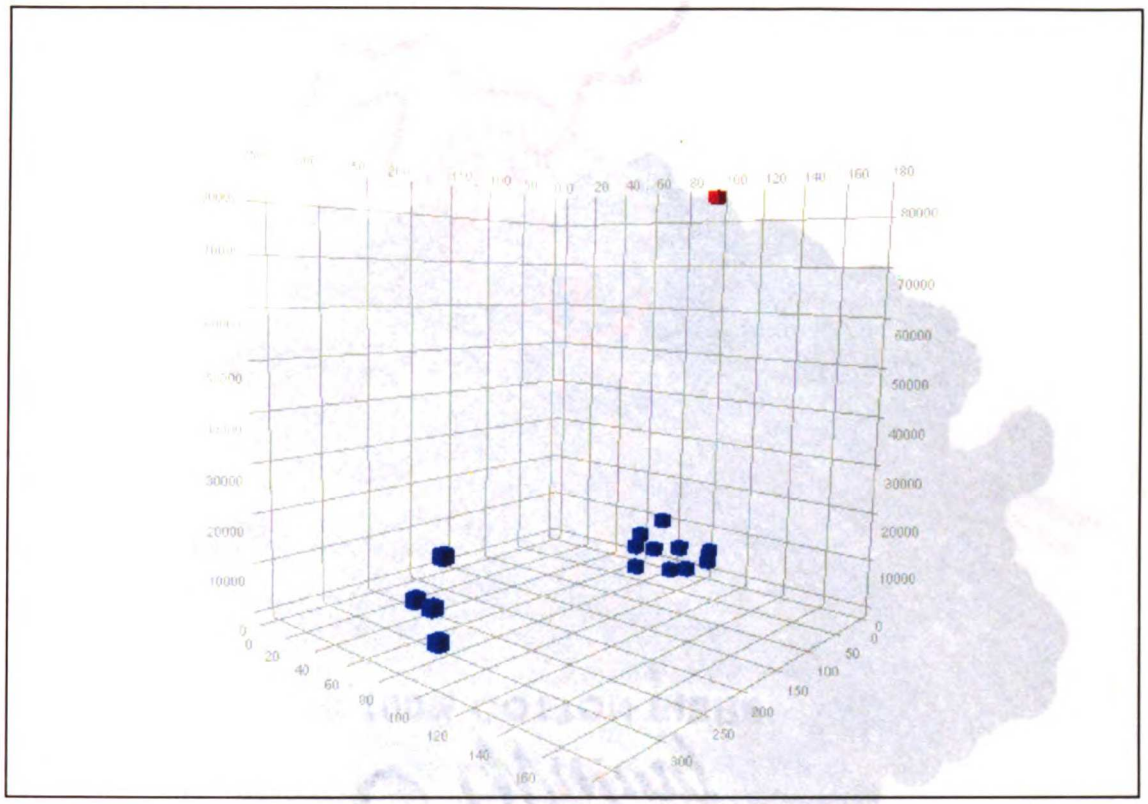


Figure 9c. The results of translating the ligand map over the receptor map, starting at the 'North' pole of the receptor, and moving in increments of ~3° along the longitude and latitude lines. Only scores above 50 are shown. The relative orientation of the ligand is kept constant as it is translated across the receptor map. The native orientation of the ligand lies at 0° longitude and 100° latitude. The scores increase as the ligand map moves closer to its native orientation.

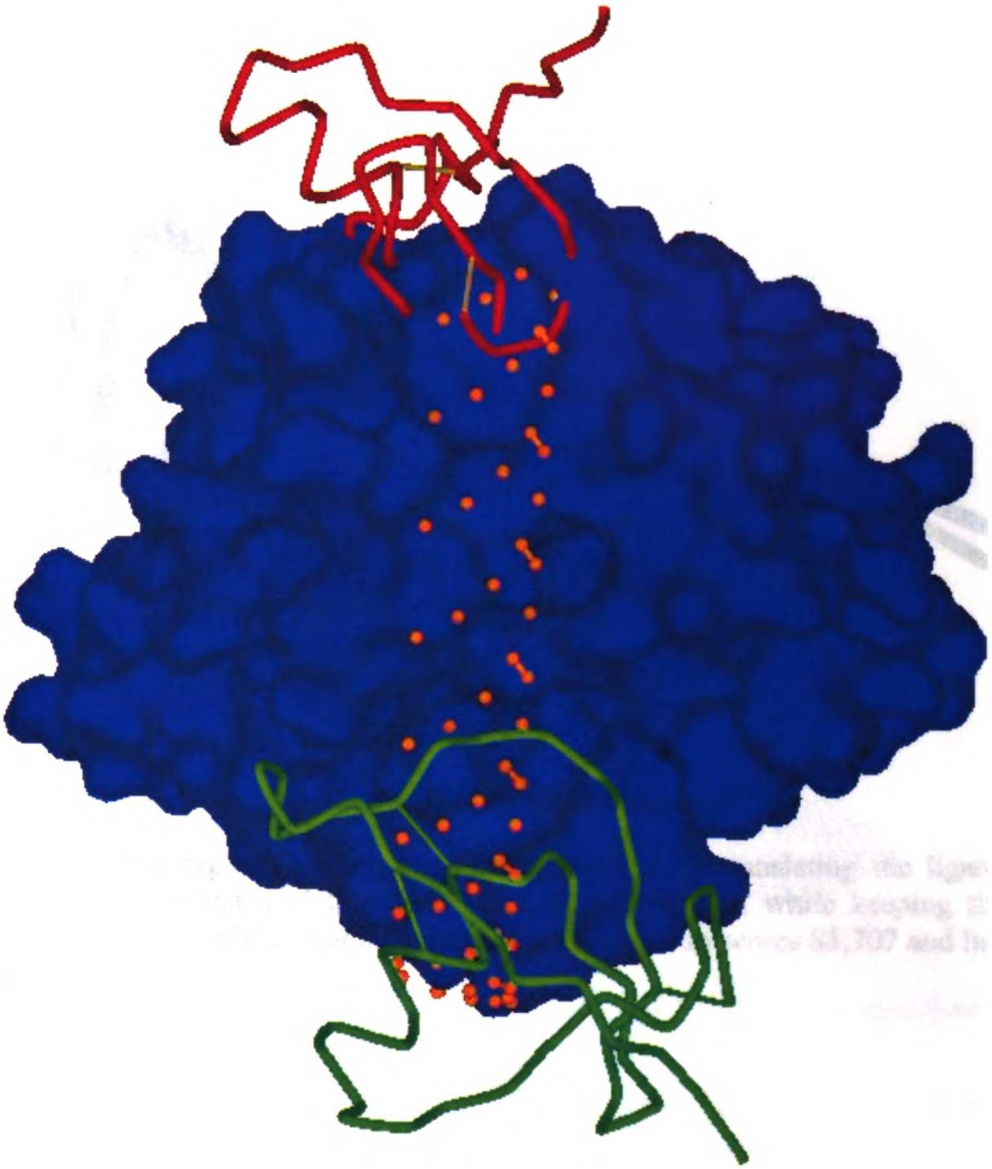


Figure 10. A depiction of the translations performed. The blue structure is the molecular surface of α-chymotrypsin. The green structure is the ligand in its native orientation. The red structure is the ligand translated to the 'North' pole of the receptor, the starting point of the translations.

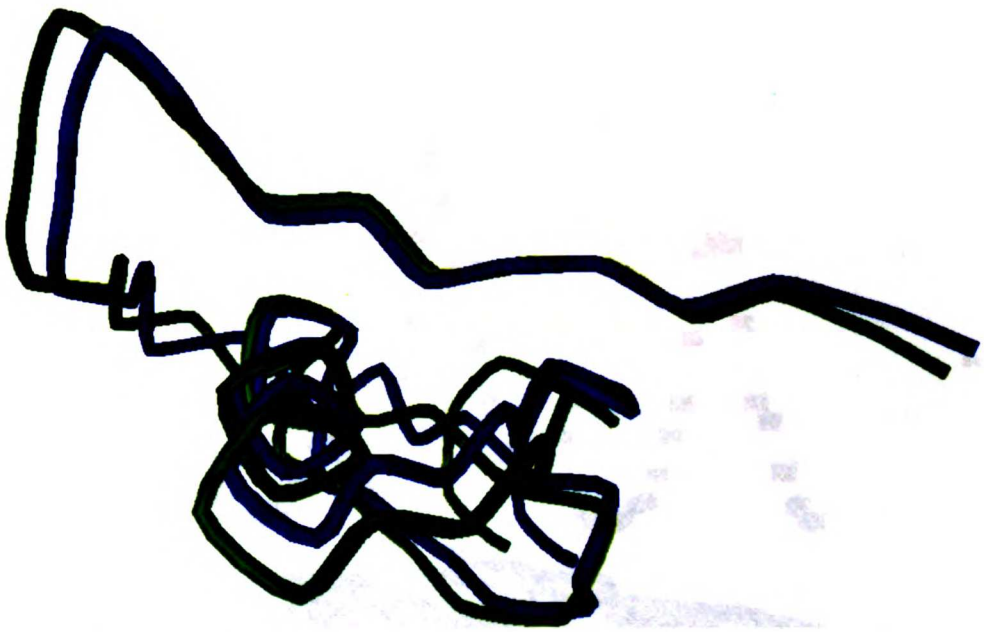


Figure 11. The highest scoring orientation obtained by translating the ligand along the longitudinal and latitudinal axes of the receptor, while keeping the relative orientation of the ligand constant. This orientation scores 85,707 and has an RMSD of 1.58 Å.

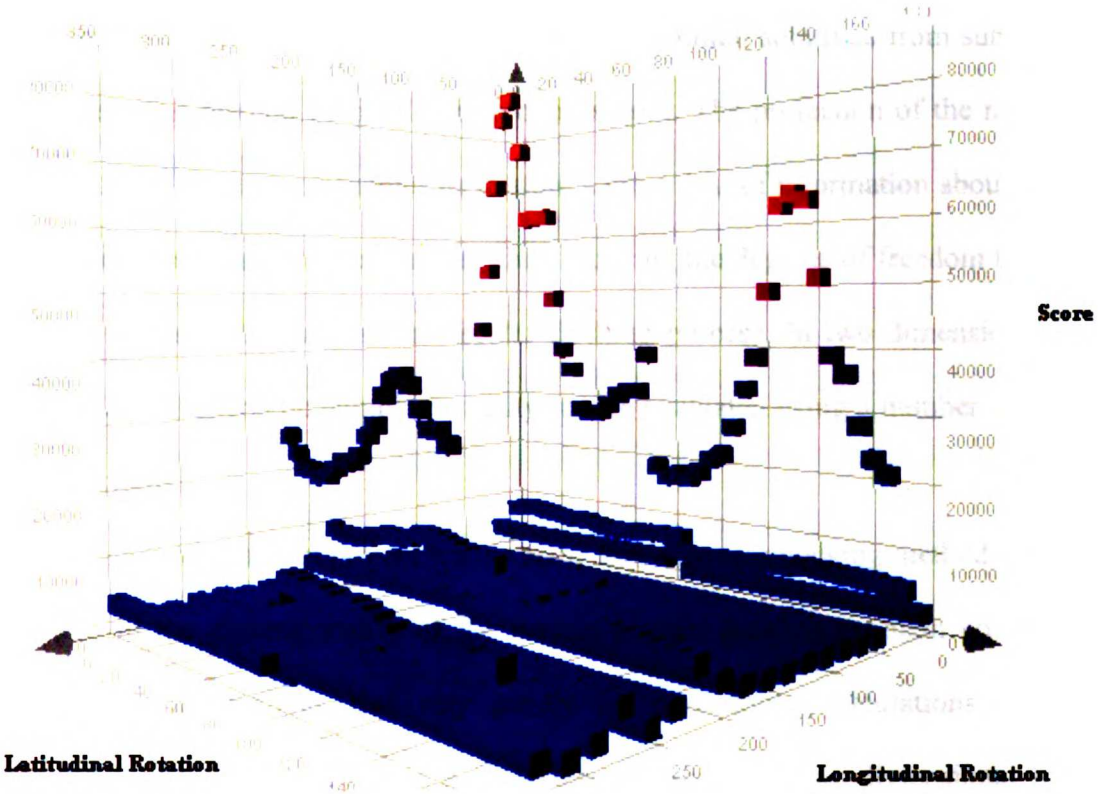


Figure 12. The results obtained from rotating the ligand about its geometric center in its native orientation relative to the receptor surface.

Discussion

Macromolecular docking has two components: a systematic search through the orientational degrees of freedom of the two objects relative to each other, and the development of a scoring function that distinguishes optimal solutions from sub or non-optimal ones. The method outlined here proposes both. The projection of the molecular surface onto two dimensions enables us to store a great deal of information about surface topologies and electrostatic information, while reducing the degrees of freedom that have to be explored as part of a macromolecular docking process. In two dimensions, both topographical and energetic complementarity can be evaluated using a number of simple scoring functions.

The goal of this pilot project was to show that a protein docking method that uses a reduction in the dimensionality of the protein-protein docking problem could yield results similar to methods that rely purely on performing calculations in three dimensions. The representation of a molecular surface as a two-dimensional object simplifies the amount of information that can be extracted from the surface. A finer two dimensional grid will store a much more complete representation of the molecular surface. The number of points on the grid is directly proportional to the radius of the sphere used to map the molecular surface, and inversely proportional to the size of the bins used to store surface data. Because points lying above a given grid section are averaged into a single point, at any given radius a smaller bin size will result in a more complete representation of the surface; the grid spacing will be smaller. Similarly, increasing the radius of the sphere will also increase the number of grid points. The amount of information stored in the representation, therefore, is also proportional to the

ratio of sphere radius to bin size. The dependency on ratio and grid increments would be reflected in a 3D→ 2D→ 3D conversion of the molecular surface. The completeness of the subsequent 3D surface would increase as the grid increments decrease.

The results indicate that the method is capable of identifying large indentations or protrusions, but only partially successful in noticing smaller 'bumps' on the surface caused by a single amino acid side chain. The scores reflect gross topographical complementarities or incongruities (as one might expect when one passes a ligand over a whole receptor surface), but are not very sensitive to local variations in structure. Specifically the method finds a large number of false positives when the ligand is held at its native position relative to the receptor and rotated about its center of mass, effectively presenting different faces to the correct docking site on the receptor. These results are consistent with the idea that a loss of information due to the transformation process would decrease the method's sensitivity to local variations in topology. Nevertheless, it is still fair to say that the project met its goal of demonstrating that a two dimensional method of molecular topologies holds some promise as an alternative to traditional protein-protein docking methods.

The results, however hopeful, are limited in their applicability for a number of reasons. Firstly, the method has only been shown to work on a single complex. In order to be truly useful as a screening tool, a method must be able to discern a protein's partner among many other proteins. Given the method's current performance, it would most likely be unable to distinguish between several structurally related proteins. Successfully differentiating between trypsin and chymotrypsin, for example, would require that the method be able to distinguish between specific amino acid residues on the molecular

surface. The method lacks the sensitivity to do so at this point in time; it should be noted that the exact nature of its insensitivity has not been fully explored, and warrants serious investigation in the future. For obvious reasons, the method, even in its current state of development, needs to be evaluated on many other complexes in order to fully characterize its strengths and limitations. A more robust test-set would include Ab/Ag, DNA/enzyme, and other complexes that vary in shape and size. Secondly, the method assumes that proteins are globular in shape. This is only true for a very small set of proteins and enzymes. The majority of proteins, and the pharmaceutically interesting proteins in particular (GPCRs, for example) are not spherical at all, or even oblate spheroids. The method would benefit greatly if it utilized a number of different shapes upon which to project the molecular surface. Some proteins would benefit from the use of a spheroid, others proteins or nucleic acid structures might be best represented by a cylindrical or conic projection. It is interesting to note that cartographers have used similar methods to achieve specific gains or losses in distortion and representation in particular projections. An idealized version of the approach outlined here might first determine the appropriate object upon which to project the molecular surface by evaluating the overall characteristics of the protein. Lastly, the method still lacks a robust search routine that finds the global optima. The results presented here are the result of brute-force evaluations of a specific subset of the search problem. I evaluated either translations of the ligand over the receptor, or rotations of the ligand in a single location, never both at once The method's speed (and therefore, its utility) is most sensitive to the degree of sampling (the number of bins) used to generate the 2D map, on the order of $O(n^2/2)$ where n is the number of bins used. The current implementation would be

impractical to use a brute-force search methodology for day-to-day database screening because it is simply too slow. Nevertheless, it is unclear whether the method would benefit most from a monte-carlo, a simplex, or a genetic algorithm-based optimizer. What is clear is that having one of them is most likely better than having nothing at all. It is not unrealistic to conceive of achieving performance ratios similar to those achieved by the implementation of the simplex optimizer in traditional docking versus the theoretical time for a brute-force docking.

Two successful methods for protein-protein docking have been developed by Shoichet and Sternberg. In a recent article, Lorber and Shoichet further explore the applicability of ensemble docking (a 3-dimensional docking method) to the protein-protein docking problem[24]. In their investigation, they find that docking proteins with flexible side-chains leads to better distinction between native and non-native orientations. Their methods achieve 1.22 Å RMSD in the best case, and 23.44 Å RMSD in the worst case, but at the cost of time. Their evaluation of bound α-chymotrypsin/inhibitor complex took 4.69 hours and resulted in an RMSD of 1.91 Å. The method outlined here took significantly less time and achieved similar results (1.58 angstrom RMSD) in 1.2 hours; it is highly probable that this method can take less time when matched with a robust optimizer. Sternberg et al. use a variation of ftDock, a Fourier transform based docking algorithm to dock proteins as well as small molecule ligands to receptors. Their method has been shown to work on numerous systems, including the α-chymotrypsin/inhibitor complex used here[25]. Recent results using a combination of ftDOCK and pairwise surface potentials yielded a 6.03 Å RMSD for the complex; there was no discussion regarding the speed of the algorithm. As indicated above, this method achieved a better result simply

on the basis of RMSD alone. The significant caveat to these results is that the method has only been evaluated on one complex. Nevertheless there is some hope that the method, when fully optimized and evaluated, will be competitive with the field's leading approaches.

As the dependence on genome derived structure-based design increases, there will be a stronger need to quickly hypothesize about the function of an unknown ORF. Threading and other structural genomics methods may be able to discern a 3-dimensional structure, but may not answer all the questions about the function of the protein. Protein-protein docking methods add a great deal of complementary information to the analysis. By designing an algorithm, and evaluating its strengths and weaknesses at some level, I hoped to lay the foundation for further exploring the value of a 2-dimensional projection-based docking method. However, there remains a great deal of work to be done on this project in order to bridge the gap between proof-of-concept and practical application. Much of that work is mentioned in the preceding paragraphs. The method's perceived strengths and weaknesses are a partly a result of performance, and partly a result of not extensively characterizing its behavior under many conditions. The next set of activities would ideally focus on completely understanding its strengths and weaknesses under computational duress. Evaluating its efficacy on different proteins, different complexes, and on non-complexed proteins or nucleic acid structures would determine the next phase of development, and set performance standards by which to gauge the success of further enhancements. With more rigorous evaluation and development, it will eventually withstand the scrutiny that a public user-base would place on it, and add to our ability to understand proteins of unknown function.

References

1. Gabb, H.A., Jackson, R.M., Sternberg, M. J.E. *J Mol Bio.* **1997**. 272, 106-120.
2. Chen, R., Weng, Z. *Proteins: Struc Func Gen.* **2002**. 47, 281 – 294.
3. Norel, R., Petrey, D., Wolfson, H.J., Nussinov, R. *Proteins: Struc Func Gen.* 1999. **36**, 307-317
4. Camacho, C.J., Vajda, S. *Proc Nat Acad Sci.* **2001**. 98, 10636-10641.
5. Jackson, R.M., Gabb, H.A., Sternberg, M.J.E. *J Mol Bio.***1998**. 276, 265-285.
6. Hendrix, D.K., Klein T.E., Kuntz, I.D. *Prot Sci.* **1999**. 8, 1010-1022.
7. Jones, S., Thornton, J.M. *J Mol Bio.* **1997**. 272, 121-132.
8. Zhang, C., Chen, J., DeLisi, C. *Proteins: Struc Func Gen.* **1999**. 34,255-267.
9. Janin, J. *Prog Bio-phys Mol Bio.* **1995**. 64,145-166.
10. Shoichet, B.K., Kuntz, I.D. *Chem Biol.* **1996**. 3,151-156.
11. Jiang, F., Kim, S. *J Mol Bio.* **1991**. 219, 79-102.
12. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesen, A.A., Aflalo, C. Wodak, S.J. 1992. *Proc Natl Acad Sci.* **1992**. 89, 2195-2199.
13. Walls, P.H., Sternberg, M.J.E. *J Mol Bio.* **1992**. 228, 277-297.
14. Helmer-Citterich, M, Tramontano, A. *J Mol Bio.* **1994**. 235, 1021-1031.
15. Shoichet, B.J., Kuntz, I.D. *J Mol Bio.* **1991**. 221,327-346.
16. Norel, R., Lin, S.L., Wolfson, H.L., Nussinov, R. *J Mol Bio.* **252**, 263-273.
17. Gilson, M.K., Honig, B. *Proteins.* **1988**. 4,7-18.
18. Vakser, I.A., Afloa, C. *Proteins.* **1994**. 20, 320-329.
19. Jackson, R.M., Sternberg, M.J.E. *J Mol Bio.* **1995**. 250, 258-275.
20. Weng, Z.P., Vajda, S., DeLisi, C. *Prot Sci.* **1996**. 5, 614-626.

21. Totrov, M., Abagyan, R. *Nature Struct Bio.* **1994.** 1, 259-263.

22. MidasPlus. *UCSF Computer Graphics Laboratory.* San Francisco, CA.

23. SYBYL, v6.5, Tripos, Inc., St. Louis, MO.

24. Lorber, D.M., Udo, M.K., Shoichet, B.K. *Protein Sci.* **2002.** 11:1393-1408.

25. Moont, G., Gabb, H.A., Sternberg, M.L. *Proteins: Struc Func Genetics.* **1999.** 35,364-373.

Conclusions

Three decades of development in structure-based design have made this science, its applications, and its practitioners vital elements of pharmaceutical and biotechnology research and development. Nevertheless there remains a great deal of further research to be done before the methods developed to date become purely predictive, instead of seeming periodic in its successes.

The preceding work in de novo design grew out of an interest in overcoming known limitations in computation, as well as a desire to provide further proof of the validity of structure-based design. The development of an algorithm capable of generating small molecule scaffolds without the use of a small molecule library treads upon a path a number of researchers had gone before. The fundamental desire to overcome the dependence on virtual libraries drove many scientists to the limits of their creativity and intellect only to conclude that the gap between the most computationally correct scaffold, and that which is truly effective and synthesizable, is enormous. These results are echoed in our own research. While the ADAPT algorithm demonstrated a strong ability to identify good scaffolds, it almost never identified the optimal solution. These results are partially a function of our use of a genetic algorithm to optimize the scaffolds, but they can also be attributed to the oversimplification of the biological and chemical methods used to optimize a chemical scaffold.

In chapter II I chose to pursue a more manual route to the de-novo design of a scaffold that targets PDZ domains. Here the algorithm was one that combined our most reliable structural-design approaches with significant medicinal chemistry insight provided by Naoaki Fujii and Kip Guy. In retrospect, the process, however contrarian it

may have been to my predilection for purely computational approaches, was successful only because we made a distinct effort to rely on consistent two-way communication of the insights gained from the computational and medicinal chemistry approaches. It validated several key ideas: (1) That computational chemistry can be used successfully to generate promising compounds, (2) given its state of development, it cannot exist in a vacuum, and (3) that a multi-faceted approach to structure-based design has a high probability of success.

Chapter III details early proof-of-concept work on an approach to increase the efficiency of protein-protein docking. Although it does not quite fit into the de-novo design paradigm, it stems from an inherent desire to make protein-protein docking a more accessible and useful approach to understanding protein-function. The results indicated that the approach did indeed meet its objective of proving that a 2D docking method could yield meaningful results. However, in order to extend the utility of the method, many more test cases will have to be performed in the near future.

There is a great deal of fundamental research that needs to be done in the future. Several key future directions that research should venture into are the further development of better energy scoring functions, as well as the design and implementation of new optimization algorithms that leverage the advances in computer hardware to make these functions computationally tractable. There should be a continued focus on trying bridge the gap between a medicinal chemist's intuition, and an algorithm's ability to deliver synthetically feasible compounds. Until the latter is solved, de-novo design algorithms will ironically be dependent on other, arguably smaller, databases of atoms, chemical fragments, or training sets that introduce some bias into the design process. We

should continue push protein-protein docking towards its potential as a screening tool for proteins and enzymes of unknown functions. This, coupled with better threading algorithms, could significantly enlighten those seeking to understand unclassified proteins.

The work contained herein, as in other publications, validates the approximations, algorithms, applications, and conclusions developed over the years. Nevertheless the field still holds an unlimited amount of potential as a platform with which one might understand and simulate processes relevant to drug discovery, agricultural biotechnology, and even allied biological fields such as nanotechnology and biological warfare. The key will be our continuing efforts to leverage developments in computer science, physics, chemistry, and mathematics with fundamental insights into biological processes. We must continue to emphasize the critical relationships that exist between various scientific disciplines; the field will only progress by successfully leveraging developments that occur in parallel in various disciplines.