Molecular Docking Towards Drug Discovery: Improving Interaction Specificity

by

Daniel Andrew Gschwend

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Chemistry

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco

[redacted]

Preface

The thoughts contained in the following chapters would not have been assembled were it not for the support and encouragement provided by many friends and colleagues. I am exceptionally grateful to my advisor, Prof. Kuntz, for his continual motivation, perceptive vision, boundless generosity, and diplomatic finesse. Tack's unique ability to extract sense from the nebulae which lie beyond distant horizons has never ceased to amaze me. To witness Tack deconvolute the entire health care crisis into the span of 45 minutes, to hear Tack ramble about hyper-dimensional representations of complex objects, to observe Tack brainstorm about a "molecular awk" (or mawk, as we fondly called it) - these are inspiring snapshots indeed. We in the Kuntz group are privileged to have such talent at our fingertips. I am further indebted to Tack for sanctioning scientific undertakings of "questionable basis," and for putting up with my incorrigible procrastination. Finally, I am grateful for Tack's economic prowess, which enabled me to travel to meetings in Boston, Buffalo, Chicago, and Switzerland; ensured that a workstation sat on every desk, however small; and financed a seemingly endless supply of Post-It™ notes to organize my graduate career (hats off to 3-M Corp. for being so damn ingenious).

The unfolding of this dissertation would have been incomplete without help from Profs. Fred Cohen, Peter Kollman, and Dan Santi - I thank them for their generous assistance and guidance throughout. Teri Klein is deserving of my gratitude for advice and encouragement on numerous occasions. Paul McCloskey was instrumental in coordinating the release of DOCK 3.5 and a pleasure to work with.

I am particularly indebted to Andy Good for his eternal optimism, motivation, generosity, and confidence. Andy's pragmatic perspective was always refreshing. He was the ideal coworker and the best of friends. I will miss our Canadian-rule eight-ball and table tennis marathons. Many thanks are due Fiona McPhee for her lively spirit and reassuring nature. I owe Brian Shoichet tremendously for early inspiration and for instilling in me the activation energy required for wet-lab entry. Connie Oshiro frequently and unselfishly offered advice and shared in system administration woes; Todd Ewing taught me to question everything. I am grateful to Elaine Meng, Dennis Benjamin, Rob Cerpa, and Diana Roe for encouragement from the very beginning. The entire Kuntz group deserves credit for making UCSF a welcome and pleasurable environment within which to do science.

A graduate career without buttressing by family would be unimaginable. I am indebted my mother, Kathrin, for her enthusiasm and everlasting faith; to my father, Heinz, for motivation and direction; and to my brother, Dominik (to whom I can never again say "it ain't rocket science"), for making me proud. The affection and recognition offered by my in-laws, Tom and Mary, have been a continual source of reinforcement. I can hardly express the gratitude due my wife, Michele, for her unconditional support and understanding.

Dissertation Abstract

Molecular Docking Towards Drug Discovery: Improving Interaction Specificity

Daniel A. Gschwend

The ever-increasing rate at which structural information is procured has had a profound impact on our ability to combat disease. Detailed three-dimensional snapshots offer exquisite insights into the molecular recognition events which govern all biological processes. An understanding of how molecules associate offers a window for chemotherapeutic intervention and consequently the possibility for modulating disease states. One area of the growing field of computational chemistry focuses on the identification of agents which bind specifically to a macromolecular target. This objective presents two fundamental challenges which form the basis for this dissertation: locating agents which are *potent* - those that bind tightly, and locating agents which are *selective* - those that bind the desired site preferentially to others.

One computational method for locating novel agents involves scanning a database of pre-existing structures for those which exhibit complementarity to the target. Strategies for "molecular docking" are embodied within a spectrum of models for ligand binding, bounded by the canonical Lock-and-Key and Induced Fit paradigms. I explore the potential of a rigid model in exploiting species specificity and of a tolerant model in predicting absolute ligand binding affinity.

The ability of structure-based drug discovery to address receptor specificity is verified through the identification of novel, selective inhibitors of dihydrofolate reductase from the opportunistic fungal pathogen, *Pneumocystis carinii.* Differential design methodology has enabled the discovery not only of nearly a dozen novel structural frameworks which bind the microbial enzyme in preference to the human variant, but also of one agent which displays *in vitro* potency and selectivity rivaling those of a common therapeutic.

Automated design protocols examine thousands of putative receptor-ligand configurations and demand rapid feedback on quality of association. The calibration of an empirical scoring scheme against over one hundred affinities for experimentally observed complexes has led to a model capable of reproducing observed binding free energies to within 1.7 kcal/mol. Emphasis has been placed on accuracy in predictions, robustness over structural diversity, and speed of evaluation. The resulting tools are likely to be of general use for assessing potency in structure-based drug design.

Table of Contents

List of Tables

Chapter 3

Chapter 4

Chapter 5

Appendix A

Appendix C

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 1.

Introduction

It comes as no surprise that the rapidly increasing amount of structural information available to the pharmaceutical community has attracted particular interest from the computational chemistry arena. Computational chemists, most being avid programmers, enjoy the challenge of encoding rules into algorithmic 1's and 0's. Rules necessarily demand information, and, in theory anyway, the quality of these rules generally exhibits some degree of proportionality to the quantity of information at hand. So, the fruit of the blossoming field of structural biology has enchanted those who would make rules - rules about molecular association. Projecting into the future, then, there will come a day when we truly understand what exactly goes on between a receptor and a ligand within the vast brew which comprises a living organism. But it is already apparent that pharmaceutical science will not simply halt when this time arrives, for new obstacles will present themselves, such as how best to ***use*** this knowledge. Science is, after all, the process of overcoming obstacles in a (more-or-less) systematic way.

The hurdle which now lies directly ahead is how, given structures detailed to one ten-billionth of a meter, can we design a small molecule to bind tightly and specifically to a macromolecular target? Dozens of approaches have been set forth over the last two decades to address various aspects of this remarkably complex problem. Broadly speaking, there are two basic recipes for locating active biomolecules: we can assemble them from atoms or groups of atoms, or we can locate them among a library of pre-existing molecules. Both methods have characteristic advantages and disadvantages. The major drawback of building molecules, particularly in an academic setting where synthetic resources may be limited, is that proposed compounds must actually be *made.* Synthesis is demanding of time, money, and often specialized expertise. These daunting prospects were stimuli enough to avert the Kuntz group towards door #2. (It would be more accurate to phrase it such that this choice was made not out of fear of synthesis, but rather out of aspirations for instant feedback.) The possibility of selecting among commercially available compounds for novel, bioactive agents was exciting indeed. Originating with the work of Renee DesJarlais, the Kuntz group has since pioneered the searching of chemical databases for pre-existing small molecules which might bind to a target receptor of known three-dimensional structure. Central to "molecular docking," to which it is commonly referred, are notions of complementarity: what sticks to what?

The ability to evaluate complementarity impinges not only upon molecular docking, but on all of structure-based drug design. All of what appears in this dissertation revolves around issues related to scoring of putative receptor-ligand associations. In the course of my graduate career, I have explored a spectrum of docking strategies varying in the level of stringency in scoring. As Chapter 2 details, these strategies are embodied within the canonical models for ligand binding: Lock-and-Key and Induced Fit. Chapter 2 represents the proceedings of research I presented at the 36th Annual Buffalo Medicinal Chemistry Symposium and has been accepted for publication in the *Journal of Molecular Recognition.*

Written as a hybrid review/research article, Chapter 2 previews and unifies the work discussed in Chapters 3, 4, and 5. Chapter 3 has been submitted in modified form to the *Journal of Medicinal Chemistry*, an abridged version of Chapter 4 has been accepted into the *Journal of Computer-Aided Molecular Design*, and Chapter 5 is being extended for submission at a later date.

At one end of the spectrum of docking strategies lie the very exacting requirements of a search for selective agents. Chapter 3 documents an investigation directed at identifying therapeutically relevant, species-specific enzyme inhibitors. In many cases, locating an inhibitor which binds the desired target is not enough - the often-overlooked but equally important attribute of a clinical candidate is selectivity over related targets so that cross-reactivity may be avoided. Dale Bodian in the group introduced technology which enabled emphasis to be placed on differing areas between two receptors. While a significant step forward, these methods necessitated gross distinguishing features, such as a unique pocket, which could be capitalized upon. Unfortunately, these opportunities are seldom as common as one would like. Often two receptors much be distinguished which possess only subtly divergent structures, demanding more stringent differentiation schemes. The enzyme dihydrofolate reductase, while historically an immensely successful antimicrobial target, presents precisely such a challenge in combating the opportunistic pathogen *Pneumocystis carinii*. Chapter 3 demonstrates the surprisingly successful use of scoring optimization tools as post-docking filters in locating novel, selective anti-*Pneumocystis* candidates. Differential optimization permits a systematic bias enabling the selection of compounds likely to bind one target but not another.

It makes sense that including score optimization *into* the docking process would improve our capacity to exploit subtle structural features. Not only would this enhance the retrieval of selective agents, but more generally, strengthens our adherence to Lock-and-Key docking by imposing stricter criteria on complementarity. But at what expense? Optimization is a resource-intensive piece of technology - how would this affect docking performance? Unexpectedly, as Chapter 4 explains, the incorporation of on-the-fly optimization into the docking process garners a net *favorable* return on efficiency. In the most comprehensive analysis of configurational sampling yet, the tight relationship between sampling and scoring unfolds.

On the other end of the spectrum of docking strategies lies the generality afforded by a tolerant, universally-applicable scoring function. As molecular docking by design presents thousands of putative ligands to a receptor which has never seen them before, a bit of clemency in deciding how they might interact is warranted. What is required is an evaluation function capable of estimating (rapidly, no less) free energies of binding for a structurally diverse set of molecular arrangements. This is the brass ring for of all of structure-based drug design. Rather than borrow a scoring method from another branch of computational chemistry, I set out to devise an evaluation scheme which was parameterized on the very values we seek: binding affinities of small-molecule ligands for macromolecules. Chapter 5 discusses the development of an empirical scoring function calibrated against the largest set of binding affinities reported to date. Through careful interaction characterization and statistical analysis, a working model capable of reproducing observed affinities to within 1.7 kcal/mol has been derived. The predictive ability of the model, while validated by statistical metrics, remains to be verified in a practical setting. It is clear, however, that the omission of entropic terms in assessing interaction strengths seriously dampens the true potential of molecular docking.

My graduate expedition has journeyed through Lock-and-Key docking, through species-specificity, through score optimization, and through "alternative" scoring functions; all the while I sat behind my workstation (not entirely - I did do a handful of *real* experiments!) amidst applications work and methodology development. Working on real-world, therapeutically relevant problems was stimulating. Developing code useful to the group and to the scientific community was satisfying. Perhaps most enlightening, however, was the research with which I conclude my tenure here at UCSF: the study of over a hundred structurally diverse molecular assemblies we call proteins. Nature has concealed a wealth of information within the confines of molecular recognition snapshots. The incorporation into structure-based design strategies of empirically-derived evaluation schemes which directly take advantage of this information (such as of the flavor outlined in Chapter 5) have the potential to vastly improve our understanding and the quality of lead discovery.

Chapter 2.

Molecular Docking Towards Drug Discovery

Daniel A. Gschwend, Andrew C. Good,[†] and Irwin D. Kuntz[*]

Department of Pharmaceutical Chemistry,

University of California, San Francisco, CA 94143-0446

[†] Current address: Rhône Poulenc Rorer, Dagenham Research Centre JB4-0, Dagenham, Essex RM10 7XS, United Kingdom.

[*] Author to whom correspondence should be addressed.

ABSTRACT

Fueled by advances in molecular structure determination, tools for structure-based drug design are proliferating rapidly. Lead discovery through searching of ligand databases with molecular docking techniques represents an attractive alternative to high-throughput random screening. The size of commercial databases imposes severe computational constraints on molecular docking, compromising the level of calculational detail permitted for each putative ligand. We describe alternative philosophies for docking which effectively address this challenge. With respect to the dynamic aspects of molecular recognition, these strategies lie along a spectrum of models bounded by the Lock-and-Key and Induced-Fit theories for ligand binding. We explore the potential of a rigid model in exploiting species specificity and of a tolerant model in predicting absolute ligand binding affinity. Current molecular docking methods are limited primarily by their ability to rank docked complexes; we therefore place particular emphasis on this aspect of the problem throughout our validation of docking strategies.

Keywords: molecular docking; empirical scoring schemes; species specificity; DOCK

INTRODUCTION

Overview

Molecular recognition is a problem fundamental to structural biology. The interaction of molecules, be they macromolecules or small ligands, is a prerequisite for nearly all biological events. Specific modulation of these interactions has been the ambition of medicinal chemists for over a century. To gain more rapid access to therapeutic agents, we must not only understand, but be able to predict, the structural details of recognition events. A precise understanding of the basis for complementarity would allow us to venture predictions for purposes of drug design. The inaccuracy of such predictions generally parallels the divergence in the nature of interactions thought to be involved - the brass ring of the field is the quantitative assessment of affinity among structurally unrelated ligands. It is important to bear in mind, however, that true measures of affinity can only be inferred when proper geometries among the components have been established. In broad terms, the prediction of a molecular recognition event embodies two not altogether independent obstacles: the generation of appropriate geometries, and the assessment of complementarity. Although we narrow our discussion to that of small molecule ligands, the general principles are extensible to macromolecular ligands as well.

Structure-based drug design

Structural information is critical to an analysis of molecular recognition events. The experimental determinations which give rise to such data lie at the heart of the structure-

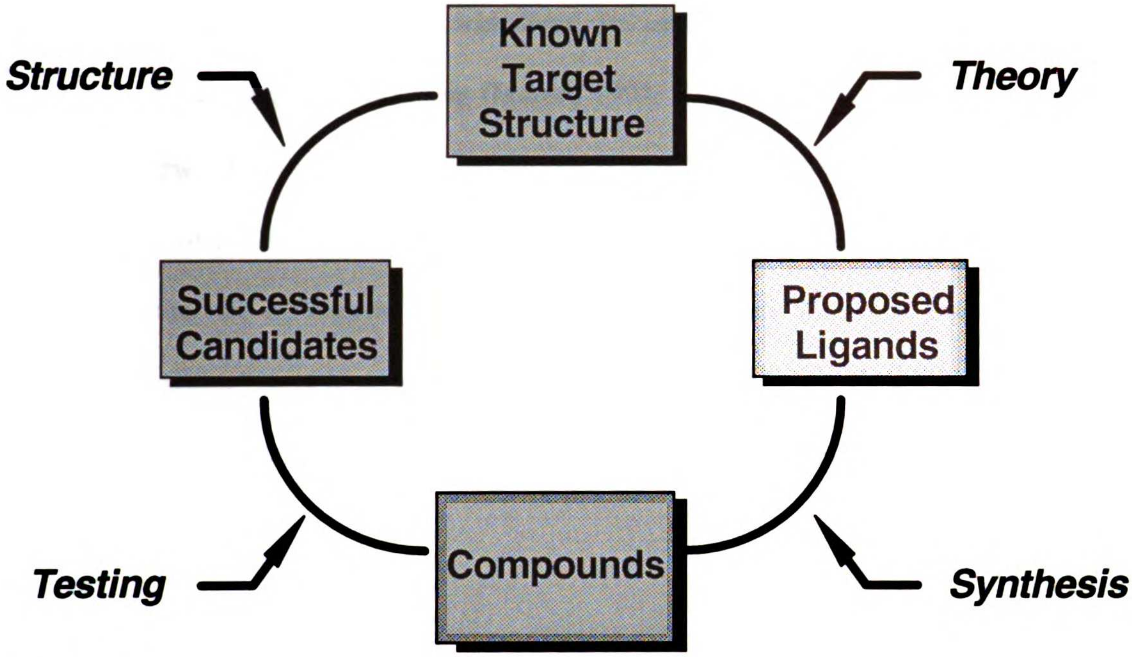


Figure 1. Structure-based drug design paradigm.
The figure emphasizes the cyclic and multidisciplinary aspects of this type of project.

based drug design cycle presented in Figure 1. The application of theoretical principles results in the proposal of putative ligands that are subsequently synthesized and tested. Biological data and receptor-ligand complex determination help to refine working hypotheses about complementarity (note that we use the term receptor in the non-classical sense to encompass any biological macromolecule that will bind ligands). The repeated application of the cycle constitutes incremental optimization of an initial bioactive compound, or lead. Tools which identify lead compounds themselves are of particular interest for acquiring chemically diverse starting points for optimization. Such diversity at the outset is valuable in maximizing the array of possibilities further downstream when pharmacokinetic and toxicological complications inevitably arise. There are a variety of computational techniques which may be useful in lead discovery in the context of detailed receptor information (Kuntz, 1992; Greer *et al.*, 1994; Kuntz *et al.*, 1994; Guida, 1994, Lybrand, 1995).

Two of the purest forms, stimulated by rapid advances in molecular structure determination, are database searching (Martin, 1992; Good & Mason, 1995) and structure generation (Lewis & Leach, 1994). The former selects ligands complementary to a receptor from a library of pre-existing compounds, while the latter attempts to create ligands tailor-made to fit the site of interest ("*de novo* design"). We here focus on the former - the molecular "docking" problem, which we define as the prediction of the observed (native) orientation of two interacting components given detailed three-dimensional information of each independently.

Molecular docking

Molecular docking attempts to arrange molecules in favorable configurations by matching complementary features (for a review of approaches, see Blaney and Dixon, 1993). This is a difficult task because there are many ways in which complex molecules can be associated. The problem is further complicated by an exponential dependence on molecule size, so that the number of possible configurations explodes when docking involves biological macromolecules such as proteins or nucleic acid polymers. An exhaustive computational analysis of configuration space is not tractable (Kuhl *et al.*, 1984; Connolly, 1986; Wang, 1991; Kuntz *et al.*, 1994), especially for database searching. Current docking methodologies thus invoke either geometric- or energy-based schemes to guide configurational sampling (Kuntz *et al.*, 1994), the former relying upon the matching of topographical features and the latter upon optimization along a potential energy surface of some kind. As alluded to earlier, however, configurational sampling is only half of the problem. The ranking of each configuration by some metric of complementarity constitutes the other major hurdle.

Complementarity

Complementarity can be assessed in many ways (see, for example, Shoichet & Kuntz, 1991). A configuration may be evaluated by its agreement with an input query or on its own merits, such as by a score independent of the method in which the docked complex was generated. One of the earliest conceptualizations of complementarity was the pairing of knobs and holes in packed α-helices (Crick, 1953). Somewhat more recently, these ideas were formulated into an algorithm for molecular docking by Connolly (1986), with variable success. Lin *et al.* (1994) have extended this formulation to the use of sparse critical points in a highly efficient solution to the docking problem (Fischer *et al.*, 1995). The use of surface complementarity has long been a fashionable scheme for guiding docking analyses (Wodak & Janin, 1978; Greer & Bush, 1978; Connolly, 1986; Jiang & Kim, 1991; Wang, 1991; Katchalski-Katzir *et al.*, 1992; Bacon & Moult, 1992; Helmer-Citterich & Tramontano, 1994; Norel *et al.*, 1994). Other mechanisms include compatibility assessments of individual atom contacts (Kuntz *et al.*, 1982; Lawrence & Davis, 1992; Shoichet & Kuntz, 1993) and extend to methods targeting specific interactions such as scoring by simplified electrostatic representations (Walls & Sternberg, 1992), satisfaction of hydrogen bonding constraints (Smellie *et al.*, 1991; Kasinos *et al.*, 1992; Yamada & Itai, 1993), or hydrophobic complementarity (Meng *et al.*, 1994; Vakser & Aflalo, 1994). Molecular mechanics force-fields remain extremely popular for the evaluation of docked complexes (Goodsell & Olson, 1990; Meng *et al.*, 1992; Hart & Read, 1992; Lawrence & Davis, 1992; Yamada & Itai, 1993; Miller *et al.*, 1994), while empirical schemes have met with renewed attention in recent years (Bohacek & McMartin, 1992; Bohacek & McMartin, 1994; Böhm, 1994a,b; Åqvist *et al.*, 1994; Warshel *et al.*, 1994).

Computational issues

The computational demands imposed by the very goals of molecular docking severely constrain the level of detail permitted in various aspects of the study. Computational screening aims to scan a 3-D database containing thousands or even millions of compounds in the time frame of days to weeks on a workstation. This leaves only on the order of a few seconds or less for an analysis of each putative ligand. Necessarily, these conditions impose serious limitations as to the thoroughness of each calculation. The typical tradeoff that results is one of speed *versus* storage: with unlimited physical memory, we may afford to sacrifice storage space for speed; conversely, with more realistic physical limitations, we must sacrifice speed for the sake of efficiency in storage. Approximations are therefore unavoidable. The most common simplifications include assumption of inflexible ligands and receptors, neglect of solvation effects, and use of crude scoring systems. It remains a challenge to formulate an interaction evaluation scheme which is both efficient and accurate. The computational constraints defined by molecular docking objectives establish a framework for deriving effective strategies.

Docking strategies

Molecular recognition events are dynamic processes. Any attempt to simulate such a process must come to terms with the kinetics and equilibria of molecules in solution. Each method must at the outset state which effects will be considered and which approximations will be made. Thus, there are a number of philosophies about how molecular docking might be carried out and which assumptions are in order. As will be described in more detail, these strategies span a spectrum of models bounded by the Lock-and-Key and Induced Fit theories for ligand binding. We explore two such strategies and examine scoring enhancements which further validate each as an appropriate model for molecular docking studies. The underlying motif of the work described here addresses our ability to rank docked complexes. Rapid methods now exist for carrying out efficient configurational sampling (Shoichet *et al.*, 1992; Lawrence & Davis, 1992; Norel *et al.*, 1994; Kuntz *et al.*, 1994), so we adjust our focus to the evaluation phase of a docking analysis. The attention that this aspect of the docking problem merits can not be underestimated in light of our successes with the combinatorial challenge. How we evaluate docked complexes has immediate repercussions on our estimation of what defines an optimal configuration, and hence, how well we mimic the physical process of molecular recognition. Proper ranking is essential for reaping the benefits of a molecular docking analysis, not only for the potential in drug discovery, but also for gaining thermodynamic insights into binding events. Thermodynamic estimates, in particular, will require more than just correct relative rankings among distinct binding modes; here, accurate gauges of absolute affinity may be necessary. The latter is, in fact, the prime directive of the field of structure-based drug design: can we predict with any certainty how tightly one molecule will bind to another?

LOCK AND KEY DOCKING

Rigid approaches

That molecular recognition events can be highly specific interactions is not new to medicinal chemists. Analogies to a "lock-and-key" concept to describe these processes were first put forth a century ago by Fischer (1894) and by Ehrlich (1909) (see also Lichtentaler, 1994). This model entails a precise matching of immutable components; the implications for molecular docking are that we may approximate the receptor and the ligand as rigid molecules. This greatly simplifies the docking problem by reducing the number of degrees of freedom from several thousand to only six. Along the spectrum of strategies for the simulation of dynamic processes, the lock-and-key concept lies at one extreme. What value does such a simplified model offer for molecular docking studies?

We envision the lock-and-key docking model as follows. A discrete conformation of the receptor has been observed experimentally. Discrete, reasonably low-energy conformations of potential ligands exist in the molecular database to be searched. Can one find *exact* fits between these pre-existing conformations? It is possible that the individual conformations will be sufficiently populated in solution that a binding event can occur with a resulting stabilization of the complex. It is upon these assumptions that molecular docking under the lock-and-key philosophy relies. These hypotheses warrant caution, but experience has taught us that this model can be quite informative.

The assumption of a rigid receptor is often less severe than one might think. Proteins are generally observed to behave as rigid entities, as studies of complexed and uncomplexed crystal structures indicate (Janin & Chothia, 1990). Although large conformational changes upon complexation have been illustrated (Miller *et al.*, 1989; Schulz *et al.*, 1990; Van Duyne *et al.*, 1991), backbone movement is typically restricted to less than 1 Å (Cherfils & Janin, 1993). The well-established prevalence of sidechain motion will present a challenge for all docking methodologies. That macromolecular plasticity (Koshland, 1971) defeats lock-and-key docking has yet to be shown. The practical application of DOCK, one of the first automated molecular docking programs (Kuntz *et al.*, 1982; DesJarlais *et al.*, 1988; Shoichet *et al.*, 1992; Meng *et al.*, 1992), has resulted in numerous successes under the simple rigid-body docking model (DesJarlais *et al.*, 1990; Kerwin *et al.*, 1991; Shoichet *et al.*, 1993; Ring *et al.*, 1993; Bodian *et al.*, 1993; Rutenber *et al.*, 1993). Albeit a tremendous simplification, it is apparent that the assumption of an inflexible receptor has significant value associated with it.

One of the most formidable tasks in drug design is that of obtaining specificity in interactions over related receptors. Species specificity is a common instance of this general theme. Differentiating functionally identical and structurally very similar targets requires extremely sensitive technology. We now push the envelope of the lock-and-key concept by attempting to distinguish, at an atomic level, two enzymes whose discrimination continues to frustrate modern medicinal chemistry: an example of exquisite similarity among phylogenetically distinct species is that of dihydrofolate reductase (DHFR) from humans and that from the pathogenic fungus *Pneumocystis carinii* (Edman *et al.*, 1988).

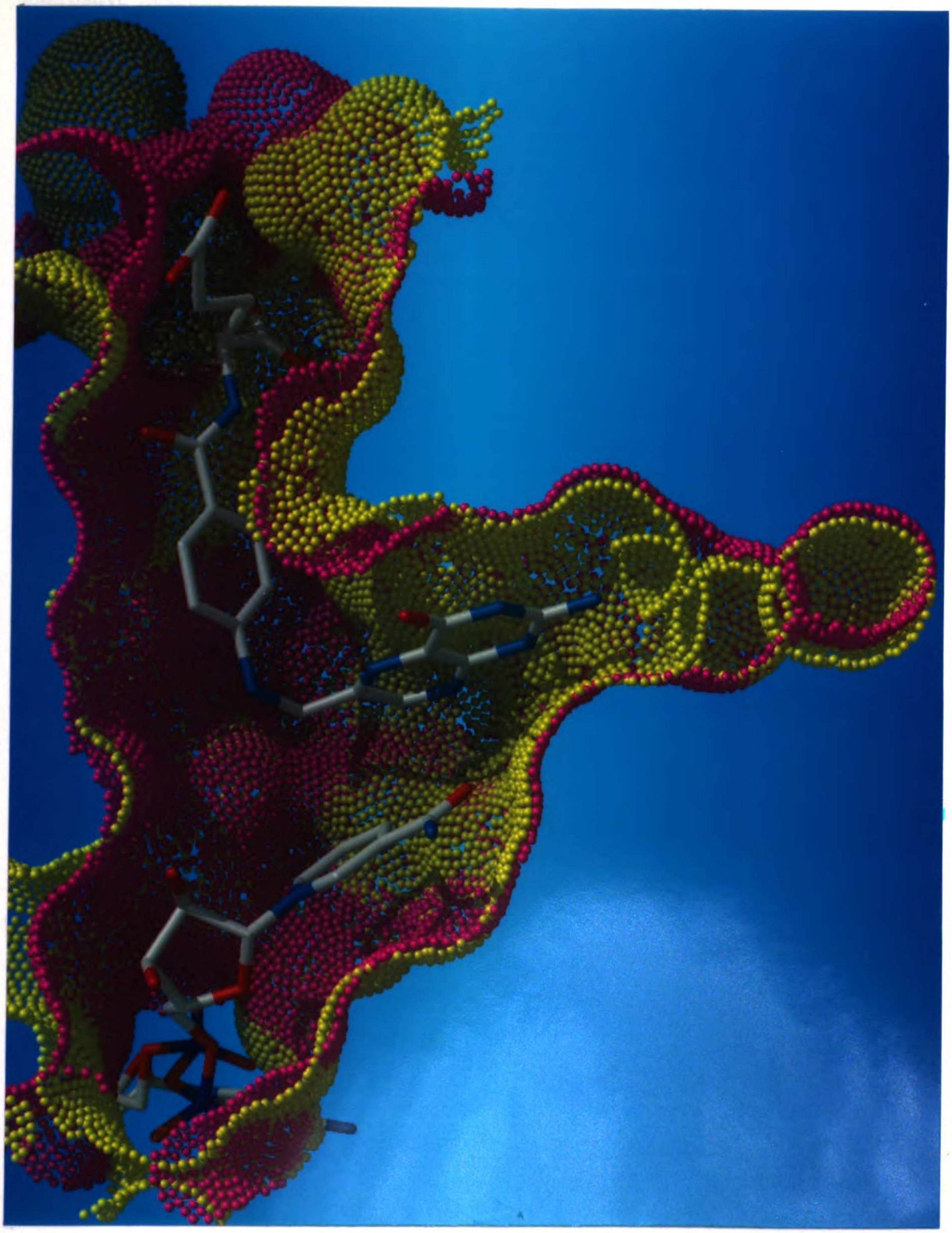
Pneumocystis, *a fungal opportunist*

Pneumocystis carinii harmlessly infects nearly all humans, but upon reactivation of latent infection by immunodeficiency can induce a disease state characterized by a crippling pneumonia (Murray & Mills, 1990; Bartlett & Smith, 1991). Not surprisingly, this opportunist is the principal agent of morbidity and mortality in HIV-infected persons (Murray & Mills, 1990; Mills & Masur, 1990). Without chemoprophylaxis, 60 to 85% of AIDS patients eventually will be afflicted by *P. carinii* pneumonia, and 25% will die from it (Walzer *et al.*, 1974; Mills, 1986; Kovacs & Masur, 1988; Justice *et al.*, 1989; Bartlett & Smith, 1991). Agents in a variety of mechanistically distinct classes are being explored, but the most successful of these approaches thus far have been the antifolates and DNA-replication antagonists. Co-trimoxazole and pentamidine isethionate are the most widely prescribed preparations for therapy and prophylaxis of *Pneumocystis carinii* pneumonia (Kovacs & Masur, 1988; Murray & Mills, 1990; U.S. Public Health Service, 1993; Gallant *et al.*, 1994).

Unfortunately, these treatments are plagued with adverse reactions (Walzer *et al.*, 1974; Jaffe *et al.*, 1983; Gordin *et al.*, 1984; Kovacs *et al.*, 1984; Wharton *et al.*, 1986; Allegra *et al.*, 1987). The frequency of such reactions is not surprising in the case of co-trimoxazole: the dihydropteroate synthetase (DHPS) inhibitor component, sulfamethoxazole, is quite toxic (Masur, 1992) and the DHFR inhibitor component, trimethoprim, is weak and non-selective (Edman *et al.*, 1989; Margosiak *et al.*, 1993). In fact, it is almost general that clinically relevant DHFR inhibitors are selective for *human* DHFR (Margosiak *et al.*, 1993). The lack of selectivity and resulting side effects of antifolates is likely a direct consequence of the similarity between host and pathogen DHFR.

Based on published sequence alignments (Blakley, 1984; Edman *et al.*, 1989), *P. carinii* DHFR displays highest similarity with that of vertebrates. Sequence identities of 35-40% and homologies near 70% are observed. High-resolution crystal structures of human (Davies *et al.*, 1990) and *P. carinii* (Oefner *et al.*, 1991) DHFR confirm only minor differences. Within the folate binding pocket, there are only six non-identical residues. The active site aspartate common to bacterial and protozoan DHFR is replaced with glutamate in *P. carinii*, as in all vertebrate DHFRs. Further drug design complications arise from the fact that the *Pneumocystis* active site is nearly universally smaller than the human active site, making exploitation of unique pockets impossible. The extensive similarity of the active site molecular surfaces (Connolly, 1983a,b) is depicted in Figure 2.

Figure 2. Active site superimposition of DHFR.
(following page) The molecular surface of *Pneumocystis carinii* DHFR is shown in magenta; the molecular surface of human DHFR is shown in yellow. Structures were aligned by superimposition of 55 active site α-carbons (0.55 Å r.m.s.). A cross-section of the active site is depicted, showing substrate (folate) and cofactor (NADPH) colored by atomic identity.



The lack of differentiation by mainstream agents highlights the urgency for alternative molecular frameworks. Thus, molecular docking is ideally suited to this task. The lock-and-key model, applied in stringent fashion, has been used to discern minute differences between the host and pathogen receptors. We here summarize the application of the DOCK screening process toward the therapeutically important problem of identifying novel, selective anti-*Pneumocystis* agents. This work will be published in greater detail elsewhere [Gschwend *et al.*, 1995 (Chapter 3)].

DOCK

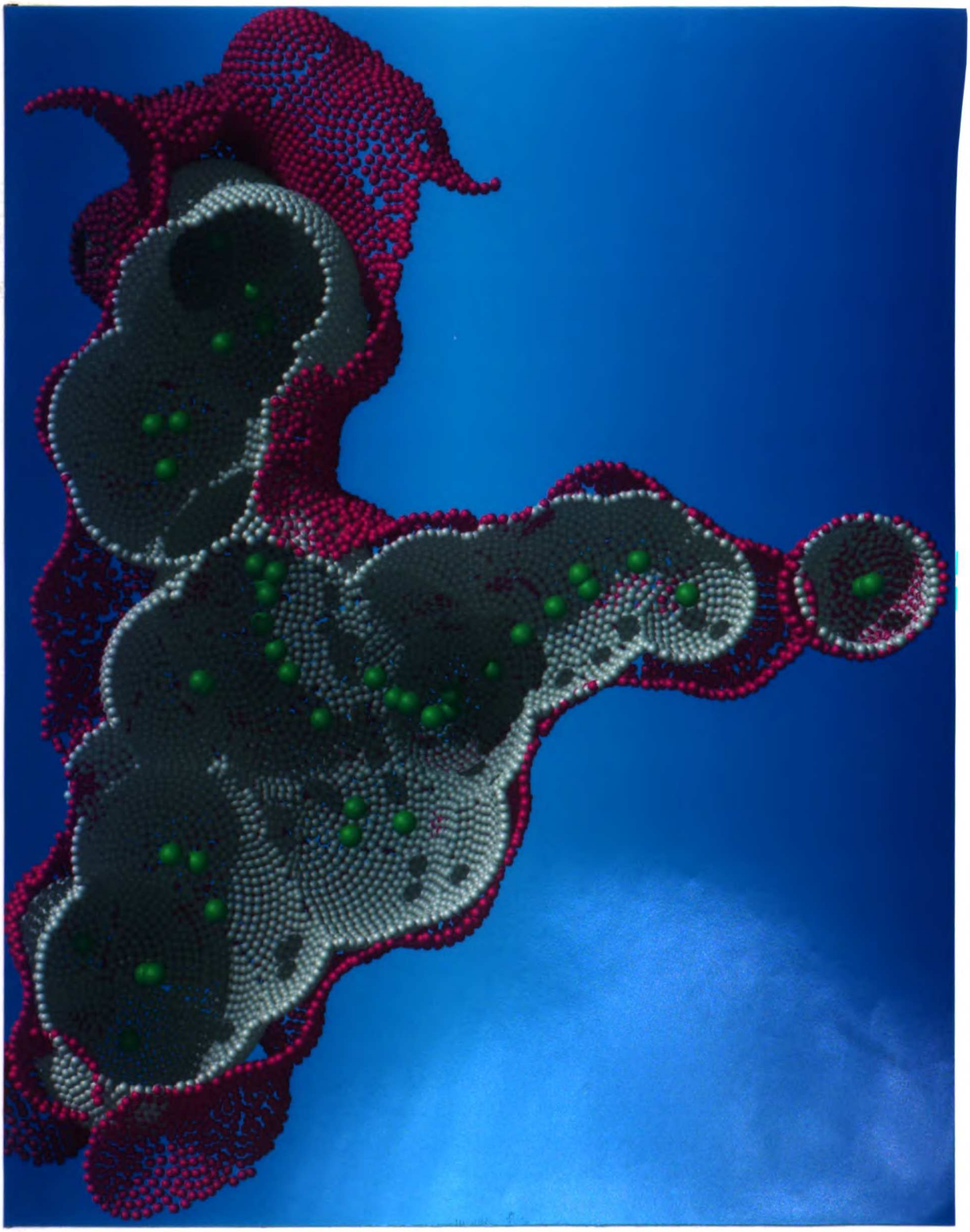
The DOCK suite of programs, like other molecular docking packages, is designed to identify putative ligands complementary to a receptor of known 3-D structure. The details of the method have been described previously (Kuntz *et al.*, 1982; Shoichet *et al.*, 1992; Meng *et al.*, 1992) - only an overview is given here. By filling the receptor site with overlapping spheres of varying sizes, a negative image capturing the bumps and grooves of the region of interest is generated. A 3-D database is searched (DesJarlais *et al.*, 1988) for molecules whose interatomic distances match the inter-sphere-center distances. Each compound is evaluated in thousands of orientations in the active site by an approximate molecular mechanics interaction energy (Meng *et al.*, 1992). The best-scoring compounds presumably exploit multiple geometric and/or chemical properties of the receptor site and are thus of considerable interest as inhibitor candidates. The DOCK procedures have been tested through studies of crystallographic complexes (Kuntz *et al.*, 1982; Shoichet & Kuntz, 1991; Meng *et al.*, 1992; Shoichet *et al.*, 1992; Shoichet & Kuntz, 1993; Meng *et al.*, 1993). The experimental geometries are associated with the best-scoring orientations, generally within 1 Å r.m.s. deviation. More importantly, compounds that have radically different structures from known inhibitors are often found.

Method

The human [Protein Data Bank (Bernstein *et al.*, 1977) entry 1dhf (Davies *et al.*, 1990)] and *P. carinii* (Oefner *et al.*, 1991) DHFR structures were aligned by superimposing active site α-carbons. The Fine Chemicals Directory (FCD3D v.89.2, MDL Information Systems, Inc., San Leandro, CA), a database of commercially available compounds (now called the Available Chemicals Directory), was screened with DOCK version 3.0 (Meng *et al.*, 1992) against *Pneumocystis* DHFR. Figure 3 illustrates the complementarity of the sphere description used to characterize the target site and perform the docking. An average of 19,000 orientations was examined for each of 53,328 compounds. The over one billion total configurations investigated at a rate of 800 per second (Silicon Graphics PI 4D/35) attest to the speed of the DOCK program. Roughly the top 5% (2,700) top force-field scoring ligands were retained for further analysis. Each of these was then subject to a quasi-Newton rigid-body minimization (Meng *et al.*, 1993) to optimize the intermolecular interactions of its best scoring orientation. This refinement was carried out independently in the context of *both P. carinii* and human DHFR active sites. Thus, the starting configuration in *Pneumocystis* DHFR was determined by DOCK, while the starting configuration in human DHFR was determined by the structural alignment. The differential, optimized force-field scores were used as an indication of species selectivity.

Figure 3. Docking spheres.

(following page) The molecular surface of *Pneumocystis carinii* DHFR is shown in magenta, sphere centers used in docking are shown as small green balls, and the collective surface of the sphere description is illustrated in white. Note the shape complementarity between the surface of the sphere description and the surface of the receptor.



A manual screening process of the computationally selected 2,700 compounds ensued. Filters were introduced to target deficiencies in the DOCK scoring scheme, such as solvation effects and neglect of conformational entropy. Hits were visually evaluated for fit using the MidasPlus graphics package (Ferrin *et al.*, 1988). As only a finite number of compounds can be assayed, chemical diversity screens with the aid of the MACCS-II 3D software (MDL Information Systems, Inc., San Leandro, CA) were employed. Substructure clustering enabled the selection of only the structurally most dissimilar compounds for biological characterization. Finally, practical concerns including solubility, reactivity, toxicity and commercial availability were addressed.

Results

Forty structurally distinct compounds were assayed for activity against *Pneumocystis carinii* dihydrofolate reductase. Of these, nearly half showed significant inhibition, greater than 20% at an inhibitor concentration of 100 μM. Roughly one quarter demonstrated IC_{50} values at or better than 100 μM. Seven of the more potent compounds against *P. carinii* DHFR were assayed against human DHFR for specificity. All were selective for the pathogenic isozyme, as illustrated in Figure 4. The most potent compound, which inhibits *P. carinii* DHFR with an IC_{50} of 7 μM, shows 25-fold selectivity. An analysis of the DOCK-predicted mode of binding for this inhibitor attributes this differentiation to contact with four of the six non-identical residues in the active site.

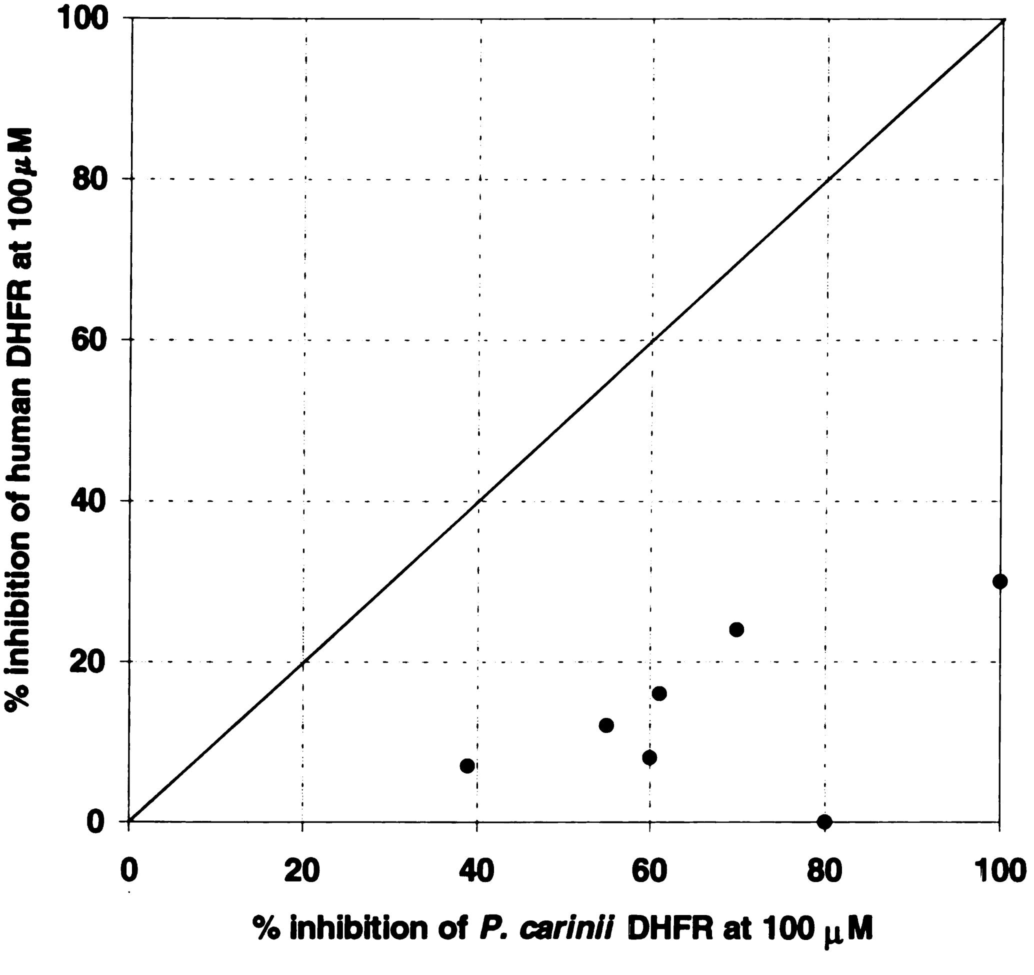


Figure 4. Selectivity plot for novel DHFR inhibitors.

Percent inhibition of DHFR at an inhibitor concentration of 100 μM is shown for each of seven assayed ligands. The diagonal line represents the absence of selectivity.

It is important to put this seemingly minor species-specificity into perspective. Trimethoprim, the DHFR inhibitor component of the most widely prescribed preparation for *P. carinii* infection, is a weak inhibitor and exhibits essentially no preference for the pathogenic enzyme, while all other clinically relevant therapeutics show modest to great selectivity for the *human* enzyme (Margosiak *et al.*, 1993). An analysis of progress in the antifolate literature indicates that even a 10-fold preference for *P. carinii* DHFR is a relatively rare occurrence. The 7-μM inhibitor discovered with the molecular docking approach described above is a lead; it is clearly not a drug. However, in the context of the extreme difficulty with which selective anti-*Pneumocystis* agents are identified, this lead may prove an advantageous starting point toward therapeutic usefulness. Finally, it is not inconsequential that the molecular framework of this agent bears no resemblance to any previously established antifolate.

Macroscopic correlations

Computational strategies for structure-based drug discovery offer a valuable alternative to the costly and time-consuming process of random screening (Kuntz, 1992). Coupled with a database of commercially available compounds, such as the ACD, programs like DOCK can provide extremely rapid access to novel leads. However, because of the many approximations underlying the search and scoring engines (*e.g.* neglect of solvation terms, rigidity of ligand and receptor, discretized scoring), DOCK can not be expected to yield predictions of a quantitative nature. Rather, we prefer to value DOCK as a "macroscopic correlator" of binding affinity and interaction score. Even in the daunting task of species-specificity, macroscopic correlations when applied in sequence can, as demonstrated here, confer a powerful tool.

In the method previewed here, the rigid-body minimization acts as the selectivity filter. An optimization of this type as a post-docking utility has been shown to improve agreement with experimentally determined binding modes (Meng *et al.*, 1993). Of the 50,000 compounds in the database, an enrichment for agents which inhibit *P. carinii* DHFR was achieved with DOCK (the first macroscopic correlation). Subsequently, the differential optimization in the context of both isozymes offered resolution along an additional dimension for these remaining compounds. By choosing structures which score highly in *Pneumocystis* DHFR and poorly in human DHFR (the second macroscopic correlation), an enrichment for agents selective against host DHFR has now been accomplished.

To summarize, our simple implementation of the lock-and-key model has proven extremely effective even in the compounded task of searching for selective agents. The optimization used here does not counteract the lock-and-key model. It offers a jiggling of the rigid components with respect to one another, allowing a more exact match to be located. The refinement in essence strengthens our adherence to the lock-and-key model by providing a more stringent scoring scheme for the evaluation of docked complexes. As dictated by the exquisite structural similarities within the species specificity problem addressed here, such a stringency is paramount in discerning atomic level differences.

New technologies

An even stricter adherence to the lock-and-key model can readily be envisioned. The optimization of intermolecular interactions described above took the form of a post-DOCK refinement: thousands of orientations of a ligand with respect to its receptor were generated, the best-scoring configuration was identified, and finally the fit of this one optimal configuration was refined. It would be more faithful to the lock-and-key model and less biased in approach if *every* orientation of the ligand was optimized, rather than only one configuration deemed best by an unrefined score. Rankings among configurations as gauged by pre- and post-refinement scores differ, sometimes dramatically (data not shown). It therefore makes sense to harness the power of minimization as a post-docking scoring tool directly in the evaluation phase of docking. This computationally demanding advance has been accomplished with an unexpected performance *increase* and incorporated into DOCK version 3.5. Despite advances in computational resources which make features such as on-the-fly optimization yet more palatable, the time spent in refinement is still large when compared with the time spent sampling configuration space. If one could judiciously reduce the number of orientations actually optimized, however, the refinement bottleneck might be dissipated.

Given the large number of spatially distributed descriptors and atoms involved in molecular docking, it is not surprising that there are many ways of pairing them which give rise to similar geometric orientations. This is obviously the result of over-sampling in certain regions, but without which some binding modes would be under-sampled or even overlooked. In the absence of refinement, this over-sampling provides a sort of rigid-body minimization itself. A better way to optimize local interactions is to find only one orientation per mode of binding and energy-minimize that orientation, while never again paying close attention to similar orientations. By removing so-called "degenerate" configurations, many non-informative minimizations are avoided. Progress towards this goal with a technique we refer to as "degeneracy checking, as well as the specifics of on-the-fly force-field score optimization, will be published in greater detail elsewhere [Gschwend & Kuntz, 1995 (Chapter 4)].

Summary

Our experiences with lock-and-key docking have been encouraging. We here have previewed the discovery of a novel, selective enzyme inhibitor under this model, and append our application to a growing list of DOCK successes. The use of DOCK to pursue selective leads of therapeutic interest had not yet been reported. Thus, this marks the first attempt to push the lock-and-key model to the limit of differentiation at an atomic level. Detection of minute differences between structurally similar targets requires a refined scoring procedure. The force-field score, comprised of Coulombic electrostatic and Lennard Jones van der Waals terms, is by nature very sensitive to exact atomic positions. This conforms well with our interpretation of the lock-and-key model for molecular docking, outlined previously, which demands stringency in the evaluation phase to locate exact matches. Heightened resolution is achieved by rigid-body optimization, which allows a configuration to exploit optimal local interactions and thereby accentuate subtle differences between targets. As gauged by the broad success in locating selective agents in the face of few distinguishing features, it appears this technology is quite powerful. The incorporation of on-the-fly refinement into the docking process can only enhance our ability to detect optimal fits. The methodology enhancements and species-specificity results reinforce our view of the lock-and-key model as a valid strategy for molecular docking.

INDUCED FIT MODEL

Flexible approaches

The lock-and-key model for protein-ligand binding can not explain all aspects of enzyme specificity (Koshland, 1994). For example, the function of ligands which modulate enzyme activity but do not participate directly in catalysis could not be defined (Koshland, 1971). Observations such as these led to the proposal of a modified theory, the induced-fit theory (Koshland, 1958), which maintains that ligands induce changes in protein structure *before* a suitable fit can occur. The ideas of induced-fit effects and macromolecular plasticity find mounting support as structural and mechanistic details of molecular recognition events are elucidated (see Jorgensen, 1991, and references therein).

The hand-shaking that occurs between receptor and ligand upon binding is difficult to simulate computationally. The inclusion of conformational degrees of freedom in addition to the six orientational degrees of freedom exhibited by rigid objects results in a configurational explosion. Molecular dynamics simulations offer assistance in the local exploration of conformational flexibility, but become intractable in a molecular docking context when a single starting position can not be assumed. There have been many approaches to tackling this massively complicated docking problem. One method which juxtaposes discrete ligand *and* receptor conformations has been reported (Leach, 1994). Numerous approaches, while ignoring receptor mobility, do treat ligand flexibility - these include energy-based methods (Goodsell & Olson, 1990; Hart & Read, 1992), genetic algorithms (Judson *et al.*, 1994; Jones *et al.*, 1995; Oshiro *et al.*, 1995), distance geometry (Ghose & Crippen, 1985), descriptor-based methods (Yamada & Itai, 1993), fragment-based techniques (DesJarlais *et al.*, 1986; Leach & Kuntz, 1992), and the independent docking of discrete, pre-generated ligand conformations (Miller *et al.*, 1994). With the exception of the last method [which scales as the number of conformations per ligand, typically of order 10 (Miller *et al.*, 1994)] and those methods which are not trivially automated (*e.g.* fragment-based techniques), these approaches for incorporating flexibility require (justifiably) roughly 50- to 5,000-fold longer execution times than an efficient rigid-body docking method. For single-molecule docking studies, where we can afford a more detailed analysis, this sacrifice is entirely acceptable. For purposes of drug discovery, however, such penalties become prohibitive.

Goals of molecular docking

Molecular docking for drug discovery aims to scan a database of compounds for ligands which exploit some aspect of complementarity to the receptor of interest. An attempt to simulate molecular recognition is made for ligands which the receptor has never "seen" before. This point deserves some clarification. The receptor structure designated for docking has been observed experimentally or modeled by homology either in an unbound or a bound state. Thus, the receptor exists in a pre-defined conformation, one possibly molded to a particular ligand. It would behoove docking studies if the receptor were allowed to respond to the presence of each putative ligand. A plasticity on the part of the receptor would permit formation of improved interactions, thereby offering a fairer gauge of the compound's potential as a true ligand. Short of introducing explicit flexibility and suffering a severe performance penalty, we wonder whether it is possible to manifest some aspects of conformational flexibility in the docking process.

Soft docking

An *implicit* breathing on the part of the receptor (and the ligand) can be introduced via a tolerant evaluation function. For example, a softer scoring potential permits slight atomic interpenetrations without penalty, in effect implying a resolving conformational change. The idea of so-called "soft docking" is not new. This concept hails from protein-protein docking investigations in which structures of unbound components are docked to reproduce the observed complexed structure (Wodak & Janin, 1978; Shoichet & Kuntz, 1991; Jiang & Kim, 1991; Walls & Sternberg, 1992). The success of such methods hinges upon a local insensitivity that fosters conformational shifts upon complexation. Tolerance is brought about either by simplified geometric representations (Wodak & Janin, 1978; Jiang &

Kim, 1991) or by imprecise scoring schemes (Shoichet & Kuntz, 1991; Walls & Sternberg, 1992). We can apply these concepts, in a somewhat more constrained manner, towards addressing dynamic aspects of small-molecule recognition.

Scoring philosophy

The scoring function we seek for drug discovery through molecular docking should exhibit four distinct qualities: 1) it should be robust over a structurally diverse set of receptor-ligand complexes; 2) it should incorporate molecular plasticity; 3) it should be easy to implement; and 4) it must be rapid to evaluate. The philosophy that this work subscribes to presupposes that, in keeping with induced-fit notions, the receptor will respond to the presence of a ligand. Thus, unfavorable interactions will be avoided, while favorable interactions will be optimized, both by concerted motion on the part of the components. These assumptions naturally compromise our ability to detect subtleties, but, as will be seen shortly, afford generality across structurally diverse receptor-ligand complexes.

Force-fields

In deriving a robust scoring function, we opt to deviate from the sensitive molecular mechanics functions of many molecular docking programs and revert to simpler, digital interaction evaluations. To a first approximation, we consider interactions as being either present or not. This implementation ensures ease of use and an insensitivity to exact local geometries, at the cost of potential accuracy. Furthermore, we avoid problematic issues confronted in using force-fields, such as partial charge computation, choice of dielectric behavior, careful assignment of atom types, and hydrogen placement. To illustrate the latter sensitivity, consider hydroxyl hydrogens (for example, on serine, threonine, and tyrosine residues). Preference for hydrogen bond geometry about these functionalities is weak (Baker & Hubbard, 1984; Thanki *et al.*, 1988; Tintelnot & Andrews, 1989), while molecular mechanics-based schemes require selection of an exact hydrogen position. Interaction strength is thus spuriously sensitive to the (typically arbitrary) placement of this hydrogen. Finally, we note that molecular mechanics is not directly parameterized to reproduce binding affinities. Force-field scores report an *enthalpy* of interaction; the quantity of interest is the *free energy* of interaction. In our experience, force-field scores are effective at identifying the optimal binding mode of a single ligand, but perform poorly at predicting even relative binding energies across a panel of ligands. Entropic contributions are likely to be fairly similar for different binding modes of one ligand, but clearly can vary substantially from one ligand to the next. It stands to reason that without the entropic half of the equation we have little hope of predicting binding affinities for structurally diverse ligands. [Successful, system-dependent examples of enthalpic correlations with binding affinity have been reported *(e.g.* Holloway *et al.*, 1995). Here, we emphasize the need for robustness across structurally unrelated ligands binding to varied receptors.]

Empirical schemes

Given the scope of the molecular docking problem, it is not unreasonable to design a scoring scheme especially suited to the task at hand - that is, evaluating strengths of interactions for a diversity of receptor-ligand complexes. There currently exist nearly as many ways of evaluating docked complexes as there are docking methods. As early researchers in the protein docking field have noted, even the simplest scoring schemes perform virtually as well as more advanced molecular mechanics treatments (Shoichet & Kuntz, 1991; Cherfils & Janin, 1993). There is thus the potential to derive an evaluation method which is not borrowed from the objectives of another branch of computational chemistry, but rather, which is parameterized to reproduce precisely the type of values we are attempting to predict.

The field of empirical scoring systems for the estimation of small-molecule binding affinities has rapidly become an active area of investigation (Bohacek & McMartin, 1992; Krystek *et al.*, 1993; Bohacek & McMartin, 1994; Böhm, 1994a,b; G.R. Marshall, personal communication; J.S. Dixon, personal communication; M.A. Murcko, personal communication; A.N. Jain, personal communication). Paralleling a QSAR study, the general procedure consists of amassing a series of receptor-ligand complexes [typically from the Protein Data Bank (Bernstein *et al.*, 1977)] with experimentally determined affinities, devising various calculable terms which describe physical interactions of interest, and attempting to obtain affinity correlations while varying coefficients for each term. Approaches vary widely in the data set composition, the terms employed in correlations, and the method in which the terms are computed. As with any correlation analysis, great care must be taken to acquire a large and diverse data set, to gauge the statistical validity of the output, to verify predictivity of proposed models, and to avoid overinterpretation of physical significance.

[To avoid duplication of some material, the reader is referred to Chapter 5 for a detailed discussion of the empirical scheme we have developed.]

Summary

Empirical schemes show great potential for benefiting database screening by molecular docking. Their simplicity by design makes them rapid to calculate and easy to implement, avoiding challenges associated with force-field implementations such as atomic parameterization and partial charge computation. The nature of the tolerant functions used in our empirical schemes makes them insensitive to exact local geometries. An implicit breathing on the part of the receptor and ligand can therefore be sanctioned. Along the spectrum of strategies for the simulation of dynamic processes, induced-fit models which address explicit flexibility on the part of the ligand and receptor are intractable for a database screening application. Our compromise toward implicit plasticity, in contrast, suffers no performance penalty. It remains to be proven whether we gain anything in a practical setting, but empirical schemes are designed to be and have shown themselves to be robust over diverse data sets. A molecular docking implementation utilizing a carefully formulated empirical scheme should be able to harness this robustness to our advantage.

OUTLOOK

Reasonable methods now exist for combining the pieces of the 3-D molecular jigsaw puzzle; we here have focused on aspects of judging whether the puzzle looks right. Given the varied ways in which one can make this judgment, there are seemingly infinite stances to be assumed for molecular docking strategies. Our flexibility becomes more limited, however, in the face of database screening applications where ligand analyses must be completed in a few seconds. We have explored two approaches here, one very exacting and one very tolerant.

The lock-and-key model has been immensely successful for molecular docking, generating leads in an array of biological systems (Kuntz, 1992). In our experience with enzymes, typical hit rates at the micromolar level range between 2 and 20%. Even in the compounded problem of addressing species specificity, we have demonstrated an ability to locate novel, selective leads. These accounts are encouraging, and validate the lock-and-key model as a strategy for drug discovery through molecular docking. We have extended our interpretation of the lock-and-key model by introducing rigid-body optimization into the docking process. This technology strengthens our adherence to the goal of identifying exact matches, which epitomizes lock-and-key docking.

Our initial steps toward capturing induced-fit effects into a docking strategy appear promising. Empirical schemes incorporating implicit plasticity herald a generality not seen with sensitive molecular mechanics-based approaches. The robustness over diverse structural arrangements embraces the presentation of thousands of molecular skeletons to a receptor which is not explicitly allowed to respond. This challenge is the essence of molecular docking for drug discovery: can we gauge the affinity of an arbitrary ligand for a given receptor? Ultimately, only methods which address the whole of the Gibbs free energy equation will prevail. Amidst the wake of vast quantities of detailed structural information now becoming available, it is imperative for comprehending molecular recognition and for pursuing structure-based drug design that we master the subtleties of complementarity.

ACKNOWLEDGMENTS

This work was funded by NIH grants GM-39552 (G. L. Kenyon, principal investigator) and GM-31497 (I.D.K.). D.A.G. acknowledges support from an ACS Division of Medicinal Chemistry Fellowship. Molecular graphics images were produced using the MidasPlus program from the Computer Graphics Laboratory, University of California, San Francisco (supported by NIH RR-1081). We thank C. Oefner for providing the coordinates for *P. carinii* DHFR, Tripos Associates for supplying the SYBYL modeling package, and MDL Information Systems for MACCS-3D and the FCD and ACD structural databases. This work has been accepted for publication in the *Journal of Molecular Recognition*, Copyright © 1995 by John Wiley & Sons Ltd.

REFERENCES

Allegra, C.J., Chabner, B.A., Tuazon, C.U., Ogata-Arakaki, D., Baird, B., Drake, J.C., Simmons, J.T., Lack, E.E., Shelhamer, J.H., Balis, F., Walker, R., Kovacs, J.A., Lane, H.C. and Masur, H. (1987). Trimetrexate for the treatment of *Pneumocystis carinii* pneumonia in patients with the acquired immunodeficiency syndrome. N. Engl. J. Med. 317: 978-985.

Åqvist, J., Medina, C. and Samuelsson, J.-E. (1994). A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **7**: 385-391.

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

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

Bartlett, M.S. and Smith, J.W. (1991). *Pneumocystis carinii*, an opportunist in immunocompromised patients. *Clin. Microbiol. Rev.* **4**: 137-149.

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

Blakley, R.L. (1984). "Dihydrofolate reductase" in Folates and Pteridines, R. Blakley and S. Benkovic ed., Wiley, New York, NY.

Blaney, J.M. and Dixon, J.S. (1993). A good ligand is hard to find: Automated docking methods. *Persp. Drug. Disc. Des.* **1**: 301-319.

Bodian, D.L., Yamasaki, R.B., Buswell, R.L., Stearns, J.F., White, J.M. and Kuntz, I.D.

(1993). Inhibition of the fusion-inducing conformational change of influenza hemagglutinin by benzoquinones and hydroquinones. *Biochemistry* **32**: 2967-2978.

Bohacek, R.S. and McMartin, C. (1992). Definition and display of steric, hydrophobic, and hydrogen-bonding properties of ligand binding sites in proteins using Lee and Richards accessible surface: Validation of a high-resolution graphical tool for drug design. *J. Med. Chem.* **35**: 1671-1684.

Bohacek, R.S. and McMartin, C. (1994). Multiple highly diverse structures complementary to enzyme binding sites: Results of extensive application of a *de novo* design method incorporating combinatorial growth. *J. Am. Chem. Soc.* **116**: 5560-5571.

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

Böhm, H.-J. (1994b). On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three-dimensional structure. *J. Comput.-Aided Mol. Design* **8**: 623-632.

Cherfils, J. and Janin, J. (1993). Protein docking algorithms: Simulating molecular recognition. *Curr. Opin. Struc. Biol.* **3**: 265-269.

Connolly, M.L. (1983a). Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**: 709-713.

Connolly, M.L. (1983b). Analytical molecular surface calculation. *J. Appl. Cryst.* **16**: 548-558.

Connolly, M.L. (1986). Shape complementarity at the hemoglobin α1β1 subunit interface. *Biopolymers* **25**: 1229-1247.

Crick, F.H.C. (1953). The packing of α-helices: Simple coiled coils. *Acta Cryst.* **6**: 689-697.

Davies, J.F., Delcamp, T.J., Prendergast, N.J., Ashford, V.A., Freisheim, J.H. and Kraut, J. (1990). Crystal structures of recombinant human dihydrofolate reductase complexed with folate and 5-deazafolate. *Biochemistry* **29**: 9467-9479.

DesJarlais, R., Sheridan, R.P., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R. (1986). Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.* **29**: 2149-2153.

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

DesJarlais, R.L., Seibel, G.L., Kuntz, I.D., Ortiz de Montellano, P.R., Furth, P.S., Alvarez, J.C. DeCamp, D.L. Babé, L.M. and Craik, C.S. (1990). Structure-based design of nonpeptide inhibitors specific for the human immunodeficiency virus 1 protease. *Proc. Natl. Acad. Sci. U.S.A.* **87**: 6644-6648.

Edman, J.C., Kovacs, J.A., Masur, H., Santi, D.V., Elwood, H.J. and Sogin, M.L. (1988). Ribosomal RNA sequence shows *Pneumocystis carinii* to be a member of the Fungi. *Nature* **334**: 519-522.

Edman, J.C., Edman, U., Cao, M., Lundgren, B., Kovacs, J.A. and Santi, D.V. (1989). Isolation and expression of the *Pneumocystis carinii* dihydrofolate reductase gene. *Proc. Natl. Acad. Sci. U.S.A.* **86**: 8625-8629.

Ehrlich, P. (1909). Über den jetzigen Stand der Chemotherapie. *Chem. Ber.* **42**: 17-47.

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

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

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Chem. Ber.* **27**: 2985-2993.

Gallant, J.E., Moore, R.D. and Chaisson, R.E. (1994). Prophylaxis for opportunistic infections in patients with HIV infection. *Ann. Intern. Med.* **120**: 932-944.

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

Good, A.C. and Mason, J.S. (1995). Three-dimensional structure database searches. *Rev. Comp. Chem.* in press.

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

Gordin, F.M., Simon, G.L., Wofsy, C.B. and Mills, J. (1984). Adverse reactions to trimethoprim-sulfamethoxazole in patients with the acquired immunodeficiency syndrome. *Ann. Intern. Med.* **100**: 495-499.

Greer, J. and Bush, B.L. (1978). Macromolecular shape and surface maps by solvent exclusion. *Proc. Natl. Acad. Sci. U.S.A.* **75**: 303-307.

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

Gschwend, D.A., Sirawarapom, W., Santi, D.V. and Kuntz, I.D. (1995). Specificity in structure-based drug design: Identification of a novel, selective inhibitor of *Pneumocystis carinii* dihydrofolate reductase. Submitted.

Gschwend, D.A. and Kuntz. I.D. (1995). Orientational sampling and rigid-body minimization in molecular docking, revisited: On-the-fly optimization and degeneracy removal. Submitted.

Guida, W.C. (1994). Software for structure-based drug design. *Curr. Opin. Struc. Biol.* **4**: 777-781.

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

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

Holloway, M.K., Wai, J.M., Halgren, T.A., Fitzgerald, P.M.D., Vacca, J.P., Dorsey, B.D., Levin, R.B., Thompson, W.J., Chen, L.J., deSolms, S.J., Gaffin, N., Ghosh, A.K., Giuliani, E.A., Graham, S.L., Guare, J.P., Hungate, R.W., Lyle, T.A., Sanders, W.M., Tucker, T.J., Wiggins, M., Wiscount, C.M., Woltersdorf, O.W., Young, S.D., Darke, P.L. and Zugay, J.A. (1995). *A priori* prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *J. Med. Chem.* **38**: 305-317.

Jaffe, H.S., Abrams, D.i., Ammann, A.J., Lewis, B.J. and Golden, J.A. (1983). Complications of co-trimoxazole in treatment of AIDS-associated *Pneumocystis carinii* pneumonia in homosexual men. *Lancet* **2**: 1109-1111.

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

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

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

Jorgensen, W.L. (1991). Rusting of the lock and key model for protein-ligand binding. *Science* **254**: 954-955.

Judson, R.S., Jaeger, E.P., and Treasurywala, A.M. (1994). A genetic algorithm based method for docking flexible molecules. *Theochem - J. Mol. Struc.* **114**: 191-206.

Justice, A.C., Feinstein, A.R. and Wells, C.K. (1989). A new prognostic staging system for the acquired immunodeficiency syndrome. *N. Engl. J. Med.* **320**: 1388-1393.

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

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U.S.A.* **89**: 2195-2199.

Kerwin, S.M., Kuntz, I.D. and Kenyon, G.L. (1991). The design of a DNA-binding compound using an automated procedure for screening potential ligands. *Med. Chem. Res.* **1**: 361-369.

Koshland, D.E., Jr. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **44**: 98-104.

Koshland, D.E., Jr. (1971). Molecular basis of enzyme catalysis and control. *Pure Appl. Chem.* **25**: 199-133.

Koshland, D.E., Jr. (1994). The key-lock theory and the induced fit theory. *Angew. Chem. Int. Ed. Engl.* **33**: 2375-2378.

Kovacs, J.A., Hiemenz, J.R., Macher, A.M., Stover, D., Murray, H.W., Shelhamer, J., Lane, H.C., Urmacher, C., Honig, C., Longo, D.L., Parker, M.M., Natanson, C., Parrillo, J.E., Fauci, A.S., Pizzo, P.A. and Masur, H. (1984). *Pneumocystis carinii* pneumonia: A comparison between patients with the acquired immunodeficiency syndrome and patients with other immunodeficiencies. *Ann. Intern. Med.* **100**: 663-671.

Kovacs, J.A. and Masur, H. (1988). *Pneumocystis carinii* pneumonia: therapy and prophylaxis. *J. Infect. Dis.* **158**: 254-259.

Krystek, S., Stouch, T. and Novotny, J. (1993). Affinity and specificity of a serine endopeptidase - protein inhibitor interactions: Empirical free energy calculations based on x-ray crystallographic structures. *J. Mol. Biol.* **234**: 661-679.

Kuhl, F.S., Crippen, G.M., and Friesen, D.K. (1984). *J. Comp. Chem.* **5**: 24.

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

Kuntz, I.D. (1992). Structure-based strategies for drug design and discovery. *Science* **257**: 1078-1082.

Kuntz, I.D., Meng, E.C., and Shoichet, B.K. (1994). Structure-based molecular design. *Acc. Chem. Res.* **27**: 117-123.

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

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

Lewis, R.A. and Leach, A.R. (1994). Current methods for site-directed structure generation. *J. Comput.-Aided Mol. Design.* **8**: 467-475.

Lichtenthaler, F.W. (1994). 100 Years "Schlüssel-Schloss-Prinzip": What made Emil Fischer use this analogy? *Angew. Chem. Int. Ed. Engl.* **33**: 2364-2374.

Lin, S.L., Nussinov, R., Fischer, D. and Wolfson, H.J. (1994). Molecular surface representations by sparse critical points. *Proteins* **18**: 94-101.

Lybrand, T.P. (1995). Ligand-protein docking and rational drug design. *Curr. Opin. Struc. Biol.* **5**: 224-228.

Margosiak, S.A., Appleman, J.R., Santi, D.V. and Blakley, R.L. (1993). Dihydrofolate reductase from the pathogenic fungus *Pneumocystis carinii*: Catalytic properties and interaction with antifolates. *Arch. Biochem. Biophys.* **305**: 499-508.

Martin, Y.C. (1992). 3D database searching in drug design. *J. Med. Chem.* **35**: 2145.

Masur, H. (1992). Prevention and treatment of *Pneumocystis* pneumonia. *N. Engl. J. Med.* **327**: 1853-1860.

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

Meng, E.C., Gschwend, D.A., Blaney, J.M. and Kuntz, I.D. (1993). Orientational sampling and rigid-body minimization in molecular docking. *Proteins* **17**: 266-278.

Meng, E.C., Kuntz, I.D., Abraham, D.J. and Kellogg, G.E. (1994). Evaluating docked complexes with the HINT exponential function and empirical atomic hydrophobicities. *J. Comput.-Aided Mol. Design* **8**: 299-306.

Miller, M., Schneider, J., Sathyanarayana, B.K., Toth, M.V., Marshall, G.R., Clawson, L., Selk, L., Kent, S.B. and Wlodawer, A. (1989). Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 Å resolution. *Science* **246**: 1149-1152.

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

Mills, J. (1986). *Pneumocystis carinii* and *Toxoplasma gondii* infections in patients with AIDS. *Rev. Infect. Dis.* **8**: 1001-1011.

Mills, J. and Masur, H. (1990). AIDS-related infections. *Sci. Amer.* **263**: 50-57.

Murray, J.F. and Mills, J. (1990). Pulmonary infectious complications of human immuno-deficiency virus infection. Part II. *Am. Rev. Respir. Dis.* **141**: 1582-1598.

Norel, R., Lin, S.L., Wolfson, H.J. and Nussinov, R. (1994). Shape complementarity at protein-protein interfaces. *Biopolymers* **34**: 933-940.

Oefner, C., Winkler, F. and D'Arcy, A. (1991). Structure determination of *P. carinii* dihydrofolate reductase by X-ray crystallography. Unpublished results.

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

Ring, C.S., Sun, E., McKerrow, J.H., Lee, G.K., Rosenthal, P.J., Kuntz, I.D. and Cohen, F.E. (1993). Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. U.S.A.* **90**: 3583-3587.

Rutenber, E., Fauman, E.B., Keenan, R.J., Fong, S., Furth, P.S., Ortiz de Montellano, P.R., Meng, E., Kuntz, I.D., DeCamp, D.L., Salto, R., Rosé, J.R., Craik, C. and Stroud, R.M. (1993). Structure of a non-peptide inhibitor complexed with HIV-1 protease: Developing a cycle of structure-based drug design. *J. Biol. Chem.* **268**: 15343-15346.

Schulz, G.E., Müller, C.W. and Diederichs, K. (1990). Induced-fit movements in adenylate kinases. *J. Mol. Biol.* **213**: 627-630.

Shoichet, B.K. and Kuntz, I.D. (1991). Protein docking and complementarity. *J. Mol. Biol.* **221**: 327-346.

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

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

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

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

Thanki, N., Thornton, J.M. and Goodfellow, J.M. (1988). Distributions of water around amino acid residues in proteins. *J. Mol. Biol.* **202**: 637-657.

Tintelnot, M. and Andrews, P. (1989). Geometries of functional group interactions in enzyme-ligand complexes: Guides for receptor modelling. *J. Comput.-Aided Mol. Design* **3**: 67-84.

U.S. Public Health Service Task Force on Anti-*Pneumocystis* Prophylaxis in Patients with HIV Infection (1993). Recommendations for prophylaxis against *Pneumocystis carinii* pneumonia for persons infected with HIV. *J. Acquir. Immun. Defic. Syndr.* **6**: 46-55.

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

Van Duyne, G.D., Standaert, R.F., Karplus, R.F., Schreiber, S.L. and Clardy, J. (1991). Atomic structure of FKBP-FK506, an immunophilin-immunosuppressant complex. *Science* **252**: 839-842.

Walls, P.H. and Sternberg, M.J.E. (1992). A new algorithm to model protein-protein recognition based on surface complementarity: Application to antibody-antigen docking. *J. Mol. Biol.* **228**: 277-297.

Walzer, P.D., Perl, D.P., Krogstad, D.J., Rawon, P.G. and Schultz, M.G. (1974). *Pneumocystis carinii* pneumonia in the United States. *Ann. Intern. Med.* **80**: 83-93.

Wang, H. (1991). Grid-search molecular accessible surface algorithm for solving the protein docking problem. *J. Comp. Chem.* **12**: 746-750.

Warshel, A., Tao, H., Fothergill, M. and Chu, Z.-T. (1994). Effective methods for estimation of binding energies in computer-aided drug design. *Israel J. Chem.* **34**: 253-256.

Wharton, J.M., Coleman, D.L., Wofsy, C.B., Luce, J., Blumenfeld, W., Hadley, W.K., Ingram-Drake, L., Volberding, P.A. and Hopewell, P.C. (1986). Trimethoprim-sulfamethoxazole or pentamidine for *Pneumocystis carinii* pneumonia in the acquired immunodeficiency syndrome. *Ann. Intern. Med.* **105**: 37-44.

Wodak, S. and Janin, J. (1978). Computer analysis of protein-protein interactions. *J. Mol. Biol.* **124**: 323-342.

Yamada, M. and Itai, A. (1993). Development of an efficient automated docking method. *Chem. Pharm. Bull.* **41**: 1200-1202.

Chapter 3.

Specificity in Structure-Based Drug Design: Identification of a Novel, Selective Inhibitor of *Pneumocystis carinii* Dihydrofolate Reductase[†]

Daniel A. Gschwend,[‡] Worachart Sirawaraporn,[§] Daniel V. Santi,[‡] and

Irwin D. Kuntz[*‡]

Departments of Pharmaceutical Chemistry and of Biochemistry and Biophysics,

University of California, San Francisco, CA 94143,

and

Department of Biochemistry, Faculty of Science,

Mahidol University, Bangkok 10400, Thailand

[†] Part of this work appears in preliminary form elsewhere [Gschwend *et al.*, 1995 (Chapter 2)].

[‡] University of California.

[§] Mahidol University.

[*] Author to whom correspondence should be addressed.

ABSTRACT

The concern for specificity is an important but unfortunately often-overlooked aspect of structure-based drug design. The ability to selectively modulate biochemical processes without interfering with related systems is crucial to the success of chemotherapy. Distinguishing between related targets in different organisms is another example within this theme. *Pneumocystis carinii* is a fungal opportunist which causes a crippling pneumonia in immunocompromised individuals and continues to frustrate modern medicinal chemistry. We report on the application of computational molecular docking techniques for identification of novel inhibitors of *P. carinii* dihydrofolate reductase (DHFR) that are selective against the human isozyme. The Fine Chemicals Directory, a database of commercially available compounds, was screened with the DOCK program suite. We have introduced a post-docking refinement directed at discerning subtle structural and chemical features and acting as an indicator of species specificity. Of forty compounds predicted to exhibit anti-*Pneumocystis* DHFR activity, each of novel chemical framework, thirteen (33%) show IC_{50} values better than 150 μM in an enzyme assay. These inhibitors were further assayed against human DHFR: ten (77%) bind preferentially to the fungal enzyme. The most potent compound identified is a 7 μM-inhibitor of *P. carinii* DHFR and displays 25-fold selectivity. This agent exhibits a number of appealing properties which might make it a suitable candidate for further investigation. The ability of molecular docking methods to locate selective inhibitors reinforces our view of structure-based drug discovery as a valuable strategy, not only for identifying lead compounds, but also for addressing more complex issues concerning receptor specificity.

INTRODUCTION

Successful chemotherapeutic treatment depends upon the exploitation of biochemical differences between host and pathogen. An ideal drug is harmful to an invader without being harmful to the host. The success of selective toxicity hinges upon the inhibition of a biochemical process vital to the pathogen's survival. Such processes present a spectrum of targets with varying risks of host toxicity. Genuinely unique biochemical systems are seldom available. Generally, differences between host and pathogen are more subtle, forcing distinction between functionally equivalent targets. The work presented here focuses on the discrimination of structurally similar entities with the intent of designing species-specific drugs.

Structural information is critical to understanding differences between functionally equivalent targets. Atomic coordinates often form a basis for directed drug design, as knowledge of the topography and chemistry within the active site allows tailoring of specific interactions. The DOCK suite of programs (Kuntz *et al.*, 1982; Shoichet *et al.*, 1992; Meng *et al.*, 1992) is one computational method that uses the three-dimensional (3-D) structure of a receptor to identify ligands of complementary shape. DOCK has been used to identify leads in a number of diverse systems (Kuntz, 1992). However, DOCK has not been explicitly applied to the discovery of selective inhibitors of therapeutic interest. To explore this avenue, a clinically relevant system has been chosen for which high-resolution structural information exists for a variety of species.

The regeneration of reduced folate cofactors required for nucleotide biosynthesis is performed by dihydrofolate reductase (DHFR). Despite its universally crucial role, this enzyme commonly exhibits low sequence conservation. This potential for selective drug design has not passed unnoticed: DHFR is thoroughly characterized with respect to structural biology across species (Kraut & Matthews, 1987; Morrison, 1991). At least a dozen crystal structures are available spanning all combinations of bound and unbound states. Thousands of inhibitors with varying potencies and selectivities have been identified (Blaney *et al.*, 1984). Thus far DHFR has proven a rewarding target for selective inhibition, resulting in numerous antibacterial, antiprotozoal, and antineoplastic agents (Schweitzer *et al.*, 1990). However, such success has been elusive in treating *Pneumocystis carinii* infections.

Pneumocystis carinii is a pathogen harmlessly infecting nearly all humans, but upon reactivation of latent infection by immunodeficiency can induce a disease state characterized by a crippling pneumonia (Murray & Mills, 1990; Bartlett & Smith, 1991). Not surprisingly, this opportunist is the principal agent of morbidity and mortality in HIV-infected persons (Murray & Mills, 1990; Mills & Masur, 1990). Without chemoprophylaxis, 60 to 85% of AIDS patients eventually will be afflicted by *P. carinii* pneumonia, and 25% will die from it (Walzer *et al.*, 1974; Mills, 1986; Kovacs & Masur, 1988; Justice *et al.*, 1989; Bartlett & Smith, 1991). Much of the research towards chemotherapy stems from treatments found effective against similar pathogens, as the metabolic pathways in *Pneumocystis* are largely uncharacterized. The organism's phylogeny is therefore quite important. Mounting evidence places *P. carinii* among the fungi (Edman *et al.*, 1988; Edman *et al.*, 1989; Stringer *et al.*, 1989; Pixley *et al.*, 1991; Ypma-Wong *et al.*, 1992; Belfield *et al.*, 1993; Furlong *et al.*, 1994), yet the atypical cell membrane composition (Kaneshiro *et al.*, 1989; Furlong *et al.*, 1994) appears to thwart common antifungal drugs (Bartlett *et al.*, 1994b). Agents in a variety of other mechanistically distinct classes are being explored, including antifolates (Allegra *et al.*, 1987b; Kovacs *et al.*, 1989; Margosiak *et al.*, 1993), DNA-replication antagonists (Walzer *et al.*, 1988; Tidwell *et al.*, 1990; Dykstra & Tidwell, 1991; Fishman *et al.*, 1993; Dykstra *et al.*, 1994; Walzer *et al.*, 1994), inhibitors of polyamine biosynthesis (Lipschik *et al.*, 1991; Saric & Clarkson, 1994), compounds which interfere with cell wall integrity (Schmatz *et al.*, 1990; Powles *et al.*, 1994; Yasuoka *et al.*, 1995), growth-stunting iron chelators (Weinberg, 1994), inhibitors of pyrimidine biosynthesis (Falloon *et al.*, 1991; Artymowicz & James, 1993; Ittarat *et al.*, 1995), microtubule disrupting agents (Bartlett *et al.*, 1994a), and sodium channel blockers (Shaw *et al.*, 1994). The most successful of these approaches thus far have been the antifolates and DNA-replication antagonists.

Co-trimoxazole and pentamidine isethionate are the most widely prescribed preparations for therapy and prophylaxis of *Pneumocystis carinii* pneumonia (Kovacs & Masur, 1988; Murray & Mills, 1990; U.S. Public Health Service, 1993; Gallant *et al.*, 1994). Unfortunately, these treatments are plagued with adverse reactions. Fifty percent of patients using pentamidine and 65% using co-trimoxazole will suffer major side effects (Walzer *et al.*, 1974; Jaffe *et al.*, 1983; Kovacs *et al.*, 1984; Gordin *et al.*, 1984; Wharton *et al.*, 1986). On average, one in four patients will suffer reactions severe enough to force discontinuation of treatment (Allegra *et al.*, 1987a). The frequency of such reactions is not surprising in the case of co-trimoxazole: the dihydropteroate synthetase (DHPS) inhibitor component, sulfamethoxazole, is quite toxic (Masur, 1992) and the DHFR inhibitor component, trimethoprim, is weak and non-selective (Edman *et al.*, 1988; Margosiak *et al.*, 1993). In fact, it is almost general that clinically relevant DHFR inhibitors are selective for *human* DHFR (Margosiak *et al.*, 1993). Derivitization of mainstream antifolates to improve selectivity for *P. carinii* DHFR has also been largely unsuccessful. Trimetrexate (Queener, 1991; Gangjee *et al.*, 1994; Rosowsky *et al.*, 1993), piritrexim (Rosowsky *et al.*, 1993; Gangjee *et al.*, 1994), and triazine (Rosowsky *et al.*, 1995) analogues fail to display any selectivity toward *Pneumocystis*, while some 2,4-diaminopteridine and 2,4-diaminoquinazoline derivatives (Queener, 1991;

Rosowsky *et al.*, 1995) exhibit favorable selectivities only on the order of 2 - to 20-fold. The lack of selectivity and resulting side effects of antifolates is likely a direct consequence of the similarity between host and pathogen DHFR.

Based on published sequence alignments (Blakley, 1984; Edman *et al.*, 1989), *P. carinii* DHFR displays highest similarity with that of vertebrates. Sequence identities of 35-40% and homologies near 70% are observed. High-resolution crystal structures of human (Davies *et al.*, 1990) and *P. carinii* (Oefner *et al.*, 1991) DHFR confirm only minor differences. Figure 2 in Chapter 2 depicts the extensive surface similarity in the active site region of the two enzymes. Within the folate binding pocket, there are in fact only six non-identical residues. The active site aspartate common to bacterial and protozoan DHFR is replaced with glutamate in *P. carinii*, as in all vertebrate DHFR. Further drug design complications arise from the fact that the *Pneumocystis* active site is nearly universally smaller than the human active site, making exploitation of unique pockets impossible. Despite the structural similarities of the target isozymes, there are differences between host and pathogen that can work to our advantage. The stronger binding affinity of substrate for human DHFR relative to *P. carinii* DHFR suggests that antifolates will be able to compete more effectively with dihydrofolate in the pathogen (Margosiak *et al.*, 1993). *Pneumocystis* also demonstrates an inability to salvage pre-formed reduced folates from the environment (Allegra *et al.*, 1987b), thus amplifying the intrinsic selectivity of any antifolate.

The lack of differentiation by mainstream agents highlights the urgency for alternative molecular frameworks. The derivitization of compounds which exhibit weak potency, unfavorable selectivity, poor uptake, or otherwise toxic effects is one approach for obtaining a clinically useful agent. We will adopt a complementary approach: the identification of novel chemical skeletons from which to initiate new routes of optimization.

Such leads frequently offer an entirely different spectrum of pharmacological properties. We here present the application of the DOCK screening process toward the therapeutically important problem of identifying novel, selective anti- *Pneumocystis* agents (Gschwend, 1995).

METHODS

Structural Preparation

The *P. carinii* DHFR structure used in this study was the ternary crystal complex with folate (a poor substrate) and NADPH (cofactor) solved to 1.8 Å resolution by Oefner *et al.* (1991). Substrate and cofactor, as well as all crystallographically observed water molecules, were removed from the structure. The human crystal structure employed was the binary complex with substrate [Protein Data Bank (Bernstein *et al.*, 1977) entry 1DHF] solved to 2.3 Å resolution published by Davies *et al.* (1990). Of the two molecules in the human DHFR unit cell, Davies *et al.* deem the B-chain more reliable, so the A-chain was removed. The substrate and all waters were also deleted. A structural alignment based on the sequence alignment of Edman *et al.* (1989) was performed using alpha carbons of all active site residues present in both structures (55 residues yielding a root-mean-square (rms) deviation of 0.56 Å).

Docking Overview

The DOCK suite of programs is designed to identify putative ligands complementary to a receptor of known 3-D structure. The details of the method have been described previously (Kuntz *et al.*, 1982; Shoichet *et al.*, 1992; Meng *et al.*, 1992) - only an overview is given here. By filling the receptor site with overlapping spheres of varying sizes

(Kuntz *et al.*, 1982), a negative image capturing the bumps and grooves of the region of interest is generated. A 3-D database is searched (DesJarlais *et al.*, 1988) for molecules whose interatomic distances match the inter-sphere-center distances. Each compound is evaluated in thousands of orientations in the active site by an approximate molecular mechanics interaction energy (Meng *et al.*, 1992). The best-scoring compounds presumably exploit multiple geometric and/or chemical properties of the receptor site and are thus of considerable interest as inhibitor candidates. The DOCK procedures have been tested through studies of crystallographic complexes. The experimental geometries are associated with the best-scoring orientations, generally within 1 Å rms deviation. More importantly, compounds that have radically different structures from known inhibitors are often found (Kuntz, 1992).

Docking Analysis

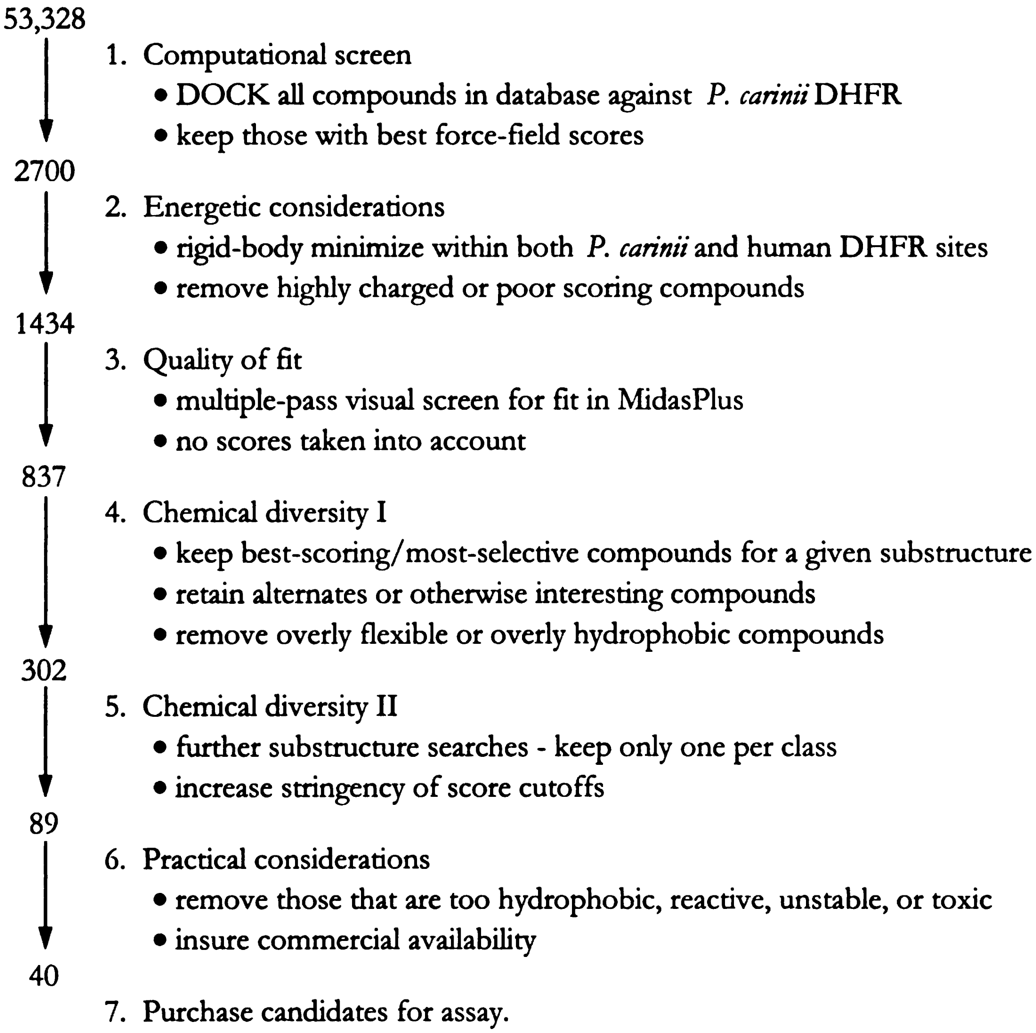
A molecular surface was generated for the *P. carinii* active site using Connolly's MS algorithm (Connolly, 1983a,b). The resulting surface formed the basis for the SPHGEN (Kuntz *et al.*, 1982) calculation, which produced a set of spheres characterizing the topology of the target site. As the length of the docking calculation depends combinatorially on the number of spheres (Shoichet *et al.*, 1992), the sphere cluster was manually trimmed to a minimal set spanning the folate binding pocket and extending to the nicotinamide-ribose portion of the cofactor groove. Spheres representing the remainder of the cofactor pocket were removed to avoid retrieving compounds which might bind exclusively to the cofactor binding region and hence bind non-specifically to multiple targets in the body. Seventy-four spheres thus defined the targeted site , as illustrated with their collective molecular surface in

Figure 3 of Chapter 2. SPHGEN's ability to capture shape features precisely is evident in the extensive complementarity between the sphere surface and the receptor molecular surface.

A box of dimensions 25 Å × 22 Å × 14 Å encompassed the spheres and delimited the force-field scoring grid. The scoring grid was computed with CHEMGRID (Meng *et al.*, 1992) using a 0.25 Å grid spacing, a dielectric of 4 *r* (where *r* is the interatomic separation), and a generous non-bonded cutoff of 20 Å; hydrogens were added in standard geometries . Close contact limits were set at 2.3 and 2.8 Å for polar and nonpolar atoms, respectively. Results are not sensitive to the precise location of the grid (data not shown). The Fine Chemicals Directory [FCD3D v.89.2 (this database is now called the Available Chemicals Directory), MDL Information Systems, Inc., San Leandro, CA] was screened with DOCK version 3.0 (Meng *et al.*, 1992). *Dislim* was set to 1.5 Å, *nodlim* to 4, and bin parameters to 0.2, 0.0, 1.0, and 0.0 Å (*lbinsz*, *lovlap*, *sbinsz*, *sovlap*, respectively); force-field score interpolation was on. An average of 19,000 orientations was examined for each of 53,328 ligands, utilizing 350 hours of CPU time (Silicon Graphics PI 4D/35 ; Silicon Graphics, Inc., Mountain View, CA). The over one billion total configurations investigated at a rate of 800 per second attest to the speed of the DOCK program.

Database Screening

The stepwise screening of the FCD is outlined in Figure 1. Arbitrarily, the top 2700 force-field scoring compounds were saved from the DOCK run against *Pneumocystis* DHFR. Each compound was then subject to a quasi-Newton rigid-body minimization (Meng *et al.*, 1993) to optimize the intermolecular interactions of the best scoring orientation. This refinement was carried out independently in the context of *both P. carinii* and human DHFR active sites. Thus, the starting configuration in *Pneumocystis* DHFR was determined by



53,328

1. Computational screen
   - DOCK all compounds in database against *P. carinii* DHFR
   - keep those with best force-field scores

2700

2. Energetic considerations
   - rigid-body minimize within both *P. carinii* and human DHFR sites
   - remove highly charged or poor scoring compounds

1434

3. Quality of fit
   - multiple-pass visual screen for fit in MidasPlus
   - no scores taken into account

837

4. Chemical diversity I
   - keep best-scoring/most-selective compounds for a given substructure
   - retain alternates or otherwise interesting compounds
   - remove overly flexible or overly hydrophobic compounds

302

5. Chemical diversity II
   - further substructure searches - keep only one per class
   - increase stringency of score cutoffs

89

6. Practical considerations
   - remove those that are too hydrophobic, reactive, unstable, or toxic
   - insure commercial availability

40

7. Purchase candidates for assay.

Figure 1. Fine Chemicals Directory screening pipeline.
See text. The number of compounds at each stage is indicated.

DOCK, while the starting configuration in human DHFR was determined by the structural alignment discussed previously. The differential, optimized force-field scores were used as an indication of species selectivity. We term E_{pc} and E_{hum} the optimized force-field scores in *Pneumocystis* and human DHFR respectively, and the differential score $E_{scl} = E_{pc} - E_{hum}$.

A manual screening process of the computationally selected 2700 compounds ensued. Because the DOCK 3.0 force-field score (Meng *et al.*, 1992) incorporates no formal solvation terms, an attempt to counter this deficiency was made. Compounds with a net charge of -3 or less (6%) were discarded. Compounds with a net charge of ±1 or ±2 (74%) were retained if $E_{pc} \leq -40$, an intermediate cutoff chosen to force charged compounds to score better than a hypothetical neutral counterpart. Compounds which were net neutral (20%) were retained regardless of score. 1434 compounds emerged from this crude solvation filter. These compounds were examined visually on a graphics terminal using the MidasPlus package (Ferrin *et al.*, 1988) in two independent passes. No scores were taken into account in this filter - structures were examined for fit to the site and for visually appealing interactions. Compounds that were either too small or too large, or docked to the surface of the receptor, were also removed at this stage. 837 compounds passed the visualization filter.

Two chemical diversity screens were introduced. Compounds were clustered by substructural class with the aid of the MACCS-II 3D package (MDL Information Systems, Inc., San Leandro, CA). The structures in each class predicted to have either high affinity for *P. carinii* DHFR (E_{pc}) or exhibit selectivity (E_{sel}) for *P. carinii* DHFR were retained. Overly flexible or extremely hydrophobic compounds were eliminated during this filter. The 302 remaining compounds were subject to the second, more stringent, chemical diversity screen. Further substructure searches were used to select the most dissimilar compounds. At this point, only one compound per class was saved, gauged by more restrictive E_{pc} and E_{sel} cutoffs. Finally, practical considerations were applied to the 89 candidate compounds. Compounds which appeared too reactive, unstable, toxic, or insoluble were discarded. After verifying commercial availability, forty compounds were declared candidates and purchased for biological evaluation.

Chemicals

NADPH and dihydrofolic acid were purchased from Sigma. DOCK-selected candidates for assay were purchased from various chemical suppliers, as indicated in Table 1. Each compound was prepared as a stock solution in DMSO and subsequently diluted in water to provide a range of concentrations suitable for IC_{50} determination. The final concentration of DMSO in the enzymatic assay was less than 1% to avoid inhibition of the reaction by DMSO itself.

Dihydrofolate reductase assay

The spectrophotometric assay for DHFR activity is well-characterized. Activity is determined by monitoring the decrease in absorption at 340 nm, corresponding to the utilization of NADPH, at 25 °C (Hillcoat *et al.*, 1967; Sirawaraporn *et al.*, 1991). The standard assay (1 ml) contained 50 mM Tes (pH 7.0), 75 mM β-mercaptoethanol, 1 mM EDTA, 1 mg/ml bovine serum albumin, 50 μM NADPH, 20 μM dihydrofolic acid, and limiting enzyme. Dihydrofolic acid concentration was checked by UV absorption and confirmed enzymatically using $\Delta\varepsilon_{340}$ of 12,300 M^{-1} cm^{-1}. Reactions were initiated with NADPH, mixed thoroughly, and monitored for 5 minutes. No-enzyme blanks were used to control for background decomposition of NADPH and/or inhibitor. The concentration of inhibitor required to reduce DHFR activity by 50% (IC_{50}) was determined by interpolation of sigmoidal plots of percentage inhibition versus log inhibitor concentration.

Table 1. Biological evaluation of DOCK-selected inhibitor candidates.

Compounds are sorted by increasing optimized force-field score against *P. carinii* DHFR.

#	Compound	Supplier	Formal	FF-Score[a]		Inhibition[b]	
			Charge	Pc	Hu	Pc	Hu
1	5-bromo-2’-deoxyuridine-5’-mono phosphate	Sigma	-2	-57.3	-49.4	Ø	
2	Acid Blue 40	Aldrich	-1	-53.8	-47.8	77	>500[c]
3	5,5’-thiodisalicylic acid	Bader	-2	-52.5	-26.3	(13%)	
4	5-bromo-4-chloro-3-indolyl phosphate	Aldrich	-2	-50.7	-40.5	Ø	
5	Palatine Fast Black WAN	Janssen	-1	-49.1	-11.3	Ø	
6	4’-(2-thiazolylsulfamoyl) phthalanilic acid	Bader	-1	-48.1	-37.1	Ø	
7	4-(2-(2-chlorobenzoyl) acetamido benzoic acid	Bader	-1	-47.3	-32.6	(26%)	
8	Acid Red 1	Aldrich	-2	-46.8	-22.8	Ø	
9	3’,3”,5’,5”-tetraiodo phenolphthalein	Aldrich	-2	-46.7	-38.1	92	290
10	1-(4-pyridylcarbonyl)-2-(carboxy methoxyacetyl) hydrazine	Bader	-1	-46.3	-36.3	(14%)	
11	N-furfuryltetrachloro phthalanilic acid	Bader	-1	-45.6	-33.6	(12%)	
12	Benzoyl Leuco Methylene Blue	TCI	0	-45.5	-34.9	Ø	
13	Palatine Chrome Black 6BN	Aldrich	-1	-44.4	-9.5	80	140
14	2,3’,6-trichloro indophenol	TCI	-1	-44.4	-32.2	95	226
15	ethyl-4-(5-chloro-2-phenoxyphenyl ureido) benzoate	Bader	0	-44.2	-16.7	(17%)	
16	Pamoic Acid	Aldrich	-2	-43.8	277.8	96	172
17	terephthaloyl-*bis*-glycine	Riedel	-2	-43.7	-25.2	Ø	
18	2,5-*bis* (2-methoxyanilino)-3,6-dichloro-1,4-benzoquinone	Bader	0	-43.4	-35.5	(31%)	
19	N,N’-(2-bromo-6-methylphenylene) *bis* (4-methylcoumarin-7ylcarbamate)	Bader	0	-43.3	-30.8	90	60
20	Gallein	Sigma	0	-42.4	-29.9	74	228
21	3,4,5,6-tetrachloro-3’-(trifluoro methyl)-phthalanilic acid	Bader	-1	-42.3	-25.7	(>21%)	
22	3,3’,5-triiodo thyropropionic acid	Sigma	-1	-42.3	-31.9	120	>500[c]
23	3,5,6-tri-(2-pyridyl)-1,2,4-triazine	Maybridge	0	-41.1	-35.5	Ø	

#	Compound	Supplier	Formal Charge	FF-Score[a]		Inhibition[b]	
				Pc	Hu	Pc	Hu
24	N-(3-methoxyphenyl) picrylamine	Bader	0	-41.0	-33.2	100	43
25	3,3',4,4',5-pentamethoxy benzophenone	Bader	0	-40.8	-18.8	100	42
26	Threne Red Violet RH	TCI	0	-40.3	-33.3	6.9	~200[d]
27	N-(*p*-(2-benzoxazolyl) phenyl)-maleimide	Kodak	0	-40.3	-34.3	Ø	
28	1-(3-bromobenzoyl)-2-(2-napthoyl) hydrazine	Bader	0	-40.0	-33.8	Ø	
29	5'-benzoyluridine	Sigma	0	-39.9	-35.5	Ø	
30	Disperse Orange 13	Aldrich	0	-39.7	-36.0	Ø	
31	1-(4-chlorobenzyl)-1-methyl-3-phenyl-2-thiourea	Bader	0	-39.2	-32.9	Ø	
32	4,4'-dimethyl-2,2'-dihydroxy-6,6'-biquinoline	Bader	0	-39.1	-28.7	Ø	
33	4-chloro-6-fluorosulfonyl-2-(4-nitrophenyl) quinoline	Bader	0	-38.5	-25.0	41	~150[d]
34	1-(2,6-dichlorophenyl)-3-(6-methyl-3-pyridyl) urea	Bader	0	-37.2	-32.7	Ø	
35	N-(2-hydroxyphenyl)-3,4,5,6-tetrachloro phthalimide	Bader	0	-37.1	-34.1	Ø	
36	N-(4-(ethoxycarbonyl) phenyl)-2-(2,4,5-trichlorophenoxy) acetamide	Bader	0	-37.1	-34.5	Ø	
37	2-(N-(3,4-dichlorophenyl) carbamoyl) amin-6-methoxy benzothiazole	Bader	0	-36.1	-32.0	(24%)	
38	Murexide	Aldrich	0	-35.7	-32.9	Ø	
39	2,4,6-triphenoxy-*s*-triazine	Aldrich	0	-35.2	-18.4	130	~500[d]
40	2,4-*bis* (*p*-tolylthio)-1,3-dithia-2,4-diphosphetane-2,4-disulfide	TCI	0	-33.6	-40.5	Ø	

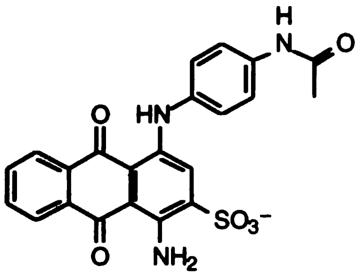
[a] Optimized force-field score (kcal/mol) against *P. carinii* (Pc) and human (Hu) DHFR

[b] Assayed inhibition against *P. carinii* (Pc) and human (Hu) DHFR. Values are μM IC_{50} values, except those in parentheses, which indicate percentage inhibition at an inhibitor concentration of 100 μM. Ø denotes that no inhibition was observed. Data are the result of at least duplicate determinations, agreeing to within 10-20%. Blank entries were unassayed.

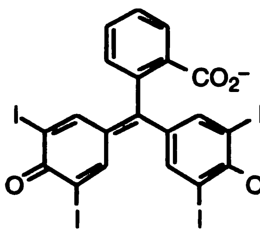
[c] No inhibition was observed at half-millimolar concentrations.

[d] IC_{50} is estimated by extrapolation due to solubility limitations.

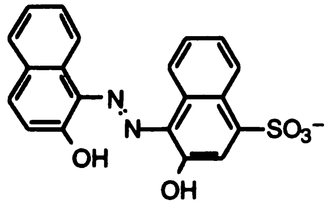
2



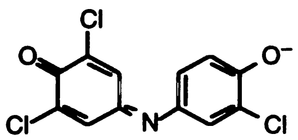
9



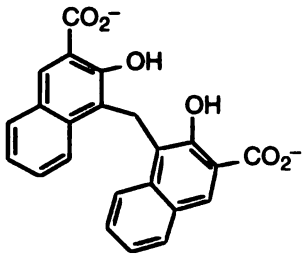
13



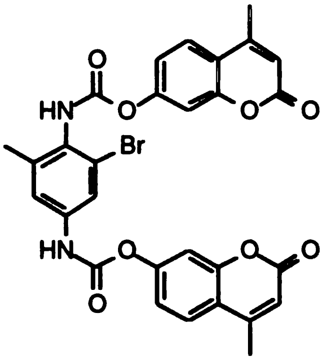
14



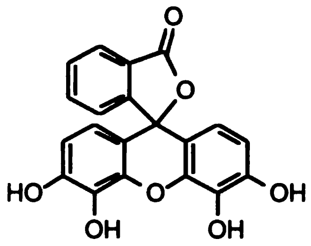
16



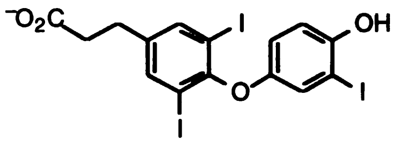
19



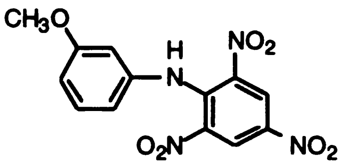
20



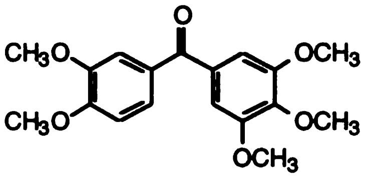
22



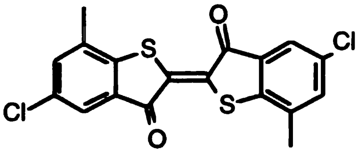
24



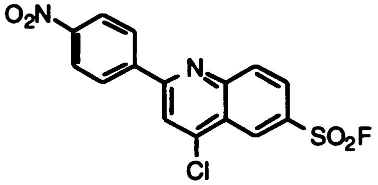
25



26



33



39

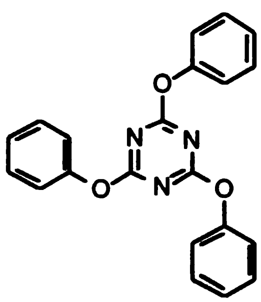


Figure 2. Structures of compounds with measured IC_{50}'s against *P. carinii* DHFR. (previous page) The numbering corresponds to that given in Table 1.

RESULTS

The forty compounds examined in the enzyme assay are listed in Table 1. Twenty-one (53%) displayed measurable inhibition against *P. carinii* DHFR at the 100 µM level, with eleven (28%) showing IC_{50}'s of 100 µM or less. Of the thirteen compounds which were potent and soluble enough to permit an IC_{50} determination against *P. carinii* DHFR (structures shown in Figure 2), ten (77%) exhibited selectivity ranging from 2- to 25-fold against human DHFR. The marked success in identifying species-specific agents is depicted graphically in Figure 3. Although the binding kinetics of these inhibitors has not been determined, their selective nature suggests that binding is not non-specific. The most potent compound (**26**) is also the most selective, having an IC_{50} of 7 µM and approximately 25-fold selectivity for *P. carinii* DHFR. DOCK's predicted mode of binding for this compound (Figure 4) places the structure in contact with four of the six non-identical residues in the active site.

The optimized DOCK force-field scores, used to select compounds for assay, are provided in Table 1. For comparison, methotrexate, a picomolar inhibitor of both *P. carinii* and human DHFR (Marogosiak *et al.*, 1993), receives a force-field score on the order of -70 kcal/mol. No apparent correlation is observed between these scores and assayed inhibition or selectivity. As detailed in the Discussion section, this was not unexpected. Also noteworthy is the bias of the force-field scores toward more highly charged molecules, evidenced by the larger proportion of such compounds higher in Table 1. This behavior

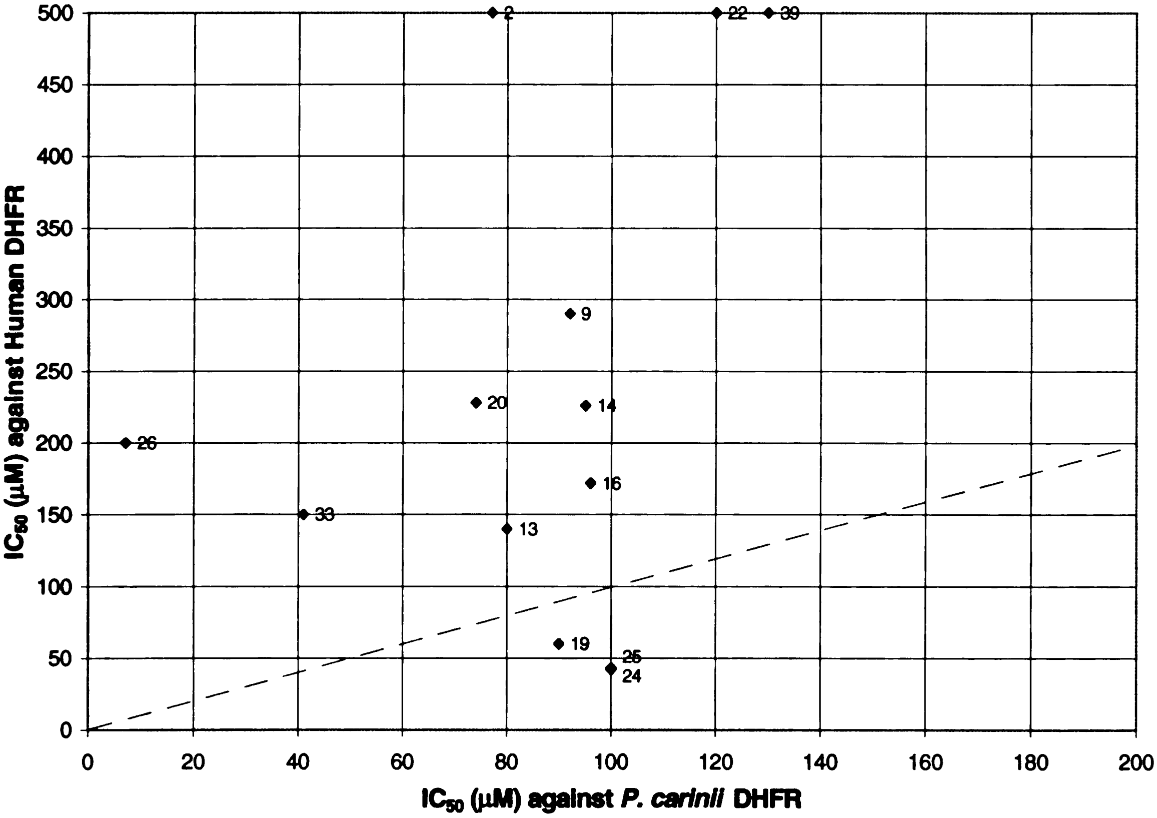


Figure 3. Species-specificity of novel *P. carinii* DHFR inhibitors.

IC_{50}'s for compounds assayed against both *P. carinii* and human DHFR are plotted. The dashed line represents the absence of selectivity; compounds falling above the line show favorable specificity. Numbers identify compounds as given in Table 1.

results from the lack of a desolvation penalty in the force-field score. Countering deficiencies like this one is the goal of the post-DOCK filtering process.

Discussion

Computational strategies for structure-based drug discovery offer a valuable alternative to the costly and time-consuming process of random screening (Kuntz, 1992). Coupled with a database of commercially available compounds, such as the FCD, programs like DOCK can provide extremely rapid access to novel leads (Gschwend *et al.*, 1995). In our experience, DOCK typically demonstrates a hit rate at the micromolar level of 2 to 20% for compounds assayed for inhibition. However, because of the many approximations underlying the search and scoring engines (*e.g.* neglect of solvation terms, rigidity of ligand and receptor, discretized scoring), DOCK can not be expected to yield predictions of a quantitative nature. Rather, we prefer to value DOCK as a "macroscopic correlator" of binding affinity and interaction score. Even in the daunting task of species-specificity, macroscopic correlations when applied in sequence can, as demonstrated here, confer a powerful tool.

In the method presented here, the rigid-body minimization acts as the selectivity filter. An optimization of this type as a post-docking utility has been shown to improve agreement with experimentally determined binding modes (Meng *et al.*, 1993; Gschwend & Kuntz, unpublished results). Of the 50,000 compounds in the database, an enrichment for agents which inhibit *P. carinii* DHFR was achieved with DOCK (the first macroscopic correlation). Subsequently, the differential optimization in the context of both isozymes offered resolution along an additional dimension for these remaining compounds. By choosing structures which score highly in *Pneumocystis* DHFR and poorly in human DHFR (the second macroscopic correlation), an enrichment for agents selective against host DHFR has now been accomplished.

The manual processing of DOCK hits is an important stage for improving the true hit rate as defined by assayed inhibition. It would be unwise to purchase the top 50 DOCK hits and take them straight to the laboratory. Several of the limitations inherent to a DOCK investigation can be addressed with prudent post-docking analysis. One of the foremost insufficiencies of the DOCK force-field score is in the treatment of solvation effects.

Charged compounds experience a severe penalty for coming out of solution. Because the present scoring scheme incorporates no formal expression for desolvation, the affinity of such ligands is likely to be overestimated. A preference for neutral compounds and more restrictive criteria for charged compounds serves to counteract this deficiency. It is readily apparent from Table 1 that charged compounds are favored by the force-field score. A breakdown of the 11 "hits" (compounds with IC $_{50}$'s $\leq$ 100 μM) reveals that six (55%) were net neutral, three (27%) had a -1 formal charge, and two (18%) had a -2 formal charge. However, when normalized by the total number of compounds tested in each charge class, roughly equal hit rates are observed: 6/23 = 26% for net neutral compounds, 3/10 = 30% for singly-charged compounds, and 2/7 = 28% for doubly-charged compounds. We would speculate, given the absence of a desolvation penalty in scoring, that selecting charged DOCK compounds for assay without increased stringency as compared with neutral compounds would result in a much lower hit rate for charged compound classes.

Scanning possible candidates visually on a graphics terminal in the context of their receptor is a valuable tool for identifying leads. At this point one can mentally evaluate the relative importance of putative interactions while at the same time introducing some "virtual" flexibility into the rigidly docked components. Interactions that look as though they could be formed in the face of local breathing warrant attention. Although crude, this approach mitigates the lack of conformational flexibility in docking. A visual analysis also permits filtering by proximity to pre-defined (*e.g.* by other known ligands, mutagenesis data, *etc.*) hot-spots in the active site. Ligands docked to the surface or outer edge of the receptor are unlikely to exhibit the same degree of specificity for this receptor as ligands docked deep within a cleft. The medicinal chemist's intuition plays a significant role in transforming a list of DOCK suggestions to a list of plausible candidates.

Much of the often subjective post-DOCK analysis could be automated within the DOCK program with appropriate technological advances. The development of a more accurate scoring function to estimate true binding affinities would obviate the need for many of the solvation- and entropy-related corrections now performed manually. This is an active area of research (Gschwend *et al.*, 1995). Until such time as accurate evaluation methods are available, pre-organizing structural databases will be useful, allowing user-specified filtering before docking is begun. Examples of possible metrics for pre-organization might include formal charge, the number of rotatable bonds, or practical concerns such as cost, reactivity, or toxicity. Features enabling the targeting of specific regions of the receptor have been introduced into DOCK 3.5 (unpublished results), bypassing a post-DOCK visual filter for this purpose.

In practice, the number of assayable compounds is finite. With the goal of discovering lead compounds, it is vital to span as chemically diverse a cross-section of candidates as possible. In this study, we have employed two-dimensional substructure searches, separating compounds into structurally distinct classes. Within each class, structures which were predicted to have the highest affinity for *P. carinii* DHFR or show the most selectivity were chosen as representatives. Overly flexible compounds were disregarded in this analysis to combat the lack of a conformational entropy term in the scoring procedure (this is reflected by the rigidity of active entities shown in Figure 2). Extremely hydrophobic compounds were also avoided to circumvent problems of solubility and of non-specific binding.

The structural and chemical diversity of the resultant hits is made apparent in Figure 2 and highlights one of the strengths of database searching techniques for drug discovery. All of these compounds, to our knowledge, have never been identified as having antifolate activity.* Each represents a unique avenue of optimization toward a clinically useful agent. Because pharmacological and toxicological complications inevitably arise in the drug development pipeline, it is advantageous to maximize the number of diverse routes for evolution at the outset. Clearly, none of the inhibitors identified in this study is a drug; only a handful will even make suitable leads. The progression of low-micromolar enzyme inhibitors discovered by DOCK to more potent and therapeutically valuable agents has been reported (Rutenber *et al.*, 1993; Shoichet *et al.*, 1993; B.K. Shoichet, personal communication; Li *et al.*, 1995).

More pertinent to assessing the relevance of novel, micromolar inhibitors is the relative weakness and lack of selectivity of several mainstream anti- *Pneumocystis* agents. Although IC_{50} values are not directly comparable, trimethoprim and pyrimethamine, for example, show IC_{50}'s in the low-micromolar range (Allegra *et al.*, 1987b; Sirawarapom *et al.*, 1991; Broughton & Queener, 1991; Queener, 1991). Furthermore, an analysis of progress in the antifolate literature indicates that even a 10-fold preference for *P. carinii* DHFR is a relatively rare occurrence. Improved specificity will be required to reduce folate-related toxicity (Blakley, 1969; Margosiak *et al.*, 1993). The novel chemical frameworks identified in this study will clearly possess distinct pharmacological profiles, but may be likely to avoid sources of antifolate toxicity which are not folate-related [*e.g.* inhibition of histamine metabolism (Duch *et al.*, 1980)]. Although the potency of the inhibitors found here is relatively weak, the above points justify investigation into their potential for novel antifolate classes. Because of the substructure searches used to categorize similar molecules, each inhibitor represents a *class* of compounds identified in the DOCK run as exhibiting

* Compounds related to **2** have been observed to interact with nucleotide-requiring enzymes, including DHFR (Beissner & Rudolph, 1978; Chambers & Dunlap, 1979).

complementarity to the receptor. Probing within the substructure class represented by each inhibitor is a rational first step towards lead optimization.

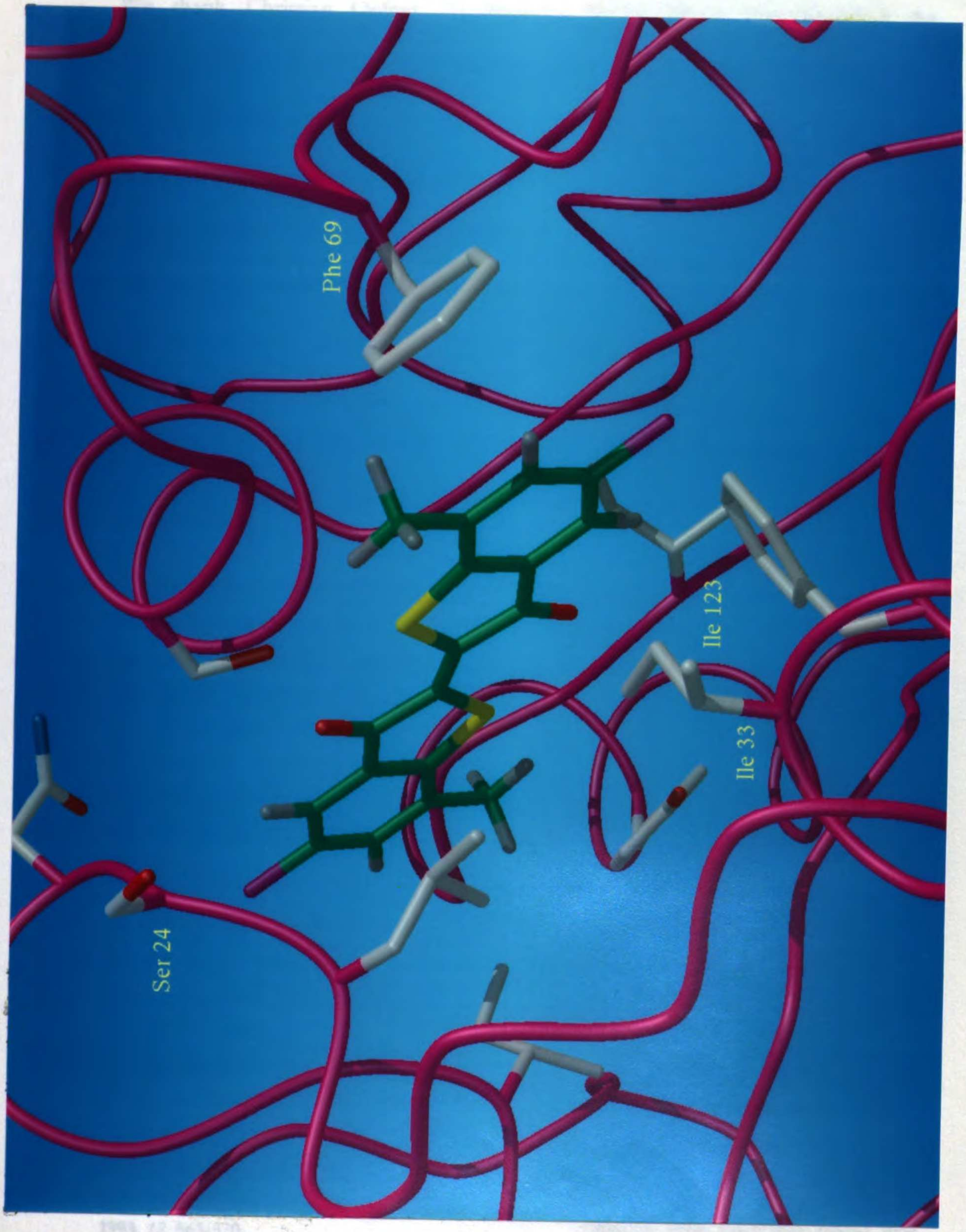
Compound **26**, on account of both its potency and selectivity, merits further exploration. From the perspective of DOCK, **26** is an ideal molecule: it is entirely rigid and possesses no formal charge. As the force-field score's lack of a conformational entropy term and overestimation of charged interactions are therefore not an issue, these attributes lend weight to the DOCK-predicted orientation for this compound. The predicted binding mode (Figure 4) entails contact with four of the six non-identical residues in the active site, offering an explanation for the 25-fold selectivity and reinforcing the plausibility of specific binding. The major hurdle for any enzyme-assay lead is the critical question of fungal uptake, which remains to be addressed in a cell-based assay. The hydrophobic nature of many of the DOCK hits (Figure 2) will be useful in this regard. For **26** in particular, compounds exhibiting structural similarity have been reported to display antifungal activity (Collier *et al.*, 1991; Klein *et al.*, 1994).

Perhaps the most effective anti-*Pneumocystis* agents will result from efforts targeting dihydropteroate synthetase (DHPS), an enzyme involved in *de novo* folate synthesis. DHPS is not present in mammalian cells, as preformed folate is acquired in the diet, and thus makes an ideal target for species-specific drug design. Sulfa drugs which target DHPS are used in therapy against *P. carinii*, yet are relatively weak inhibitors of the enzyme (Merali *et al.*, 1990; Voeller *et al.*, 1994; Hong *et al.*, 1995). The potential for the discovery of more potent agents without species-specificity issues remains substantial. To date, no structural information has been published concerning DHPS.

In the frequent absence of suitable targets unique to a pathogen, drug discovery must proceed via more stringent means by discerning similar entities. When afforded three-dimensional structural information, techniques such as molecular docking are desirable for rapid access to novel leads. We have introduced a simple methodology enhancement which expands the domain of molecular docking to encompass selective inhibition studies. A post-docking, differential refinement enables the discrimination among similar receptors in the face of few distinguishing features. In light of the success achieved in locating selective enzyme inhibitors with a post-docking methodology advance, it is logical to expect that transferring this tool directly into the docking process (Gschwend & Kuntz, unpublished results) can only amplify our ability to detect subtle features. Although there have been numerous reports of DOCK's success at identifying lead compounds in a diversity of systems (Kuntz, 1992; Gschwend *et al.*, 1995), this study represents the first attempt to locate selective inhibitors of therapeutic interest. The results further validate molecular docking as a strategy toward drug discovery, and herald favorable prospects for structure-based differential design.

Figure 4. DOCK-predicted mode of binding for 26.

(following page) Four active site residues which differ between *P. carinii* and human DHFR and which are proposed to interact with this ligand are indicated.


Phe 69
Ile 123
Ile 33
Ser 24

ACKNOWLEDGMENT

We thank Christian Oefner (Hoffman-La Roche, Basel) for providing the coordinates of *P. carinii* DHFR, and Brian Shoichet and Chris Carreras for helpful discussions. This work was funded by NIH grants GM-39552 (G. L. Kenyon, principal investigator), GM-31497 (I.D.K.), and AI-30261 (D.V.S.). D.A.G. acknowledges support from an ACS Division of Medicinal Chemistry Fellowship. The facilities of the Computer Graphics Laboratory (RR-1081 to R. Langridge) were used for part of this project. We thank Tripos Associates for supplying SYBYL , and MDL Information Systems for MACCS-3D and the FCD structural database.

REFERENCES

Allegra, C.J.; Chabner, B.A.; Tuazon, C.U.; Ogata-Arakaki, D.; Baird, B.; Drake, J.C.; Simmons, J.T.; Lack, E.E.; Shelhamer, J.H.; Balis, F.; Walker, R.; Kovacs, J.A.; Lane, H.C.; Masur, H. Trimetrexate for the treatment of *Pneumocystis carinii* pneumonia in patients with the acquired immunodeficiency syndrome. *N. Engl. J. Med.* **1987a**, *317*, 978-985.

Allegra, C.J.; Kovacs, J.A.; Drake, J.C.; Swan, J.C.; Chabner, B.A.; Masur, H. Activity of antifolates against *Pneumocystis carinii* dihydrofolate reductase and identification of a potent new agent. *J. Exp. Med.* **1987b**, *165*, 926-931.

Artymowicz, R.J.; James, V.E. Atovaquone: a new anti-*Pneumocystis* agent. *Clinical Pharmacy* **1993**, *12*, 563-570.

Bartlett, M.S.; Smith, J.W. *Pneumocystis carinii*, an opportunist in immunocompromised patients. *Clin. Microbiol. Rev.* **1991**, *4*, 137-149.

Bartlett, M.S.; Edlind, T.D.; Lee, C.H.; Dean, R.; Queener, S.F.; Shaw, M.M.; Smith, J.W. Albendazole inhibits *Pneumocystis carinii* proliferation in inoculated immunosuppressed mice. *Antimicrob. Agents Chemother.* **1994a**, *38*, 1834-1837.

Bartlett, M.S.; Queener, S.F.; Shaw, M.M.; Richardson, J.D.; Smith, J.W. *Pneumocystis carinii* is resistant to imidazole antifungal agents. *Antimicrob. Agents Chemother.* **1994b**, *38*, 1859-1861.

Beissner, R.S.; Rudolph, F.B. Interaction of Cibacron Blue 3G-A and related dyes with nucleotide-requiring enzymes. *Arch. Biochem. Biophys.* **1978,** *189,* 76-80.

Belfield, G.P.; Bauer, M.; Ross-Smith, N.; Tan, P.; Colthurst, D.R.; Tuite, M.F. EF-3: a novel fungal elongation factor with homology to *E. coli* ribosomal protein S5. *Biochem. Soc. Trans.* **1993,** *21,* 331S.

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

Blakley, R.L. *The Biochemistry of Folic Acid and Related Pteridines*; North-Holland: Amsterdam, 1969.

Blakley, R.L. Dihydrofolate Reductase. In *Folates and Pterins: Chemistry and Biochemistry of Folates*; Blakley, R.L., Benkovic, S.J., Eds.; Wiley Interscience: New York, 1984; Vol. 1, pp 191-253.

Blaney, J.M.; Hansch, C.; Silipo, C.; Vittoria, A. Structure-activity relationships of dihydrofolate reductase inhibitors. *Chem. Rev.* **1984,** *84,* 333-407.

Chambers, B.B.; Dunlap, R.B. Interaction of dihydrofolate reductase from amethopterin-resistant *Lactobacillus casei* with Cibacron blue, Blue Dextran, and Affi-Gel Blue. *J. Biol. Chem.* **1979,** *254,* 6515-6521.

Collier, P.J.; Austin, P.; Gilbert, P. Isothiazolone biocides: enzyme-inhibiting pro-drugs. *Int. J. Pharmaceutics* **1991,** *74,* 195-201.

Connolly, M.L. Analytical molecular surface calculation. *J. Appl. Cryst.* **1983a,** *16,* 548-558.

Connolly, M.L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **1983b,** *221,* 709-713.

Davies, J.F.; Delcamp, T.J.; Prendergast, N.J.; Ashford, V.A.; Freisheim, J.H.; Kraut, J. Crystal structures of recombinant human dihydrofolate reductase complexed with folate and 5-deazafolate. *Biochemistry* **1990,** *29,* 9467-9479.

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

Duch, D.S.; Edelstein, M.P.; Nichol, C.A. Inhibition of histamine-metabolizing enzymes and elevation of histamine levels in tissues by lipid-soluble anticancer folate antagonists. *Mol. Pharmacol.* **1980,** *18,* 100-104.

Dykstra, C.C.; Tidwell, R.R. Inhibition of topoisomerases from *Pneumocystis carinii* by aromatic dicationic molecules. *J. Protozool.* **1991,** *38,* 78S-81S.

Dykstra, C.C.; McClernon, D.R.; Elwell, L.P.; Tidwell, R.R. Selective inhibition of topoisomerases from *Pneumocystis carinii* compared with that of topoisomerases from mammalian cells. *Antimicrob. Agents Chemother.* **1994**, *38*, 1890-1898.

Edman, J.C.; Edman, U.; Cao, M.; Lundgren, B.; Kovacs, J.A.; Santi, D.V. Isolation and expression of the *Pneumocystis carinii* dihydrofolate reductase gene. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 8625-8629.

Edman, J.C.; Kovacs, J.A.; Masur, H.; Santi, D.V.; Elwood, H.J.; Sogin, M.L. Ribosomal RNA sequence shows *Pneumocystis carinii* to be a member of the Fungi. *Nature* **1988**, *334*, 519-522.

Falloon, J.; Kovacs, J.; Hughes, W.; O'Neill, D.; Polis, M.; Davey, R.T., Jr.; Rogers, M.; LaFon, S.; Feuerstein, I.; Lancaster, D.; Land, M.; Tuazon, C.; Dohn, M.; Greenberg, S.; Lane, H.C.; Masur, H. A preliminary evaluation of 566C80 for the treatment of *Pneumocystis* pneumonia in patients with the acquired immunodeficiency syndrome. *N. Engl. J. Med.* **1991**, *325*, 1534-1538.

Ferrin, T.E.; Huang, C.C.; Jarvis, L.E.; Langridge, R. The MIDAS display system. *J. Mol. Graphics* **1988**, *6*, 13-27.

Fishman, J.A.; Queener, S.F.; Roth, R.S.; Bartlett, M.S. Activity of topoisomerase inhibitors against *Pneumocystis carinii in vitro* and in an inoculated mouse model. *Antimicrob. Agents Chemother.* **1993**, *37*, 1543-1546.

Furlong, S.T.; Samia, J.A.; Rose, R.M.; Fishman, J.A. Phytosterols are present in *Pneumocystis carinii. Antimicrob. Agents Chemother.* **1994**, *38*, 2534-2540.

Gallant, J.E.; Moore, R.D.; Chaisson, R.E. Prophylaxis for opportunistic infections in patients with HIV infection. *Ann. Intern. Med.* **1994**, *120*, 932-944.

Gangjee, A.; Devraj, R.; McGuire, J.J.; Kisliuk, R.L.; Queener, S.F.; Barrows, L.R. Classical and nonclassical furo[2,3-d]pyrimidines as novel antifolates: synthesis and biological activities. *J. Med. Chem.* **1994**, *37*, 1169-1176.

Gordin, F.M.; Simon, G.L.; Wofsy, C.B.; Mills, J. Adverse reactions to trimethoprim-sulfamethoxazole in patients with the acquired immunodeficiency syndrome. *Ann. Intern. Med.* **1984**, *100*, 495-499.

Gschwend, D.A.; Good, A.C.; Kuntz, I.D. Molecular docking towards drug discovery. *J. Mol. Recogn.* **1995, in press.**

Gschwend, D.A. Preliminary presentation of this work: Division of Medicinal Chemistry Award Address (Paper 106), American Chemical Society Meeting, Chicago, August, 1995.

Hong, Y.L.; Hossler, P.A.; Calhoun, D.H.; Meshnick, S.R. Inhibition of recombinant *Pneumocystis carinii* dihydropteroate synthetase by sulfa drugs. *Antimicrob. Agents Chemother.* **1995**, *39*, 1756-1763.

Hillcoat, B.L.; Nixon, P.F.; Blakley, R.L. Effect of substrate decomposition on the spectrophotometric assay of dihydrofolate reductase. *J. Med. Chem.* **1967**, *21*, 178-189.

Ittarat, I.; Asawamahasakda, W.; Bartlett, M.S.; Smith, J.W.; Meshnick, S.R. Effects of atovaquone and other inhibitors on *Pneumocystis carinii* dihydroorotate dehydrogenase. *Antimicrob. Agents Chemother.* **1995**, *39*, 325-328.

Jaffe, H.S.; Abrams, D.I.; Ammann, A.J.; Lewis, B.J.; Golden, J.A. Complications of co-trimoxazole in treatment of AIDS-associated *Pneumocystis carinii* pneumonia in homosexual men. *Lancet* **1983**, *2*, 1109-1111.

Justice, A.C.; Feinstein, A.R.; Wells, C.K. A new prognostic staging system for the acquired immunodeficiency syndrome. *N. Engl. J. Med.* **1989**, *320*, 1388-1393.

Kaneshiro, E.S.; Cushion, M.T.; Walzer, P.D.; Jayasimhulu, K. Analyses of *Pneumocystis* fatty acids. *J. Protozool.* **1989**, *36*, 69S-72S.

Klein, L.L.; Yeung, C.M.; Weissing, D.E.; Lartey, P.A.; Tanaka, S.K.; Plattner, J.J.; Mulford, D.J. Synthesis and antifungal activity of 1,3,2-benzodithiazole *S*-oxides. *J. Med. Chem.* **1994**, *37*, 572-578.

Kovacs, J.A.; Masur, H. *Pneumocystis carinii* pneumonia: therapy and prophylaxis. *J. Infect. Dis.* **1988**, *158*, 254-259.

Kovacs, J.A.; Allegra, C.J.; Beaver, J.; Boarman, D.; Lewis, M.; Parrillo, J.E.; Chabner, B.; Masur, H. Characterization of de novo folate synthesis of *Pneumocystis carinii* and *Toxoplasma gondii*: potential for screening therapeutic agents. *J. Infect. Dis.* **1989**, *160*, 312-320.

Kovacs, J.A.; Hiemenz, J.R.; Macher, A.M.; Stover, D.; Murray, H.W.; Shelhamer, J.; Lane, H.C.; Urmacher, C.; Honig, C.; Longo, D.L.; Parker, M.M.; Natanson, C.; Parrillo, J.E.; Fauci, A.S.; Pizzo, P.A.; Masur, H. *Pneumocystis carinii* pneumonia: a comparison between patients with the acquired immunodeficiency syndrome and patients with other immunodeficiencies. *Ann. Intern. Med.* **1984**, *100*, 663-671.

Kraut, J.; Matthews, D.A. Dihydrofolate reductase. In *Biological Macromolecules and Assemblies*; Jurnak, F., McPherson, A., Eds.; Wiley Interscience: New York, 1984; Vol. 3, pp. 1-71.

Kuntz, I.D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078-1082.

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

Li, R.; Chen, X.; Gong, B.; Dominguez, J.N.; Davidson, E.; Kurzban, G.; Miller, R.E.; Nuzum, E.O.; Rosenthal, P.J.; McKerrow, J.H.; Kenyon, G.L.; Cohen, F.E. *In vitro* antimalarial activity of chalcones and their derivatives. *J. Med. Chem.* **1995**, in press.

Lipschik, G.Y.; Masur, H.; Kovacs, J.A. Polyamine metabolism in *Pneumocystis carinii. J. Infect. Dis.* **1991**, *163*, 1121-1127.

Margosiak, S.A.; Appleman, J.R.; Santi, D.V.; Blakley, R.L. Dihydrofolate reductase from the pathogenic fungus *Pneumocystis carinii*: catalytic properties and interaction with antifolates. *Arch. Biochem. Biophys.* **1993**, *305*, 499-508.

Masur, H. Prevention and treatment of *Pneumocystis* pneumonia. *N. Engl. J. Med.* **1992**, *327*, 1853-1860.

Meng, E.C.; Gschwend, D.A.; Blaney, J.M.; Kuntz, I.D. Orientational sampling and rigid-body minimization in molecular docking. *Proteins* **1993**, *17*, 266-278.

Meng, E.C.; Shoichet, B.K.; Kuntz, I.D. Automated docking with grid-based energy evaluation. *J. Comp. Chem.* **1992**, *13*, 505-524.

Merali, S.; Zhang, Y.; Sloan, D.; Meshnick, S. Inhibition of *Pneumocystis carinii* dihydropteroate synthetase by sulfa drugs. *Antimicrob. Agents Chemother.* **1990**, *34*, 1075-1078.

Mills, J.; Masur, H. AIDS-related infections. *Sci. Amer.* **1990**, *263*, 50-57.

Mills, J. *Pneumocystis carinii* and *Toxoplasma gondii* infections in patients with AIDS. *Rev. Infect. Dis.* **1986**, *8*, 1001-1011.

Morrison, J.F. Dihydrofolate reductase. In *A Study of Enzymes*; Kuby, S.A., Ed.; CRC Press: Boston, 1991; Vol. 2, pp. 193-226.

Murray, J.F.; Mills, J. Pulmonary infectious complications of human immunodeficiency virus infection. Part II. *Am. Rev. Respir. Dis.* **1990**, *141*, 1582-1598.

Oefner, C.; Winkler, F.; D'Arcy, A. Structure determination of *P. carinii* dihydrofolate reductase by x-ray crystallography. Unpublished results, **1991**.

Pixley, F.J.; Wakefield, A.E.; Banerji, S.; Hopkin, J.M. Mitochondrial gene sequences show fungal homology for *Pneumocystis carinii. Mol. Microbiol.* **1991**, *5*, 1347-1351.

Powles, M.A.; McFadden, D.C.; Liberator, P.A.; Anderson, J.W.; Vadas, E.B.; Meisner, D.; Schmatz, D.M. Aerosolized L-693,989 for *Pneumocystis carinii* prophylaxis in rats. *Antimicrob. Agents Chemother.* **1994**, *38*, 1397-1401.

Queener, S.F. Inhibition of *Pneumocystis* dihydrofolate reductase by analogs of pyrimethamine, methotrexate, and trimetrexate. *J. Protozool.* **1991**, *38*, 154S-157S, 1991.

Rosowsky, A.; Hynes, J.B.; Queener, S.F. Structure-activity and structure-selectivity studies on diaminoquinazolines and other inhibitors of *Pneumocystis carinii* and *Toxoplasma gondii* dihydrofolate reductase. *Antimicrob. Agents Chemother.* **1995**, *39*, 79-86.

Rosowsky, A.; Mota, C.E.; Wright, J.E.; Freisheim, J.H.; Heusner, J.J.; McCormack, J.J.; Queener, S.F. 2,4-Diaminothieno[2,3-d]pyrimidine analogues of trimetrexate and piritrexim as potential inhibitors of *Pneumocystis carinii* and *Toxoplasma gondii* dihydrofolate reductase. *J. Med. Chem.* **1993,** *36*, 3103-3112.

Rutenber, E.; Fauman E.B.; Keenan, R.J.; Fong, S.; Furth, P.S.; Ortiz de Montellano, P.R.; Meng, E.; Kuntz, I.D.; DeCamp, D.L.; Salto, R.; Rosé, J.R.; Craik, C.S.; Stroud, R.S. Structure of a non-peptide inhibitor complexed with HIV-1 protease. *J. Biol. Chem.* **1993,** *268*, 15343-15346.

Saríc, M.; Clarkson, A.B., Jr. Ornithine decarboxylase in *Pneumocystis carinii* and implications for therapy. *Antimicrob. Agents Chemother.* **1994,** *38*, 2545-2552.

Schmatz, D.M.; Romancheck, M.A.; Pittarelli, L.A.S.; R.E., Fromtling, R.A.; Nollstadt, K.H.; Vanmiddlesworth, F.L.; Wilson, K.E.; Turner, M.J. Treatment of *Pneumocystis carinii* pneumonia with 1,3-β-glucan synthesis inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **1990,** *87*, 5950-5954.

Schweitzer, B.I.; Dicker, A.P.; Bertino, J.R. Dihydrofolate reductase as a therapeutic target. *FASEB J.* **1990,** *4*, 2441-2452.

Shaw, M.M.; Kleyman, T.R.; Bartlett, M.S.; Smith, J.W. Sodium transport inhibitors decrease proliferation of *Pneumocystis carinii* in short-term culture. *J. Euk. Microbiol.* **1994,** *41*, 110S.

Shoichet, B.K.; Bodian, D.L.; Kuntz, I.D. Molecular docking using shape descriptors. *J. Comp. Chem.* **1992,** *13*, 380-397.

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

Sirawaraporn, W.S.; Edman, J.C.; Santi, D.V. Purification and properties of recombinant *Pneumocystis carinii* dihydrofolate reductase. *Protein Expr. Purif.* **1991,** *2*, 313-316.

Stringer, S.L.; Stringer, J.R.; Blase, M.A.; Walzer, P.D.; Cushion, M.T. *Pneumocystis carinii*: sequence from ribosomal RNA implies a close relationship with fungi. *Exp. Parasitol.* **1989,** *68*, 450-461.

Tidwell, R.R.; Jones, S.K.; Geratz, J.D.; Ohemeng, K.A.; Bell, C.A.; Berger, B.J.; Hall, J.E. Development of pentamidine analogues as new agents for the treatment of *Pneumocystis carinii* pneumonia. *Ann. NY Acad. Sci.* **1990,** *616*, 421-441.

U.S. Public Health Service Task Force on Anti-*Pneumocystis* Prophylaxis in Patients with HIV Infection. Recommendations for prophylaxis against *Pneumocystis carinii* pneumonia for persons infected with HIV. *J. Acquir. Immun. Defic. Syndr.* **1993,** *6*, 46-55.

Voeller, D.; Kovacs, J.; Andrawis, V.; Chu, E.; Masur, H.; Allegra, C. Interaction of *Pneumocystis carinii* dihydropteroate synthase with sulfonamides and diaminodiphenyl sulfone (Dapsone). *J. Infect. Dis.* **1994**, *169*, 456-459.

Walzer, P.D.; Foy, J.; Runck, J.; Steele, P.; White, M.; Klein, R.S.; Otter, B.A.; Sundberg, R.J. Guanylhydrazones in therapy of *Pneumocystis carinii* pneumonia in immunosuppressed rats. *Antimicrob. Agents Chemother.* **1994**, *38*, 2572-2576.

Walzer, P.D.; Kim, C.K.; Foy, J.; Linke, M.J.; Cushion, M.T. Cationic antitrypanosomal and other antimicrobial agents in the therapy of experimental *Pneumocystis carinii* pneumonia. *Antimicrob. Agents Chemother.* **1988**, *32*, 896-905.

Walzer, P.D.; Perl, D.P.; Krogstad, D.J.; Rawon, P.G.; Schultz, M.G. *Pneumocystis carinii* pneumonia in the United States. *Ann. Intern. Med.* **1974**, *80*, 83-93.

Weinberg, G.A. Iron chelators as therapeutic agents against *Pneumocystis carinii. Antimicrob. Agents. Chemother.* **1994**, *38*, 997-1003.

Wharton, J.M.; Coleman, D.L.; Wofsy, C.B.; Luce, J.; Blumenfeld, W.; Hadley, W.K.; Ingram-Drake, L.; Volberding, P.A.; Hopewell, P.C. Trimethoprim-sulfamethoxazole or pentamidine for *Pneumocystis carinii* pneumonia in the acquired immunodeficiency syndrome. *Ann. Intern. Med.* **1986**, *105*, 37-44.

Yasuoka, A.; Oka, S.; Komuro, K.; Shimizu, H.; Kitada, K.; Nakamura, Y.; Shibahara, S.; Takeuchi, T.; Kondo, S.; Shimada, K.; Kimura, S. Successful treatment of *Pneumocystis carinii* pneumonia in mice with benanomicin A (ME1451). *Antimicrob. Agents Chemother.* **1995**, *39*, 720-724.

Ypma-Wong, M.F.; Fonzi, W.A.; Sypherd, P.S. Fungus-specific translation elongation factor 3 gene present in *Pneumocystis carinii. Infect. Immun.* **1992**, *60*, 4140-4145.

Chapter 4.

Orientational Sampling and Rigid-Body Minimization in Molecular Docking, Revisited:

On-the-fly Optimization and Degeneracy Removal

Daniel A. Gschwend and Irwin D. Kuntz*

Department of Pharmaceutical Chemistry,

University of California, San Francisco, CA 94143-0446

* Author to whom correspondence should be addressed.

ABSTRACT

Strategies for computational association of molecular components entail a compromise between configurational exploration and accurate evaluation. Following the work of Meng *et al.* [*Proteins* 17 (1993) 266], we investigate issues related to sampling and optimization in molecular docking within the context of the DOCK program. An extensive analysis of diverse sampling conditions for six receptor-ligand complexes has enabled us to evaluate the tractability and utility of on-the-fly force-field score minimization, as well as the method for configurational exploration. We find that the sampling scheme in DOCK is extremely robust in its ability to produce configurations near to those which are experimentally observed. Furthermore, despite the heavy resource demands of refinement, the incorporation of a rigid-body, grid-based simplex minimizer directly into the docking process results in a docking strategy which is more efficient at retrieving experimentally observed configurations than docking in the absence of optimization. We investigate the capacity for further performance enhancement by implementing a degeneracy checking protocol aimed at circumventing redundant optimizations of geometrically similar orientations. Finally, we present methods which assist in the selection of sampling levels appropriate to desired result quality and available computational resources.

***Keywords*:** molecular recognition; configurational sampling; ligand docking; structure-based drug design

INTRODUCTION

Molecular recognition is a problem fundamental to structural biology. The interaction of molecules, be they macromolecules or small ligands, is a prerequisite for nearly all biological events. Specific modulation of these interactions has been the ambition of medicinal chemists for over a century. To gain more rapid access to therapeutic agents, we must not only understand, but be able to predict, the structural details of recognition events. The prediction of the observed orientations of two interacting components is known as the "docking problem."

There exist many computational approaches to the docking problem [1,2], but each must accomplish two principal tasks: sampling and evaluation. The task of sampling relates to the exploration of the large number of configurations varying in the relative geometry of the components. The task of evaluation refers to the ranking of each configuration by some metric. These seemingly independent phases of docking are in fact closely linked. Without an accurate evaluation scheme, the native configuration can not be recognized even when it has been sampled. Conversely, without adequate sampling, even the most accurate evaluation scheme can not recognize the native configuration if it has not been generated. The molecular docking problem in particular is further complicated by the thousands of degrees of freedom available to interacting atomic assemblies. Even when constraining the components to only six translational and rotational degrees of freedom, the docking problem is a difficult one because there are still myriads of possible configurations. Heuristics must be invoked to direct sampling and ensure computational tractability.

We previously have reported a descriptor-based rigid-body method (DOCK) to address the molecular docking problem [3-5]. More recently, an investigation into orientational sampling issues was undertaken [6]. That study presented the juxtaposition of sampling with optimization: are known binding modes retrieved more effectively with intensive sampling alone or with modest sampling and a post-docking refinement? The favorable effects of rigid-body minimization as a post-docking tool were clearly evident - steric clashes were resolved, scores were improved significantly, and experimentally observed geometries were reproduced more accurately. Unfortunately, the implementation was impractically slow. In this paper, we describe an enhancement to the minimization method, achieving nearly a 50-fold increase in speed. This accelerated rate now permits incorporation of the refinement directly into the docking process. Every configuration generated can be optimized in the context of the receptor, thus capturing the power of minimization as a post-docking scoring tool in the evaluation phase of docking. We shall also show that on-the-fly minimization improves *sampling*, further supporting the close relationship between sampling and scoring.

Despite advances in computational resources which make features such as on-the-fly optimization more palatable, the time spent in the refinement is still large when compared with the time spent sampling. If one could judiciously reduce the number of orientations actually optimized, however, the refinement bottleneck might be dissipated. We describe progress toward this goal with a technique we refer to as "degeneracy checking." Given the large number of spatially distributed descriptors and atoms involved in molecular docking, it is not surprising that there are many ways of pairing them which give rise to "similar" geometric orientations. This is obviously the result of over-sampling in certain regions. In the absence of refinement, this over-sampling provides a sort of rigid-body minimization itself. A better way to optimize local interactions is to find only one orientation per "family" (*i.e.* mode of binding) and energy-minimize that orientation, while never again paying close attention to further orientations generated in that family. By removing so-called "degenerate" configurations, many non-informative minimizations are avoided.

Following the work of Meng *et al.* [6], this paper delves further into issues related to sampling and refinement in molecular docking. We investigate the tractability and utility of on-the-fly optimization, with and without coupling to a degeneracy checking protocol. The current sampling scheme used in DOCK is evaluated in light of these data.

Table 1. Test systems.

PDB entry	Resol. (Å)	Receptor	Docked ligand	Ligand Atoms[a]	Receptor Spheres
1gst	2.2	glutathione S-transferase	glutathione	20	114
2gbp	1.9	D-galactose/D-glucose binding protein	β-D-glucose	12	75
3cpa	2.0	carboxypeptidase A	glycyl-L-tyrosine	17	44
3dfr	1.7	*L. casei* dihydrofolate reductase	methotrexate	33	72
4dfr	1.7	*E. coli* dihydrofolate reductase	2,4-diamino-6-methyl pteridine	13	86
6rsa	2.0	ribonuclease A	uridine 3'-phosphate	21	47

[a] Number of non-hydrogen ligand atoms.

METHODS

Test Systems

Six well-determined structures of ligand-receptor complexes available in the Brookhaven Protein Data Bank [7] were selected for analysis (Table 1): 1gst (glutathione S-transferase: glutathione [8]), 2gbp (D-galactose/D-glucose binding protein: β-D-glucose [9]), 3cpa (carboxypeptidase A: glycyl-L-tyrosine [10]), 3dfr (*L. casei* dihydrofolate reductase: methotrexate [11]), 4dfr (*E. coli* dihydrofolate reductase: methotrexate [11]), 6rsa (ribonuclease A: uridine vanadate [12]). The 2gbp, 3cpa, 4dfr, and 6rsa systems have been used in previous investigations of sampling [6] and scoring issues [5], as has the 3dfr system [4,13]. For reasons noted in earlier work [5], the docked ligands for the 4dfr and 6rsa systems differ from the complexed ligands; they are 2,4-diamino-6-methylpteridine and uridine 3'-phosphate, respectively. The 1gst complex has proven a difficult one to reproduce with the current site characterization, so we introduce it as a stringent test of methods.

Preparation for docking for all systems was carried out as described previously [5] - we give only an overview here. For each system, all water molecules and ions were removed and the ligand and receptor were separated. A molecular surface of the receptor binding pocket was computed with MS [14]. The program SPHGEN [3] was used to generate a negative image of the binding pocket by filling the molecular surface with overlapping spheres of varying sizes. The number of spheres generated for docking is given in Table 1. Hydrogens were added to both ligand and receptor in standard geometries. CHEMGRID [5] was used to generate the force-field scoring grid (0.30 Å resolution). DOCK force-field scores are approximate intermolecular interaction enthalpies, comprised of a 6-12 Lennard

Jones van der Waals term and a Coulombic electrostatics term. Van der Waals parameters and partial atomic charges were derived as before [5]. For electrostatic calculations, a 10.0-Å cutoff and dielectric function of $\varepsilon = 4r$ were used, where r is the interatomic distance.

Force-field Score Optimization

A rigid-body minimizer, affecting only the six intermolecular rotational and translational degrees of freedom, was incorporated directly into the DOCK scoring scheme. The simplex technique of Nelder and Mead [15] is employed, with slight modifications in the convergence treatment. Because the simplex method requires no derivatives, it lends itself to optimization on a jagged potential surface. The function that is minimized is the grid-based force-field score of Meng *et al.* [5]. One change to the standard DOCK force-field scoring van der Waals parameter file was also required, however - polar hydrogens were given a small (0.6 Å) non-zero radius. This was necessary to prevent the minimizer from taking advantage of the large electrostatic attraction that would result from a charged, volumeless hydrogen approaching an oppositely charged nucleus. Construction of the initial simplex allowed up to 1.0 Å translation and 0.5 degrees of rotation. Minimization convergence is treated in a two-stage fashion. Convergence within a simplex occurs when upper and lower bounds concur within 0.2 kcal/mol. Completion of a simplex signals a restart, initiating a new simplex. The minimization is deemed complete when a restarted simplex fails to reduce the force-field score by more than 1.0 kcal/mol. Other parameter values for simplex construction or convergence criteria resulted in slower and/or premature convergence (data not shown).

Explicit comparisons between the simplex minimizer and the quasi-Newton method published previously [6] were carried out using the stand-alone programs DOCKMIN_SIM and

DOCKMIN_DFP (distributed with DOCK 3.5) where the effect of optimization could be isolated. For each system, one DOCK run at an intermediate sampling level was chosen in which between 400 and 600 orientations were written. These output orientations were subject to stand-alone minimization. Performance was assessed for both minimization techniques in each of two modes: continuum (using exact interatomic distance calculations) and grid-based (using pre-calculated interaction scores). Stand-alone minimization was performed with default parameters. Trilinear interpolation [5] was utilized for all grid-based force-field scoring.

Degeneracy Checking

Problem Description

Degeneracy checking aims to remove geometrically similar orientations of the ligand to reduce the number of time-consuming minimizations. To be maximally efficient, such a protocol must operate without knowledge of atomic coordinates, as placement of the ligand into the context of the receptor (the "orienting" phase [2,16]) requires a significant investment of CPU resources. The removal of degenerate configurations after orienting would be much less advantageous than removal before this time-intensive step. The difficulty then lies in deciphering where in the active site an orientation lies based solely on the sphere-atom pairings involved in the match. Furthermore, the degeneracy checking algorithm must be able to perceive when the *same geometry* has been produced with *different sphere-atom pairings.* Consider the simple model depicted in Figure 1, with **E** representing the "receptor," **F** the "ligand," and circled points the atoms and spheres to be matched. Using a three-node match, one can superimpose **F** onto **E** by the pairings **b3, c4, d5**. However, the

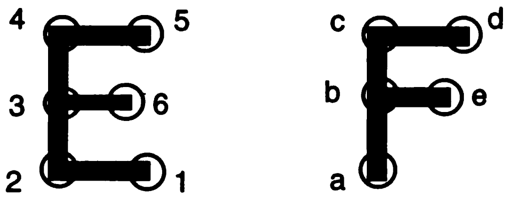


Figure 1. Hypothetical, two-dimensional degeneracy checking example.
See text. **E** represents the receptor, **F** the ligand; spheres are numbered and atoms are lettered.

matching of **a2, e6, d5** produces the identical geometric orientation. Given the latter pairing, the algorithm must recognize that this will generate an orientation degenerate with the former.

Degeneracy Check

When a unique orientation is found (*e.g.* the very first match), the new procedure records the nearest sphere to *every* atom in the ligand – this information is, of course, dependent on the orientation relative to the receptor. Every subsequent orientation must be checked for degeneracy, *i.e.* has this geometry been seen before? To avoid wasting considerable time orienting the ligand with respect to the receptor, only knowledge concerning the sphere-atom pairings involved in the match may be used for assessing degeneracy. A simple check to see if all pairings occurred simultaneously in a previous unique match imparts the answer. Note that the nearest sphere to *every* atom in the ligand for unique matches must be recorded to allow detection of similar geometries produced by different pairings.

Information Storage and Retrieval

Simply saving a list of all matches for each possible sphere-atom pairing would require allotting a tremendous amount of memory (particularly since Fortran does not support dynamic memory allocation). With the current dimensions in DOCK, this brute force approach would require 500 (*maxlig*, the maximum number of ligand atoms) by 120 (*maxpts*, the maximum number of receptor spheres) by at least 10,000 unique matches at four bytes per integer, or a 2.4Gb array. As much of this array would be empty, methods for compacting it can be devised. We choose hashing, employing open addressing with double hashing as described by Knuth [17]. A hash table allows the storage of only non-zero elements of the 3-dimensional array mentioned above, with clever methods of retrieving information given a hash code. The hash code is a function of one sphere-atom pairing and dictates where in the table matches containing this pair can be found. Thus, given one sphere-atom pairing, one can quickly retrieve all other orientations which contained this pairing.

Sensitivity Reduction: virtual spheres

All that is required for differentiating orientations is a small set of way points in the active site. Here, a way point is merely a geometric descriptor which signals the occupancy by the ligand of a particular portion of the active site volume. A typical active site will be represented by on the order of 50-100 spheres, an excess for such a simple task. Each way point describes a particular volume within the site, the size of which is generally inversely proportional to the number of way points. A ligand orientation is described by the way points its atoms "see." The more way points used, the more discerning the algorithm will be in differentiating active site volume: fewer degenerate orientations will be removed because more matches will be considered unique. The goal here is the opposite: to reduce the number of orientations passed through to the minimizer. The number of way points can be reduced easily by clustering the sphere set ("true spheres") to generate a reduced set of "virtual" spheres. We average all neighboring (within some distance *vsph*) true spheres into one virtual sphere with a single-linkage clustering algorithm. This creates an even distribution of way points throughout docking space. It is these reduced virtual spheres, rather than full set of receptor spheres, that are used only in the degeneracy checking process. The nearest virtual sphere to each point on a cubic lattice is stored for rapid access during degeneracy assessment, analogous to the utilization of a force-field scoring grid for interaction evaluation.

Degeneracy Stringency: wobble

Another method for increasing the number of degenerate orientations removed is to tolerate error in comparing the sphere-atom pairings with those in unique matches. This feature is termed "wobble" (borrowing the term from codon mismatch in protein synthesis), as a non-zero number of "mistakes" is permitted in the degeneracy check. The predictable effect of introducing wobble is to increase the number of degenerate orientations because binding modes are smeared out over a larger volume.

Safety net

Because the first orientation in a family is deemed the representative of a particular binding mode, the depiction of this binding mode is highly dependent on the quality of this orientation. All future orientations in this family will be considered degenerate to the initial member. If the quality (*e.g.* force-field score) is very poor, then this binding mode is unfairly represented. It would be beneficial to afford popular binding modes renewed chances at locating an optimal representative. The parameter *degenerate_save_interval* dictates how often a degenerate orientation must be found in a given family before orienting and minimizing another member. This feature has the desirable effect of smoothing sampling over all binding modes.

Configurational Sampling

DOCK version 3.5 was run in single mode for all docking studies. The matching algorithm for generating ligand orientations remains unchanged from that in DOCK 2.0 [4]. Because the method for orientation generation defines how configurational sampling is done, we summarize the matching algorithm. Ligand orientations are produced by matching all distances among (a minimum of) four non-hydrogen ligand atoms to complementary distances among receptor spheres. Distances are first computed relative to a seed "node", a node being any one sphere-atom pairing. All possible combinations of ligand atoms with receptor spheres are employed as seed nodes. Ligand atoms and receptor spheres are then placed into bins based upon the distance from their counterpart in the seed (first) node. As four nodes are required to form a match (a *clique*), the three bins furthest from the ligand atom of the seed node are explored. For each of these three bins, every atom in the bin is paired with every sphere in the corresponding receptor sphere bin. The second node of the growing clique is thus drawn from the first (furthest) bin. Every sphere-atom pairing from this bin results in a possible second node, as only one distance need be complementary and the complementarity of this distance is guaranteed by the bin architecture. The third and fourth nodes are sought in the second and third furthest bins from the seed node. However, with each node beyond the second, additional complementarity checks must be made to insure that new nodes are compatible with *all* existing nodes, not just the seed node. Compatibility in DOCK's matching algorithm is defined as agreement of two distances to within some tolerance.

The number of configurations (matches) generated and thus the level of sampling performed is under user control through five parameters (all in units of Ångstroms). In addition to the matching tolerance, the user controls the ligand bin size, receptor bin size, ligand bin overlap, and receptor bin overlap. Enlarging bin sizes results in a greater number of atoms or spheres per bin, and a corresponding combinatorial expansion in possible matches. The overlap parameters smooth the discrete nature of the bin architecture and increase sampling by merging portions of neighboring bins. Of the thousands of matches typically generated for a DOCK run, only a subset is written out subject to user-specified score cutoffs. Here, all orientations were examined that had negative (*i.e.* favorable) force-field scores. To insure that timing results were unbiased by slow I/O routines, coordinates for acceptable matches were never written to disk.

Performance Evaluation

To obtain a clear picture of the impact of new features related to sampling, it is vital to examine performance over a diverse array of sampling parameters. In contrast to prior investigations related to sampling issues [4-6,13], where at most a handful of different sampling levels were examined for a particular system, here we examine DOCK's ability to reproduce experimentally observed complexes over a continuum of sampling conditions. Rather than arbitrarily choose a few select combinations of bin parameters, we opt to vary two sampling parameters independently in discrete increments over a large range. Our method is as follows. The two parameters to be varied are the bin size and bin overlap. We set both the ligand bin size and receptor bin size equal to the variable bin size. Similarly, we set both the ligand bin overlap and receptor bin overlap equal to the variable bin overlap. Finally, we set the matching distance tolerance to be equal to the *sum* of the bin size and the bin overlap. The dependence of the distance tolerance on the bin parameters insures that all distance compatibility assessments for growing cliques are made with similar stringency.

Bin sizes and bin overlaps ranged in increments of 0.05 Å from 0.05 Å to between 0.40 Å and 1.00 Å. In general, bin parameters were no longer incremented when runtimes began to exceed several minutes. This protocol led to a few hundred individual single mode DOCK runs per system, enabling a statistically significant analysis of result quality versus CPU time. For each DOCK run, a record was kept of the number of matches attempted, the best force-field score obtained, the root-mean-square (rms) deviation of the orientation having the best force-field score to the experimentally observed orientation (hydrogens were not included), and the amount of CPU time invested. All acceptable matches were formed from exactly four nodes and tolerated no more than two bad contacts.

For evaluation of new technology, three sets of runs as described above were performed for each system: once using traditional DOCK without new features, once with on-the-fly force-field score minimization, and once with on-the-fly force-field score minimization coupled to degeneracy removal. Data were transformed into a success- *versus*-effort format as follows. Effort was quantified in two ways: by the number of matches attempted, and by the amount of CPU time required. Success was also measured in two ways: by whether the rms deviation of the best force-field scoring orientation was within 1.0 Å of the observed mode, and by whether the best force-field score obtained was within some cutoff (typically 5 kcal/mol) about the global minimum. The global minimum force-field score was taken as the best force-field score seen by *any* of the DOCK runs for that system. Thus, this extremum represents the best among no fewer than several million configurations. Effort is binned on a logarithmic scale: within each effort bin, a probability of success was computed by dividing the number of successful DOCK runs in the bin into the total number of DOCK runs falling in the bin. A seven-point moving average was used to smooth plots.

Sampling Robustness

To assess whether failure by DOCK to reproduce experimentally observed geometries generally results from deficiencies in sampling or in scoring, we isolated the effects from sampling. By removing scoring restrictions and analyzing only agreement in Cartesian space between docked orientations and the observed binding mode, the precision of the sampling algorithm is revealed. A set of DOCK runs with sampling level varied as described above was thus performed in which all orientations within 2.5 Å rms deviation from the experimentally observed configuration were written out, regardless of force-field score.

Hardware

All calculations were carried out on a Silicon Graphics 200MHz R4400 Indigo2 workstation (Silicon Graphics, Inc., Mountain View, CA) with 128Mb of physical memory.

RESULTS

Sampling Robustness

The ability of DOCK's sampling algorithm to locate the experimentally observed binding mode is illustrated in Figure 2. For each system, those sampling levels (indicated by the number of matches attempted) which produced an orientation within 2.5 Å rms deviation, regardless of score, are plotted. It can be seen that for all receptor-ligand complexes explored here, the matching algorithm is robust enough to find the native configuration. With the exception of the 1gst system (Figure 2a), a few hundred to a thousand matches are sufficient to locate an orientation within 1.0 Å rms deviation. This point highlights the robust nature of the sphere description and matching algorithm used in DOCK. Having demonstrated that the sampling method is adequate, it thus becomes a task for scoring schemes to recover the native mode as the optimal configuration.

Figure 2, a-f. Rms deviation vs. matches tried.

(following pages) The best rms deviation to the experimentally observed configuration seen, regardless of force-field score is plotted as a function of number of matches attempted. Each point represents a single DOCK run with distinct sampling parameters.

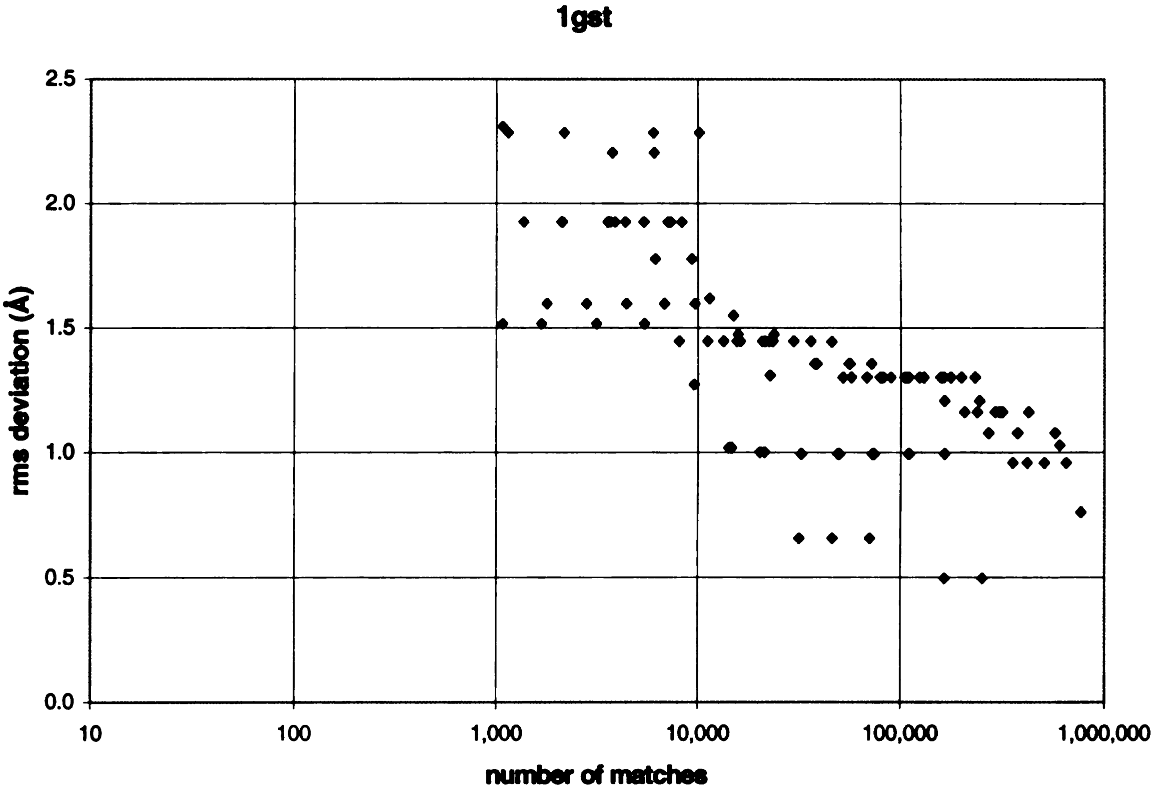


Figure 2a.

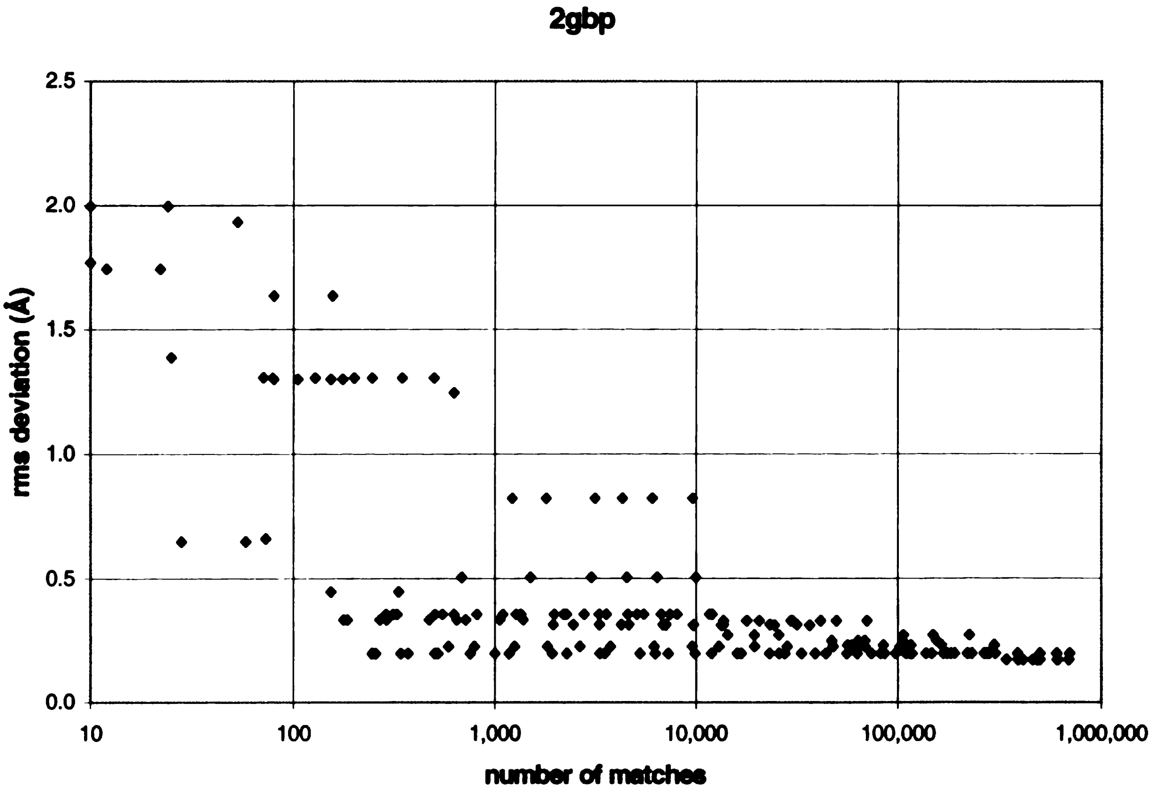


Figure 2b.

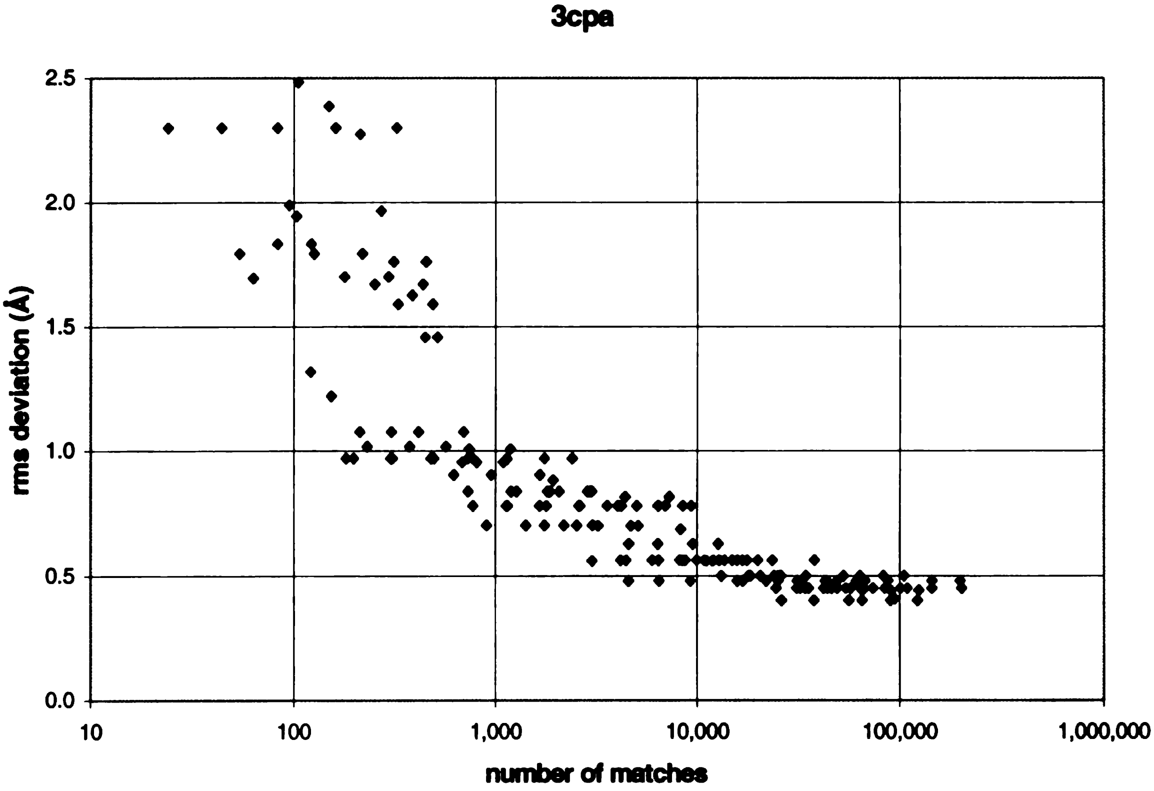


Figure 2c.

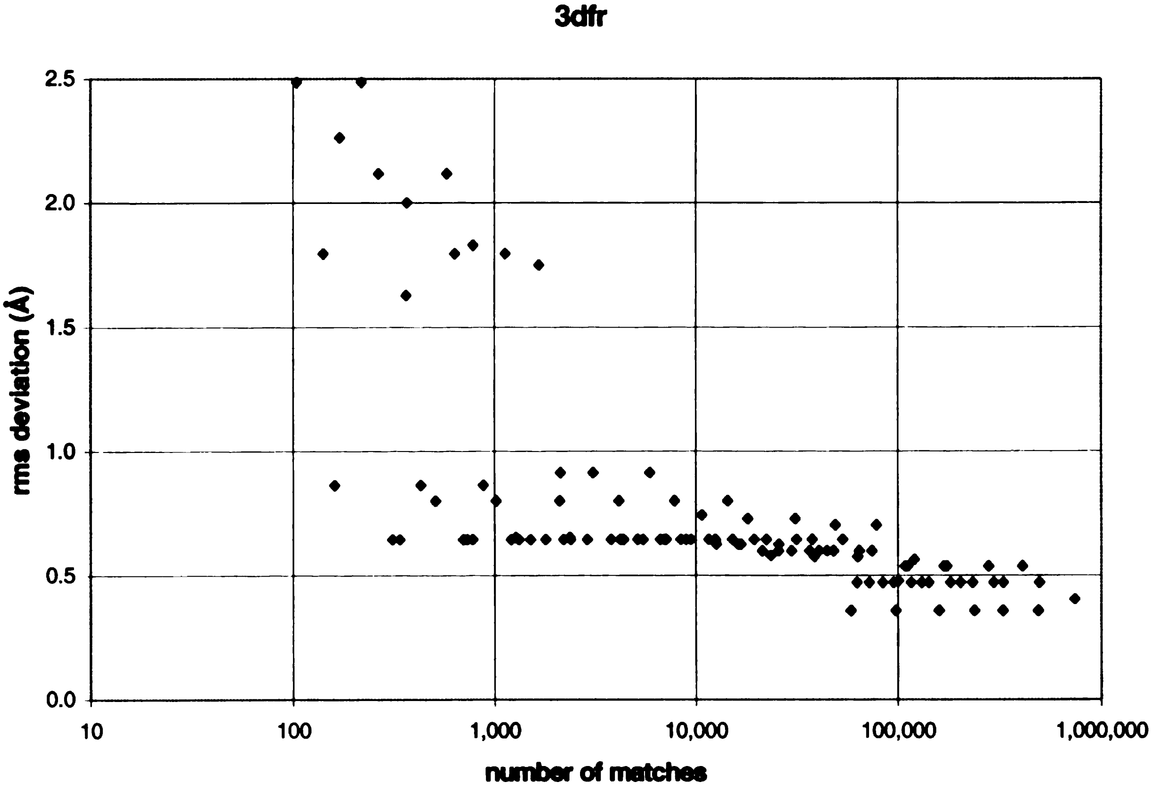


Figure 2d.

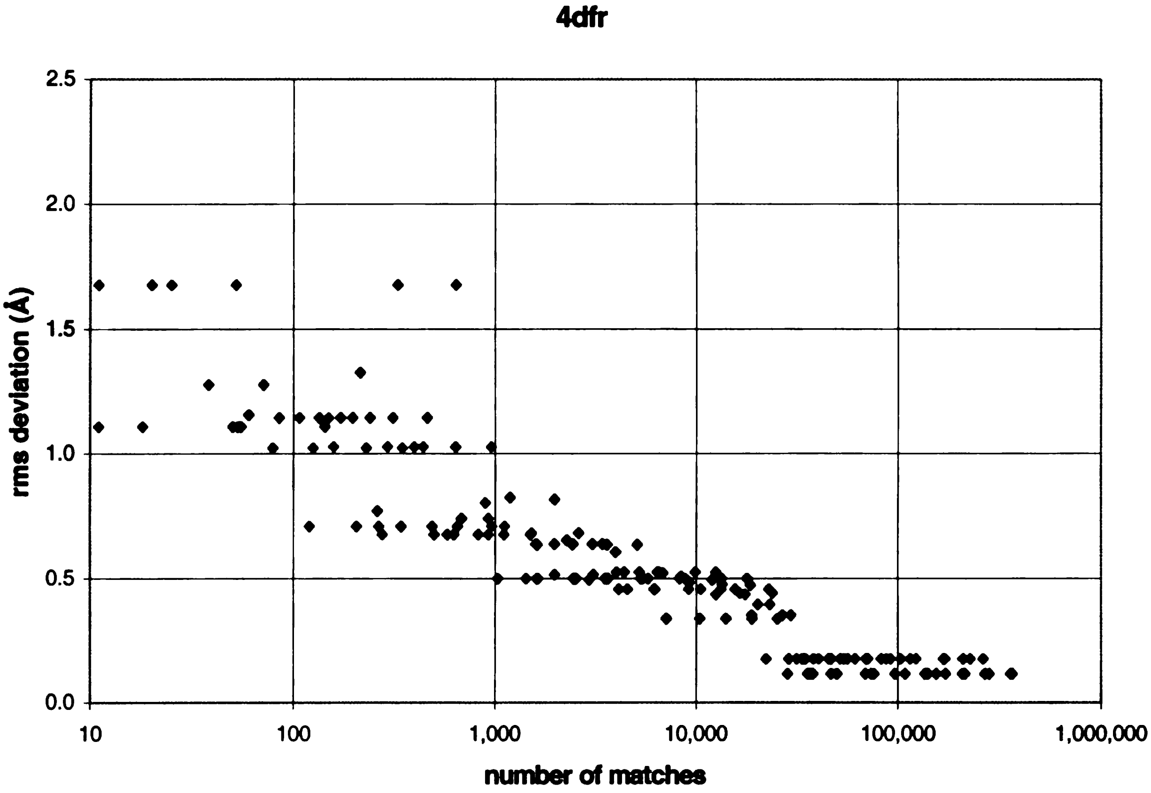


Figure 2e.

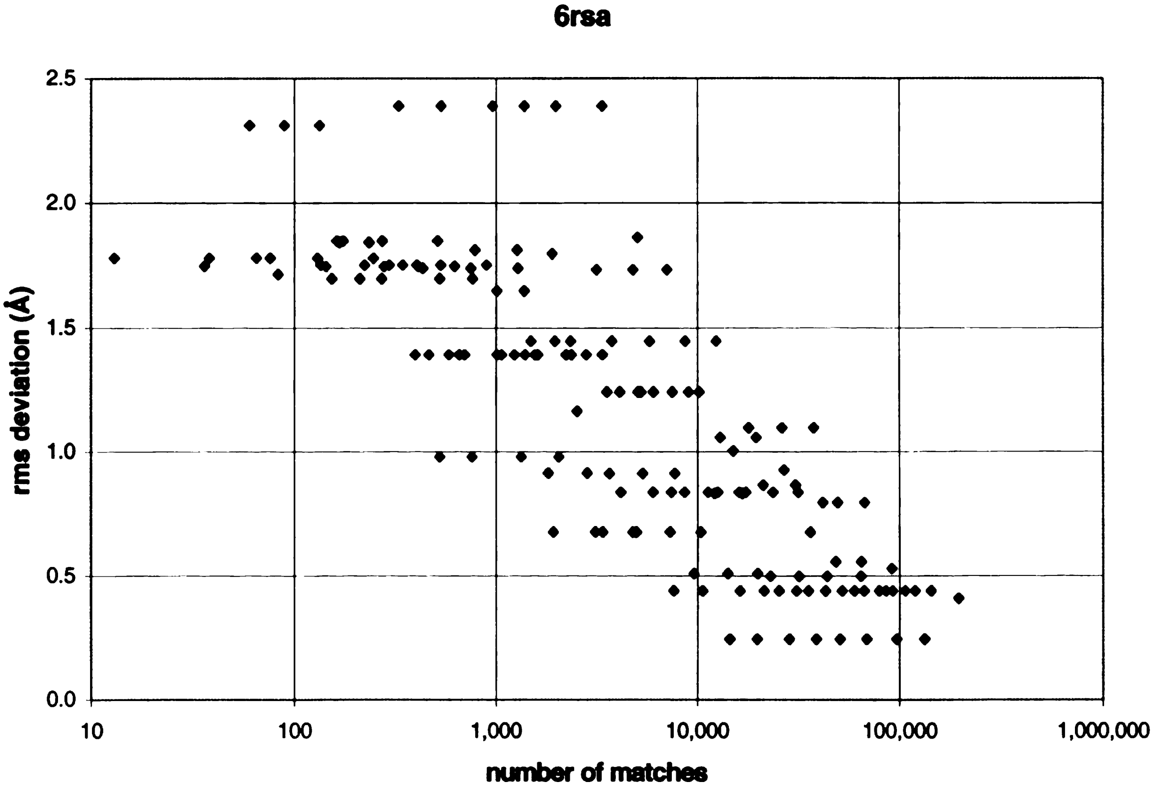


Figure 2f.

Minimizer Performance

To verify that a fast rigid-body optimization suitable for incorporation into DOCK could operate as effectively as the more resource-intensive method explored by Meng *et al.* [6], we compared the ability of minimization techniques to refine pre-existing DOCK output. Table 2 juxtaposes the performance of the grid-based simplex minimizer with that of the continuum-mode quasi-Newton Davidon-Fletcher-Powell (DFP) [18] method described previously [6]. Approximately 500 orientations obtained from an intermediate sampling level DOCK run for each system were subject to post-DOCK optimization. Between 30- and 75-fold faster operation is achieved by implementing the simplex using pre-calculated interaction scores on a lattice. The near-unit slopes and reasonably high correlation coefficients between the optimized scores indicate result quality is both balanced and comparable. The offset favoring the continuum DFP by 1-2 kcal/mol is attributable to the use of exact interatomic distances rather than trilinear interpolation among pre-calculated grid scores. Convergence radii for the two minimization techniques are of similar magnitudes. We take as a measure of convergence radius, or the capacity to pull distant structures into a local minimum, the rms deviation occurring during minimization. The simplex operating in continuum mode and the grid-based DFP demonstrated performance intermediate to the two methods presented in Table 2 (data not shown).

Table 2. Performance comparison of minimization methods.

System	CPU time per ligand (sec.)		rms deviation (Å)[a]		Correlation[b]
	continuum DFP	grid simplex	continuum DFP	grid simplex	
1gst	3.10	0.070	0.00 to 3.77	0.10 to 5.31	$r^2 = 0.77$
			1.04 ± 0.63	1.09 ± 0.63	$y = 0.88x - 2.30$
2gbp	1.47	0.039	0.00 to 3.26	0.11 to 2.70	$r^2 = 0.80$
			0.76 ± 0.46	0.81 ± 0.44	$y = 0.97x - 0.83$
3cpa	2.93	0.062	0.00 to 3.93	0.00 to 5.78	$r^2 = 0.86$
			0.91 ± 0.67	0.95 ± 0.72	$y = 1.01x - 0.76$
3dfr	3.92	0.115	0.00 to 4.77	0.07 to 3.57	$r^2 = 0.95$
			1.14 ± 0.76	1.09 ± 0.65	$y = 1.01x - 1.03$
4dfr	2.72	0.037	0.00 to 1.95	0.00 to 2.10	$r^2 = 0.87$
			0.50 ± 0.29	0.49 ± 0.28	$y = 1.00x - 1.30$
6rsa	2.02	0.064	0.00 to 5.14	0.00 to 6.56	$r^2 = 0.88$
			1.37 ± 0.98	1.39 ± 0.96	$y = 0.99x - 1.28$

[a] rms deviation from starting position is given as minimum and maximum values, followed by average ± standard deviation; hydrogens were not included in calculations. Values represent minimization of approximately 500 DOCK output orientations for each system.

[b] Correlations of continuum DFP force-field scores (y) *versus* grid simplex force-field scores (x).

On-the-fly Optimization and Degeneracy Checking

The performance impact of on-the-fly force-field score optimization and degeneracy checking was gauged via success-*versus*-effort analyses in a three-stage approach. A set of runs with variable sampling parameters was performed for DOCK in "native" mode (without any new technology), for DOCK with force-field score minimization, and for DOCK with force-field score minimization coupled to the degeneracy checking protocol. The range of sampling parameters, number of DOCK runs, and total configurations generated for each set are enumerated in Table 3. A grand total in excess of 2,500 DOCK runs covering a wide range of sampling conditions has allowed a comprehensive analysis of tradeoffs between configurational exploration and rigid-body optimization.

[Footnotes to Table 3]

[a] "native" refers to DOCK runs in which neither force-field score minimization nor degeneracy checking was used. "min" refers to DOCK runs in which force-field score minimization was used without degeneracy checking. "min+deg" refers to DOCK runs in which force-field score minimization was used in conjunction with degeneracy checking.

[b] Increments of 0.05 Å were used within these ranges.

[c] The number of DOCK runs examined is in some cases less than the bin ranges would indicate for three possible reasons: runtimes began to exceed several minutes, convergence at 100% in the success-versus-effort plots had been reached, or the maximum number of allowable unique matches for degeneracy checking had been exceeded.

[d] Minimum force-field score (kcal/mol) observed over all DOCK runs for each system.

[e] For bin sizes of 0.55 to 1.00 in the 3cpa native DOCK runs, bin overlaps ranged only from 0.55 to 1.00, hence only 300 runs resulted. This was an effort to obtain more high-sampling runs.

[f] Degeneracy parameters: *wobble* = 2, *vsph* = 1.5, *degenerate_save_interval* = 10.

[g] Degeneracy parameters: *wobble* = 2, *vsph* = 2.0, *degenerate_save_interval* = 25.

Table 3. Sampling conditions explored in methodology evaluation.

System	DOCK features[a]	Bin size range (Å)[b]	Bin overlap range (Å)[b]	DOCK runs[c]	total # matches	matches/ second	global minimum[d]
1gst	native	0.05 - 0.50	0.05 - 1.00	200	18,833,108	2647	-49.037
	min	0.05 - 0.40	0.05 - 0.80	128	2,421,887	83	
	min+deg[f]	0.05 - 0.50	0.05 - 1.00	128	1,783,514	626	
2gbp	native	0.05 - 0.50	0.05 - 1.00	200	14,270,188	4187	-24.538
	min	0.05 - 0.50	0.05 - 0.50	100	306,144	52	
	min+deg[f]	0.05 - 0.50	0.05 - 1.00	126	956,378	281	
3cpa	native	0.05 - 1.00	0.05 - 1.00	300[e]	28,582,470	2566	-47.188
	min	0.05 - 0.50	0.05 - 0.50	100	117,713	78	
	min+deg[f]	0.05 - 0.50	0.05 - 1.00	162	1,597,427	626	
3dfr	native	0.05 - 0.40	0.05 - 0.80	128	7,136,487	2863	-70.945
	min	0.05 - 0.40	0.05 - 0.40	64	125,421	326	
	min+deg[g]	0.05 - 0.40	0.05 - 0.80	111	2,616,774	1882	
4dfr	native	0.05 - 0.50	0.05 - 1.00	200	6,207,365	2354	-33.916
	min	0.05 - 0.50	0.05 - 0.50	100	180,282	36	
	min+deg[f]	0.05 - 0.50	0.05 - 1.00	152	1,951,826	293	
6rsa	native	0.05 - 0.50	0.05 - 1.00	200	2,834,980	1731	-66.003
	min	0.05 - 0.50	0.05 - 0.50	100	68,953	68	
	min+deg[f]	0.05 - 0.50	0.05 - 1.00	171	1,493,590	596	

Table 3 footnotes are given on previous page.

Figure 3 illustrates, for each of the six receptor-ligand systems, the probability of locating an orientation with a force-field score within 5 kcal/mol of the global minimum as a function of the number of matches attempted. It is readily apparent that the use of force-field score minimization consistently outperforms native DOCK in this respect. This is to be expected: both methods generate the identical orientations but the former is afforded an optimization of intermolecular interactions, an operation which can only improve results. In the limit of ideality, coupling to a degeneracy checking protocol would show identical behavior to minimization alone, when effort is measured by number of matches. This is because on a match-per-match basis, force-field score optimization on its own defines the maximal envelope of result quality. In actuality, we see that our degeneracy checking method, although in some cases (3cpa, 6rsa) reasonably close to the outer envelope, generally falls intermediate to DOCK with and without minimization. Note that in two systems examined here (1gst, 3dfr), native DOCK is completely unable to locate an orientation close to the global minimum in the absence of refinement, even when sampling on the order of one million configurations. Plots of success in placing the best force-field scoring orientation within 1 Å rms deviation of the experimentally observed configuration as a function of number of matches tried parallel nearly identically the force-field score success plots in Figure 3 (data not shown).

Figure 3, a-f. Force-field score success vs. matches tried.

(following pages) The probability of locating an orientation having a force-field score within 5 kcal/mol of the global minimum is plotted as a function of number of matches attempted. "dock" represents native DOCK, "min" represents DOCK with on-the-fly minimization, "mindeg" represents DOCK with on-the-fly minimization and degeneracy checking. The absence of a curve for native DOCK in some systems indicates that no successful run ever occurred.

Figure 4, a-f. Force-field score success vs. CPU time.

(following pages) The probability of locating an orientation having a force-field score within 5 kcal/mol of the global minimum is plotted as a function of CPU seconds required. The key is as given in the legend for Figure 3.

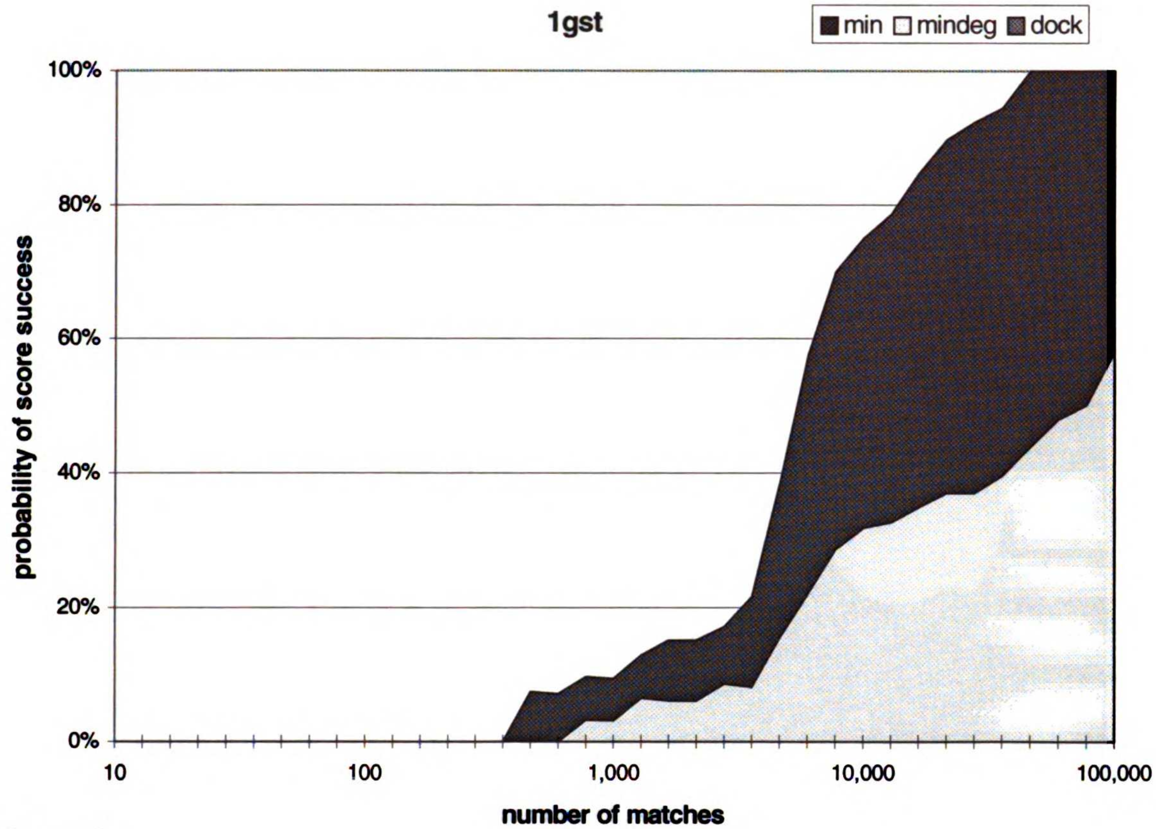


Figure 3a.

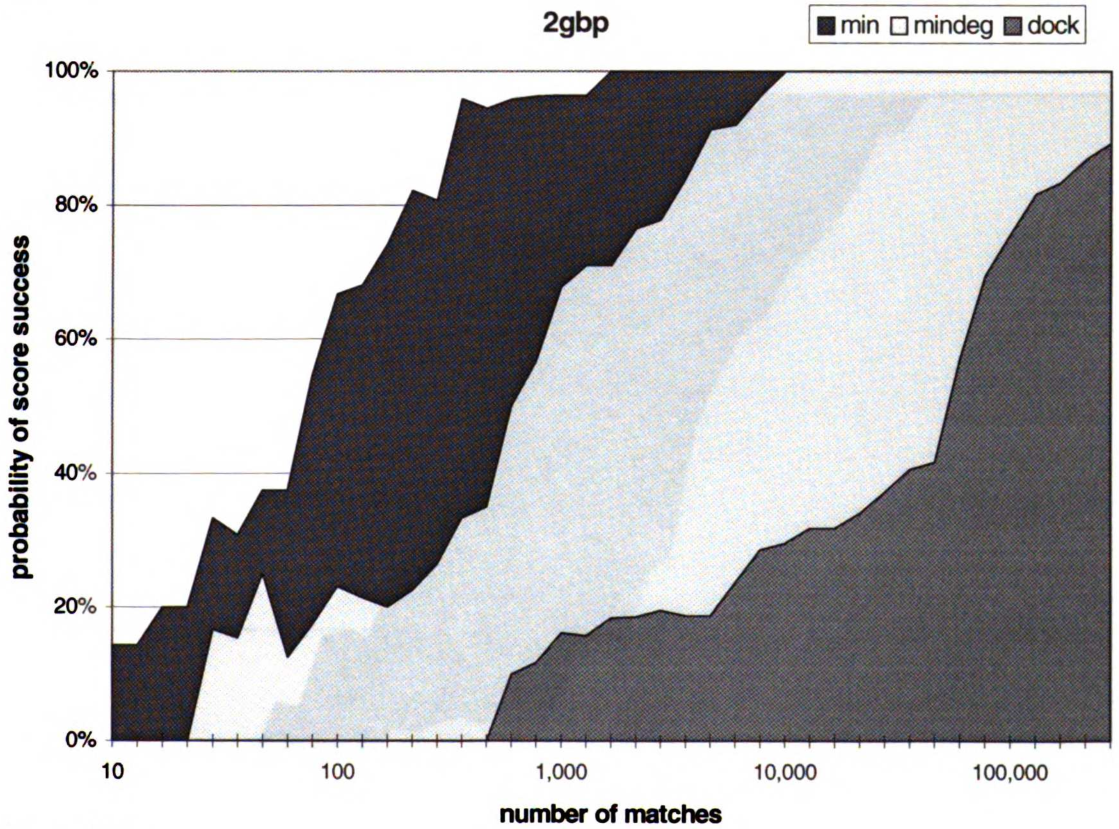


Figure 3b.

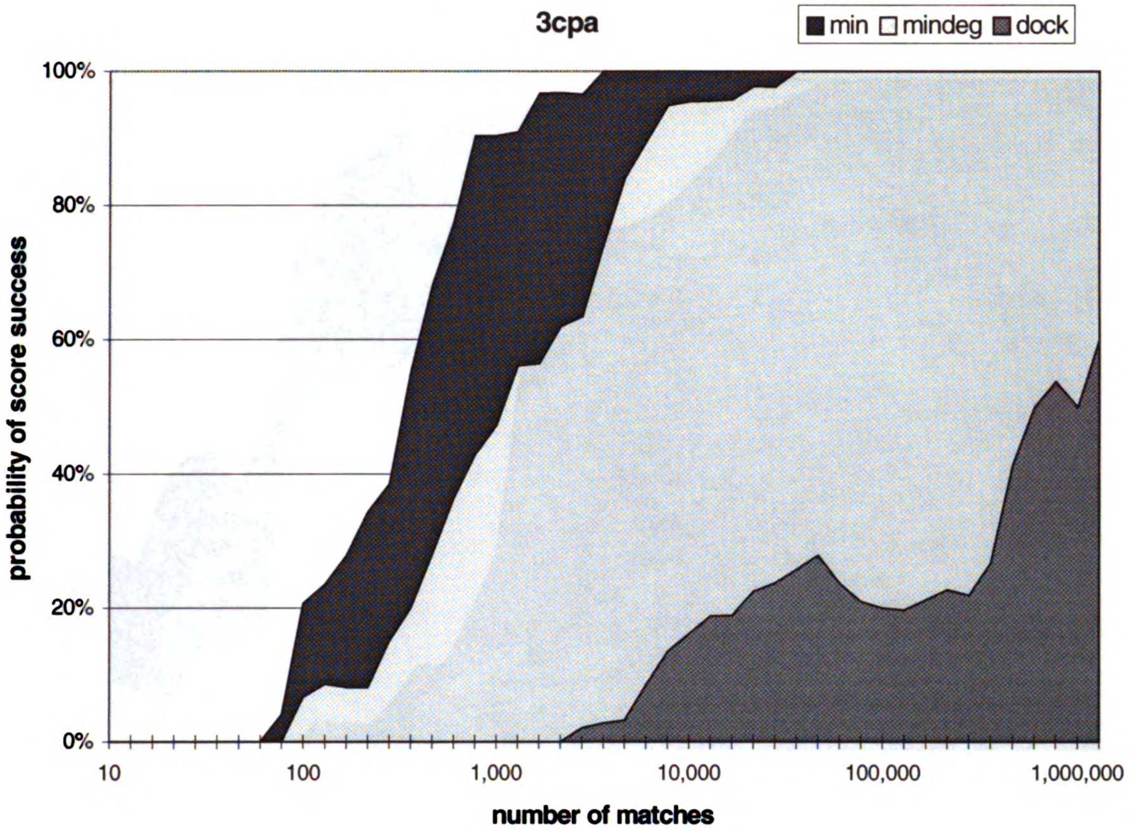


Figure 3c.

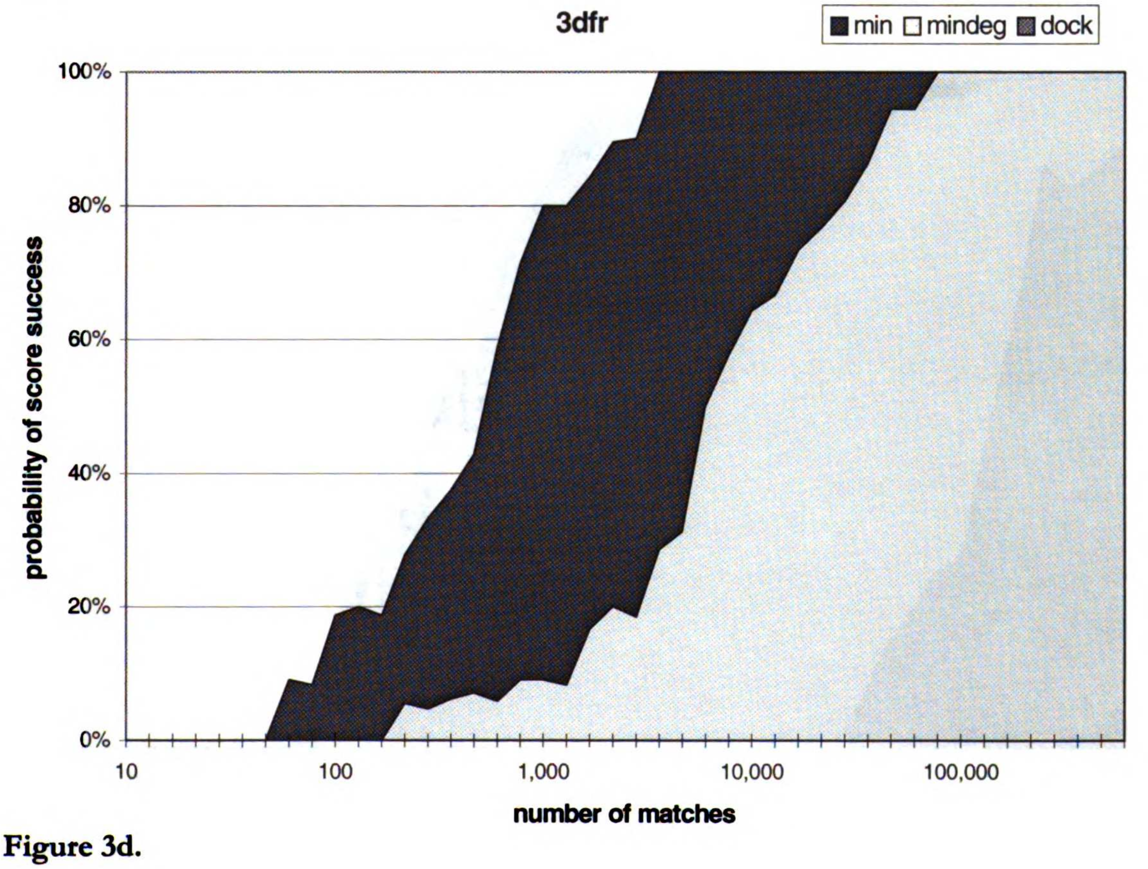


Figure 3d.

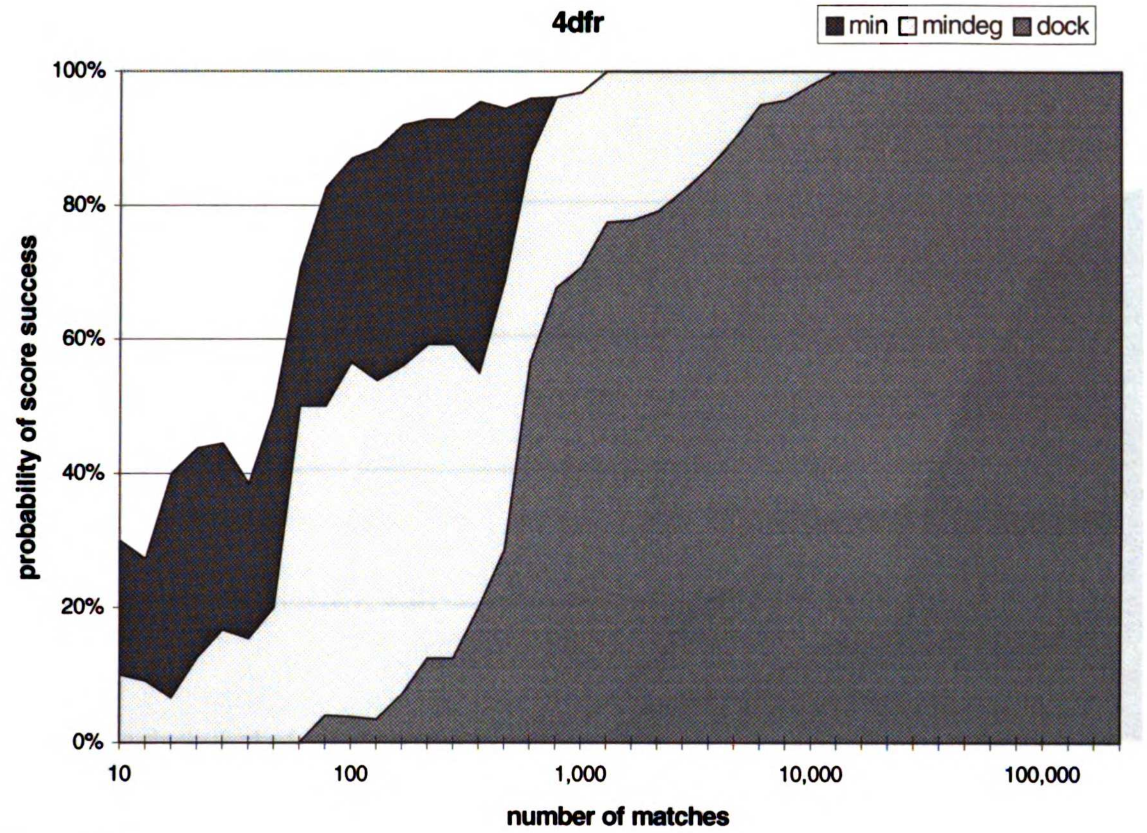


Figure 3e.

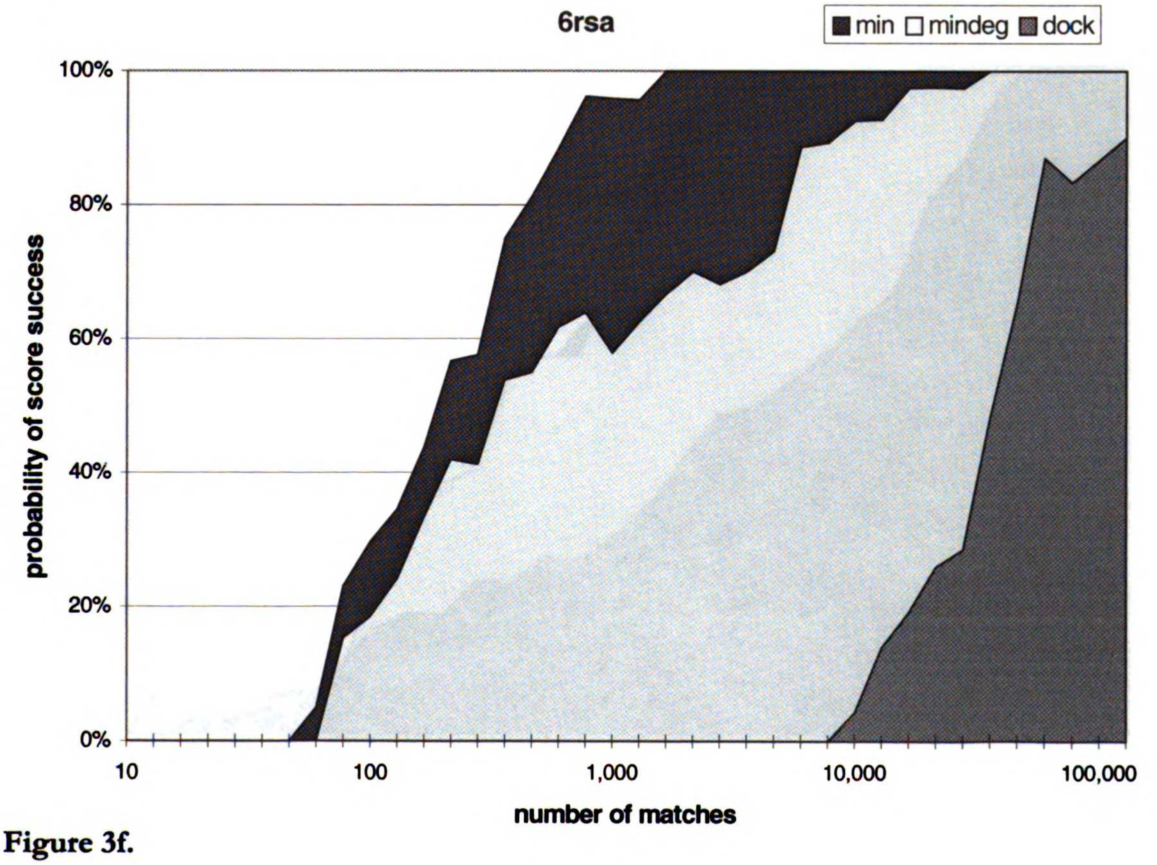


Figure 3f.

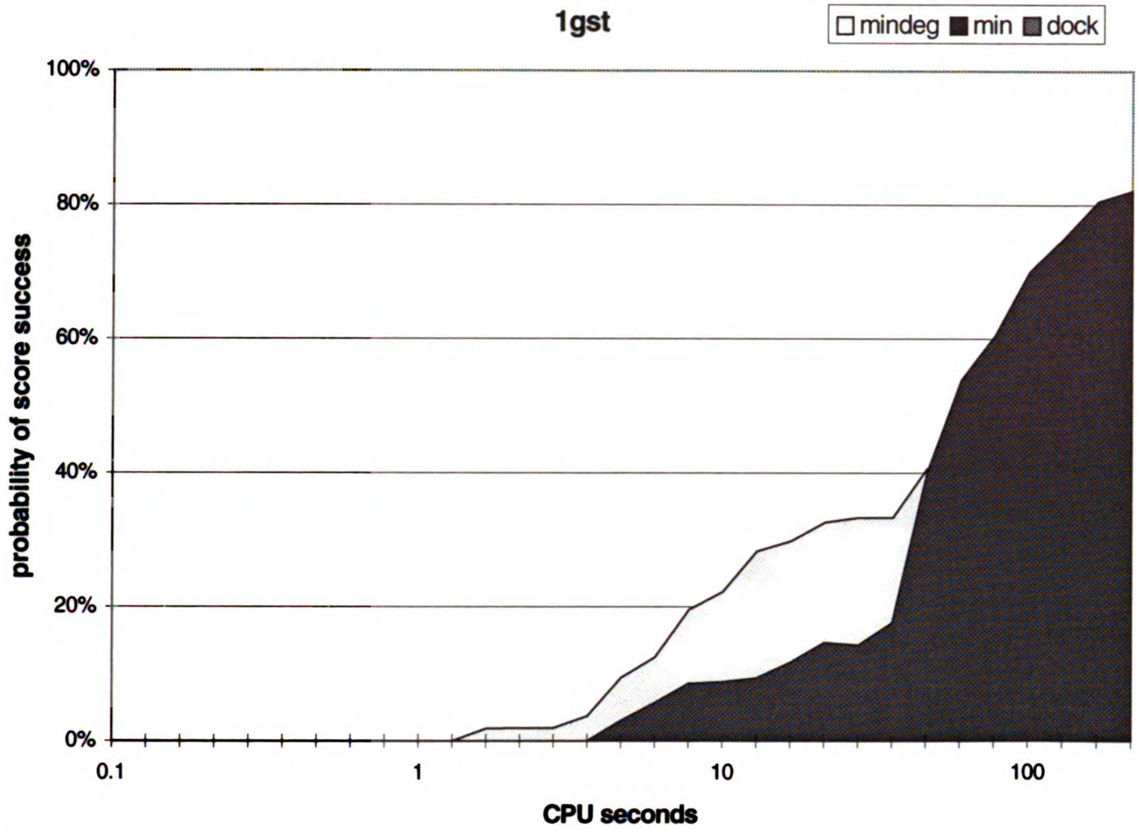


Figure 4a.

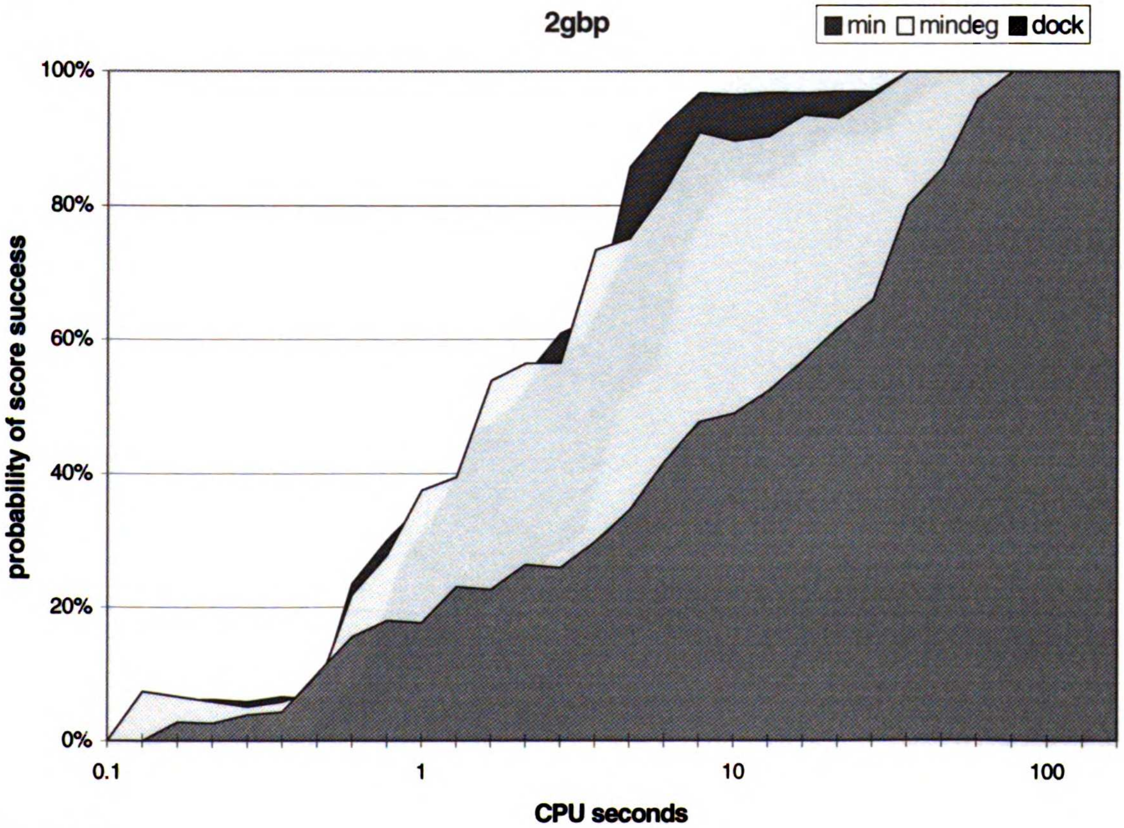


Figure 4b.

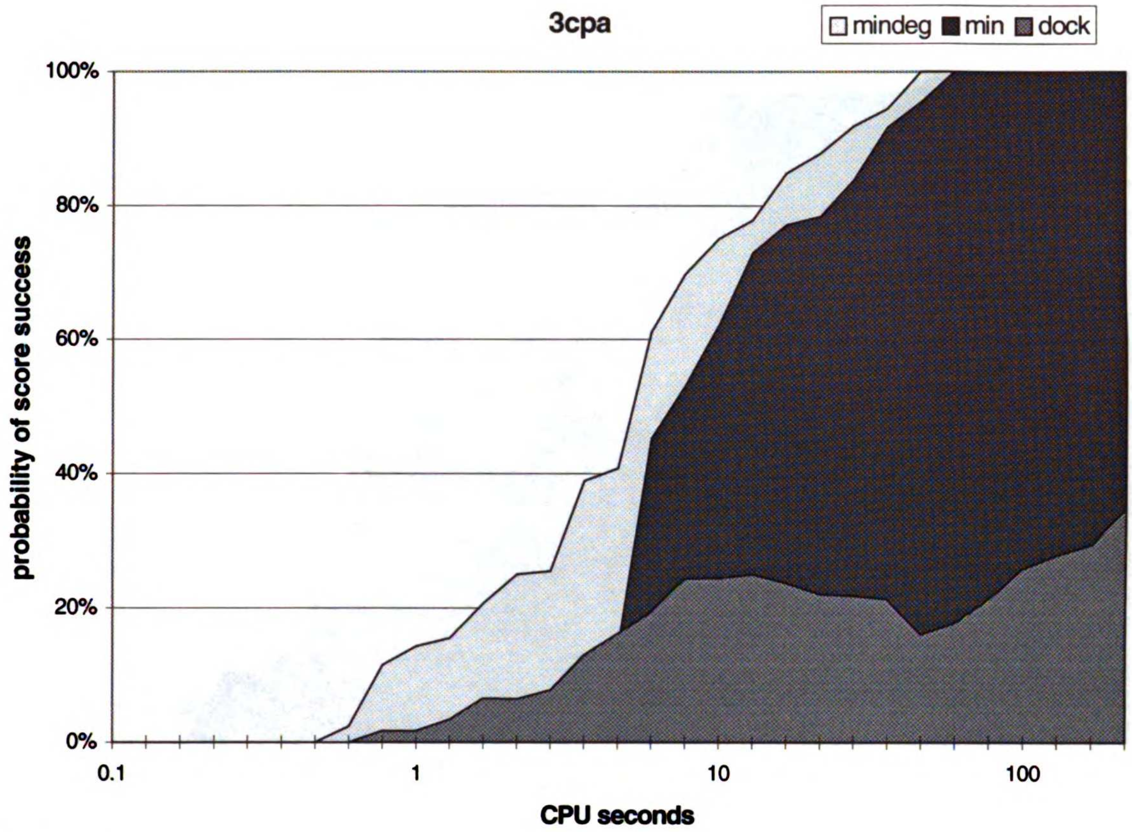


Figure 4c.

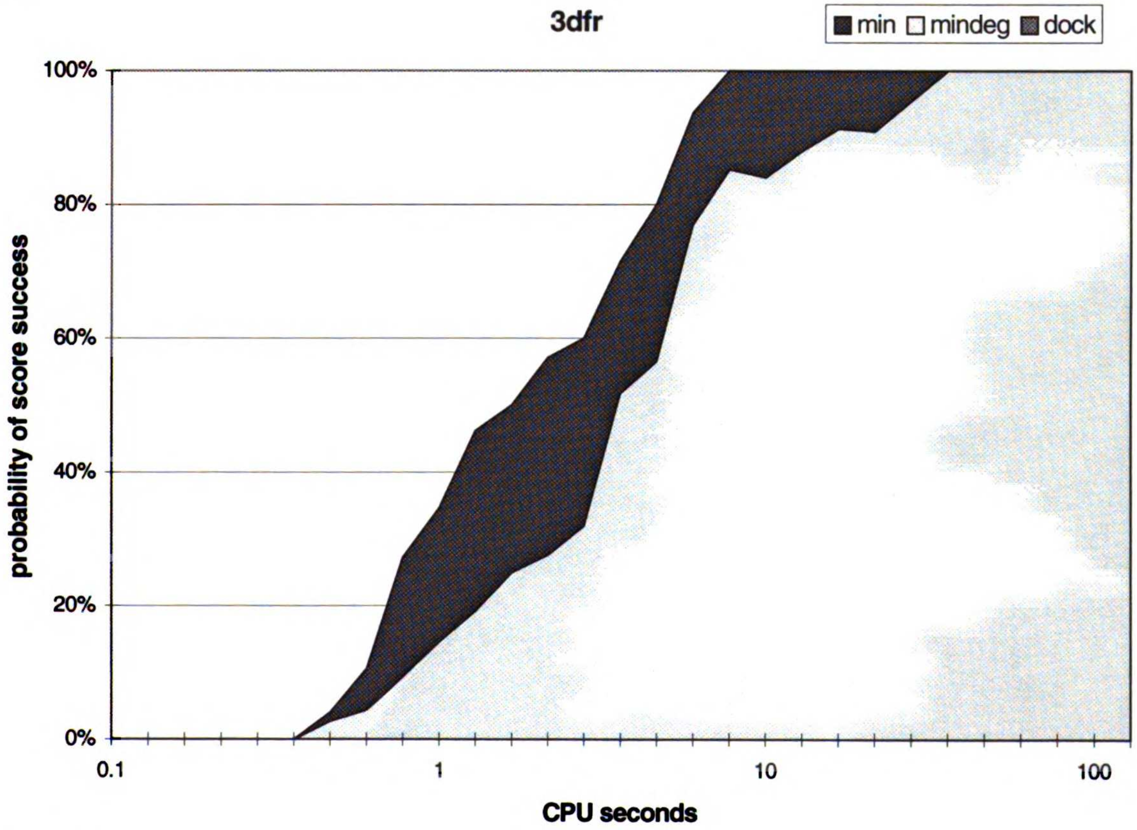


Figure 4d.

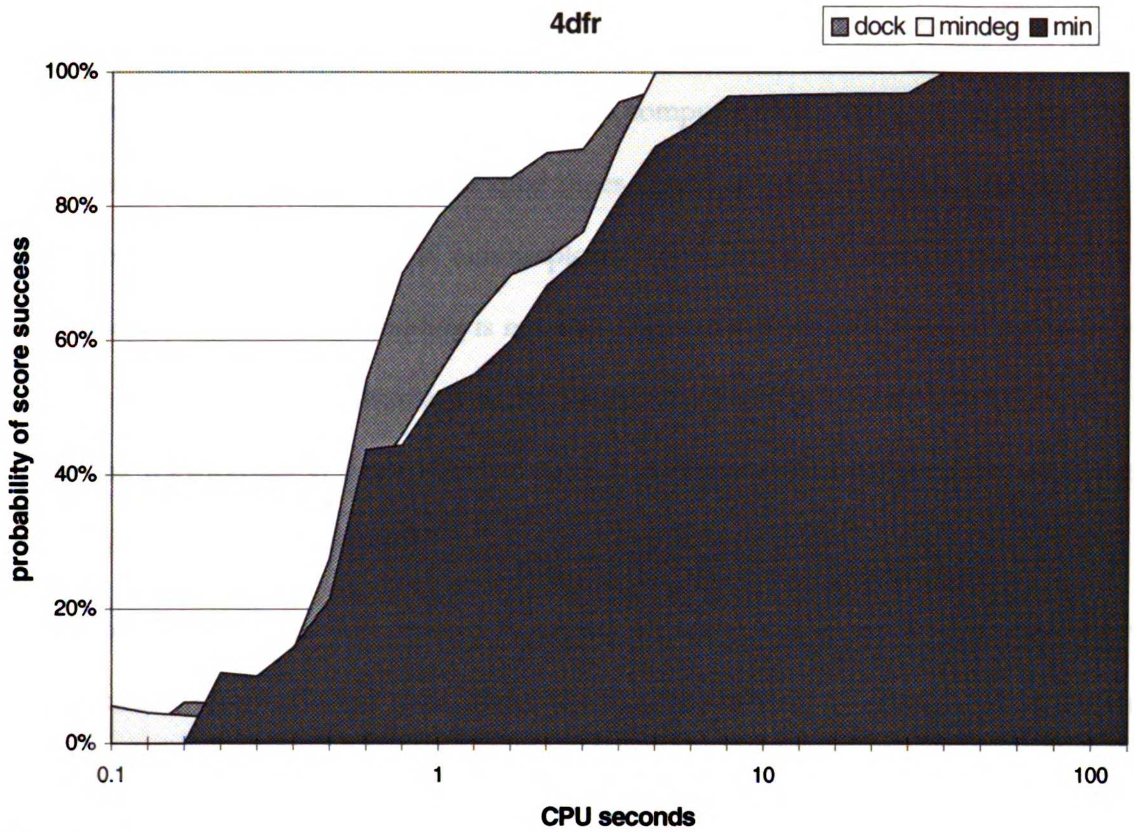


Figure 4e.

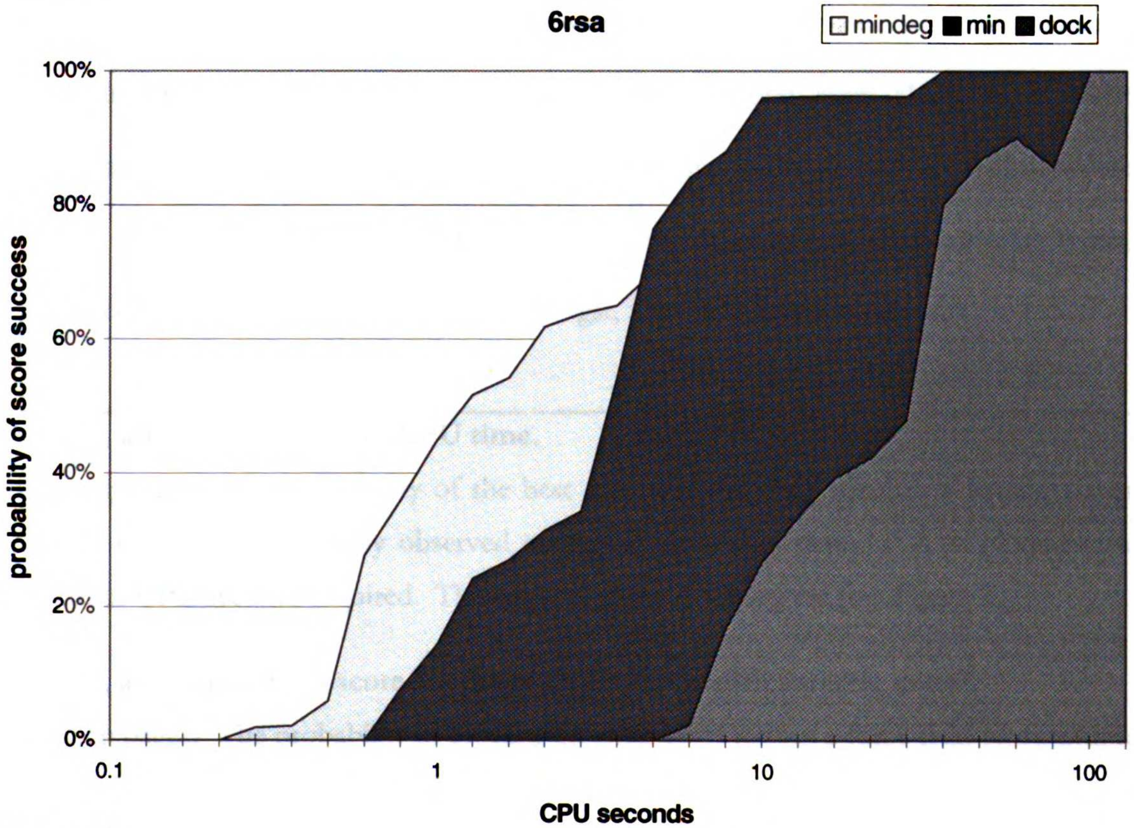


Figure 4f.

In practice, however, the primary concern for molecular docking is not how many configurations are examined, but rather how much computer time is required. Because each optimization takes on average one hundred times longer to carry out than a single score evaluation (data not shown), DOCK runs employing force-field score minimization are likely to become intractable unless sampling is reduced. But can sampling be reduced sufficiently to counteract this great disadvantage while maintaining high-quality solutions? Figure 4 depicts the transformation from effort measured in numbers of configurations to effort gauged by computational demands.

Excepting only the 4dfr system (Figure 4e), we see that using on-the-fly optimization is dramatically more efficient than native DOCK at arriving at near-global-minimum solutions, despite the much higher per-match resource requirements (Table 3). The implementation of the degeneracy checking protocol, while equally superior to native DOCK, does not display as dramatic improvements when compared with minimization alone. In one case (6rsa) we see significant gains, in two cases (1gst, 3cpa) slight improvements, in two cases (2gbp, 4dfr) no difference, and in one case (3dfr) slightly worse behavior. Degeneracy checking generally manifests its advantages at lower sampling levels, as evidenced by the early successes seen in the 1gst, 3cpa, and 6rsa complexes.

Figure 5, a-f. Rms success vs CPU time.
(following pages) The probability of the best force-field-scoring orientation having a rms deviation to the experimentally observed configuration of less than 1.0 Å is plotted as a function of CPU seconds required. The key is as given in the legend for Figure 3.

Figure 6, a-f. Force-field score success vs. CPU time with variable cutoff.
(following pages) The probability of locating an orientation having a force-field score within a variable cutoff of the global minimum is plotted as a function of CPU seconds required. The cutoff in kcal/mol is given in the key. Curves apply to native DOCK only.

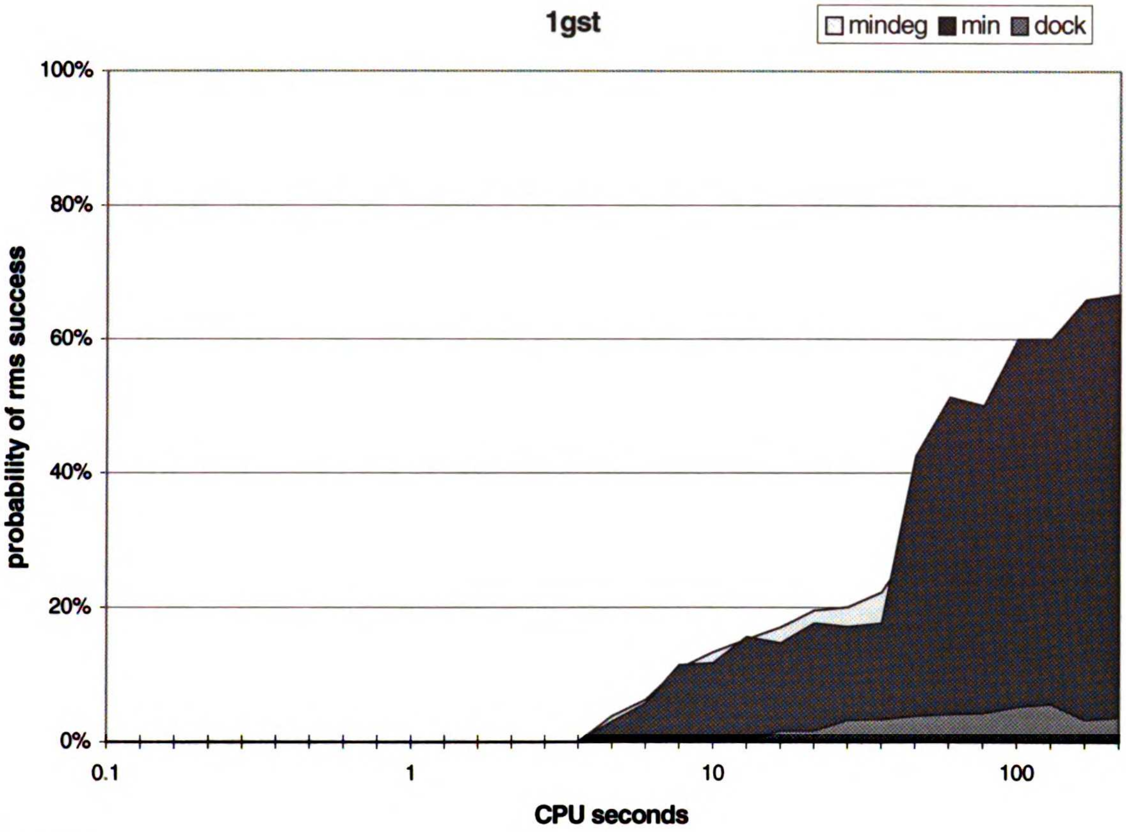


Figure 5a.

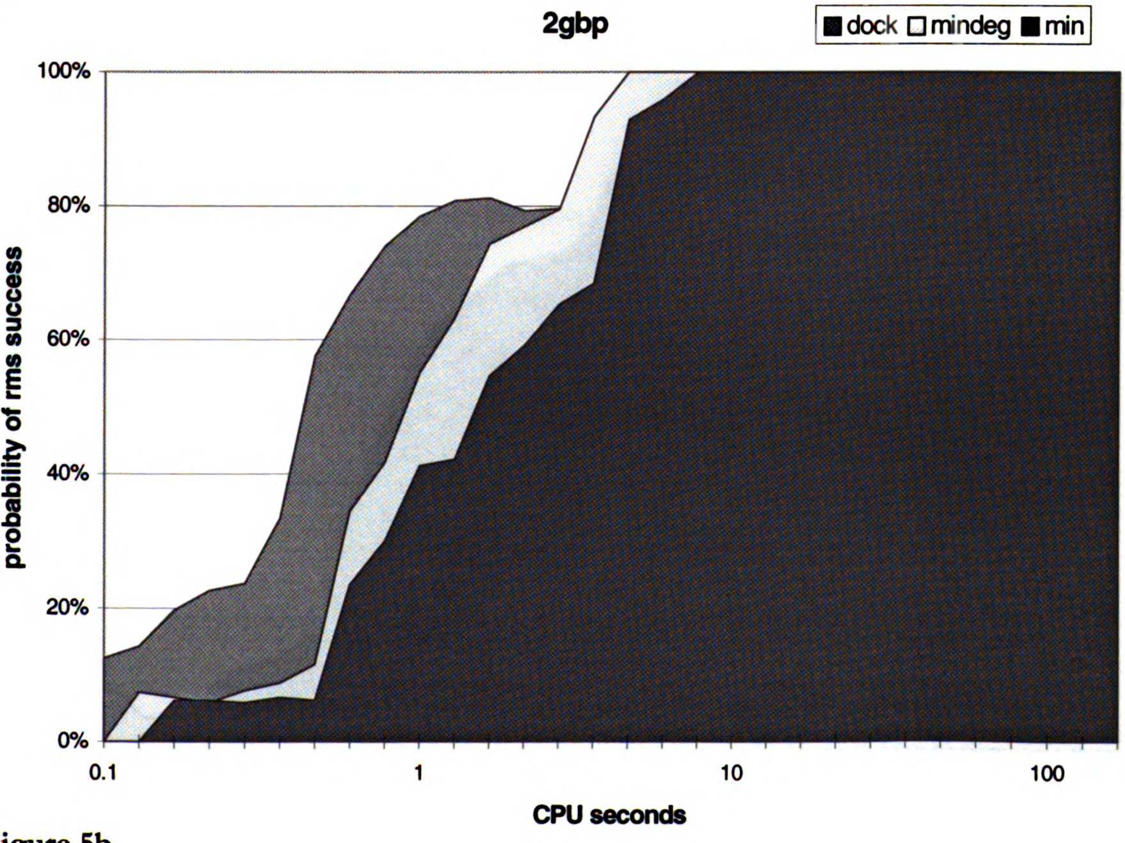


Figure 5b.

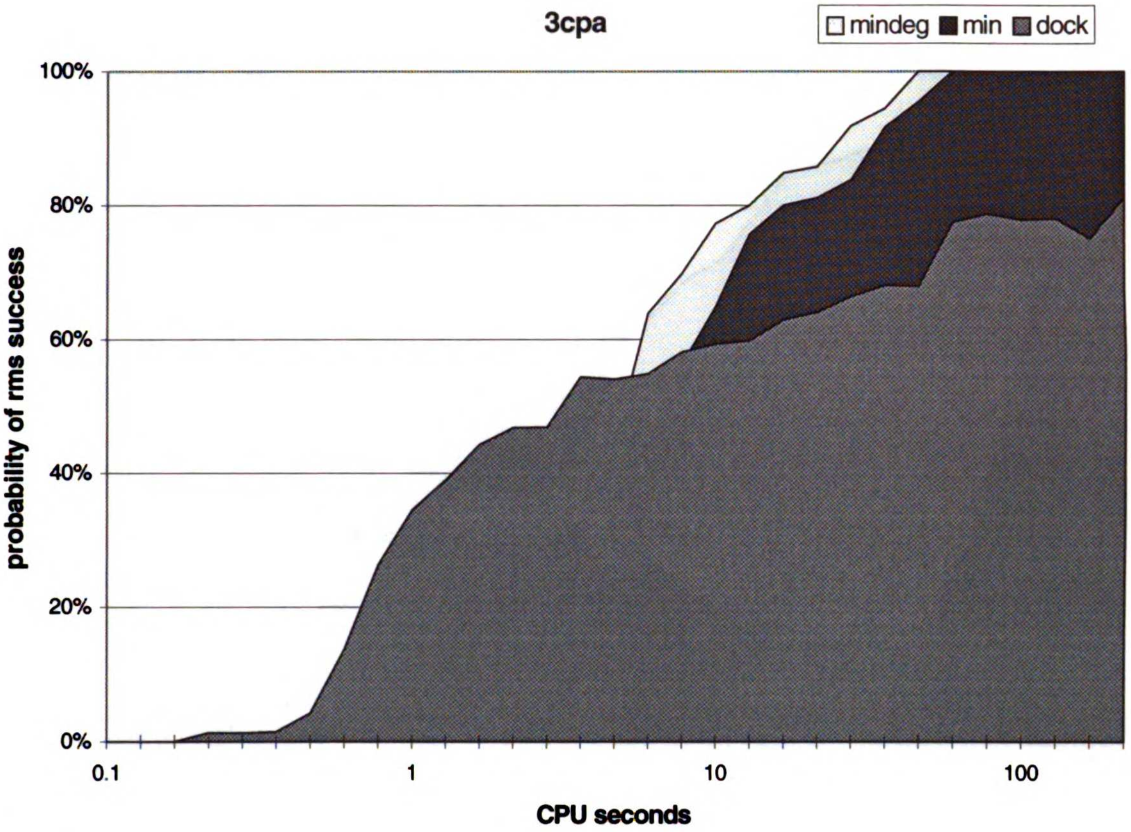


Figure 5c.

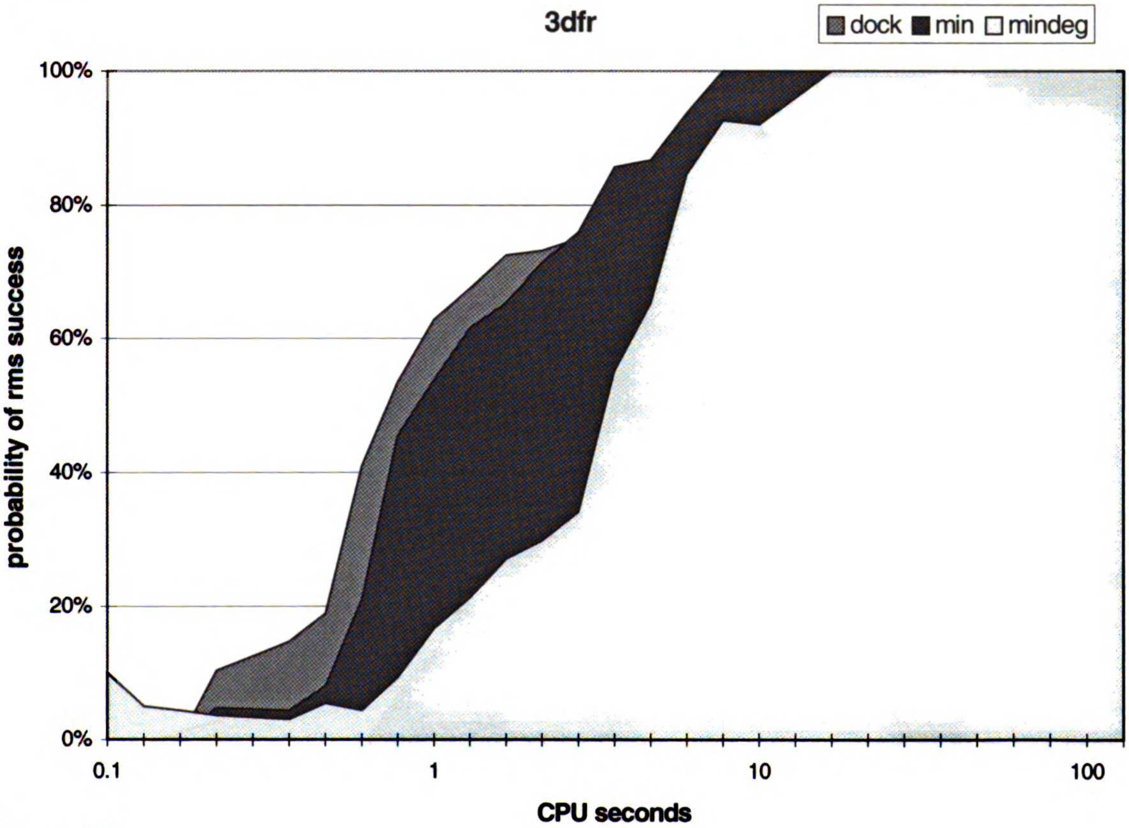


Figure 5d.

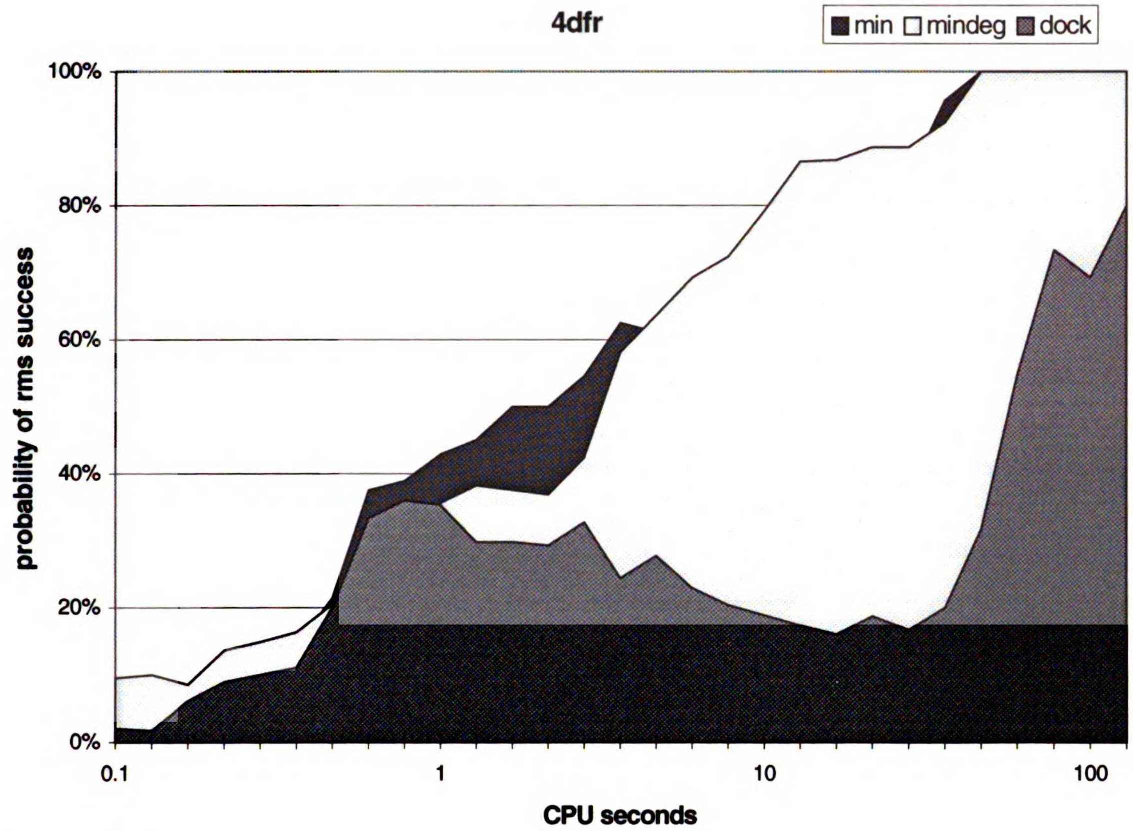


Figure 5e.

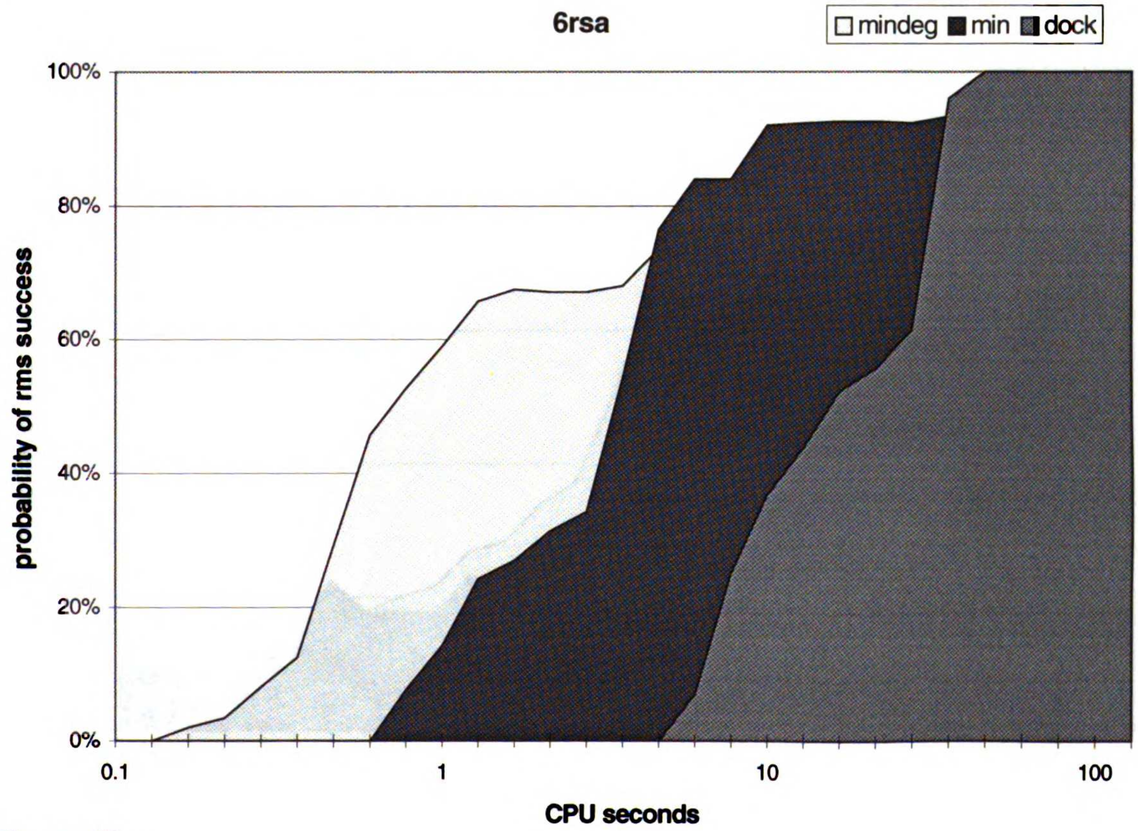


Figure 5f.

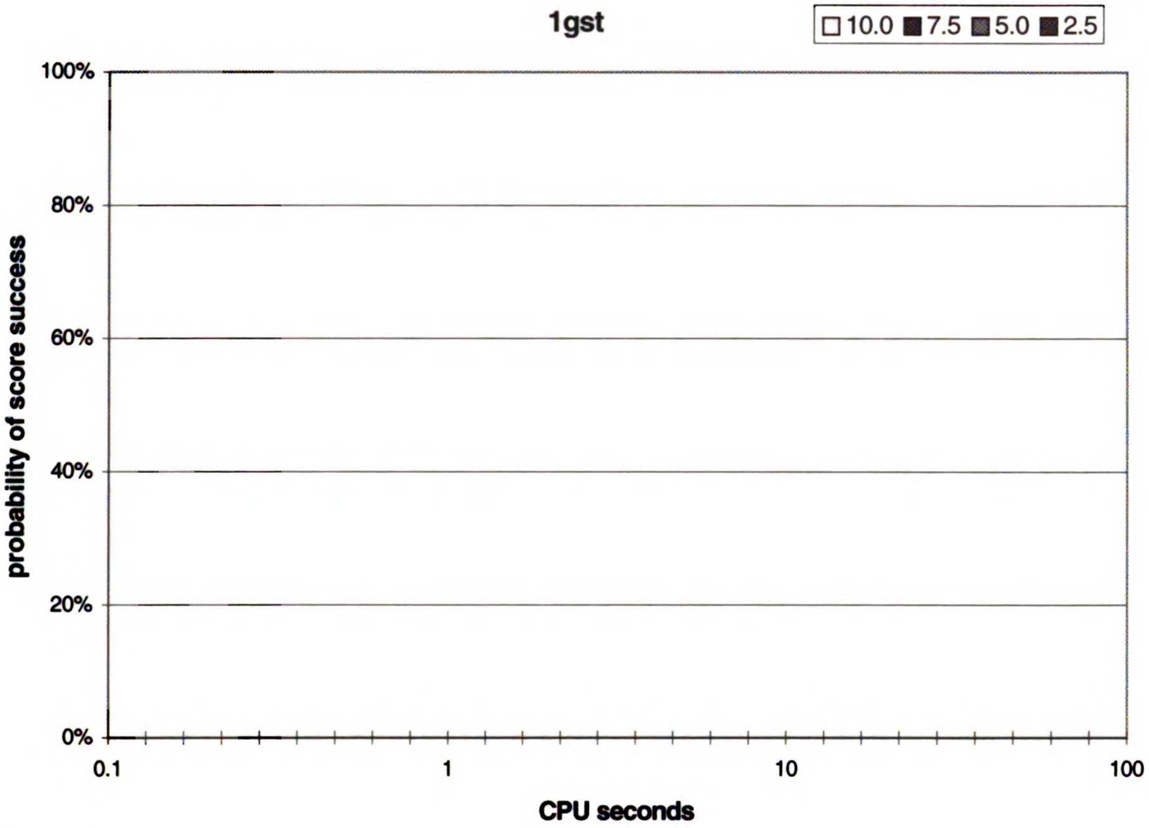


Figure 6a.

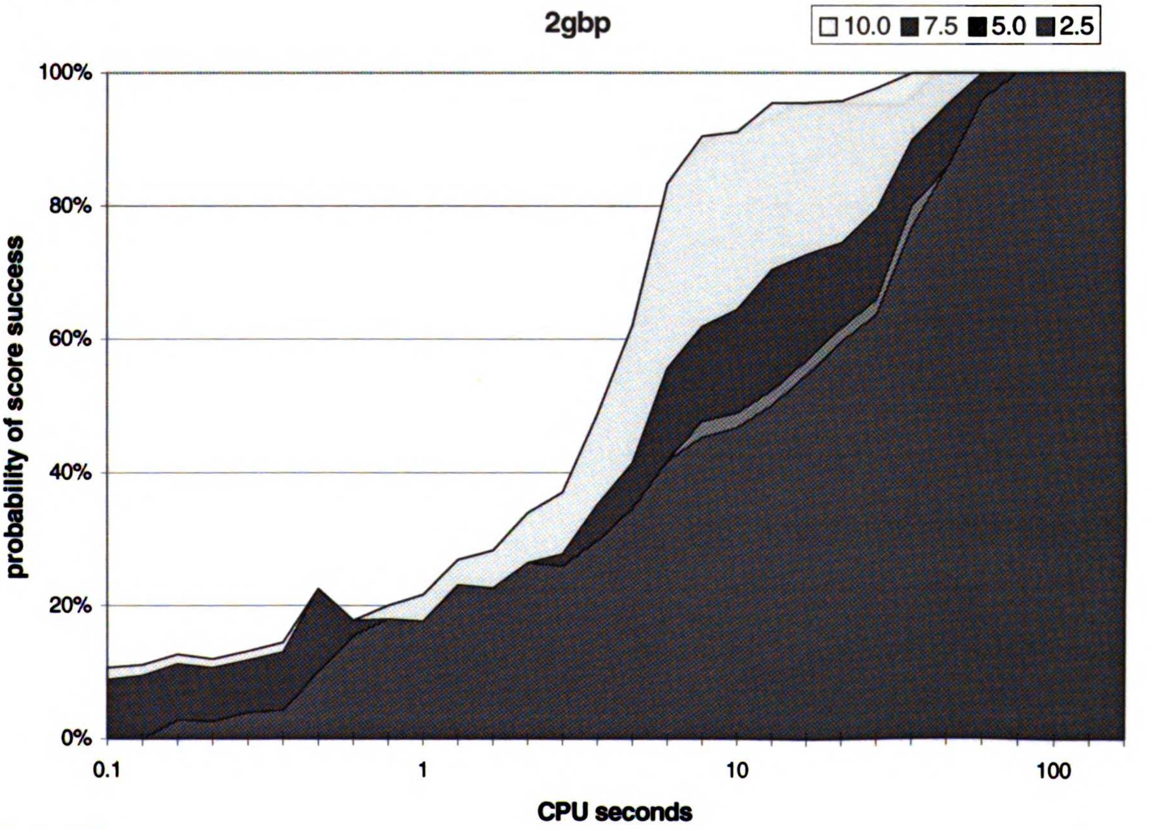


Figure 6b.

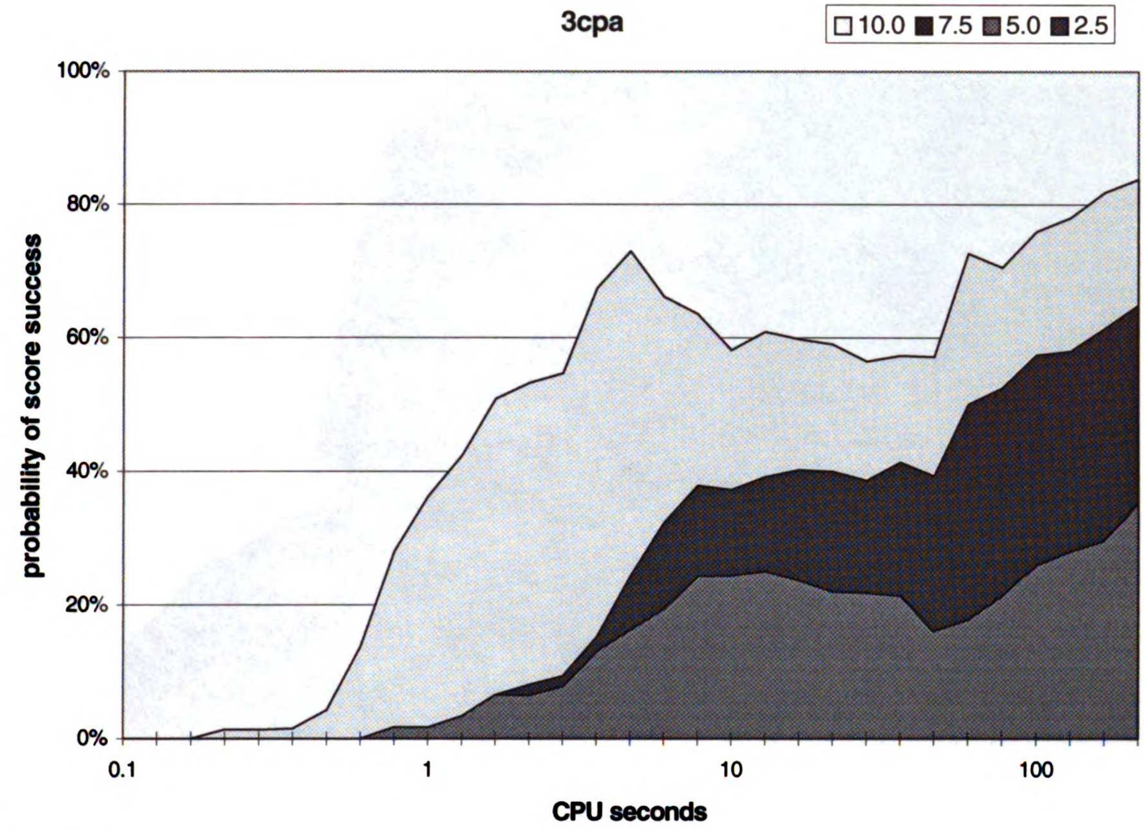


Figure 6c.

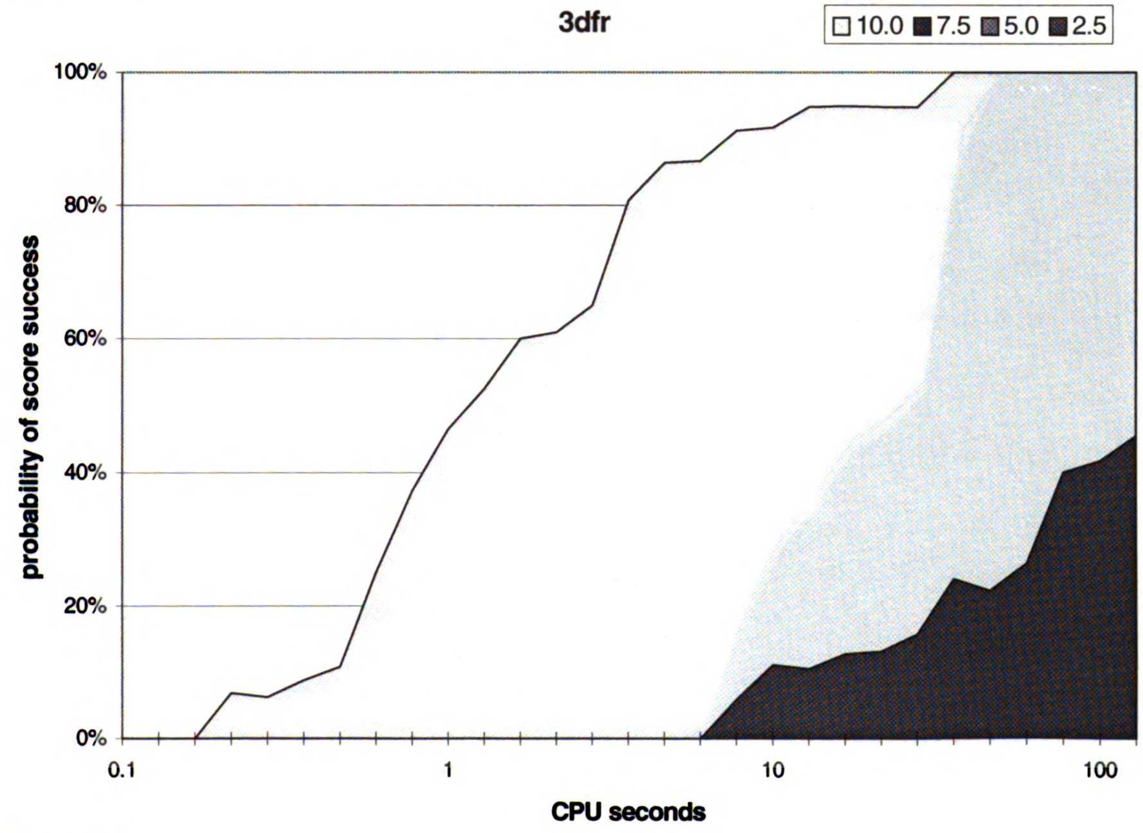


Figure 6d.

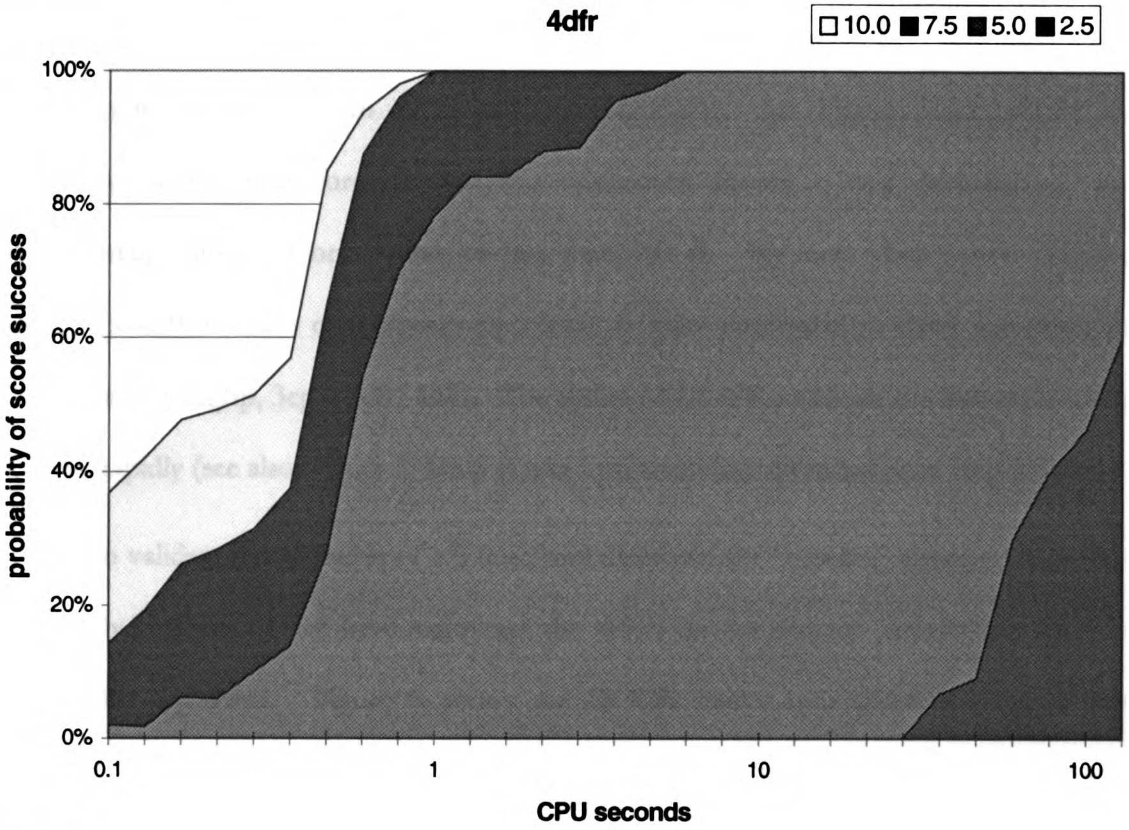


Figure 6e.

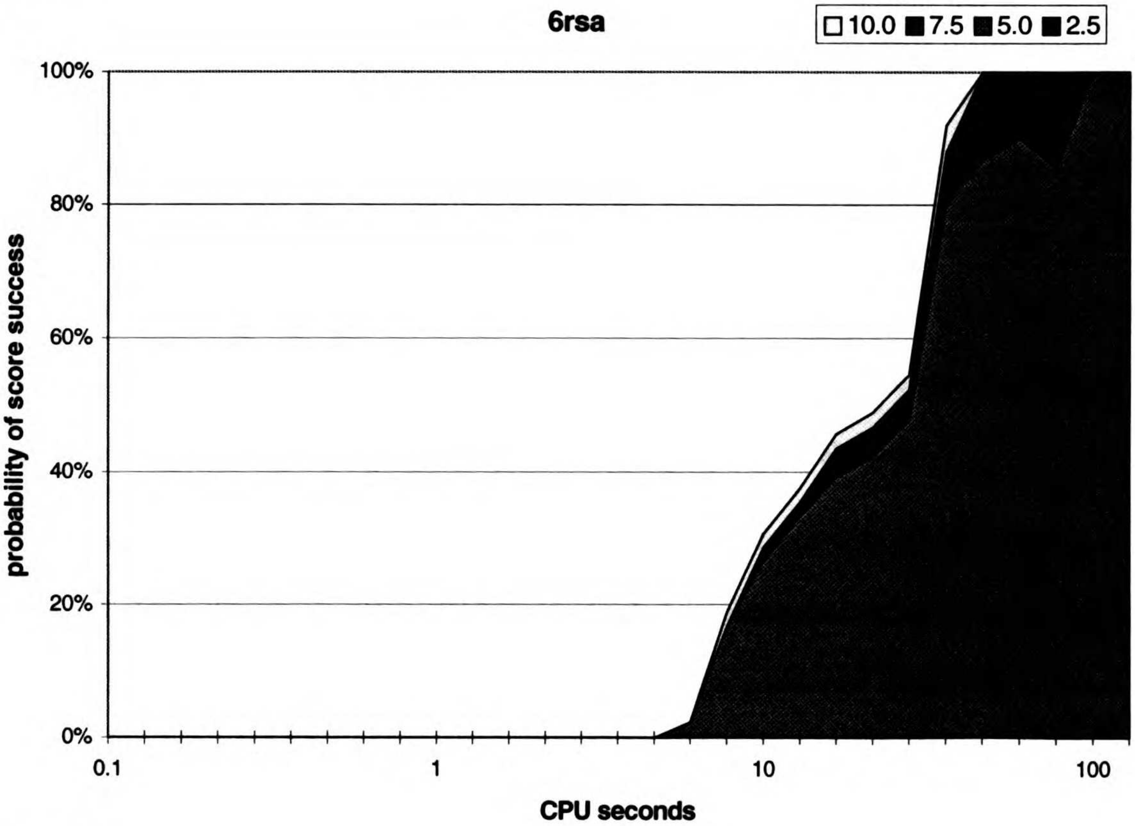


Figure 6f.

Although not applicable to database searches, a common metric in evaluating docked complexes is the similarity to an observed configuration. In Figure 5 we present the probability of the best force-field-scoring orientation having a rms deviation to the experimentally observed orientation of less than 1.0 Å. We note that native DOCK, although it suffers from convergence problems, is quite successful at short run-times in several systems (2gbp, 3cpa, 3dfr, 4dfr). The ability of DOCK to locate the known binding mode so rapidly (see also Figure 2) hints at why implementing optimization is so powerful.

To validate the selection of a 5 kcal/mol threshold for "success" about a f orce-field score global minimum, we have examined the effect on the success- *versus*-effort plots of varying this threshold. Figure 6 shows the DOCK native runs plotted using success thresholds of 2.5, 5.0, 7.5, and 10.0 kcal/mol. In all systems but 2gbp it is apparent that a 2.5 kcal/mol cutoff is too stringent for a fair comparison with minimization. The 5.0 kcal/mol and 7.5 kcal/mol envelopes look similar in the 2gbp, 3cpa, 4dfr, and 6rsa systems, indicating that a plateau has been reached. 5.0 kcal/mol is a reasonable upper limit on the noise in making comparisons among different ligands in a database scan. The 10.0 kcal/mol threshold is too tolerant for a sensible comparison, particularly given that this value represents 20-40% of the global minimum for the majority of the test cases (Table 3). The analogous series of envelopes for minimization with and without degeneracy checking are nearly constant across the entire 2.5 - 10.0 kcal/mol range (data not shown).

One can envision a simple alternative to introducing force-field score optimization into the docking process: merely performing a stand-alone minimization on the output of a native DOCK run. Given the negligible cost of a single grid-based simplex refinement (Table 2), this could conceivably be an efficient method for improving results. We have entertained this possibility in four of the test systems, and compare post-DOCK minimization to native DOCK and DOCK with on-the-fly optimization in Figure 7. In two systems (2gbp, 3cpa) post-DOCK minimization is actually the most effective method for short run-times, but displays convergence problems as runtimes lengthen, particularly in the case of 3cpa (and also 1gst). A shortcoming of such a method is illustrated in the 6rsa system, where post-DOCK minimization is barely an improvement over native DOCK. Possible explanations for why this behavior is likely to be a common instance are taken up in the Discussion.

Figure 7, a-d. Force-field score success vs. CPU time.

(following pages) The probability of locating an orientation having a force-field score within 5 kcal/mol of the global minimum is plotted as a function of CPU seconds required. "dock" represents native DOCK, "min" represents DOCK with on-the-fly minimization, "postmin" represents native DOCK with stand-alone grid-based simplex minimization performed on the output.

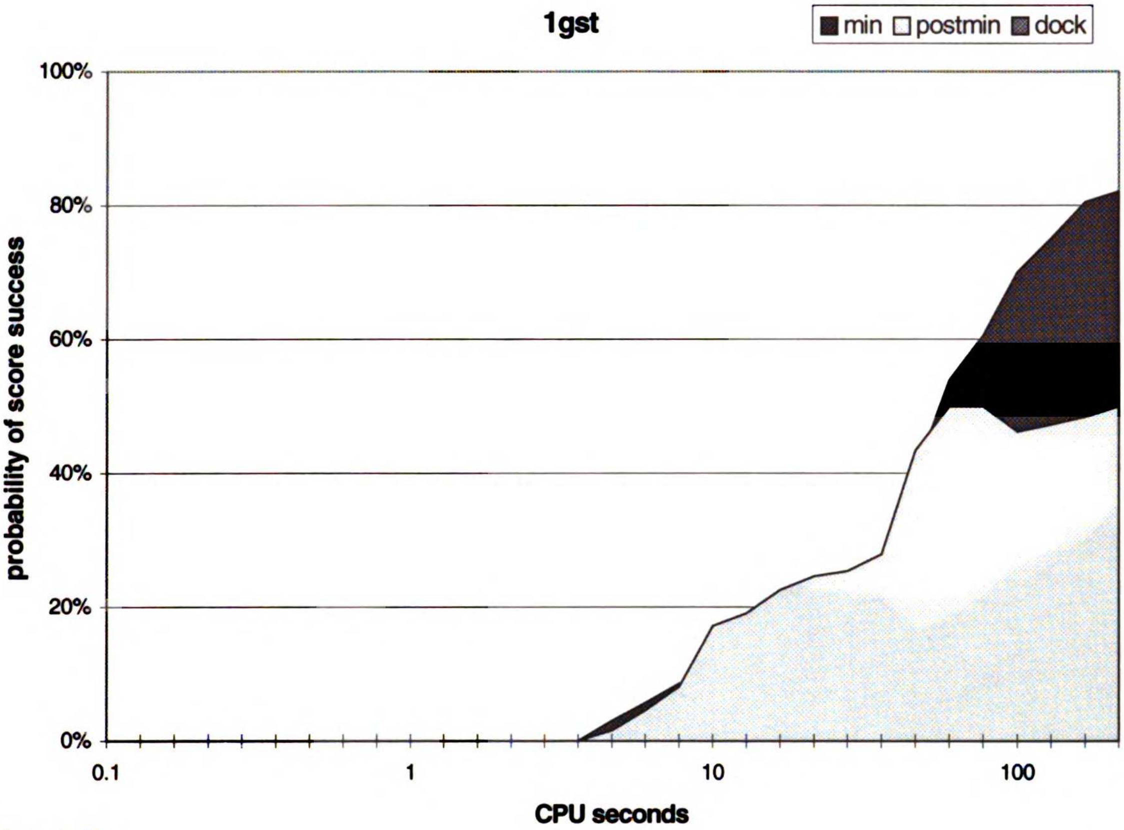


Figure 7a.

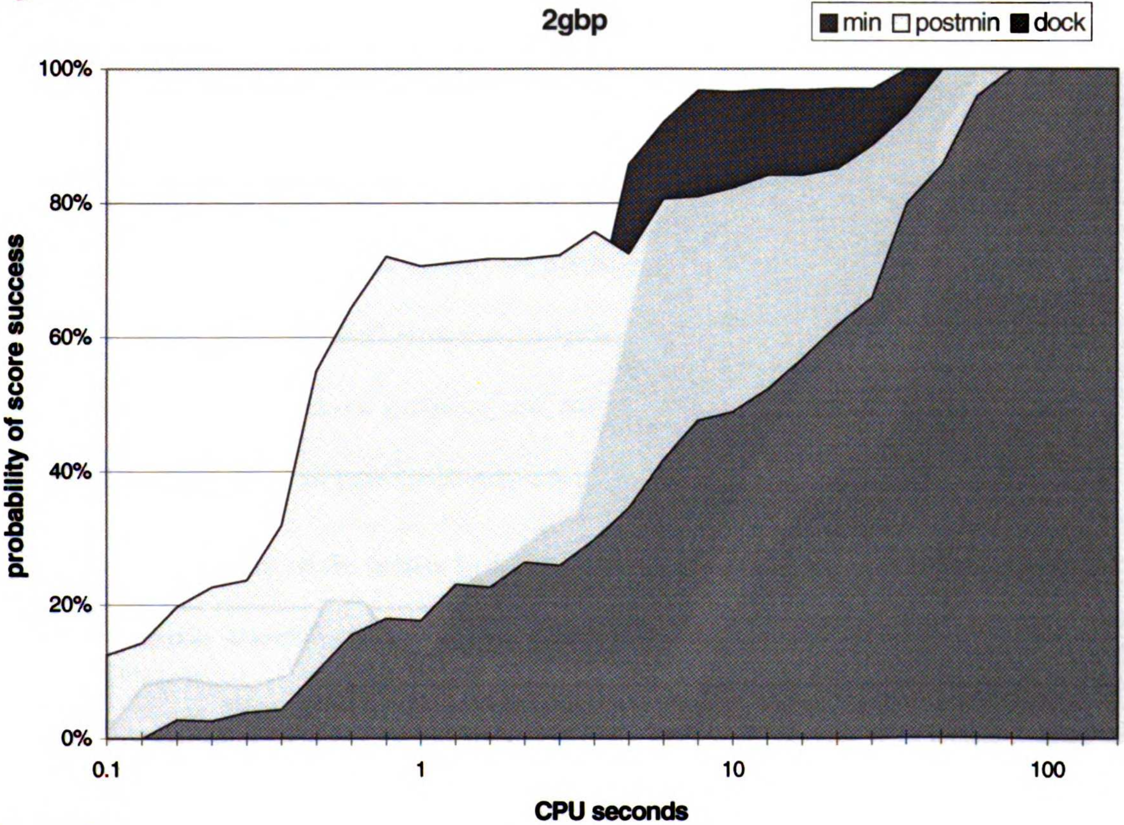


Figure 7b.

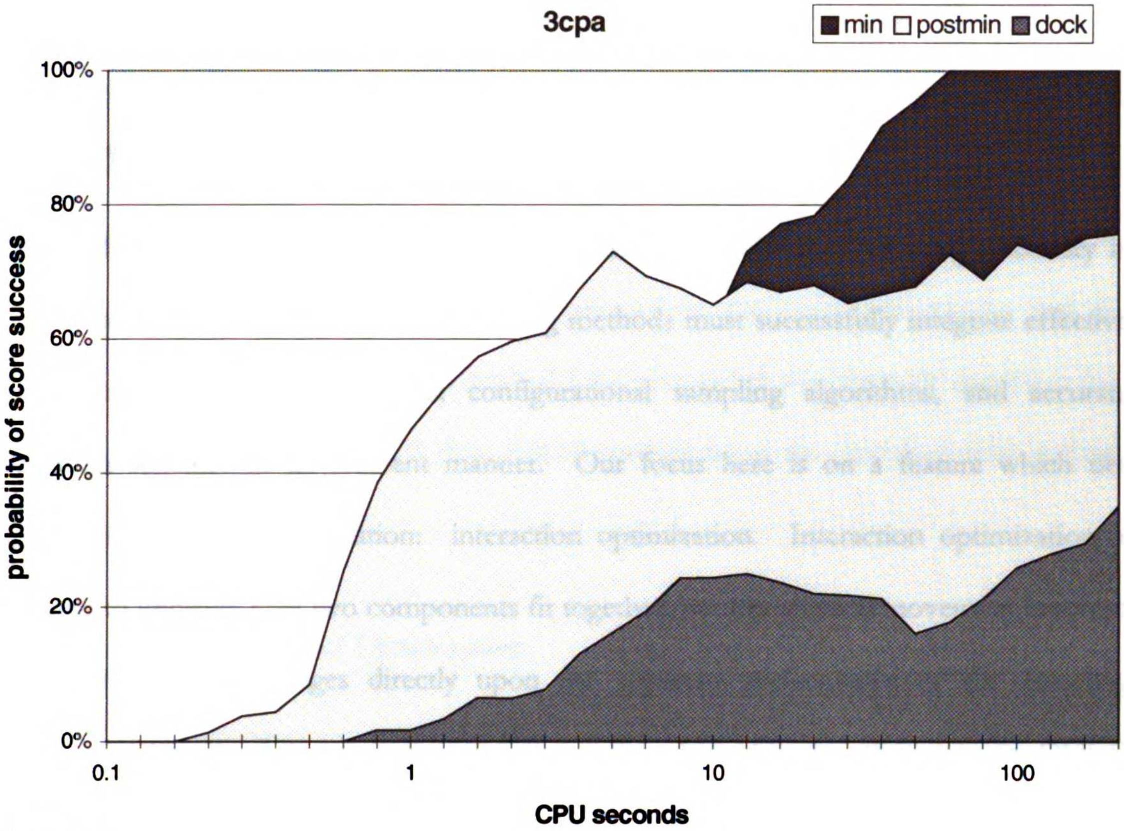


Figure 7c.

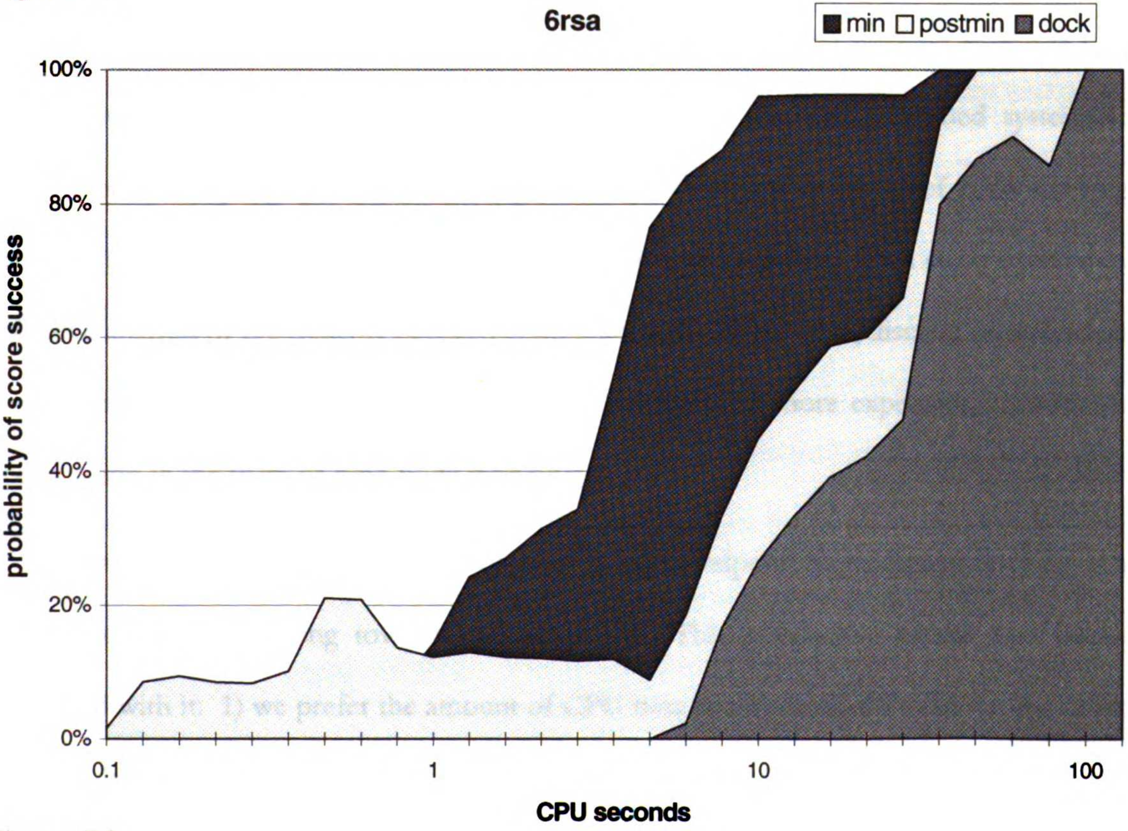


Figure 7d.

DISCUSSION

Perspective

Molecular docking has become an increasingly popular tool for drug discovery in recent years [19]. To be truly useful, docking methods must successfully integrate effective site description techniques, robust configurational sampling algorithms, and accurate evaluation schemes in an efficient manner. Our focus here is on a feature which ties together sampling and evaluation: interaction optimization. Interaction optimization is designed to improve how two components fit together, but the physical movement involved in the refinement impinges directly upon the apparent performance of the sampling algorithm. Thus, our investigation into the utility of rigid-body refinement in DOCK necessarily probes configurational search methods.

Interaction optimization is not new to automated molecular docking methods [20-23]. However, to our knowledge, this article represents the first published systematic exploration of sampling space for a docking method. We analyze in excess of 2,500 docking runs, not simply an arbitrary slice of the vast configurational universe. This study enables an objective analysis of the tradeoff between computationally inexpensive, discrete optimization in the form of configurational sampling and the considerably more expensive, continuous optimization in the form of rigid-body refinement.

Our assessment of the results is colored by our standpoint on molecular docking as a tool for database searching toward lead discovery. This perspective carries two biases associated with it: 1) we prefer the amount of CPU time spent per ligand to be on the order of seconds, not minutes; and 2) we rank binding modes and ligands by interaction scores, not rms deviations to observed configurations. The latter point implies that efforts should be directed toward locating the global minimum in a scoring function, not necessarily toward identifying a known binding mode. *We make the assumption that the experimentally observed orientation is in fact at the global minimum.* It is therefore the task of scoring function developers to insure coincidence between the global optimum of the evaluation scheme and the observed mode. For all six systems studied here, the global minimum of the force-field score developed by Meng *et al.* [5] does indeed correspond to the crystallographic solution (to within 0.5 Å rms deviation).

Robustness in Sampling and Optimization

The grid-based simplex minimizer introduced here displays close to a 50-fold average speed increase over the quasi-Newton method used in the previous investigation [6], with no loss in accuracy. This dramatic improvement has enabled the incorporation of refinement into the docking process, albeit still at considerable computational expense when compared with the speed of matching or force-field scoring alone (Table 3). We note that native DOCK processes about 2000-3000 matches per second, while DOCK with on-the-fly minimization only about 2% of that. The actual cost of one minimization is 100 times that of a single force-field score evaluation (data not shown), but the full effect of this penalty is not realized within DOCK because not all orientations are minimized (only those which pass the bad contacts filter). We see in Figure 4 the nearly across-the-board ability of on-the-fly optimization to not only counteract this handicap, but significantly surpass native DOCK in efficiently locating low-energy solutions. Why should this be so?

The compelling plots presented in Figure 2 speak to the robust nature of the sphere description and matching algorithm currently implemented in DOCK. The fact that the sampling method can readily retrieve configurations extremely close to the experimentally observed configuration indicates that failure to identify this mode as optimal lies with scoring and not with sampling. The effect of minimization, then, is to salvage the many orientations generated near the crystallographic mode which would otherwise be thrown out due to steric clashes with the receptor. *Optimization allows maximal use to be made of all information provided by the matching algorithm.* We expect on-the-fly optimization to benefit database searches most by rescuing ligands for which the proper binding mode is sampled but for which no low energy orientations can be found in the absence of refinement. Two such examples appear in this work, 1gst and 3dfr (Figure 4 and Figure 6), and their recovery underscores the utility of on-the-fly optimization. The tolerance of a non-zero number of bad contacts within DOCK is imperative to taking full advantage of the potential of minimization as a rescue device.

Degeneracy Removal

The degeneracy checking protocol described here has met with mixed success. Although typically 90% of orientations are deemed degenerate and not examined further, this savings under the current implementation does not significantly outweigh the cost of assessing degeneracy. The advantages are manifested primarily at shorter runtimes, as evidenced in the 1gst, 3cpa, and 6rsa systems (Figure 4). This capacity will find use in database searching applications when CPU resources are quite limited, as not all ligands are likely be sampled adequately with the same set of sampling parameters.

The judicious selection of fewer orientations for optimization is obviously a compromise between refining all and refining none. By refining all orientations, resources are spent insuring each orientation is within a local minimum, not sampling the vast configurational universe (akin to a depth-first search). Conversely, by refining no orientations, resources are spent exploring configuration space without particular regard to the quality of each orientation (a breadth-first search). Refinement is relatively expensive computationally and configurational exploration inexpensive, so the optimal tradeoff comes when configuration space is thinly but evenly sampled with refined orientations. The early advantages evidenced with the degeneracy removal protocol at short runtimes are the result of exactly this tradeoff. At longer runtimes when inexpensive configurational sampling is more intense, minimization alone generally performs at least as well as when coupled with degeneracy removal.

We believe the largest hurdle in devising a more successful degeneracy removal protocol lies in the selection of a representative for each binding mode. In this work, we choose the first orientation found in a binding mode as that family's "parent" for assessing degeneracy. If this orientation should be a poor representative, further orientations in that family will nonetheless be considered degenerate and thrown out, regardless of how they might have scored. The *degenerate_save_interval* alleviates this bias to some extent, but functions as a crutch rather than a solution.

There are many degeneracy parameters to be varied, but their effects have not been examined systematically here. In preliminary exploration, we find that *vsph* of 1.5 to 2.0 Å for creating virtual spheres, *wobble* of 2, and *degenerate_save_interval* of 10 to 25 appear to offer a reasonable compromise between speed and accuracy. Although a hash table is used to reduce the memory requirements of storing information about unique matches, the memory demands of degeneracy checking are still quite steep. When the hash table begins to fill, retrieval from the table also becomes more costly and performance begins to degrade.

Hence, we again advocate the use of degeneracy checking for low to medium sampling levels only.

Prospects for Post-DOCK Optimization

The appropriate control experiment for the introduction of on-the-fly minimization entails performing a DOCK run without minimization and subsequently optimizing the output in the same fashion. In this way, we reveal the benefits imparted by minimizing all DOCK orientations as opposed to minimizing only the best unoptimized orientation. The most obvious danger of selecting only the lowest-energy unoptimized orientation is that other orientations may lie higher in energy but in a deeper well, so that upon optimization these other orientations would have finished lower in energy. This possibility is borne out by the shuffling of pre- and post-optimization force-field scores (data not shown). Surprisingly, the 2gbp and 3cpa systems perform quite well at short runtimes, but along with 1gst begin to suffer from convergence problems as runtimes lengthen.

The convergence problems shown in the 1gst and 3cpa systems and the lack of improvement seen in the 6rsa system (Figure 7) are likely to be common occurrences for the following reasons. Finding an orientation in the observed binding mode is a necessary but not a sufficient condition for obtaining a force-field score near the global minimum after optimization. Because on-the-fly optimization refines *every* orientation, it is afforded the luxury of the chance that *any* of the orientations near the observed binding mode (Figure 2) will refine near to the global minimum in force-field score. In contrast, DOCK without on-the-fly optimization has available only *one* orientation deemed best by an unoptimized force-field score, with the additional constraint that this one orientation must be in the observed binding mode (Figure 5). DOCK without on-the-fly optimization therefore gets *at most* one chance to refine an orientation into the global minimum if post-docking optimization is performed. Thus, it is to be expected that situations such as that displayed by 6rsa will occur frequently. Nonetheless, we see that performing a post-DOCK optimization is in all cases superior, and in many cases substantially so, to performing a native DOCK run without any refinement whatsoever.

Matching Algorithm Discontinuities

A disconcerting consequence of the bin architecture for ligand-site matching is that results obtained at a low level of sampling are not guaranteed to be a subset of results obtained at a higher level of sampling. This point has been noted previously [6]. Although in general this is not the case, this artifact can lead to strange behavior, particularly when examining arbitrary slices of sampling parameters. The analysis of hundreds of DOCK runs for each system in this study enables us to collect statistically significant success probabilities and bypass much of the problem. One will note, however, that the plots in Figure 2 through Figure 7 do not display monotonic functions: the jagged nature of these curves is the result of the discontinuity arising from the bin architecture. Fortunately, the physical convergence of orientations into local minima by on-the-fly minimization mitigates the severity of this artifact.

Sampling Guidelines

One of the most instructive findings from the great number of DOCK runs examined is insight into the amount of sampling required to obtain a desired probability of success. The success-*versus*-effort plots carry a great deal of information, and can be used as guidelines for performing DOCK runs appropriate to available resources. For instance, one might be interested in performing a large database search where each ligand would be allotted the minimum resources to obtain 100% success. In this case, one might calibrate sampling conditions to expend an average of 10 CPU seconds per ligand (or on the order of 1000 to 3000 matches; Figure 3). In another example, one might be interested in analyzing a small database with the assurance that each ligand was well into the 100% success plateau. For this case, one might calibrate sampling conditions to expend 100 CPU seconds per ligand. It would be reasonable to construct a success- *versus*-effort plot for a known ligand, if available, for performance gauges customized to the system being studied. In this manner, the success-*versus*-effort plots provide a valuable mechanism for setting sampling levels in molecular docking.

CONCLUSIONS

We have coupled a fast and effective grid-based, rigid-body simplex minimizer with the robust configurational sampling algorithm used in DOCK to allow on-the-fly force-field score optimization in a tractable manner. This coupling, despite the heavy resource demands of refinement, results in a docking strategy which is computationally more efficient at retrieving experimentally observed configurations than docking in the absence of optimization. In some cases, only with the use of on-the-fly optimization could the observed binding mode be identified as the global minimum in the scoring function. On-the-fly optimization salvages poor orientations which would otherwise be discarded, thus making maximal use of information afforded by the sampling algorithm. The removal of geometrically similar orientations to circumvent redundant optimizations is a tradeoff between expensive refinement and inexpensive sampling - our implementation shows mixed success, but with greatest potential at short per-ligand runtimes. Finally, while not as effective as on-the-fly optimization, it is clearly wiser to perform a post-docking optimization than none at all. We find that success- *versus*-effort plots for gauging docking performance lend valuable insight into the setting of sampling levels for the inevitable compromise between result quality and computational resources.

ACKNOWLEDGMENTS

We thank Todd Ewing, Andrew Good, Elaine Meng, and Brian Shoichet for numerous insightful discussions. This work was funded through NIH grants GM-31497 (I.D.K.) and GM-39552 (G. L. Kenyon, principal investigator). D.A.G. gratefully acknowledges support from an ACS Division of Medicinal Chemistry Fellowship. Force-field score optimization and degeneracy checking features are available in DOCK version 3.5. The DOCK suite of programs is implemented in Fortran and accessible through I.D.K.

REFERENCES

1. Blaney, J.M. and Dixon, J.S., Persp. Drug Discov. Design, 1 (1993) 301.
2. Kuntz, I.D., Meng, E.C. and Shoichet, B.K., Acc. Chem. Res., 27 (1994) 117.
3. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
4. Shoichet, B.K., Bodian, D.L. and Kuntz, I.D., J. Comp. Chem., 13 (1992) 380.
5. Meng, E.C., Shoichet, B.K. and Kuntz, I.D., J. Comp. Chem., 13, (1992) 505.
6. Meng, E.C., Gschwend, D.A., Blaney, J.M. and Kuntz, I.D., Proteins, 17 (1993) 266.
7. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R. Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.

8. Ji, X, Zhang, P., Armstrong, R.N. and Gilliland, G.L., Biochemistry, 31 (1992) 10169.

9. Vyas, N.K., Vyas, M.N. and Quiocho, F.A., Science, 242 (1988) 1290.

10. Rees, D.C. and Lipscomb, W.N., Proc. Natl. Acad. Sci. U.S.A., 80 (1983) 7151.

11. Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin , R.C. and Kraut J., J. Biol. Chem., 257 (1982) 13650.

12. Borah, B. Chen, C.-W., Egan, W., Miller, M., Wlodawer, A. and Cohen, J.S., Biochemistry, 24 (1985) 2058.

13. Shoichet, B.K. and Kuntz, I.D., Protein Eng., 7 (1993) 723.

14. Connolly, M.L., Science, 221 (1983) 709.

15. Nelder, J.A. and Mead, R., Computer J., 7 (1965) 308.

16. Ferro, D.R. and Hermans, J., Acta Crystallogr. Sect. A, 33 (1977) 345.

17. Knuth, D.E. The Art of Computer Programming, Vol. 3, Addison-Wesley, Menlo Park, California, 1973, pp. 506-549.

18. Fletcher, R., Practical Methods of Optimization, Interscience, New York, 1960.

19. Gschwend, D.A., Good, A.C. and Kuntz, I.D., J. Mol. Recogn., in press.

20. Goodsell, D.S. and Olson, A.J., Proteins, 8 (1990) 195.

21. Hart, T.N. and Read, R.J., Proteins, 13 (1992) 206.

22. Yamada, M. and Itai, A., Chem. Pharm. Bull., 41 (1993) 1200.

23. Miller, M.D., Kearsley, S.K., Underwood, D.J. and Sheridan, R.P., J. Comput.-Aided Mol. Design, 8 (1994) 153.

Chapter 5.

Prediction of Absolute Binding Affinity of Ligands for Macromolecular Receptors of Known Three-Dimensional Structure[†]

Daniel A. Gschwend, Andrew C. Good,[‡] and Irwin D. Kuntz[*]

Department of Pharmaceutical Chemistry,

University of California, San Francisco, CA 94143-0446

[†] Part of this work appears in preliminary form elsewhere [Gschwend *et al.*, 1995 (Chapter 2)].

[‡] Current address: Rhône Poulenc Rorer, Dagenham Research Centre JB4-0, Dagenham, Essex RM10 7XS, United Kingdom.

[*] Author to whom correspondence should be addressed.

ABSTRACT

Computational methods for drug design have benefited tremendously from the burgeoning field of structure determination. The availability of high-resolution structural information has promoted innovative techniques for exploring receptor-ligand interactions. The greatest hindrance to structure-based strategies remains the inability to accurately and consistently estimate ligand binding affinities. Automated design protocols examine thousands of putative receptor-ligand configurations and demand rapid feedback on the quality of the association. Towards this goal, we present the development of an empirical scoring scheme calibrated against binding affinities for experimentally observed complexes. Emphasis is placed on accuracy in predictions, robustness in handling structural diversity, and speed of evaluation. Effective interaction descriptions coupled with an all-possible-subsets multiple linear regression analysis have led to a model capable of reproducing observed binding free energies to within 1.7 kcal/mol for a large, complex data set. The calibration data set, the largest yet reported, consists of 103 structurally diverse receptor-ligand complexes spanning over twelve orders of magnitude in binding affinity. The performance of the empirical model is contrasted with a molecular mechanics function used in a popular molecular docking package. It is crucial for evaluation methods which aim to be generally applicable in structure-based design strategies to consider both enthalpic and entropic contributions to binding free energy.

Keywords: structure-based drug design, empirical scoring schemes, interaction evaluation, binding affinity prediction, molecular docking

INTRODUCTION

The wealth of high-resolution structural data furnished by crystallographic and spectroscopic techniques has kindled structure-based drug design strategies. There are now a variety of computational techniques which may be useful towards drug discovery in the context of detailed receptor information (Kuntz, 1992; Greer *et al.*, 1994; Guida, 1994; Lybrand, 1995). In striving to identify agents which will bind to a receptor of known structure, these techniques are divided broadly amongst those which dock molecules and those which build them. Docking methods scan databases of pre-existing compounds for complementary ligands (Blaney & Dixon, 1993; Kuntz *et al.*, 1994; Good & Mason, 1995); building ("*de novo* design") methods create ligands tailored to the site of interest (Lewis & Leach, 1994). While core technology is well-established, *i.e.* configurational sampling for the former and molecular assembly for the latter, each approach manifests characteristic weaknesses: docking methods are limited by the diversity of the compound library, while building methods suffer from concerns regarding synthetic feasibility.

All structure-based approaches, however, are limited by the accuracy with which the affinity of proposed ligands can be gauged. Correct relative ranking of putative ligand-receptor associations is prerequisite to a useful strategy for drug design. Hence, it is the evaluation scheme which scores interactions between components that now commands the most attention. Scoring functions must be rapidly evaluable, as docking and building strategies typically consider thousands of ligand-receptor complexes. Complementarity itself may be evaluated in many ways (Cherfils & Janin, 1993; Gschwend *et al.*, 1995). One of the most popular methods for assessing small-molecule binding, ushered in by the early work of Goodford (1985), employs a molecular mechanics force-field. More recently, empirical schemes have met with significant interest (Bohacek & McMartin, 1992; Horton & Lewis, 1992; Krystek *et al.*, 1993; Bohacek & McMartin, 1994; Böhm, 1994a,b; Vajda *et al.*, 1994; G.R. Marshall, personal communication; A.N. Jain, personal communication; M.A. Murcko, personal communication). The free energy perturbation (FEP) methods are currently the most rigorous and most accurate for determining relative binding free energies (Beveridge & DiCapua, 1989; Kollman & Merz, 1990; Straatsma & McCammon, 1992). Despite various approximations geared towards performance enhancements (Gerber *et al.*, 1993; Åqvist *et al.*, 1994; Warshel *et al.*, 1994), these techniques remain restricted by staggering computational demands and to small molecular systems, precluding their use for screening thousands of ligands of varying chemical framework.

Our goal in this study is to derive an empirical scoring function that can rapidly estimate affinities over a structurally diverse array of receptor-ligand complexes. By rapid, we desire that several evaluations be performed in one second, not one evaluation in several minutes or hours. This requirement is dictated by the vast number of arrangements which must be considered within the molecular docking and *de novo* design paradigms. These structure-based design tools are intended to generate many unnatural associations expressly so that novel, potent binding agents can be discovered. To engender an ability to cope with such foreign molecular combinations, the evaluation function should be calibrated against a large and complex data set. Thus, rather than borrow a functional description parameterized against an endpoint different from that which interests us, we seek to derive a function designed to estimate absolute binding free energies for use specifically in automated structure-based drug design techniques.

In particular, we aim to deviate from molecular mechanics-based functions. Molecular mechanics has been parameterized to reproduce internal properties of small molecules, such as dipole moments, torsional barriers, and heats of formation (Clark, 1985).

When used for assessing intermolecular interactions, "force-field scores" report an enthalpy of interaction, while the quantity of most interest in structure-based drug design is a free energy of interaction. In our experience, force-field scores are effective at identifying the optimal binding mode of a single ligand (Meng *et al.*, 1992; Gschwend & Kuntz, unpublished results), but perform poorly at predicting even relative binding energies across a panel of ligands. Entropic contributions are likely to be fairly similar for different binding modes of one ligand, but clearly can vary substantially from one ligand to the next. Without the entropic half of the equation we have little hope of predicting binding affinities for structurally diverse ligands. It is noteworthy, however, that successful, system-specific examples of enthalpic correlations with binding affinity have been reported (*e.g.* Holloway *et al.*, 1995). Here, we emphasize the need for robustness across structurally unrelated ligands binding to varied receptors.

To compensate for the omission of entropic contributions by molecular mechanics, several researchers have augmented the standard description with empirical terms (Novotny *et al.*, 1989; Wilson *et al.*, 1991; Krystek *et al.*, 1993; Vajda *et al.*, 1994). While this has appeared useful, we nevertheless choose to dispose of traditional electrostatic and van der Waals representations for several reasons. First, we avoid problematic issues such as selection of partial charge set and choice of dielectric behavior, both of which remain subjective yet can have profound effects on results. Second, we bypass the need for hydrogen placement. To illustrate, consider hydroxyl hydrogens - on serine, threonine, and tyrosine residues, for example. Preference for hydrogen bond geometry about these functionalities is weak (Baker & Hubbard, 1984; Thanki *et al.*, 1988; Tintelnot & Andrews, 1989), while molecular mechanics-based schemes require selection of an exact hydrogen position. Interaction strength is thus spuriously sensitive to the (typically arbitrary) placement of this hydrogen. Finally, and perhaps most importantly, the over-sensitivity to precise atomic position, rooted in the steep van der Waals potential, undermines the robustness of molecular mechanics-based functions. A softer potential favors generality across diverse systems, though possibly at the expense of accuracy in details.

Automated structure-based design methods seek ligands which exploit some aspect of complementarity to the receptor of interest. An attempt to simulate molecular recognition is made for ligands which the receptor has never encountered (Gschwend *et al.*, 1995). The receptor is modeled in a pre-defined conformation, frequently one molded to a particular ligand, yet it would be beneficial for design strategies if the receptor were allowed to respond to the presence of each putative ligand. An implicit breathing on the part of interacting components can be introduced by a tolerant evaluation function. For example, a soft scoring potential might permit slight atomic interpenetrations without penalty, in effect implying a resolving conformational change. The concept of so-called "soft docking" hails from protein-protein docking investigations in which structures of unbound components are docked to reproduce the observed complexed structure (Wodak & Janin, 1978; Shoichet & Kuntz, 1991; Jiang & Kim, 1991; Walls & Sternberg, 1992). The success of such methods hinges upon a local insensitivity that fosters conformational shifts upon complexation. By adopting some of these ideas, we aim to introduce generality in the scoring function's ability to predict binding affinities.

As early researchers in the protein docking field have noted, even the simplest scoring schemes perform virtually as well as more advanced molecular mechanics treatments (Shoichet & Kuntz, 1991; Cherfils & Janin, 1993). There is thus the potential to derive an evaluation method which is not borrowed from the objectives of another branch of computational chemistry, but rather, which is parameterized to reproduce precisely the type of values we are attempting to predict. In the same vein as Böhm's work (1994a), our empirical evaluation scheme is derived by calibration against experimentally determined affinities of ligand-macromolecule complexes for which structural information is available. The general procedure consists of amassing a series of receptor-ligand complexes [typically from the Protein Data Bank (Bernstein *et al.*, 1977)] with known affinities, devising various calculable terms which describe physical interactions of interest, and attempting to obtain affinity correlations while varying coefficients for each term. Approaches vary widely in the data set composition, the terms employed in correlations, and the method in which the terms are computed. We use restrictive, pre-defined criteria for selecting complexes to comprise a calibration set which is significantly larger than any yet reported: nearly 150 complexes form our basis set. The use of multiple linear regression with an all-possible-subsets protocol enables careful analysis of the relative importance of each proposed contribution to affinity.

The absolute assessment of ligand-receptor affinity remains one of the greatest challenges for computational chemistry. Theoretically-rigorous, resource-intensive methods such as FEP can in the best cases estimate binding energies to within one kcal/mol of experiment (Beveridge & DiCapua, 1989; Kollman & Merz, 1990), and only for systems of limited complexity. We would be foolish to believe that simple empirical schemes, with resource requirements many orders of magnitude smaller, could supplant such methods. What we seek is simply a guide for rapidly screening huge numbers of diverse ligand-receptor associations generated by automated structure-based design strategies.

METHODS

Data Set

All ligand-receptor complexes analyzed in this study have been obtained from the Protein Data Bank (PDB) (Bernstein *et al.*, 1977). A list of 237 complexes from the PDB for which experimental affinity data have been determined, generously provided by Keske & Dixon (unpublished results), was combined with the list complexes taken from the PDB used by Böhm (1994a). The consolidated list was stripped of complexes which were either unrefined or model-built, or which contained covalently-bound, incompletely-modeled, or macromolecular ligands. Affinity data and experimental conditions for the assay and for structure determination were located in the literature for each of the 144 remaining complexes. Affinity data vary considerably in assay methods, measurement error, and type of affinity reported; a handful are IC_{50} values, but most represent K_i, K_d, or K_m determinations. We present the affinity data as pK_i ($-\log K_i$) or pIC_{50} values. Over twelve orders of magnitude in binding affinity are spanned by the data set. The structural data are diverse (64 different receptors are represented) and of high quality (85% of the complexes are solved to 2.5 Å resolution or better). Of 144 complexes, 126 display associated water structure. (See Appendix C for the complete complex listing - note that, as discussed below, not all complexes in this listing were used in the calibration data set.)

Preparation of Receptor-Ligand Complexes

Ligand

Each ligand was separated from the remainder of the complex and further processed with the Sybyl modeling package (version 6.1; Tripos Associates, St. Louis, MO). Atoms were assigned appropriate types and hydrogens added in standard geometries. After correcting formal charges, partial atomic charges were computed by the method of Gasteiger and Marsili (Gasteiger & Marsili, 1980, 1981; Marsili and Gasteiger, 1980). Note that we define ligand in a structural, not a functional, sense: in some cases the "ligand" to which the affinity refers is a cofactor.

Receptor

Solvent molecules, if present, were extracted and saved as a separate entity. All remaining atoms (*i.e.* save for those in the solvent or in the ligand in question), including metal ions, glycosylation sites, and other ligands and/or cofactors, were treated as the receptor. Hydrogens were added in standard geometries.

Evaluation of Ligand-Receptor Affinity: Molecular Mechanics

The molecular mechanics method of interaction evaluation (Meng *et al.*, 1992) used in the DOCK molecular docking program (Kuntz *et al.*, 1982; Shoichet *et al.*, 1992) was applied to each ligand-receptor complex to gauge the performance of proposed empirical schemes. This force-field score, an approximation to intermolecular interaction enthalpy, is comprised of Lennard-Jones van der Waals and Coulombic electrostatics terms (Meng *et al.*, 1992). To alleviate contacts in the experimental structure deemed unfavorable by the force-field score, each ligand was subject to a quasi-Newton rigid-body optimization as described by Meng *et al.* (1993). The DOCKMIN_DFP minimization program, distributed with DOCK 3.5, was run in continuum mode with default parameters and a 4 *r* dielectric (where *r* is the interatomic separation). Resulting optimized force-field scores were used as one estimate of ligand-receptor binding affinity.

Evaluation of Ligand-Receptor Affinity: Empirical Scoring

Overview

The empirical evaluation of affinity proceeds through the calculation of many terms describing interactions or properties of the ligand-receptor system. These properties are used in multilinear regression analysis to select terms which contribute most strongly to observed affinity. Thus, many more terms are computed than appear in the final functional form. The following sections outline the major phases in deriving an empirical affinity-prediction model: 1) evaluation of pairwise intermolecular interactions, 2) assessment of surface area burial, 3) lattice implementation, 4) calculation of interaction-independent terms, and 5) model refinement.

Evaluation of Pairwise Intermolecular Interactions

Labeling. All ligand, receptor, and solvent atoms are first assigned a chemical label which will be used in assessing interactions. These labels are derived from Sybyl atom types. Each atom receives only one of ten possible labels: hydrophobe (sulfur, phosphorous, silicon, halogens, and non-aromatic carbon atoms not adjacent to a charged atom), aromatic (aromatic carbon and nitrogen atoms), acceptor (hydrogen bond acceptors), donor (hydrogen bond donors), polar (hydrogen bond acceptors and donors), imidazole (nitrogens in imidazole rings), plus (positively charged atoms, not including monatomic cations), minus (negatively charged atoms), water, cation (monatomic cations, *e.g.* Ca^{2+}, Zn^{2+}, Mg^{2+}). Atoms are initially assigned generic labels which are refined by detecting progressively more specific functional groups. Functional groups perceived include ether, aniline, hydroxyl, imidazole, guanidyl, amidine, carboxylate, nitro, sulfoxide, sulfone, sulf(on)ate, and phosph(on)ate. Hydrogens receive the same label as their parent atom. Throughout the labeling procedure, formal charges in accord with physiological pH are assigned to functional groups and split among component atoms (*e.g.* carboxylate oxygens each receive half of a negative formal charge, guanidyl nitrogens each receive one-third of a positive formal charge).

Charge Smoothing. The strength of interaction between a charged group on the ligand and a charged sidechain on a receptor will be dependent on whether the sidechain neighbors other charged residues on the receptor. We introduce such receptor polarization effects to capture qualitatively some aspects of short-range electrostatics. All charged sidechains on the receptor within typical hydrogen-bonding distance of an oppositely charged sidechain are demoted from charged status. That is, the component atoms (for example, carboxylate oxygens of an aspartate sidechain and guanidyl nitrogens and hydrogens of a neighboring arginine sidechain) would be re-assigned an uncharged chemical label indicative of their hydrogen bonding capabilities alone. The neighbor-defining distance is set by the user.

Interaction Evaluation. A matrix, supplied by the user, indicates an interaction type associated with each possible pairing among the ten labels. As an example, donor-acceptor, polar-acceptor, and polar-donor pairings might provide a minimal set of neutral hydrogen bonds and therefore would all assigned the same interaction type. The matrix used in this work is shown in Table 1. Each interaction type is attributed a cutoff distance which defines the interacting step function: an interatomic separation less than this value receives unit contribution, while interatomic separations greater than this value receive zero contribution. Ligand and receptor non-hydrogen atoms are examined pairwise for contributions appropriate to the interaction type defined by their atom labels. The evaluation protocol proceeds in three stages.

The first stage entails assessing interactions among charged and hydrogen bonding atoms. A simple tally of the number of contacts for each interaction type involving these atoms (codes A - F in Table 1) is maintained. Interactions between two charged atoms (codes A, B) are modulated by the product of the formal charges on the component atoms (*e.g.* a carboxylate oxygen interacting with a guanidyl nitrogen would contribute -0.500 × 0.333 = -0.167 interaction units). To extract a number of specific hydrogen bonding terms which might contribute to affinity, more extensive analysis is performed.

Table 1. Interaction matrix.

Numbers represent atom labels and letters indicate interaction types as given in the accompanying key.

Atom Label	code	0	1	2	3	4	5	6	7	8	9	10
unassigned	0	J	J	J	J	J	J	J	J	J	J	J
hydrophobe	1	J	H	H	I	I	I	I	I	I	I	I
aromatic	2	J	H	G	I	I	I	I	I	I	I	I
acceptor	3	J	I	I	D	C	C	E	F	C	E	C
donor	4	J	I	I	C	D	C	F	E	C	F	C
polar	5	J	I	I	C	C	C	E	E	C	E	C
plus	6	J	I	I	E	F	E	B	A	E	B	E
minus	7	J	I	I	F	E	E	A	B	E	A	A
water	8	J	I	I	C	C	C	E	E	C	E	C
cation	9	J	I	I	E	F	E	B	A	E	B	E
imidazole	10	J	I	I	C	C	C	E	A	C	E	C

Interaction code	Interaction type
A	charge-charge attractive
B	charge-charge repulsive
C	hydrogen bond
D	acceptor-acceptor/donor-donor clash
E	charged hydrogen bond
F	charged acceptor-acceptor/donor-donor clash
G	aromatic
H	hydrophobic
I	hydrophobic-polar clash
J	unassigned with anything

For each pair of atoms which could potentially form a hydrogen bond, the interacting geometry is analyzed. The angular dependence of hydrogen bond strength is gauged typically by the deviation from linearity of the two heavy atoms and intervening hydrogen. Unfortunately, crystallographic methods provide no direct information concerning hydrogen position. Although the position of many protein hydrogens is well-defined (*e.g.* amide protons), there is ambiguity, as discussed earlier, around functionalities such as hydroxyl groups. To circumvent this difficulty, hydrogen bonds involving poorly-defined hydrogens are gauged by two angles involving only non-hydrogen atoms. These two angles are computed among X:D-A and D-A:X atoms (D = donor, A = acceptor), where X represents the adjacent heavy-atom providing the best possible angular geometry. There are

Table 2. Angular classes used for evaluation of hydrogen bond geometry.

Class	Angle Bounds (degrees)	Donor Functionalities	Acceptor Functionalities
ideal	0 - 45[a]	amide nitrogen secondary sp^3-nitrogen	
narrow	75 - 165[b]	hydroxyl	secondary sp^2-nitrogen ether carboxylate
wide	75 - 180[b]	primary sp^3-nitrogen primary sp^2-nitrogen	hydroxyl carbonyl phosph(on)ate sulf(on)ate sulfoxide, sulfone

[a] Angle is measured as D:H - A (D = donor, A = acceptor).

[b] Angle is measured as X:D - A or X:A - D (D = donor, A = acceptor, X = adjacent heavy atom).

three classes, “ideal”, “narrow”, and “wide”, which define the angular tolerance about the interacting atom. These angular classes and the functionalities to which they apply are given in Table 2. Geometry is classified as either “good”, “bad”, or “none” according to the angular class of the interacting atom. Angles which fall within the bounds listed in Table 2 are deemed good. Angles which do not deviate more than 15 degrees from the bounds are deemed bad. Both angles must be good for the hydrogen bond geometry to be deemed good overall; if either of the angles is bad, the hydrogen bond geometry is also deemed bad. In the case of angles involving water molecules or monatomic cations where no adjacent non-hydrogen atoms exist, only one angle is used in the geometry analysis. A tally of the number of good hydrogen bonds, bad hydrogen bonds, charged hydrogen bonds (good hydrogen bonds where only one partner is charged), and hydrogen bonds to water is thus kept.

The second stage of the evaluation protocol examines atoms which were not involved in an interaction in the first stage (interaction codes G - I in Table 1). Atoms which participated in hydrogens bonds can not, for example, be penalized at this stage for clashes with hydrophobic atoms. Polar atoms are permitted at most one clash with hydrophobic atoms (code I). Interactions between hydrophobic atoms and aromatic atoms or other hydrophobic atoms (code G) are based on surface area. That is, for every such interacting pair, the solvent accessible surface area of each atom and it’s attached hydrogens, if any, is summed. We add the constraint, however, that the surface area for any atom may be included only once.

In the third and final stage of the evaluation protocol, after all interacting atoms have been paired, terms relating to the collective state of paired or unpaired atoms are assessed. The number of hydrogen bonds left unsatisfied in the complexed state is evaluated. The atomic formal charges buried (where buried is defined as being involved in an interaction) on the part of the ligand and receptor are summed. The number of single bonds buried in the ligand and receptor is also tallied in an effort to gauge a conformational entropy penalty.

Assessment of Buried Surface Area

The use of surface area as a measure for solvation effects has considerable precedent (Hermann, 1972; Chothia, 1974; Reynolds *et al.*, 1974; Eisenberg & McLachlan, 1986; Sharp *et al.*, 1991). We compute the amount of solvent-accessible surface area buried upon complexation by methods described in Appendix A. Buried surface area is subdivided based on atomic label and molecule of origin (receptor or ligand), giving rise to twenty possible terms. Surface area calculations are performed with a 1.0 dot/Å ^{2} density.

Lattice Implementation

Like the force-field score computation, the empirical interaction evaluation can benefit from pre-computing certain terms and storing them on a lattice. As angular information needs to be calculated for classifying hydrogen bonds, gauging hydrogen bond strength using a lattice is difficult. We may still garner significant performance enhancements by simply storing the identify of every receptor atom near to each lattice point. This implementation allows the examination of only *nearby* receptor atoms rather than *all* receptor atoms when assessing the interactions a ligand atom makes with the receptor ("nearby" is defined as the maximum interaction radius, usually about 4.0 Å).

Storing a list of nearby receptor atoms for every lattice point requires a great amount of memory. As Fortran does not support dynamic memory allocation, arrays must be dimensioned at compile-time, not run-time. The need for pre-dimensioning forces sizing of arrays based on worst-case behavior. One can, however, implement a simple innovation to permit dimensioning based on average-case behavior. Rather than storing an array of nearby atoms for each lattice point, one can catenate the list of nearby atoms for all lattice points into one array. Only one array need therefore be pre-dimensioned. By maintaining pointers into this array for lattice point, the receptor atoms near any lattice point are readily retrieved.

The lattice implementation of interaction evaluation bypasses many needless distance calculations, thus greatly speeding up evaluation time. We sacrifice memory for the sake of this efficiency. By altering the storage format, we can cut the high memory requirements roughly in half. Considerably more troublesome than evaluating pairwise interactions on a lattice, however, is the assessment of surface area burial on a lattice. Expedient methods for determining the amount of solvent accessible surface area buried upon complexation using a lattice implementation are described in Appendix A.

Calculation of Interaction-Independent Terms

In addition to the affinity terms which depend on intermolecular interaction, several interaction-independent properties are used in model selection. Ligand volume is included as a regressor, as it might conceivably be proportional to the entropic bonus for removing solvent molecules from the binding pocket. Several alternative representations of ligand conformational entropy are invoked, including the number of rotatable bonds in the ligand and the log of the number of conformations of the unbound ligand. The number of ligand conformations is estimated in three ways: theoretically, based simply on a factorial expansion of rotatable bonds; energetically, using the Sybyl systematic search feature; and rule-based, using Chem-X (Chemical Design Ltd.). The log of ligand molecular weight has been shown to correlate with the loss of rotational and translational entropy upon complexation (Williams *et al.*, 1991). Finally, because experimentally determined structures frequently are obtained under conditions which differ from those of the affinity assay, the structure pH and assay pH are used as additional regressors.

Model Refinement

All of the terms described are introduced into the multilinear regression package to be outlined shortly. Only regressors which significantly contribute to a proposed model should be included; the optimal model represents a tradeoff between predictivity and the number of terms involved. The best models which can be derived will therefore result from a subset and/or combination of the many descriptors characterized above. The combination of terms avoids overly complex representations of minor effects and simplifies interpretations. True orthogonalization of all terms is difficult, so introducing contention between competing descriptions may illuminate which better represents the desired interaction. Descriptors have been defined to capture forces deemed important for ligand binding. It is important to bear in mind, however, that whether the terms in a proposed model sustain the *intended* physical meaning is a matter of considerable debate.

The manual optimization of adjustable parameters such as the interaction cutoff distances and hydrogen bonding angles was carried out by iteratively performing the interaction evaluation and examining models with regression analysis. Note that in this process, the parameters to be optimized were assumed to behave independently, which clearly may be a questionable assumption. Improved methods for optimization at this stage, however, can only increase model performance. The most time-consuming step of the interaction evaluation is reading receptor and ligand coordinates and surface areas off the disk. To facilitate the optimization cycle, the interaction evaluation program was rewritten to sacrifice memory for efficiency. By reading all data into memory only once and entering a command mode, the user can make changes to the parameter files, evaluate interactions, and perform regression analysis, all without exiting the program and having to re-read data off disk. While seemingly a trivial innovation, the amount of memory required for storing atomic, coordinate, and surface area data for receptor and ligand of more than one hundred complexes is tremendous; conservative methods of data storage must be devised.

Regression Analysis: GREMLIN

The Fortran program GREMLIN (an anagram for Multiple LINear REGression) has been written to facilitate data analysis. This program computes linear regressions against the observed data for all possible subsets of input regressors using Gaussian elimination. Thus, given ten input descriptors, 2^{10} or 1024 regressions result (each regressor has the option of being included or excluded). The use of a constant is optional. Other features of the analysis package include sorting of data by various metrics, crude graphing abilities, versatile methods for regression filtering and retrieval, cross-validation, analysis of variance, and computation of Kendall's tau statistic *.

To run GREMLIN, the user provides an input file containing all of the calculated and observed data for each system and for each independent variable, a short description and optionally intended coefficient sign. Regressions which generate coefficients with signs different from those specified are tagged and can be filtered out easily. This feature is useful for imposing preconceived notions upon a particular term (one might not unreasonably insist, for example, that burial of nonpolar surface area be favorable; regressions which fortuitously indicate the opposite can be discarded). After statistics have been computed for

* Kendall's τ statistic is defined (Press *et al.*, 1988) as $\frac{\#\,correct - \#\,incorrect}{n(n-1)/2}$. For all $n(n-1)/2$ pairs of n observed data values, the observed ranking is compared with the predicted ranking for the two values. If the ranking is the same, *#correct* is incremented; if the ranking is opposite, *#incorrect* is incorrect. τ ranges from -1.0 (all rankings opposite) to +1.0 (all rankings correct).

all regressions, the user may page through sorted regressions sequentially or search for regressions utilizing particular regressors or number of regressors. Analyzing the variables involved in the best models as judged by r^2-value quickly lends insight as to which descriptors are the most useful. To assess whether models will have any value outside the data set, model predictivity is gauged by leave-one-out cross-validation. In this process, each observation (in this case, receptor-ligand complex) is sequentially left out of the regression and predicted based on the remaining observations. Cross-validation measures how well a model predicts data not used in model construction. In our experience, a cross-validated r^2 less than 90% of the fitted r^2 indicates bias of the model towards particular observations, warning of limited predictivity. Kendall's τ statistic gauges the ability to *rank* data rather than the ability to reproduce it. For automated structure-based strategies such as molecular docking and *de novo* design, the capacity to rank output is often sufficient to guide discovery efforts. Following Jain *et al.* (1994), we report τ for each model as a useful metric for activity prediction methods.

Hardware

All calculations were carried out on a Silicon Graphics 200MHz R4400 Indigo2 workstation (Silicon Graphics, Inc., Mountain View, CA) with 128Mb of physical memory.

RESULTS

Preliminary results toward empirical scoring schemes for automated structure-based design strategies follow. We present performance over a data set consisting of 103 receptor-ligand complexes for the molecular mechanics scoring function currently in use in the

DOCK program (Meng *et al.*, 1992) and for an empirical scoring scheme. Only 103 of the 126 water-containing complexes were included due to parameterization difficulties encountered with setting up molecular mechanics evaluation.

The poor performance of the force-field score in estimating absolute binding affinity is revealed in Table 3. The fitted and cross-validated r^2 values are very low and reflect standard errors in excess of two log units. Model A, which represents the typical implementation of the force-field score in the DOCK package, is illustrated in Figure 1. Virtually the same binding energy is predicted regardless of what is observed experimentally. Allowing electrostatics and van der Waals terms to vary independently improves the model (B), in effect implying that a $4r$ dielectric is sub-optimal. The four-fold scaling of van der Waals to electrostatics coefficients indicates that perhaps a $16\ r$ dielectric would be more appropriate. This tendency to reduce the contribution of electrostatics argues that our characterization of electrostatic interactions is insufficient. Note that electrostatics on its own is completely useless in predicting affinity (a negative q^2 is in fact possible).

Table 3. Force-field score regression models.

	Term Coefficients[a]				Regression Statistics[b]					
Model	**vdw**	**ele**	**total ff**	**const**	**n**	**r^2**	**s**	**F**	**q^2**	**τ**
A			-0.0417	4.0534	2	0.191	2.187	23.8	0.131	0.305
B	-0.0852	-0.0205		3.1406	3	0.316	2.020	23.1	0.252	0.409
C	-0.0823			3.5832	2	0.283	2.058	39.9	0.250	0.401
D		-0.0143		5.7640	2	0.016	2.411	1.7	-0.038	0.017

[a] All coefficients are in pK_i units. Coefficients indicate contributions to each regression model by the following terms: *vdw* = van der Waals component of the force-field score; *ele* = electrostatic component of the force-field score; *total ff* = sum of van der Waals and electrostatics components; *const* = constant term in regression.

[b] Statistics are as follows: *n* = number of adjustable parameters; r^2 = fitted r^2; *s* = standard error of fitted model (log units); *F* = fitted model significance given by F-ratio; q^2 = cross-validated r^2; τ = Kendall's tau statistic.

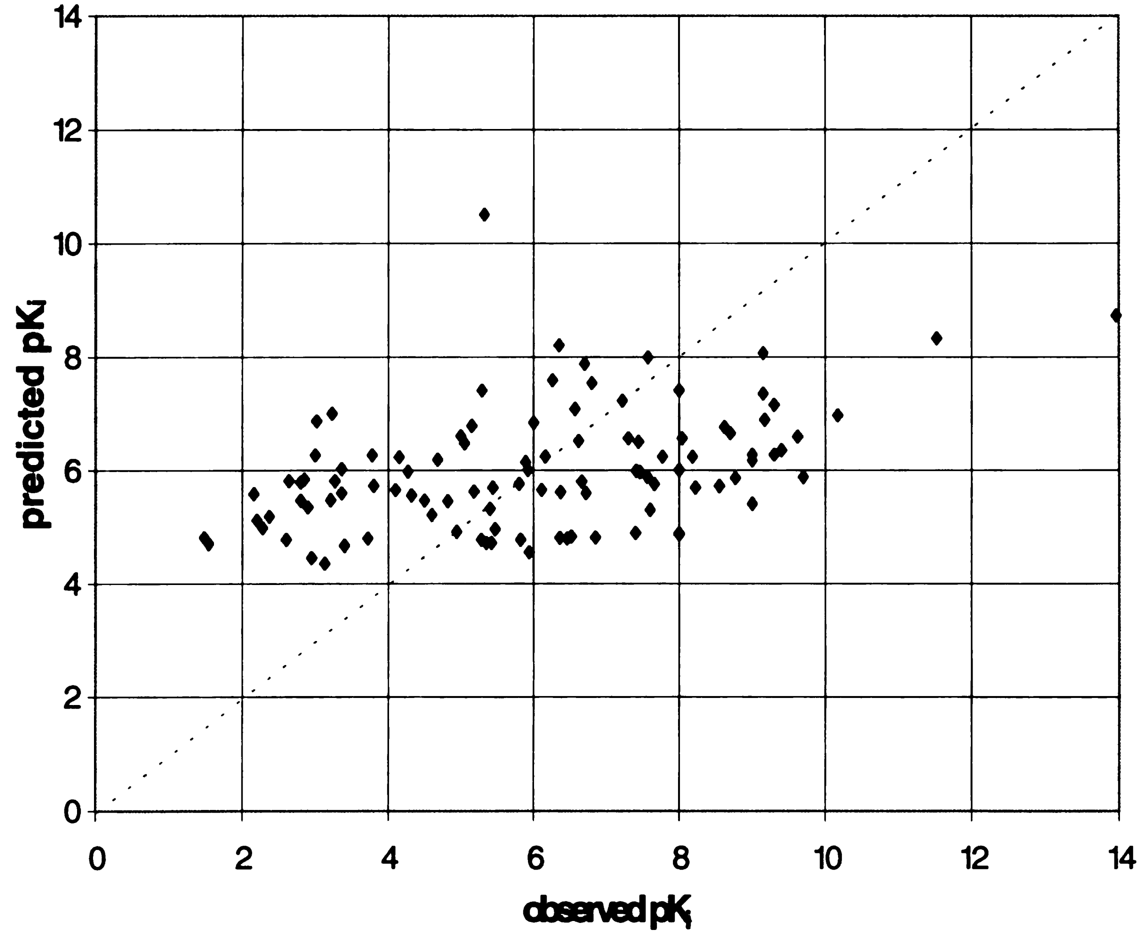


Figure 1. Binding energy predictions: minimized DOCK force-field score.
Predicted *vs.* observed affinity is plotted for 103 receptor-ligand complexes using model A (Table 3). The diagonal line has unit slope and represents ideality. r ^{2}=0.191, q^2=0.131, s=2.19 (2.97 kcal/mol), 2 adjustable parameters. Force-field scores were optimized with the quasi-Newton method described in Meng *et. al.* (1993).

The ability to reproduce absolute binding affinities with a preliminary empirical scheme is considerably better. Table 4 presents models involving the use of eight descriptors which have been found to contribute to the best empirical regression solutions. Model E utilizes all of these descriptors, while subsequent models (F - M) illustrate the effect of dropping each term independently. The amount of non-polar solvent accessible surface area of the ligand buried upon complexation is the most significant term, while the number of good-geometry hydrogen bonds and aromatic contacts also play a large role. The least significant of the terms shown in Table 4 is the entropic penalty measured by the number of receptor single bonds immobilized upon complexation - note that the sign of the coefficient varies when other terms are left out. The number of formal charges buried is also a minor contributor, yet consistently has appeared in the best regression models. The effect of leaving out both of these terms is given in model N, which, because it involves the fewest number of terms but maintains predictivity, garners the highest F significance statistic. Our experience with this data set and thousands of resultant regression models leads us to believe that the number of buried formal charges is a contribution worth retaining. Thus, we have settled on regression M, with a standard error of 1.7 kcal/mol, as the working model. Predicted versus observed affinities obtained using this model are plotted in Figure 2. The statistical significance and 95% confidence intervals for terms in model M are excellent, as Table 5 indicates.

Table 4. Empirical scoring scheme regression models.

	Term Coefficients[a]									Regression Statistics[b]					
Model	lnpSA	burchg	gChg	gHB	bHB	wHB	arom	rfroz	const	n	r^2	s	F	q^2	τ
E	0.0073	-0.2885	1.0900	0.2409	-0.6185	-0.2047	0.3383	-0.0801	3.9952	9	0.758	1.239	36.9	0.713	0.679
F		-0.4028	1.2124	0.2239	-0.6987	-0.2699	0.3108	*0.1093*	5.1629	8	0.564	1.655	17.6	0.466	0.548
G	0.0077		0.6660	0.1998	-0.6594	-0.2194	0.3322	-0.0637	3.4279	8	0.730	1.302	36.8	0.689	0.655
H	0.0077	*0.0548*		0.2271	-0.6405	-0.1717	0.3472	-0.0834	3.3544	8	0.673	1.434	27.9	0.616	0.627
I	0.0070	-0.0468	0.9896		-0.4650	-0.1771	0.3041	*0.0007*	4.4946	8	0.621	1.544	22.2	0.558	0.579
J	0.0077	-0.3349	1.1204	0.2113		-0.1902	0.3218	-0.0893	3.7145	8	0.690	1.395	30.2	0.636	0.643
K	0.0080	-0.3316	0.9708	0.2271	-0.5812		0.3354	-0.1179	3.6266	8	0.700	1.372	31.7	0.650	0.642
L	0.0069	-0.2636	1.1344	0.2171	-0.5592	-0.2006		-0.0248	3.9550	8	0.624	1.537	22.5	0.549	0.579
M	0.0061	-0.2596	1.0966	0.2167	-0.6327	-0.2275	0.3145		3.8302	8	0.745	1.267	39.6	0.701	0.668
N	0.0067		0.7061	0.1835	-0.6676	-0.2367	0.3133		3.3400	7	0.722	1.316	41.5	0.684	0.652

[a] All coefficients are in pK_i units. Coefficients indicate contributions to each regression model by the following terms: *lnpSA* = ligand nonpolar solvent accessible surface area buried upon complexation; *burchg* = number of formal charges on receptor and ligand buried upon complexation; *gChg* = number of salt bridges formed with good hydrogen bond geometry; *gHB* = number of hydrogen bonds formed with good geometry; *bHb* = number of hydrogen bonds formed with bad hydrogen bond geometry; *wHB* = number of hydrogen bonds to water molecules; *arom* = number of aromatic-aromatic contacts; *rfroz* = number of single bonds on the receptor immobilized upon complexation; *const* = constant term in regression. For definitions of "good" and "bad" geometry, see text.

[b] Statistics are as follows: *n* = number of adjustable parameters; r^2 = fitted r^2; *s* = standard error of fitted model (log units); *F* = fitted model significance given by F-ratio; q^2 = cross-validated r^2; τ = Kendall's tau statistic.

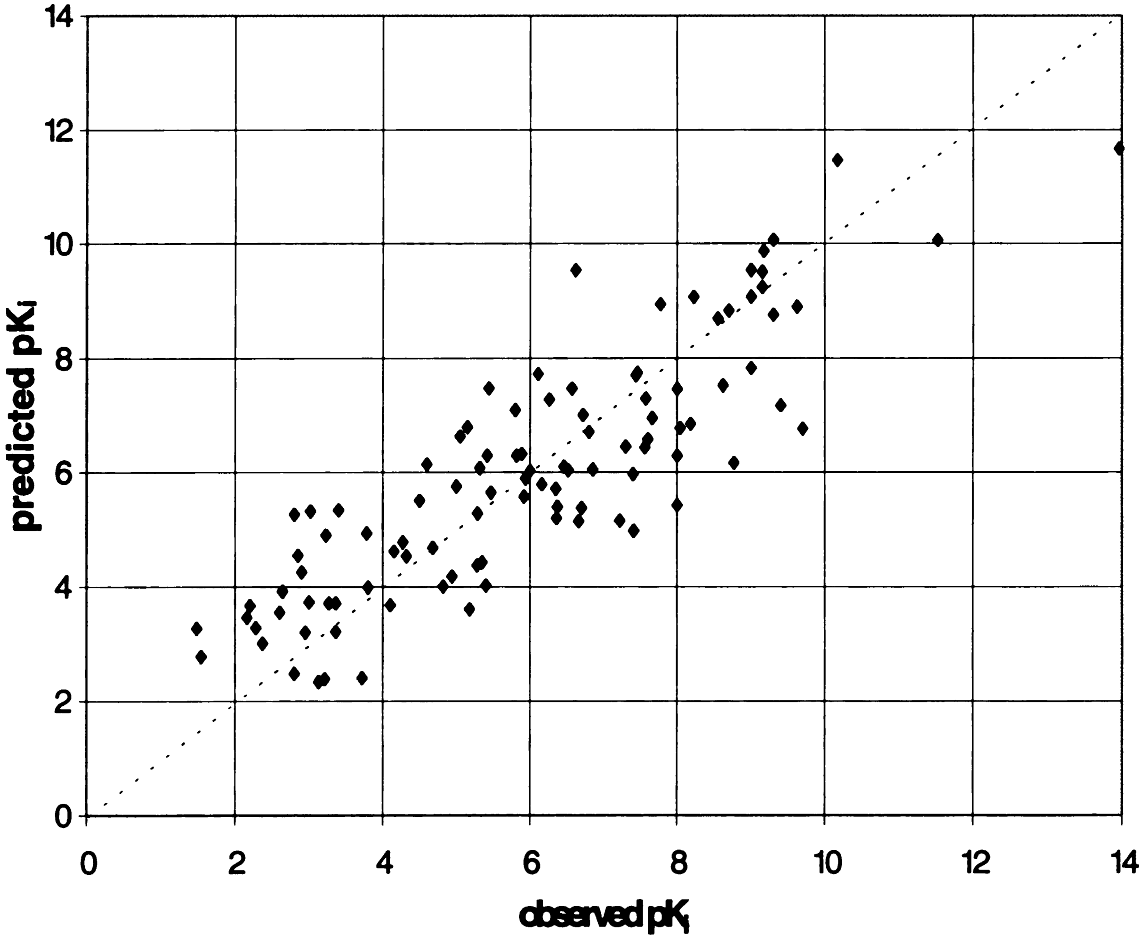


Figure 2. Binding energy predictions: empirical scheme.

Predicted vs. observed affinity is plotted for 103 receptor-ligand complexes using model M (Table 4). The diagonal line has unit slope and represents ideality. r^2=0.745, q^2=0.701, s=1.27 (1.72 kcal/mol), 8 adjustable parameters.

Table 5. Significance of coefficients for empirical regression model M (Table 4).

Term[a]	Coefficient	Standard Error	P-value
lnpSA	0.0061	0.0007 (11.0 %)	1.4×10^{-14}
burchg	-0.2596	0.0884 (34.1 %)	4.2×10^{-3}
gChg	1.0966	0.1929 (17.6 %)	1.4×10^{-7}
gHB	0.2167	0.0319 (14.7 %)	9.7×10^{-10}
bHB	-0.6327	0.1227 (19.4 %)	1.4×10^{-6}
wHB	-0.2275	0.0429 (18.8 %)	7.3×10^{-7}
armarm	0.3145	0.0467 (14.9 %)	1.3×10^{-9}
constant	3.8302	0.3933 (10.3 %)	6.0×10^{-16}

[a] Term descriptions are as given in Table 4.

DISCUSSION

The juxtaposition of analyses differing greatly in the number of adjustable parameters, while statistically reconcilable, nevertheless merits a healthy suspicion. Our justification for contrasting the two methods is only to evaluate the current DOCK scheme and highlight direction for improvement. We do not imply that molecular mechanics is fundamentally flawed, but rather that it is incomplete. The omission of entropic and solvation effects undeniably contributes to the poor performance of the force-field score at predicting binding affinities across diverse receptor-ligand complexes. The empirical regression illustrated in Figure 2, in contrast, does include terms intended to capture both enthalpic and entropic effects. Moreover, this scheme, unlike the force-field score, has been derived explicitly to reproduce binding affinities, so it's improved performance is not unexpected. The evaluation of the empirical scheme, while not as efficient as the force-field score, is fairly rapid. About ten evaluations per second are possible, but as the current code computes many additional terms not used in the regression model, this rate could conceivably be increased by an order of magnitude.

The descriptors comprising the empirical model can be equated with physical principles. We prefer to interpret the model in terms of solubilities in different media, *i.e.* an aqueous phase and a receptor-bound phase. Every molecule will have its affinity for a receptor modulated by one or more terms. Functionalities on the ligand which hydrogen bond with water, for example, will increase solubility in the aqueous phase, thus decreasing affinity for the receptor (the negative *wHB* term). Charged moieties will also favor the aqueous state, unless a salt bridge can be made with the receptor (negative *burchg* and positive *gChg* terms, respectively). The formation of poor-quality hydrogen bonds between the ligand and the receptor will once again favor the solvated state (negative *bHB* term), where hydrogen bonds can optimally be satisfied. However, should good-quality hydrogen bonds between ligand and receptor be made, these are likely to be stronger than hydrogen bonds to water due to full unit occupancies of receptor atoms, thus favoring a bound configuration (positive *gHB*). The additional favorably entropy of displacing water molecules on the formation of receptor-ligand hydrogen bonds may contribute to the *gHB* term. The removal of hydrophobic surface area from water must entropically also be favorable - this is reflected by the positive *lnpSA* term [two notes regarding this term deserve mention: first, the coefficient is much smaller than the others because this term is in units of square Ångstroms of surface area; and second, the sum of receptor and ligand nonpolar surface areas is generally preferable, but removing the receptor portion greatly simplifies and accelerates the computations (see Appendix A)]. Finally, the placement of aromatic groups on the ligand adjacent to similar groups on the receptor results in a significant decrease in aqueous solubility.

The empirical scheme presented here compares favorably with similar investigations by other researchers. Böhm (1994a) reported a standard error of 1.38 log units and cross-validated r^2 of 0.696 for a set of 45 complexes using five adjustable parameters; Marshall (personal communication) reported a standard error of 1.15 log units and cross-validated r^2 of 0.72 for a set of 52 complexes using in excess of ten parameters through partial-least-squares (PLS) analysis. Our model with a cross-validated r^2 of 0.701 and standard error of 1.27 log units with only eight adjustable parameters over a data set consisting of more than 100 complexes is very encouraging. Our analysis finds strength in a calibration data set more than twice as large as any yet reported. In contrast to the work of Böhm (1994a) and of G.R. Marshall (personal communication), our data set is sparsely populated with instances of numerous ligands binding to the same receptor. It is our opinion that this introduces some bias with respect to the types of interactions being represented and reduces the overall complexity of the problem as it applies to screening molecular databases. Clearly, both approaches are valid and informative, but we prefer to stress diversity in our analysis. With a compilation of nearly 150 complexes, we now are afforded the luxury of subdividing the data set into distinct structural classes. For example, an analysis of all peptidic ligands or of all hydrophobic ligands permits the isolation of specific molecular features which are contributing to interaction strength with different receptors. Conversely, it is important that a proposed model perform well in reproducing trends amongst different ligands binding to the same receptor.

In deriving scoring functions, it is crucial to consider interactions relative to a solvated medium, not a vacuum. For example, we believe that, as a large fraction of hydrogen bonds are satisfied in aqueous solution (Kuntz, 1971; Williams *et al.*, 1991), we should penalize for hydrogen bonds *lost* upon complexation (Fersht *et al.*, 1985) rather than reward for hydrogen bonds "gained" (Jorgensen, 1989; Sali *et al.*, 1991). The use of water information has and will play a critical role in obtaining a complete representation of molecular recognition events. Despite much work on the prediction of water binding sites (Danziger & Dean, 1989; Pitt *et al.*, 1993), cleverness will be required to rapidly determine sites for an optimal arrangement of receptor-ligand water-mediated hydrogen bonds.

We have completed a reasonable first step toward rapid and accurate empirical affinity prediction schemes for use in structure-based design strategies. Many issues remain to be resolved, leaving much room for future work. The predictivity of proposed models must be verified on data not included in the calibration set. Leave-one-out cross-validation is a helpful diagnostic tool for monitoring predictivity, but its usefulness becomes limited when the data set reaches the size it has here. Larger test sets of order twenty complexes will be required to obtain practical estimates of predictive ability.

The incorporation of empirical schemes into molecular docking and *de novo* design programs introduces additional complexities. Because automated design strategies generate many poor molecular arrangements, empirical schemes must be appropriately parameterized for negative interactions - steric clashes and neighboring like charges, for example. This is difficult due to the under-representation of such effects in experimentally observed structures. Our regression model also presupposes information regarding water structure (a significant contribution: term *wHB* in Table 4). This vastly complicates the evaluation phase when *each* of thousands of putative ligand-receptor configurations must be solvated. A further complication arises because empirical scoring schemes are calibrated against experimentally observed configurations. Most, if not all, configurations produced by automated methods deviate from the observed mode, so the ability to retrieve this mode from a deviant configuration must be manifest in the scoring scheme. One might envision calibrating the scoring function against not just one observed configuration of the ligand, but against several which differ slightly in translation and rotation. This might introduce sufficient softness in the evaluation scheme so as to permit recognition of a sub-optimal configuration as the observed binding mode. Finally, we propose the following experiment as a practical measure of any scoring function's utility for molecular docking methods. For a receptor-ligand complex data set such as that described in this work, extract all ligands and create a structural database. To increase the stringency of the test, this ligand database can be supplemented with "random" compounds from one of the many commercially available structural databases. The ligand database is then docked against each receptor in the data set, saving the best-scoring ligand. The evaluation scheme which most often pairs a receptor with its cognate ligand, preferentially to all other ligands in the database, is likely to be a useful scoring method for molecular docking applications.

Finally, we remark that even the best scoring scheme will not achieve perfect accuracy. Resource-intensive calculations such as free energy perturbation rarely achieve accuracies much better than 1 kcal/mol (Beveridge & DiCapua, 1989; Kollman & Merz, 1990), and moreover, experimental affinity determinations are error-prone and subject to variability in method of ascertainment. However, rapid evaluation methods capable of predicting binding affinities to within 1 to 2 kcal/mol will prove immensely useful in structure-based drug design applications.

CONCLUSIONS

The greatest hindrance to structure-based drug design is the inability to accurately and consistently estimate the affinity of ligands for a receptor. There are nearly as many ways of assessing molecular interactions as there are design strategies. This study presents the development of an empirical scoring scheme for use in automated design strategies with emphasis on robustness over structurally diverse molecular arrangements, accuracy in absolute binding affinity prediction, and speed of evaluation. A model, calibrated against a complex set of diverse structural data, has been derived using effective interaction descriptions and statistical analysis to reproduce observed binding affinities to within 1.7 kcal/mol. This model performs considerably better than the molecular mechanics function used in the DOCK molecular docking suite. Interaction evaluation methods which manifest both enthalpic and entropic contributions to binding affinity will display great potential in drug discovery efforts.

ACKNOWLEDGMENTS

We thank Todd Ewing and Peter Kollman for insightful discussions, and Jonathan Keske and Scott Dixon for an initial list of affinity data. This work was funded through NIH grants GM-31497 (I.D.K.) and GM-39552 (G. L. Kenyon, principal investigator). D.A.G. gratefully acknowledges support from an ACS Division of Medicinal Chemistry Fellowship. We thank Tripos Associates for generously making the SYBYL package available to us.

REFERENCES

Åqvist, J.; Medina, C.; Samuelsson, J.-E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385-391.

Baker, E.N.; Hubbard, R.E. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **1984**, *44*, 97-179.

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

Beveridge, D.L.; DiCapua, F.M. Free energy via molecular simulation. Application to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431-492.

Blaney, J.M.; Dixon, J.S. A good ligand is hard to find: Automated docking methods. *Persp. Drug. Disc. Des.* **1993**, *1*, 301-319.

Bohacek, R.S.; McMartin, C. Definition and display of steric, hydrophobic, and hydrogen-bonding properties of ligand binding sites in proteins using Lee and Richards accessible surface: Validation of a high-resolution graphical tool for drug design. *J. Med. Chem.* **1992**, *35*, 1671-1684.

Bohacek, R.S. and McMartin, C. Multiple highly diverse structures complementary to enzyme binding sites: Results of extensive application of a de novo design method incorporating combinatorial growth. *J. Am. Chem. Soc.* **1994**, *116*, 5560-5571.

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

Böhm, H.-J. On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three-dimensional structure. *J. Comput.-Aided Mol. Design* **1994b**, *8*, 623-632.

Cherfils, J.; Janin, J. Protein docking algorithms: simulating molecular recognition. *Curr. Opin. Struct. Biol.* **1993**, *3*, 265-269.

Chothia, C. Hydrophobic bonding and accessible surface area in proteins. *Nature* **1974**, *248*, 338-339.

Clark, T. *A Handbook of Computational Chemistry*; Wiley-Interscience: New York, 1985.

Danziger, D.J.; Dean, P.M. Automated site-directed drug design: a general algorithm for knowledge acquisition about hydrogen-bonding regions at protein surfaces. *Proc. R. Soc. Lond. B.* **1989**, *236*, 101-113.

Eisenberg, D.; McLachlan, A.D. Solvation energy in protein folding and binding. *Nature* **1986**, *319*, 199-203.

Fersht, A.R.; Shi, J.; Knill-Jones, J.; Lowe, D.M.; Wilkinson, A.J.; Blow, D.M.; Brick P.; Carter, P.; Waye, M.M.Y.; Winter, G. Hydrogen bonding and biological specificity analysed by protein engineering. *Nature* **1985**, *314*, 235-238.

Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219-3288.

Gasteiger, J.; Marsili, M. *Organ. Magn. Reson.* **1981**, *15*, 353.

Gerber, P.R.; Mark, A.E.; van Gunsteren, W.F. An approximate but efficient method to calculate free energy trends by computer simulation: Application to dihydrofolate reductase - inhibitor complexes. *J. Comput.-Aided Mol. Design* **1993**, *7*, 305-323.

Good, A.C.; Mason, J.S. Three-dimensional structure database searches. *Rev. Comp. Chem.* **1995**, in press.

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

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

Gschwend, D.A.; Good, A.C.; Kuntz, I.D. Molecular docking towards drug discovery. *J. Mol. Recogn.* **1995**, in press.

Guida, W.C. Software for structure-based drug design. *Curr. Opin. Struc. Biol.* **1994**, *4*, 777-781.

Hermann, R.B. Theory of hydrophobic bonding. II. The correlation of hydrocarbon solubility in water with solvent cavity surface area. *J. Phys. Chem.* **1972,** *76,* 2754-2759.

Holloway, M.K.; Wai, J.M.; Halgren, T.A.; Fitzgerald, P.M.D.; Vacca, J.P.; Dorsey, B.D.; Levin, R.B.; Thompson, W.J.; Chen, L.J.; deSolms, S.J.; Gaffin, N.; Ghosh, A.K.; Giuliani, E.A.; Graham, S.L.; Guare, J.P.; Hungate, R.W.; Lyle, T.A.; Sanders, W.M.; Tucker, T.J.; Wiggins, M.; Wiscount, C.M.; Woltersdorf, O.W.; Young, S.D.; Darke, P.L.; Zugay, J.A. *A priori* prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *J. Med. Chem.* **1995,** *38,* 305-317.

Horton, N.; Lewis, M.L. Calculation of the free energy of association for protein complexes. *Protein Sci.* **1992,** *1,* 169-181.

Jain, A.N.; Koile, K.; Chapman, D. Compass: Predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994,** *37,* 2315-2327.

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

Jorgensen, W.L. Free energy calculations: A breakthrough for modeling organic chemistry in solution. *Acc. Chem. Res.* **1989,** *22,* 184-189.

Kollman, P.A.; Merz, K.M., Jr. Computer modeling of the interactions of complex molecules. *Acc. Chem. Res.* **1990,** *23,* 246-252.

Krystek, S.; Stouch, T.; Novotny, J. Affinity and specificity of a serine endopeptidase - protein inhibitor interactions: Empirical free energy calculations based on x-ray crystallographic structures. *J. Mol. Biol.* **1993,** *234,* 661-679.

Kuntz, I.D. Hydration of macromolecules. *J. Am. Chem. Soc.* **1971,** *92,* 514-516.

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

Kuntz, I.D. Structure-based strategies for drug design and discovery. *Science* **1992,** *257,* 1078-1082.

Kuntz, I.D.; Meng, E.C.; Shoichet, B.K. Structure-based molecular design. *Acc. Chem. Res.* **1994,** *27,* 117-123.

Lewis, R.A.; Leach, A.R. Current methods for site-directed structure generation. *J. Comput.-Aided Mol. Design.* **1994,** *8,* 467-475.

Lybrand, T.P. Ligand-protein docking and rational drug design. *Curr. Opin. Struc. Biol.* **1995,** *5,* 224-228.

Marsili, M.; Gasteiger, J. *Croat. Chem. Acta* **1980,** *53,* 601.

Meng, E.C.; Gschwend, D.A.; Blaney, J.M.; Kuntz, I.D. Orientational sampling and rigid-body minimization in molecular docking. *Proteins* **1993,** *17,* 266-278.

Meng, E.C.; Shoichet, B.K.; Kuntz, I.D. Automated docking with grid-based energy evaluation. *J. Comp. Chem.* **1992,** *13*, 505-524.

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

Press, W.H.; Flannery, B.P.; Teukolsky, S.A.; Vetterling, W.T. *Numerical Recipes in C*; Cambridge University Press: Cambridge, 1988.

Reynolds, J.A.; Gilbert, D.B.; Tanford, C. Empirical correlation between hydrophobic free energy and aqueous cavity surface area. *Proc. Natl. Acad. Sci. U.S.A.* **1974,** *71*, 2925-2927.

Sali, D.; Bycroft, M.; Fersht, A. In Techniques in Protein Chemistry; Villafranca, J.J., Ed.; Academic Press: New York, NY, 1991; Vol. 2, pp. 295-303.

Sharp, K.A.; Nicholls, A.; Fine, R.F.; Honig, B. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science* **1991,** *252*, 106-109.

Shoichet, B.K.; Kuntz, I.D. Protein docking and complementarity. *J. Mol. Biol.* **1991,** *221*, 327-346.

Shoichet, B.K.; Bodian, D.L.; Kuntz, I.D. Molecular docking using shape descriptors. *J. Comp. Chem.* **1992,** *13*, 380-397.

Straatsma, T.P.; McCammon, J.A. *Annu. Rev. Phys. Chem.* **1992,** *43*, 407-435.

Thanki, N.; Thornton, J.M.; Goodfellow, J.M. Distributions of water around amino acid residues in proteins. *J. Mol. Biol.* **1988,** *202*, 637-657.

Tintelnot, M.; Andrews, P. Geometries of functional group interactions in enzyme-ligand complexes: Guides for receptor modeling. *J. Comput.-Aided Mol. Design* **1989,** *3*, 67-84.

Vajda, S.; Weng, Z.; Rosenfeld, R.; DeLisi, C. Effect of conformational flexibility and solvation on receptor-ligand binding free energies. *Biochemistry* **1994,** *33*, 13977-13988.

Walls, P.H.; Sternberg, M.J.E. A new algorithm to model protein-protein recognition based on surface complementarity: Application to antibody-antigen docking. *J. Mol. Biol.* **1992,** *228*, 277-297.

Warshel, A.; Tao, H.; Fothergill, M.; Chu, Z.-T. Effective methods for estimation of binding energies in computer-aided drug design. *Israel J. Chemistry* **1994,** *34*, 253-256.

Williams, D.H.; Cox, J.P.L.; Doig, A.J.; Gardner, M.; Gerhard, U.; Kaye, P.T.; Lal, A.R.; Nicholls, I.A.; Salter, C.J.; Mitchell, R.C. Toward the semiquantitative estimation of binding constants: Guides for peptide-

peptide binding in aqueous solution. *J. Am. Chem. Soc.* **1991,** *113,* 7020-7030.

Wilson, C.; Mace, J.E.; Agard, D.A. Computational method for the design of enzymes with altered substrate specificity. *J. Mol. Biol.* **1991,** *220,* 495-506.

Wodak, S.; Janin, J. Computer analysis of protein-protein interactions. *J. Mol. Biol.* **1978,** *124,* 323-342.

Appendix A.

Evaluation of Buried Solvent-Accessible Surface Area Using a Lattice

BACKGROUND

Biological processes occur in an aqueous environment. The energetics of solvent-solute interactions determine the behavior of molecular associations. The most common characterization of the extent to which a molecule can interact with solvent involves the solvent accessible surface area (SASA) (Lee & Richards, 1971). The SASA has been defined succinctly by Chothia (1975): "For a given atom it is defined as the area over which the center of a water molecule can be placed while retaining van der Waals' contact with that atom and not penetrating any other atom." Accessible surface areas correlate well with hydrophobic free energies (Hermann, 1972; Chothia, 1974; Reynolds *et al.*, 1974; Eisenberg & McLachlan, 1986; Sharp *et al.*, 1991) and merit exploration for incorporation into evaluation functions for structure-based design strategies. While the construction of a SASA is easy, the rapid evaluation of changes in area between two states of a molecular system is not trivial. To be useful for automated design strategies, this evaluation must be capable of being performed many times per second - herein lies the challenge. I here describe an efficient lattice-based method for computing the amount of solvent accessible area buried upon molecular complexation. This method may make a useful addition for addressing solvation issues within the existing molecular mechanics force-field (Meng *et al.*, 1992) or within empirically derived scoring schemes such as that described in Chapter 5.

METHOD

Surface area generation

A Fortran program, entitled "access," has been written to generate surface areas of molecules or SPHGEN (Kuntz *et al.*, 1982) sphere clusters. The algorithm is simple. For each atom, a net of evenly spaced points is placed on a sphere having a radius equal to the sum of the atom's van der Waals' radius and the solvent radius. After laying such a net around each atom, all points internal (closer than a distance equal to the sum of an atom's van der Waals' radius and the probe radius) are removed. Each point is assigned an associated surface area. The computation is reasonably quick: to generate a surface having a density of 5.0 dots/Å^2 using a solvent radius of 1.4Å requires only about 0.1 sec (SGI 200MHz R4400 Indigo2) for a typical ligand and about 10 sec for a 20kD protein. The program accepts both PDB and Sybyl MOL2 (Tripos Associates, St. Louis, MO, 63117) formats and can be instructed to use either Sybyl, Amber (Weiner *et al.*, 1984), or MS (Connolly, 1983) atomic radii. Solvent radius and surface density are also under user control. The output surface format is the same as that of the UCSF MS implementation.

Evaluation of buried surface area on a lattice

Input

Coordinates and pre-generated solvent accessible surfaces are required for both ligand and receptor (practical implementations, such as that described in Chapter 5, can generate the ligand surface on-the-fly without significant penalty). A density of 1.0 dots/Å^2 is sufficiently dense for excellent results. MS atomic radii were used for all calculations in this work.

Lattice Construction

For every point on a lattice spanning the region of interest (see Meng *et al.*, 1992), a list is stored of all nearby receptor surface points. "Nearby" is defined as within the a distance equal to the sum of ligand probe atom and solvent radii. A lattice is saved for several ligand probe atom radii, generally ranging from 1.2 to 2.2 Å in 0.2 Å increments to cover the common range of atomic radii. Thus, using these parameters, six lattices would be constructed, each containing a list of nearby receptor surface points but varying in the radius of the probing atom. As more than one lattice is retained in memory, the density of the lattice must be reduced. A lattice resolution of 1.0 Å is sufficient. Implementing lists as described above is very memory intensive. I use the method detailed under "Lattice Implementation" in Chapter 5 to reduce memory requirements.

An additional lattice, the occlusion grid, is constructed for the receptor. This lattice is of higher density (typically 0.3 Å resolution) and contains binary values indicating whether lattice points lie inside or outside of the receptor SASA.

Lattice-Based Burial Evaluation

The evaluation of buried surface area is now simple, given the appropriately constructed lattices. To assess surface area burial for an arbitrary configuration and/or conformation of the ligand, the ensuing protocol is followed.

Receptor SASA burial. For every atom in the ligand, the lattice with the probe radius nearest to the radius of the atom is employed. The lattice point nearest the ligand atom is located, and the list of all nearby receptor surface points is extracted. All nearby receptor surface points are then considered buried by this ligand atom. After processing all ligand atoms, the surface areas of all receptor points that were buried by this orientation of the ligand are summed, giving rise to the amount of receptor surface area buried upon complexation.

Ligand SASA burial. For each ligand surface point, a simple check on the occlusion grid reveals whether the point (and its associated surface area) lie within the SASA of the receptor. If so, this portion of ligand surface area is considered buried. A summation over all buried ligand surface points gives rise to the amount of ligand surface area buried upon complexation.

"Actual" Burial Evaluation

The amount of surface area buried upon complexation is equal to the sum of the surface areas of the receptor and ligand minus the surface area of the complex. This total area of burial can be partitioned into area lost by the receptor and by the ligand. To gauge the accuracy of the lattice-based evaluation of buried surface area, results were compared with answers obtained by computing accessible surface areas for the receptor, ligand, and receptor-ligand complex using the access program. Note that these "actual" buried areas use exact distances but do *not* represent analytical surface area calculations.

Hardware

Calculations were performed on a SGI 150MHz R4400 Challenge with 256Mb of physical memory.

RESULTS

Table 1 compares three methods for computing surface area burial for each of seven test receptor-ligand systems: actual calculations using a high and low density surface and lattice-based using a low density surface. Differences between the use of high and low density surfaces in actual assessments of surface areas amount to only on the order of 1%. For a fair comparison, the low-density lattice-based evaluation is compared with the low-density actual calculation. The errors introduced by using the lattice-based method are approximately 1-3%. However, the lattice-based evaluation can be performed *several hundred times per second*, making it at least four orders of magnitude faster than the actual method (data not shown), which involves computing and subtracting the surface area of the entire complex from the sum of receptor and ligand surface areas.

[Footnotes to Table 1]

[a] Systems are: pcdhfr/flt = *P. carinii* dihydrofolate reductase and folate; 2gbp/glc = D-galactose/D-glucose binding protein and β-D-glucose; 3cpa/yg = carboxypeptidase A and glycyl-L-tyrosine; 4dfr/mtx = *E. coli* dihydrofolate reductase and methotrexate; 6rsa/urp = ribonuclease A and uridine phosphate; 1fkf/fk5 = FK506 binding protein and FK506; 3cla/clm = chloramphenicol acetyltransferase and chloramphenicol. With the exception of the pcdhfr/flt system (see Chapter 3), all are available in the Protein Data Bank (Bernstein *et al.*, 1977).

[b] Methods for computation are as follows: act 5.0 = actual solvent accessible areas determined by the access program using a surface density of 5.0 dots/Å^2. act 1.0 = actual solvent accessible areas determined by the access program using a surface density of 1.0 dots/Å^2. grid 1.0 = lattice-based estimation using a surface density of 1.0 dots/Å^2. Error is given as the percent error between grid 1.0 and act 1.0 methods.

[c] In units of Å^2.

[d] Total buried area is the sum of receptor and ligand accessible areas less the complexed accessible area.

Table 1. Comparison of surface-area burial estimation methods.

		Molecular Accessible Area[c]			Total	Contribution[c] of	
System[a]	Method[b]	receptor	ligand	complex	buried[c,d]	ligand	receptor
pcdhfr/flt	act 5.0	10882.8	700.6	10589.6	993.8	632.4	361.4
	act 1.0	10931.0	701.0	10637.0	995.0	629.0	366.0
	grid 1.0				989.0	627.0	362.0
	error				0.6 %	0.3 %	1.1 %
2gbp/glc	act 5.0	13151.8	347.2	13093.8	405.2	346.8	58.4
	act 1.0	13221.0	347.0	13164.0	404.0	344.0	60.0
	grid 1.0				405.0	346.0	59.0
	error				0.2 %	0.6 %	1.7 %
3cpa/yg	act 5.0	12089.4	462.4	11945.6	606.2	446.0	160.2
	act 1.0	12175.0	456.0	12030.0	601.0	437.0	164.0
	grid 1.0				607.0	438.0	169.0
	error				1.0 %	0.2 %	3.0 %
4dfr/mtx	act 5.0	8705.6	709.2	8535.8	879.0	548.0	331.0
	act 1.0	8745.0	716.0	8588.0	873.0	543.0	330.0
	grid 1.0				884.0	545.0	339.0
	error				1.3 %	0.4 %	2.7 %
6rsa/urp	act 5.0	7110.4	455.2	7014.4	551.2	337.2	214.0
	act 1.0	7088.0	465.0	7001.0	552.0	342.0	210.0
	grid 1.0				568.0	343.0	225.0
	error				2.9 %	0.3 %	7.1 %
1fkf/fk5	act 5.0	6049.0	1036.2	6151.6	933.6	542.0	391.6
	act 1.0	6061.0	1059.0	6150.0	970.0	565.0	405.0
	grid 1.0				963.0	561.0	402.0
	error				0.7 %	0.7 %	0.7 %
3cla/clm	act 5.0	10856.6	483.8	10744.0	596.4	361.2	235.2
	act 1.0	10925.0	483.0	10816.0	592.0	359.0	233.0
	grid 1.0				602.0	361.0	241.0
	error				1.7 %	0.6 %	3.4 %

The lattice-based method offers the further advantage that surface area burial can be attributed to specific atoms. This makes for trivial implementation of atomic solvation parameter-based methods such as that pioneered by Eisenberg & McLachlan (1986). One should be aware, however, that although errors for total molecular surface area burial, as presented in Table 1, may be small, errors on an atomic basis may still be large. Large per-atom errors may be opposite in direction, thus masking their presence in the total molecular area. Naturally, if effective atomic solvation parameter methods are desired, accurate atomic surface area changes are required. For the 332 total ligand atoms in the seven systems studied, the atomic errors between lattice-based and actual methods were distributed as follows: 91% showed exact agreement, 5.7% differed by 1 Å^2, 1.8% by 2 Å^2, and 1.2% by 3 Å^2. (The analogous analysis for receptor atoms is slightly more involved, as the many internal receptor atoms which do not have any surface area dominate the distribution). Thus, the lattice-based scheme is quite accurate at reproducing changes in both atomic and molecular surface areas.

Summary

A highly efficient lattice-based method has been developed for quantifying surface area changes that occur upon receptor-ligand complexation. Surface area changes computed by explicit non-lattice-based calculations are reproduced precisely, at an atomic and molecular level. The algorithm, while memory-intensive, is capable of several hundred evaluations per second and is dependent on both ligand configuration and conformation. These features make it amenable to incorporation into automated structure-based drug design packages. The ability to compute accessible surface area differences between molecular systems afford improved assessments of solvation effects in evaluating receptor-ligand interactions.

REFERENCES

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535.

Chothia, C.H. (1974). *Nature* **248**, 338.

Chothia, C.H. (1975). *Nature* **256**, 705.

Connolly, M.L. (1983). *Science* **221**, 709.

Eisenberg, D. and McLachlan, A.D. (1986). *Nature* **319**, 199.

Hermann, R.B. (1972). *J. Phys. Chem.* **76**, 2754.

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E. (1982). *J. Mol. Biol.* **161**, 269.

Lee, B. and Richards, F.M. (1971). *J. Mol. Biol.* **55**, 379.

Meng, E.C., Shoichet, B.K. and Kuntz, I.D. (1992). *J. Comp. Chem.* **13**, 505.

Reynolds, J.A, Gilbert, D.B. and Tanford, C. (1974). *Proc. Natl. Acad. Sci. U.S.A.* **71**, 29257.

Sharp, K.A., Nicholls, A., Fine, R.F. and Honig, B. (1991). *Science* **252**, 106.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P. (1984). *J. Am. Chem. Soc.* **106**, 765.

Appendix B.

MDL SDFile to Tripos MOL2 Database Conversion

BACKGROUND

The advent of scoring schemes that require detailed information concerning atomic hybridization raises a concomitant need for more accurate automated atom typing in structural databases used for docking. The empirical scheme developed in Chapter 5, for example, necessitates a knowledge not only of elemental type, but also of hybridization and charge state for each atom. The Sybyl modeling package (Tripos Associates, St. Louis, 63117), with which DOCK has had close association in the past (Meng, 1993), possesses a versatile set of atom types in the Tripos MOL2 file format. Databases prepared for force-field scoring (Meng, 1993) demanded less stringent standards for atom typing because the Sybyl atom types were eventually mapped into appropriate Amber (Weiner *et al.*, 1984) atom types used in DOCK-format databases. Nevertheless, Elaine Meng made significant strides toward a useful atom typing scheme for the conversion of Molecular Design Limited (MDL Information Systems, San Leandro, CA, 94577) structural databases into Tripos MOL2 format and ultimately into DOCK 3.0 databases (Meng, 1993).

Future versions of DOCK will in all likelihood read MOL2 format databases directly, allowing maximum use to be made of the versatility of this format and bypassing the need for yet another database format. While Sybyl has the capacity to read MDL-format databases, it does a considerably less-than-satisfactory job of interpreting atomic hybridization and charge states. I here describe an advanced conversion process which is rapid, easy to use, and most importantly, faithful in recognizing correct Sybyl atom types. This scheme will benefit scoring methods which make explicit use of atom type information and may even improve the accuracy of charge computations by providing more precise atom hybridization states.

METHODS

This conversion process begins with an MDL SDFile (structure data file) containing structures for any number of molecules. The SDFile is currently obtained by a relatively painless but terribly slow extraction using the ISIS package (MDL Information Systems, San Leandro, CA, 94577) and specially designed ISIS PL scripts. This task will not be described here, except to mention that the output should contain molecule names, registry numbers, and elemental identities, three-dimensional coordinates, connectivities and bond orders for each atom. Hydrogen atoms are occasionally present in MDL database structures, but, as their presence can not be relied upon, are ignored and added a later step. There are two major stages in the conversion process: 1) atom typing, and 2) hydrogen addition and charge computation. The first stage is performed by the Fortran program "sdf2mol2," while the second stage is performed by the Sybyl SPL program "sybdb."

sdf2mol2

Description

This program takes as input an MDL SDF database file and writes out a Sybyl multi-MOL2 file. This step types all atoms using Sybyl's fairly versatile and descriptive atom types.

Usage

```
sdf2mol2 -i SDFile -o MOL2file [-b start_at stop_at]
```

where *SDFile* is the name of the input MDL SDFile; *MOL2file* is the name of the output multi-MOL2 file; *start_at* and *stop_at* are optional bounds for starting and ending structure numbers; *e.g.* to process only the first hundred structures, use

```
sdf2mol2 -i SDFile -o MOL2file -b 1 100
```

Speed

~10 min for 100,000 structures (SGI R4400 Indigo2)

Method

0. Read in SDF structure. Obtain connectivity and bond orders. Define hybridization of each atom based on the highest bond order.

1. Search for rings. A breadth first search is used to find the smallest number of smallest rings.

2. Assign generic atom types. Atoms other than C, O, N, S, and P receive an atom type the same as the atom name - this is useful for atoms which have only one possible hybridization state (e.g. halogens) and atoms such as metals. All phosphorous atoms become P.3 as this is the only possible phosphorous atom type. Atoms C, O, N, and S get assigned types based on their hybridization as inferred from the bond orders.

E.g. doubly-bonded nitrogens become N.2, triply-bonded carbons become C.1, *etc.* Nitrogens with either zero or 4 neighbors are automatically assigned N.4.

3. Detect aromaticity. All rings which contain only X.2 or X.ar atoms are considered aromatic, subject to the following constraint (please note that the Hückel 4n+2 rule is *not* employed). Rings which have all X.2 atoms but only as a result of exocyclic double bonds, *e.g.* quinones, are not assigned as aromatic.

4. Treat specific functionalities.

   a) Carboxylate-like oxygens are typed. This includes carboxylates, sulf(on,in)ates, phosph(on,in)ates, and nitros. For purposes of this discussion, a singly-connected atom refers to one which has only one neighbor atom, regardless of the bond order of that bond. A carbon with two or more singly-connected oxygens, or, a sulfur with three or more singly-connected oxygens, *or*, a terminal sulfur with two or more singly-connected oxygens, *or*, a phosphorous with two or more singly-connected oxygens: the oxygens in these groups all considered O.co2's with single bonds. A nitrogen with two or more singly-connected oxygens is considered a nitro - the oxygens are both assigned O.2 with double-bonds, and the nitrogen is given N.pl3 status (this is the way Sybyl does it).

   b) Nitrogen functionalities. Any nitrogen alpha to an olefin is N.pl3. Any nitrogen alpha to an X=O or X=S group is considered an amide N.am.

   c) Sulfur functionalities. If the number of singly-connected oxygens is one, this is a sulfoxide (S given S.O type); if two, this is a sulfone (S given S.O2 type);

otherwise, the sulfur becomes S.2 if it has a double bond, S.3 in all other cases.

5. Consider functionalities which may depend on completion of step 4 entirely.

   a) Guanidyls and amidines. A carbon with three neighbors is considered an amidine if at least one neighbor is a C.ar and the other two neighbors are non-aromatic nitrogens. If the central carbon is in a ring, the two nitrogens may not be members of this ring. A guanidyl is any carbon with three non-aromatic nitrogen neighbors. Nitrogens in amidines and guanidyls are given N.pl3, the central carbon is assigned C.cat to insure a formal charge of +1 (again, this is the way Sybyl does it; viz. arginine). Finally, for both amidines and guanidyls, if any of the nitrogens themselves have heteroatom neighbors, this functionality is considered too electron-deficient to be charged and hence *not* an amidine or guanidyl.

6. Insure correct protonation state. Tetrahedral (N.3) nitrogens which do not have heteroatoms or olefins as neighbors are considered protonated and hence promoted to N.4.

7. Lone atoms are removed (*e.g.* single-atom counterions).

8. Atoms are renumbered sequentially and atom names made uppercase. Spaces in molecule name converted to underscores.

9. The structure is written out.

10. Return to 0.

sybdb

Description

This program is a shell script which creates a Sybyl command file which creates and runs a Sybyl SPL macro. Input is an sdf2mol2 output multi-MOL2 file, output is a multi-MOL2 file. The program removes all but the largest covalently bonded substructure, adds hydrogens, adjusts formal charges as appropriate, and computes partial charges using the method of Gasteiger & Marsili (Gasteiger & Marsili, 1980; Marsili & Gasteiger, 1980; Gasteiger & Marsili, 1981).

Usage

Customization: before the first use, please update inside the sybdb script the location of Sybyl. This will require modifying the variable TA_ROOT, which specifies the root directory for your version of Sybyl, and `sybcommand`, which stores the actual command used to access Sybyl at your site.

```
sybdb inputMOL2file outputMOL2file
```

where *`inputMOL2file`* is the output from sdf2mol2 and *`outputMOL2file`* is the cleaned up multi-MOL2 format file.

A log file called `sybdb.log` is also written which includes the name of each molecule processed, formal charges, modifications to formal charges by the script, and any other warnings that Sybyl may have generated. The file `sybdb.out` contains a record of the entire Sybyl session so that all actions may be examined.

Note 1: Due to memory limitations, you will in all likelihood need to run sybdb on multi-MOL2 files containing fewer molecules (*e.g.* 1000). Please use the accompanying "chunks" script for this purpose. This script will allow you to process a database of any size by splitting it into manageable pieces, processing each piece, then catenating all results.

Note 2: The Gasteiger-Marsili charges within Sybyl do *not* have parameters for the very common sulfoxide and sulfone functional groups. If you do not supply parameters for these types of sulfurs, charges on the sulfur and accompanying oxygens will be 0! What I have done is to copy the S.3 parameters to S.O and S.O2 so that at least *something* gets used. To do this, you should edit the file

`$TA_ROOT/sybylbase/tables/gastpar.tab`

and add the following four lines exactly as shown here:

```
S 29   2.3900  10.1400  20.6500         SO        copied from S3
P 29   0.0000   6.6000  20.6400         SO        copied from S3
S 30   2.3900  10.1400  20.6500         SO2       copied from S3
P 30   0.0000   6.6000  20.6400         SO2       copied from S3
```

You may use an altered `gastpar.tab` for temporary use only be placing it in your working directory. If no `gastpar.tab` is found in the working directory, the default file specified above will be used. For further details, see "Appendix 1: Parameter Tables: Charges" in the Sybyl Theory manual. (This is section A-1.7 beginning on page A-444 in the Sybyl 6.1 8/94 documentation.)

Speed

~3-4 hours for 100,000 structures (R4400 Indigo2)

Method

0. Setup:

   a) Add new bond parameters. A bogus N.ar-H bond type and length is assigned. Sybyl can not seem to accommodate a tertiary (and hence charged) aromatic nitrogen. It would prefer to add a hydrogen and so needs bonding information for the N.ar-H bond. By setting the bond type to `nc` (not connected), this hydrogen atom never really gets added, but the +1 formal charge is indeed now recognized. A bogus N.1-H bond type and length is also assigned, for similar reasons.

   b) Set up to use Gasteiger-Marsili pi charges (off by default).

   c) Load metal parameters. This is to insure that Sybyl does not assign dummy types to unrecognizable atoms.

1. Read in all molecules, then loop over each one as follows.

2. Remove all but the largest substructure.

3. Add hydrogens.

4. Remove any dummy atoms.

5. Rename atoms sequentially - this insures that added hydrogens will have names, as the "fillvalence" command adds hydrogens without giving them names.

6. Check for functionalities for which Sybyl incorrectly adds hydrogens. Tri-alkyl phosphines incorrectly get an additional hydrogen on the phosphorous. These hydrogens are stripped. Isocyanates (-N≡C) should be net neutral (*i.e.* +1 on nitrogen, -1 on carbon), so the hydrogen normally added to the carbon is removed.

7. Adjust formal charges. Sulf(on)ates and phosph(on)ates get their formal charge distributed evenly about the O.co2 oxygens.

8. Compute partial charges with the modified formal charges.

9. Return to 2 using next structure.

Further processing

To convert the multi-MOL2 database resulting from sybdb to a dock database, run mol2db but be sure to say *no* to charge adjustment, as this has already been done within the sybdb program.

References

Gasteiger, J. and Marsili, M. (1980). *Tetrahedron* **36**, 3219.

Gasteiger, J. and Marsili, M. (1981). *Organ. Magn. Reson.* **15**, 353.

Marsili, M. and Gasteiger, J. (1980). *Croat. Chem. Acta* **53**, 601.

Meng, E.C. (1993). Ph.D. Thesis, University of California, San Francisco.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984). *J. Am. Chem. Soc.* **106**, 765.

Appendix C.

Receptor-Ligand Complex Data Set

OVERVIEW

The following table reports the structural and affinity data used in Chapter 5 for the derivation of empirical scoring schemes. These receptor-ligand complexes are all available in the Protein Data Bank (Bernstein *et al.*, 1977). Affinities given in the table are presented as K_i values, but frequently represent K_d or K_m values, and on occasion even IC_{50} values. Two abbreviated literature references are provided for each receptor-ligand complex: one reports the structure solution and the other the affinity determination. Note that crystallization and assay *conditions* may not be reported in these papers and often appear elsewhere in the literature - these additional references have also been compiled but will not be discussed here. In rare instances where no reference for affinity determination could be found, the affinity reported by Keske & Dixon (unpublished results) or by Böhm (1994) was used. It is worth mentioning that in some cases affinities identified by Keske & Dixon or by Böhm differed, sometimes substantially, from those located in the literature. I owe great thanks to Jonathan Keske for supplying his list of affinity data, which served as a starting point for the data which follow. This data set should prove invaluable in the development of scoring tools for structure-based drug design.

Table 1 lists the source, affinity, crystallographic resolution, and literature references for all 144 complexes. The 103 complexes indicated with an asterisk were used in the model calibration discussed in Chapter 5.

REFERENCES

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535.

Böhm, H.-J. (1994). *J. Comput.-Aided Mol. Design* **8**, 243.

Table 1. Receptor-ligand complex data set.

Code	Receptor Source	Receptor	Ligand	Resol-ution Å	Affinity (-log K_d)	Affinity Reference	Assay Reference
1abe ara*	E. coli	L-arabinose binding protein	L-arabinose	1.7	6.52	J. Biol. Chem. 254:7529-7533 (1979)	J. Biol. Chem. 254:7529-7533 (1979)
1abf fuc*	E. coli	L-arabinose binding protein	D-fucose	1.9	5.42	Nature 340:404-407 (1989)	J. Biol. Chem. 258:13665-13672 (1983)
1ak3 amp	bovine heart mitochondrial matrix	adenylate kinase	AMP	1.9	3.86	Eur. J. Biochem. 93:263-270 (1979)	Eur. J. Biochem. 93:263-270 (1979)
1apb fuc*	E. coli	L-arabinose binding protein P254G	D-fucose	1.76	5.82	J. Biol. Chem. 265:16592-16603 (1990)	J. Biol. Chem. 265:16592-16603 (1990)
1apt lst*	Penicillium janthinellum	penicillopepsin	Iva-Val-Val-Lysta-OEt	1.8	9.4		
1apu sta*	Penicillium janthinellum	penicillopepsin	iva-val-val-sta-P-OEt	1.8	7.66	J. Org. Chem. 85:6268-6274 (1990)	J. Org. Chem. 85:6268-6274 (1990)
1apv fso*	Penicillium janthinellum	penicillopepsin	iva-val-val-(2,2-difluoro-3-hydrostatone)-N-methylamide	1.8	9.00	Biochemistry 31:3872-3886 (1992)	Biochemistry 31:3872-3886 (1992)
1apw fsi*	Penicillium janthinellum	penicillopepsin	iva-val-val-difluorostatine-N-methylamide	1.8	8.00	Biochemistry 31:3872-3886 (1992)	Biochemistry 31:3872-3886 (1992)
1bap ara*	E. coli	L-arabinose binding protein P254G	L-arabinose	1.75	6.85	J. Biol. Chem. 265:16592-16603 (1990)	J. Biol. Chem. 265:16592-16603 (1990)
1cbx bzs*	bovine pancreas	carboxypeptidase A	L-benzyl succinate	2.0	6.35	J. Mol. Biol. 223:573-578 (1992)	Biochemistry 12:2070-2078 (1973)
1cla clm*	plasmid R387	chloramphenicol acetyltransferase type 3	chloramphenicol	2.34	5.28	Biochemistry 29:2075-2080 (1990)	Biochemistry 29:2075-2080 (1990)
1cps cpm*	bovine	carboxypeptidase A	CPM: [L-(-)-2-carboxy-3-phenylpropyl]methyl-sulfodiimine	2.25	6.66	J. Biol. Chem. 267:19192-19197 (1992)	J. Am. Chem. Soc. 111:4467-4472 (1989)
1csc cmc	chicken heart muscle	citrate synthase	carboxymethyl coenzyme A	1.7	7.10	Biochemistry 29:2213-2219 (1990)	Eur. J. Biochem. 120:47-52 (1981)
1csc mal	chicken heart muscle	citrate synthase	L-malate	1.7	1.62	Eur. J. Biochem. 93:505-513 (1979)	Eur. J. Biochem. 93:505-513 (1979)
1drl bio	chicken liver	dihydrofolate reductase	biopterin	2.2	5.57	Biochemistry 31:7264-7273 (1992)	J. Biol. Chem. 242:3934-3942 (1967)
1drf fol*	human recombinant	dihydrofolate reductase	folate	2.0	7.44	Biochemistry 27:3664-3671 (1988)	Biochemistry 27:3664-3671 (1988)
1dwb bnz*	human plasma	alpha-thrombin	benzamidine	3.16	2.90	Thromb. Res. 36:457-465 (1984)	Thromb. Res. 36:457-465 (1984)
1dwc mit*	human plasma	alpha-thrombin	MD-805 (agratroban): (2R,4R)-4-methyl-1-[N-alpha-[3-methyl-1,2,3,4-tetrahydro-8-quinolinyl)sulphonyl]-L-arginyl)]-2-piperidinecarboxylic acid	3.0	7.41	J. Biol. Chem. 266:20085-20093 (1991)	Biochemistry 23:85-90 (1984)
1dwd mid*	human plasma	alpha-thrombin	NAPAP: N-alpha-(2-nAphthyl-sulfonyl-glycyl)-(DL)-para-amidinophenylalanyl-piperidine	3.0	8.18	J. Biol. Chem. 266:20085-20093 (1991)	Thromb. Res. 29:635-642 (1983)
1etu gdp	E. coli	elongation factor Tu	GDP	2.9	8.52	Biochemistry 20:6265-6272 (1981)	Biochemistry 20:6265-6272 (1981)

Code	Receptor Source	Receptor	Ligand	Resolution A	Affinity (-log K_i)	Affinity Reference	Assay Reference
1stp bio	Streptomyces avidinii	streptavidin	biotin	2.6	13.46	J. Am. Chem. Soc. 114:3197-3200 (1984)	J. Am. Chem. Soc. 114:3197-3200 (1984)
1tha iit*	human plasma	transthyretin (prealbumin)	3,3'-diiodo-L-thyronine	2.0	5.35	Biochemistry 19:55-63 (1980)	Biochemistry 19:55-63 (1980)
1tlp pho*	Bacillus thermoproteolyticus	thermolysin	phosphoramidon	2.3	7.56	Arch. Biochem. Biophys. 171:727-731 (1975)	Arch. Biochem. Biophys. 171:727-731 (1975)
1tmn clt*	Bacillus thermoproteolyticus	thermolysin	N-(1-carboxy-3-phenylpropyl)-L-leucyl-L-tryptophan	1.9	7.30	Biochem. Biophys. Res. Commun. 102:963-969 (1981)	J. Biol. Chem. 255:3482-3486 (1980)
1ulb gun	human erythrocytes	purine nucleoside phosphorylase	guanine	2.75	4.40	J. Med. Chem. 36:1024-1031 (1993)	
1wgc sgc*	wheat germ	agglutinin	N-acetylneuraminyllactose	2.2	3.13	Biochemistry 21:3050-3057 (1982)	Biochemistry 21:3050-3057 (1982)
1xli thg*	Arthrobacter	D-xylose isomerase	5-thio-alpha-D-glucose	2.5	1.48	J. Mol. Biol. 212:211-235 (1990)	unpublished results : (1990)
2cpp cam*	Pseudomonas putida	cytochrome P450cam	camphor	1.63	5.94	J. Biol. Chem. 263:18842-18849 (1988)	J. Biol. Chem. 263:18842-18849 (1988)
2csc mal*	chicken heart muscle	citrate synthase	D-malate	1.7	3.36	Eur. J. Biochem. 93:505-513 (1979)	Eur. J. Biochem. 93:505-513 (1979)
2er6 h25*	chestnut blight fungus	endothiapepsin	H-256	2.0	7.22	Nature 327:349-352 (1987)	Nature 327:349-352 (1987)
2er7 h26*	chestnut blight fungus	endothiapepsin	H-261	1.6	9.15	J. Mol. Biol. 216:1017-1029 (1990)	J. Hyperten. 3:13-18 (1985)
2fnr amp	spinach	ferredoxin reductase	2'-phospho-5'-AMP	3.0	5.70	J. Biol. Chem. 261:11214-11223 (1986)	J. Biol. Chem. 261:11214-11223 (1986)
2gbp glc*	E. coli	D-galactose/D-glucose binding protein	D-glucose	1.9	7.40	J. Biol. Chem. 255:2465-2471 (1980)	J. Biol. Chem. 255:2465-2471 (1980)
2ifb plm*	rat recombinant	intestinal fatty acid binding protein	palmitate	2.0	5.44	J. Biol. Chem. 262:5931-5937 (1987)	J. Biol. Chem. 262:5931-5937 (1987)
2ldb nad*	Bacillus stearothermophilus	L-lactate dehydrogenase	NAD+	3.0	4.15	Biochemistry 27:1617-1622 (1988)	Biochemistry 27:1617-1622 (1988)
2mcp ppc	mouse	immunoglobulin MC/PC603 FAB	phosphocholine	3.1	4.70	PNAS 66:3689-3692 (1972)	PNAS 66:3689-3692 (1972)
2phh adp*	Pseudomonas fluorescens	p-hydroxybenzoate hydroxylase	ADP ribose	2.7	3.36	Biochemistry 28:7199-7205 (1989)	Biochemistry 28:7199-7205 (1989)
2phh phb*	Pseudomonas fluorescens	p-hydroxybenzoate hydroxylase	p-hydroxybenzoate	2.7	4.60	Eur. J. Biochem. 128:21-27 (1983)	Eur. J. Biochem. 128:21-27 (1983)
2pk4 aca*	human	plasminogen kringle 4	epsilon-aminocaproic acid	2.25	4.32	Biochemistry 28:1368-1376 (1989)	Biochemistry 28:1368-1376 (1989)
2r04 win	human virus	rhinovirus 14	WIN IV	3.0	6.22	PNAS 85:3304-3308 (1988)	
2rnt gpg*	Aspergillus oryzae	ribonuclease T1 K25	guanylyl-2',5'-guanosine	1.8	3.78	In: The Enzymes Ed: Boyer PD:435-465 (1982)	In: The Enzymes Ed: Boyer PD:435-465 (1982)
2sns dpt	Staphylococcus aureus	Staphylococcal nuclease	2'-deoxy-3',5'-diphosphothymidine	1.5	6.70	In: The Enzymes Ed: Boyer PD:177-200 (1971)	In: The Enzymes Ed: Boyer PD:177-200 (1971)
2tmn npl*	Bacillus thermoproteolyticus	thermolysin	N-phosphoryl-L-leucinamide	1.6	5.89	Biochemistry 18:3032-3038 (1979)	Biochemistry 18:3032-3038 (1979)
2tsc cb3*	E. coli	thymidylate synthase	CB3717	1.97	9.00	Biochem. Pharmacol. 32:3783-3790 (1983)	Biochem. Pharmacol. 32:3783-3790 (1983)

Code	Receptor Source	Receptor	Ligand	Resolution Å	Affinity (-log K_d)	Affinity Reference	Assay Reference
1fbc ahg*	porcine kidney cortex	fructose-1,6-bisphosphatase	2,5-anhydroglucitol-1,6-bisphosphate	2.6	6.26	J. Biol. Chem. 251:2963-2966 (1976)	J. Biol. Chem. 251:2963-2966 (1976)
1fbf ahm*	porcine kidney cortex	fructose-1,6-bisphosphatase	2,5-anhydromannitol-1,6-bisphosphate	2.7	6.00	J. Biol. Chem. 259:5115-5123 (1984)	J. Biol. Chem. 259:5115-5123 (1984)
1fbp amp	porcine kidney cortex	fructose-1,6-bisphosphatase	AMP	2.5	4.82	PNAS 87:5243-5247 (1990)	J. Biol. Chem. 243:4923-4926 (1968)
1fbp f6p	porcine kidney cortex	fructose-1,6-bisphosphatase	fructose-6-phosphate	2.5	0.00	PNAS 88:2989-2993 (1991)	PNAS 88:2989-2993 (1991)
1fkb rap*	E. coli	FK506 binding protein	rapamycin	1.7	9.70	J. Mol. Biol. 229:105-124 (1993)	PNAS 87:9231-9235 (1990)
1fkf fk5*	human recombinant	FK506 binding protein	FK506	1.7	8.77	PNAS 87:9231-9235 (1990)	Biochemistry 29:3813-3816 (1990)
1gpd nad	lobster	D-glyceraldehyde-3-phosphate dehydrogenase	NAD+	2.9	8.30	Biochim. Biophys. Acta 191:214-220 (1969)	Biochim. Biophys. Acta 191:214-220 (1969)
1gst gsh*	rat liver	glutathione S-transferase	glutathione	2.2	4.68	J. Biol. Chem. 267:4296-4299 (1992)	J. Biol. Chem. 267:4296-4299 (1992)
1l83 bnz*	bacteriophage T4	lysozyme C54T,C97A,L99A	benzene	1.70	3.40	Nature 355:371-373 (1992)	Nature 355:371-373 (1992)
1ldm ndh	dogfish muscle	lactate dehydrogenase	NADH	2.1	5.44	In: The Enzymes Ed: Boyer PD:193-292 (1975)	In: The Enzymes Ed: Boyer PD:193-292 (1975)
1mbi imd	sperm whale	myoglobrin	imidazole	2.0	1.66	J. Mol. Biol. 158:305-315 (1982)	J. Mol. Biol. 158:305-315 (1982)
1phd pim	Pseudomonas putida	cytochrome P450cam	2-phenyl imidazole	1.6	5.15	Biochemistry 19:3590-3599 (1980)	Biochemistry 19:3590-3599 (1980)
1phf pim	Pseudomonas putida	cytochrome P450cam	4-phenyl imidazole	1.6	4.40	Biochemistry 19:3590-3599 (1980)	Biochemistry 19:3590-3599 (1980)
1phg mty	Pseudomonas putida	cytochrome P450cam	metyrapone	1.6	7.37	Biochemistry 11:4740-4745 (1972)	Arch. Biochem. Biophys. 145:531-542 (1971)
1phh dhb	Pseudomonas fluorescens	p-hydroxybenzoate hydroxylase	3,4-dihydroxybenzoate	2.3	3.3	Proteins 21:22-29 (1995)	J. Biol. Chem. 254:6657-6666 (1979)
1phh fad	Pseudomonas fluorescens	p-hydroxybenzoate hydroxylate	FAD	2.3	7.35	Biochemistry 28:7199-7205 (1989)	Eur. J. Biochem. 128:21-27 (1983)
1ppc nap*	bovine	trypsin	NAPAP: N-alpha-(2-naphthyl-sulfonyl-glycyl)-DL-p-amidinophenylalanyl-piperidine	1.8	6.16	FEBS Lett. 287:133-138 (1991)	
1pph tap*	bovine	trypsin	3-TAPAP	1.9	5.92	FEBS Lett. 287:133-138 (1991)	
1ppl iva*	Penicillium janthinellum	penicillopepsin	iva-val-val-leu-P-(O)phe-OMe	1.7	8.55	J. Org. Chem. 85:6268-6274 (1990)	J. Org. Chem. 85:6268-6274 (1990)
1ppm zaa*	Penicillium janthinellum	penicillopepsin	Cbz-Ala-Ala-Leu-P-(O)Phe-OMe	1.7	5.80	J. Org. Chem. 85:6268-6274 (1990)	J. Org. Chem. 85:6268-6274 (1990)
1rbp ret*	human serum	retinol binding protein	retinol	2.0	6.72	Proteins 8:44-61 (1990)	Eur. J. Biochem. 65:71-78 (1975)
1rnb dgc*	Bacillus amyloliquefaciens recombinant	barnase	dGPC	1.9	0.00		
1rne cgp*	human recombinant	renin	CGP 38'560	2.4	8.70	J. Struct. Biol. 107:227 (1991)	J. Med. Chem. 31:1839-1846 (1988)
1rnt gmp*	Aspergillus oryzae	ribonuclease T1	2'-GMP	1.9	5.18	In: The Enzymes Ed: Boyer PD:435-465 (1982)	In: The Enzymes Ed: Boyer PD:435-465 (1982)
1rus phg	Rhodospirillum rubrum recombinant	rubisco	3-phosphoglycerate	2.9	3.08	Biochemistry 20:2219-2225 (1981)	Biochemistry 20:2219-2225 (1981)
1snc dpt*	Staphylococcus aureus recombinant	staphylococcal nuclease	deoxythymidine 3',5' bisphosphate	1.65	6.70	In: The Enzymes Ed: Boyer PD:177-200 (1971)	In: The Enzymes Ed: Boyer PD:177-200 (1971)

Code	Receptor Source	Receptor	Ligand	Resol-ution Å	Affinity (-log K_i)	Affinity Reference	Assay Reference
2xim xyl*	Actinoplanes missouriensis recombinant	D-xylose isomerase K253R	xylitol	2.3	2.28	Biochemistry 31:2239-2253 (1992)	Biochemistry 31:2239-2253 (1992)
2yhx otg	Saccharomyces cerevisiae	yeast hexokinase B	ortho-toluoylglucosamine	2.1	5.00	PNAS 75:4848-4852 (1978)	J. Biol. Chem. 148:117 (1943)
2ypi pga*	Saccharomyces cerevisiae	triose phosphate isomerase	2-phosphoglycolic acid	2.5	4.82		
3cla clm*	plasmid R387 recombinant	chloramphenicol acetyltransferase type 3	chloramphenicol	1.75	4.94	Biochemistry 30:3763-3770 (1991)	Biochemistry 30:3763-3770 (1991)
3cpa gyp	bovine pancreas	carboxypeptidase A	glycyl-L-tyrosine	2.0	4.00	In: Advances in Protein Chemistry Ed: Anfinsen CB, Edsall JT, Richards FM:1-47 (1971)	In: Advances in Protein Chemistry Ed: Anfinsen CB, Edsall JT, Richards FM:1-47 (1971)
3csc coa*	chicken heart muscle	citrate synthase	acetyl coenzyme A	1.9	5.15	J. Mol. Biol. 174:205-219 (1984)	Acta Chem. Scand. 17:S129-S134 (1963)
3csc mal*	chicken heart muscle	citrate synthase	L-malate	1.9	2.64	Biochemistry 30:6024-6031 (1991)	
3fx2 fmn*	Desulfovibrio vulgaris recombinant	flavodoxin	FMN	1.9	9.3		
3gap cmp*	E. coli	catabolite gene activator protein	cAMP	2.5	5.00	J. Biol. Chem. 247:2717-2722 (1972)	J. Biol. Chem. 247:2717-2722 (1972)
3pgm phg	Saccharomyces cerevisiae	phosphoglycerate mutase	phosphoglycerate	2.8	3.19	Biochem. Soc. Trans. 18:257 (1989)	Arch. Biochem. Biophys. 165:179-187 (1974)
3ptb bnz*	bovine pancreas	beta-trypsin	benzamidine	1.7	4.50	Biophys. Chem. 54:75-81 (1995)	Biophys. Chem. 54:75-81 (1995)
3tpi ivp*	bovine pancreas	trypsinogen	Ile-Val	1.9	4.27	J. Mol. Biol. 127:357-374 (1979)	J. Mol. Biol. 127:357-374 (1979)
4cla clm*	plasmid R387 recombinant	chloramphenicol acetyltransferase type 3 L160F	chloramphenicol	2.0	5.47	Biochemistry 30:3763-3770 (1991)	Biochemistry 30:3763-3770 (1991)
4dfr mtx*	E. coli	dihydrofolate reductase	methotrexate	1.7	8.62	J. Biol. Chem. 254:8143 (1979)	J. Biol. Chem. 254:8143 (1979)
4er1 pd1*	chestnut blight fungus	endothiapepsin	PD125967	2.0	6.62	Biochemistry 31:8142-8150 (1992)	Biochemistry 31:8142-8150 (1992)
4er2 pep*	chestnut blight fungus	endothiapepsin	pepstatin	2.0	9.30	Biochem. J. 289-2:363-371 (1993)	Biochem. J. 289-2:363-371 (1993)
4er4 h14*	chestnut blight fungus	endothiapepsin	H142	2.1	6.80	Biochemistry 26:5585-5590 (1987)	Biochemistry 26:5585-5590 (1987)
4fab fds	mouse	IgG kappa Fab 4-4-20	fluorescein dianion	2.7	8.05	Proteins 5:271-280 (1989)	Proteins 3:155-160 (1988)
4grl rgs*	human erythrocytes	glutathione reductase	retro-GSSG	2.4	2.20	J. Biol. Chem. 265:10443-10445 (1990)	
4hmg sia	influenza virus recombinant	hemagglutinin L226Q	sialic acid	3.0	2.55	Biochemistry 28:8388-8396 (1989)	Biochemistry 28:8388-8396 (1989)
4hvp mvt*	synthetic	HIV1 protease	MVT101	2.3	6.11	Science 246:1149-1152 (1989)	Science 246:1149-1152 (1989)
4mba imd	sea hare	myoglobin	imidazole	2.0	1.62	J. Mol. Biol. 158:305-315 (1982)	J. Mol. Biol. 158:305-315 (1982)
4mdh nad*	porcine heart	cytoplasmic malate dehydrogenase	NAD+	2.5	3.23	In: The Enzymes Ed: Boyer PD:369-395 (1975)	
4phv l70*	recombinant	HIV1 protease	L700,417	2.10	9.17	Biochem. Biophys. Res. Commun. 164:955-960 (1989)	Biochem. Biophys. Res. Commun. 164:955-960 (1989)
4sga pap*	Streptomyces griseus	proteinase A	Ace-Pro-Ala-Pro-Phe	1.8	3.27	J. Mol. Biol. 144:43-88 (1980)	Biochemistry 15:1296-1299 (1976)
4tim phg*	Trypanosoma brucei	triosephosphate isomerase	2-phosphoglycerate	2.4	2.16	J. Med. Chem. 34:2709-2718 (1991)	Eur. J. Biochem. 168:69-74 (1987)

Code	Receptor Source	Receptor	Ligand	Resol-ution Å	Affinity (-log K_i)	Affinity Reference	Assay Reference
4tln lno*	Bacillus thermoproteolyticus	thermolysin	L-leucyl-hydroxylamine	2.3	3.72	Acc. Chem. Res. 21:333-340 (1988)	Biochemistry 17:2846-2850 (1978)
4tmn fla*	Baccilus thermoproteolyticus	thermolysin	ZF-P-LA: Cbz-Phe-P-Leu-Ala	1.7	10.17	Acc. Chem. Res. 21:333-340 (1988)	Biochemistry 26:8553-8561 (1987)
4tpi vvp*	bovine pancreas	trypsinogen	Val-Val	2.2	2.80	J. Mol. Biol. 127:357-374 (1979)	J. Mol. Biol. 127:357-374 (1979)
4xia sor*	arthrobacter	D-xylose isomerase	D-sorbitol	2.3	1.54	In: The Enzymes Ed: Boyer PD:349-355 (1972)	In: The Enzymes Ed: Boyer PD:349-355 (1972)
5acn trc*	pig heart	aconitase	tricarballylic acid	2.1	2.80	Biochemistry 23:4572-4580 (1984)	Biochemistry 23:4572-4580 (1984)
5cpp adm	Pseudomonas putida	cytochrome P450cam	adamantanone	2.08	5.89	J. Am. Chem. Soc. 107:5018-5019 (1985)	
5enl phg*	Saccharomyces cerevisiae	enolase	2-phospho-D-glycerate	2.2	3.8	Biochem. Biophys. Res. Commun. 211:607-613 (1995)	Biochem. Biophys. Res. Commun. 211:607-613 (1995)
5er2 cp6*	chestnut blight fungus	endothiapepsin	CP-69,799	1.8	6.57	EMBO J. 8:2179-2188 (1989)	
5hvp pep*	recombinant	HIV1 protease	acetyl-pepstatin	2.0	7.46	FEBS Lett. 247:113-117 (1989)	FEBS Lett. 247:113-117 (1989)
5icd ict*	E. coli	isocitrate dehydrogenase	isocitrate	2.5	5.29	J. Biol. Chem. 264:20482-20486 (1989)	J. Biol. Chem. 264:20482-20486 (1989)
5ldh nal	pig heart	lactate dehydrogenase	S-Lac-NAD+	2.7	2.82	Biochemistry 17:4621-4626 (1978)	Biochemistry 17:4621-4626 (1978)
5p21 gpp*	human recombinant	ras p21 protein	GPPNP: guanosine-5'-(beta,gamma-imido) triphosphate	1.35	5.32	Science 249:169-171 (1990)	Science 249:169-171 (1990)
5sga ppy*	Streptomyces griseus	proteinase A	Ace-Pro-Ala-Pro-Tyr	1.8	2.85	J. Mol. Biol. 144:43-88 (1980)	Biochemistry 15:1295,1296-1299 (1976)
5tln ina*	Bacillus thermoproteolyticus	thermolysin	HONH-benzylmalonyl-L-alanylglycine-p-nitroanilide	2.3	6.37	Acc. Chem. Res. 21:333-340 (1988)	Biochemistry 17:2846-2850 (1978)
5tmn zgn*	Bacillus thermoproteolyticus	thermolysin	ZG-P-LL: Cbz-Gly-P-Leu-Leu	1.6	8.04	Acc. Chem. Res. 21:333-340 (1988)	Science 235:569-571 (1987)
5xia xyl*	arthrobacter	D-xylose isomerase	xylitol	2.5	2.60	In: The Enzymes Ed: Boyer PD:349-355 (1972)	In: The Enzymes Ed: Boyer PD:349-355 (1972)
6abp ara*	E. coli	L-arabinose binding protein M108L	L-arabinose	1.67	6.36	Biochemistry 30:6861-6866 (1991)	Biochemistry 30:6861-6866 (1991)
6apr iva*	Rhizopus chinensis	rhizopuspepsin	pepstatin	2.5	7.77	Proteins 13:195-205 (1992)	In: Aspartic Proteinases and Their Inhibitors. Ed: Kotska V:401-420 (1985)
6cpa zaf*	bovine pancreas	carboxypeptidase A	ZAA-P-(O)F: O-[[(1R)-[[n-phenylmethoxycarbonyl)-L-alanyl]amino]ethyl]hydroxyphosphinyl]-L-3-phenyllactate	2.0	11.52	Biochemistry 29:5546-5555 (1990)	Biochemistry 28:6294-6305 (1989)
6cpp can	Pseudomonas putida	cytochrome P450cam	camphane	1.9	4.34	J. Biol. Chem. 263:18842-18849 (1988)	J. Biol. Chem. 263:18842-18849 (1988)
6enl phg*	Saccharomyces cerevisiae	enolase	phosphoglycolic acid	2.2	3.0	Biochem. Biophys. Res. Commun. 211:607-613 (1995)	Biochem. Biophys. Res. Commun. 211:607-613 (1995)

Code	Receptor Source	Receptor	Ligand	Resolution Å	Affinity (-log K_i)	Affinity Reference	Assay Reference
6rnt amp*	Aspergillus oryzae recombinant	ribonuclease T1	2'-AMP	1.8	2.37	J. Biol. Chem. 266:15128-15134 (1991)	J. Biol. Chem. 266:15128-15134 (1991)
6tim g3p*	Trypanosoma brucei	triosephosphate isomerase	glycerol-3-phosphate	2.2	3.21	Proteins 10:50-69 (1991)	Eur. J. Biochem. 168:69-74 (1987)
6tmn zgl*	Bacillus thermoproteolyticus	thermolysin	ZG-P-(O)LL: Cbz-Gly-P-(O)-Leu-Leu	1.6	5.05	Acc. Chem. Res. 21:333-340 (1988)	Science 235:569-571 (1987)
7abp fuc*	E. coli	L-arabinose binding protein M108L	D-fucose	1.67	6.46	Biochemistry 30:6861-6866 (1991)	Biochemistry 30:6861-6866 (1991)
7acn ict	porcine heart	aconitase	isocitrate	2.0	4.31	Biochemistry 19:2358-2362 (1980)	Biochemistry 19:2358-2362 (1980)
7cat ndp*	beef liver	catalase	NADPH	2.5	8.00	PNAS 81:4343-4347 (1984)	PNAS 81:4343-4347 (1984)
7cpa fvf*	bovine pancreas	carboxypeptidase A	FVF: o-(((1R)-((N-(phenylmethoxycarbonyl)-L-phenylalanyl)amino)isobutyl)hydroxyphosphinyl)-L-3-phenyllactate	2.0	13.96	Biochemistry 30:8165-8170 (1991)	Biochemistry 30:8165-8170 (1991)
7cpp ncm	Pseudomonas putida	cytochrome P450cam	norcamphor	2.0	3.82	J. Am. Chem. Soc. 107:5018-5019 (1985)	
7dfr fol	E. coli	dihydrofolate reductase	folate	2.5	4.96	In: A Study of Enzymes, vol 2, Mechanism of Enzyme Action Ed: Kuby SA:193-226 (1991)	In: A Study of Enzymes, vol 2, Mechanism of Enzyme Action Ed: Kuby SA:193-226 (1991)
7dfr ndp	E. coli	dihydrofolate reductase	NADP+	2.5	6.10	In: A Study of Enzymes, vol 2, Mechanism of Enzyme Action Ed: Kuby SA:193-226 (1991)	In: A Study of Enzymes, vol 2, Mechanism of Enzyme Action Ed: Kuby SA:193-226 (1991)
7est tfa*	porcine pancreas	elastase	TFAP: trifluoroacetyl-L-leucyl-L-alanyl-p-trifluorometylphenylanilide	1.8	7.60	J. Mol. Recogn. 3:36-43 (1990)	J. Biol. Chem. 258:8312-8316 (1983)
7hvp jg3*	synthetic	HIV1 protease	JG-365	2.4	9.62	PNAS 87:8805-8809 (1990)	PNAS 87:8805-8809 (1990)
7tim pgh*	Saccharomyces cerevisiae	triosephosphate isomerase	phosphoglycolohydroxamate	1.9	5.40	J. Biol. Chem. 249:136-142 (1974)	J. Biol. Chem. 249:136-142 (1974)
8abp gal*	E. coli	L-arabinose binding protein M108L	D-galactose	1.49	8.00	Biochemistry 30:6861-6866 (1991)	Biochemistry 30:6861-6866 (1991)
8acn nic	bovine heart	aconitase	nitroisocitrate	2.0	7.14	Biochemistry 19:2358-2362 (1980)	
8atc pal*	E. coli	aspartate carbamoyltransferase	PALA: N-phosphonacetyl-L-aspartate	2.5	7.57	J. Biol. Chem. 246:6599-6605 (1971)	J. Biol. Chem. 246:6599-6605 (1971)
8cpa agf*	bovine pancreas	carboxypeptidase A	ZAG-P-(O)F: Cbz-Ala-Gly-P-(O)-Phe	2.0	9.15	Biochemistry 30:8165-8170 (1991)	Biochemistry 30:8165-8170 (1991)
8cpp tcm	Pseudomonas putida	cytochrome P450cam	thiocamphor	2.1	5.52	J. Biol. Chem. 263:18842-18849 (1988)	J. Biol. Chem. 263:18842-18849 (1988)
8hvp u85*	synthetic	HIV1 protease	U-85548E	2.5	9.00	J. Biol. Chem. 265:14675-14683 (1990)	Biochemistry 29:264-269 (1990)
8icd ict*	E. coli	isocitrate dehydrogenase S113E	isocitrate	2.5	3.02	Science 249:1044-1046 (1990)	Science 249:1044-1046 (1990)
8xia xys*	Streptomyces rubiginosus	D-xylose isomerase	D-xylose	1.9	2.95	In: The Enzymes Ed: Boyer PD:349-355 (1972)	In: The Enzymes Ed: Boyer PD:349-355 (1972)

Code	Receptor Source	Receptor	Ligand	Resol-ution Å	Affinity (-log K_i)	Affinity Reference	Assay Reference
9aat pmp*	chicken heart mitochondria	aspartate aminotransferase	pyridoxal-5'-phosphate	2.2	8.22	J. Mol. Biol. 225:495-517 (1992)	Biochem. Biophys. Res. Commun. 89:345-352 (1979)
9abp gal*	E. coli	L-arabinose binding protein P254G	D-galactose	1.97	8.00	J. Biol. Chem. 265:16592-16603 (1990)	J. Biol. Chem. 265:16592-16603 (1990)
9hvp a74	recombinant	HIV1 protease	A-74704	2.8	8.35	Science 249:527-533 (1990)	Science 247:954-958 (1990)
9ldt ndh	porcine muscle	lactate dehydrogenase	NADH	2.0	5.43	In: The Enzymes Ed: Boyer PD:193-292 (1975)	In: The Enzymes Ed: Boyer PD:193-292 (1975)
9ldt oxm	porcine muscle	lactate dehydrogenase	oxamate	2.0	4.74	In: The Enzymes Ed: Boyer PD:193-292 (1975)	In: The Enzymes Ed: Boyer PD:193-292 (1975)
9rub rub	Rhodospirillum rubrum recombinant	rubisco	ribulose-1,5-bisphosphate	2.6	4.70	Biochemistry 20:2219-2225 (1981)	Biochemistry 20:2219-2225 (1981)