VOLUME 27 NUMBER 5

ACCOUNTS OF CHEMICAL RESEARCH®

MAY 1994

Structure-Based Molecular Design

IRWIN D. KUNTZ,* ELAINE C. MENG, AND BRIAN K. SHOICHET†

Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, California 94143-0446

Received June 18, 1993

Creating molecules with specific properties has been a cherished goal of chemists for generations. Progress includes elegant strategies for natural product syntheses, development of intricate models for molecular recognition, and the design of novel enzyme inhibitors and therapeutical agents. Finding new drugs, in particular, is an important part of the new initiatives in health care. We will focus on discovery or design of molecules that interact with biochemical targets whose three-dimensional structures are known, a field called "structure-based" design. The idea that molecules can interact in a highly specific manner has a long history in medicinal chemistry and was well articulated by Fischer[1] and Ehrlich[2] a century ago using a "lock-and-key" analogy. We explore the usefulness of this concept by reviewing both general principles of molecular recognition and specific computer programs that pack molecules together.

The drug discovery process is complex, typically taking 10 years for new products to reach the market. The first step is finding interesting leads. Such compounds are identified through a wide variety of routes.[3] Structure-based design, spurred by the rapid advances in molecular structure determination and computational resources, is now being tested as a means of generating new pharmaceuticals.[4] The field is in its infancy, but the underpinnings of the method can be assessed and initial results can be examined. A critical assumption is that our understanding of intermolecular interactions is sufficiently advanced so that novel compounds can be proposed and optimized. A related issue is whether macromolecular plasticity[5] and "induced fit" effects[5,6] seriously compromise all "lock-and-key" models.[7]

We restrict ourselves to computational approaches to molecular design,[3,8] with particular focus on the "docking problem": placing putative ligands in configurations appropriate for interacting with the receptor. We will discuss some of the fundamental problems of representing molecular properties and analyzing interaction energies. We next outline the approximations used to treat these problems and discuss the details of prominent methods. Lastly, we summarize progress to date. Effective solutions to molecular docking have important implications for molecular recognition, materials science, and drug development.

Irwin Kuntz received his A.B. degree in chemistry from Princeton University in 1961 and his Ph.D. in physical chemistry with Melvin Calvin at the University of California, Berkeley, in 1965. He returned to the Chemistry Department at Princeton and in 1971 joined the faculty of the University of California, San Francisco, where he is now Professor of Medicinal Chemistry and Acting Director of the Molecular Design Institute. His research interests include protein structure and molecular recognition.

Elaine Meng received her B.S. degree in pharmacy at the University of Cincinnati in 1988 and her Ph.D. degree in pharmaceutical chemistry at UCSF in 1993. She is currently a postdoctoral fellow with Peter Kollman at UCSF studying molecular solvation. Dr. Meng won the 1993 Drug Information Association Award for her Ph.D. thesis.

Brian Shoichet received B.S. degrees in chemistry and history from MIT in 1985 and a Ph.D. in pharmaceutical chemistry from UCSF in 1991. He is currently a Damon Runyon–Walter Winchell Postdoctoral Fellow in the laboratory of Brian Matthews at the University of Oregon, where he is investigating the relations among protein stability, structure, and function.

† Current address: Molecular Biology Institute, University of Oregon, Eugene, OR 97403-1229.

(1) Fischer, E. *Chem. Ber.* **1894**, *27*, 2985–2993.
(2) Ehrlich, P. *Chem. Ber.* **1907**, *42*, 17.
(3) Kuntz, I. D. *Science* **1992**, *257*, 1078–1082.
(4) Reich, S. H.; Fuhry, M. A. M.; Nguyen, D.; Pino, M. J.; Welsh, K. M.; Webber, S.; Janson, C. A.; Jordan, S. R.; Matthews, D. A.; Smith, W. W.; Bartlett, C. A.; Booth, C. L. J.; Herrmann, S. M.; Howland, E. F.; Morse, C. A.; Ward, R. W.; White, J. *J. Med. Chem.* **1992**, *35*, 847–858.
(5) Koshland, D. E., Jr. *Pure Appl. Chem.* **1971**, *25*, 119–133.
(6) Schulz, G. E.; Muller, C. W.; Diederichs, K. *J. Mol. Biol.* **1990**, *213*, 627–630.
(7) Jorgensen, W. L. *Science* **1991**, *254*, 954–955.
(8) Cohen, N. C.; Blaney, J. M.; Humblet, C.; Gund, P.; Barry, D. C. *J. Med. Chem.* **1990**, *33*, 883–894.

0001-4842/94/0127-0117$04.50/0

Methodological Issues

We divide the docking problem into three components: site/ligand representation; juxtaposition of the ligand and site frames of reference; and evaluation of complementarity. There are serious mathematical issues to address. First, there is no "best" method for describing molecular shape.[9,10] Second, packing irregular objects together, the so-called "knapsack" problem,[11] has no general solution. Third, the search issues are related to an area of graph theory, the study of isomorphous subgraphs.[12] These last two problems are especially difficult,[13] and it is unlikely that general and efficient solutions will be found. Rather, progress depends on the quality of the approximations required to search an extremely large number of molecular conformations and molecular configurations.

Site and Ligand Description. Atomic coordinates for receptor macromolecules can be obtained through X-ray crystallography, nuclear magnetic resonance (NMR), and homology modeling. Structures are available through the Protein Data Bank.[14] What are typical uncertainties in atomic coordinates expected from the different structural techniques? Excellent protein structures derived from X-ray diffraction have an average uncertainty of a few tenths of an angstrom for non-hydrogen atoms,[15] with the greatest errors arising from positioning side chains into regions of very low electron density. Crystal-to-crystal differences for closely related proteins or different crystal forms are about 1.0 Å.[16] NMR-generated coordinates have precisions of 0.5–1.0 Å in the backbone region and 1.5 Å or greater in average side chain positions.[17] Homology modeling can be calibrated through structures of related proteins, with minimum errors of 0.5–1.0 Å for the backbones of highly similar sequences and much larger (and uncertain) side chain errors in loop regions.[18] Thermal displacements in macromolecules have both harmonic and anharmonic components, and they average a few tenths of an angstrom at room temperature within domains.[19] Larger conformational deformations between domains can also occur.[20] Structural models accurate to 1.0–2.0 Å have been useful for structure-based design.[3]

Site representations can be as simple as the atomic coordinates of the receptor. A crucial question is treatment of hydrogen atoms.[21] Alternatively, there are derived quantities such as the van der Waals surface,[22] the molecular surface,[23,24] the solvent-accessible surface,[23] or the extra radius surface.[8] The molecular surface is differentiable, whereas the van der Waals surface and the solvent-accessible surface contain cusps. Site volume can be defined if a site boundary is identifiable. The space available for ligand binding may also be characterized by points and the distances among them,[25] or by bond vectors targeting receptor groups of interest.[26] Physical and chemical information may be associated with surface or site points; scalar properties such as the electrostatic potential, vector properties such as the hydrophobic moment, and discrete classifications such as polar versus nonpolar have all been employed (see below).

Ligand descriptions closely parallel site descriptions. Structures for approximately 100 000 organic and inorganic compounds have been determined experimentally using X-ray or neutron crystallography and are available in the Cambridge Structure Database.[27] A list of purchasable compounds, the Available Chemicals Directory (formerly called the Fine Chemical Directory), is distributed by MDL Information Systems, Inc., San Leandro, CA. Compounds that have been synthesized are cataloged in the Chemical Abstracts Registry.[28] To obtain approximate three-dimensional coordinates for compounds not included in the experimental database, one can turn to programs such as systematic search,[29] MM3,[30] CONCORD,[31] distributed by Tripos Associates, St. Louis, MO, WIZARD,[32] and COBRA.[33] These programs use a combination of force fields and heuristic rules to generate atomic coordinates. Some of them produce only a single conformation, while others generate several low-energy conformations. Experimental accuracy for small molecules is very high, typically better than 0.1 Å. The major uncertainty is the preferred conformation in the receptor environment.

Juxtaposition of the Ligand and Site Frames of Reference. The objective of molecular docking is to obtain the lowest free energy structure(s) for the receptor–ligand complex. Among the many reasons for doing this are (1) searching a database of putative ligands and ranking them in terms of their interaction energies with a particular receptor; (2) calculating the differential binding of a ligand to two different macromolecular receptors; (3) studying the geometry of a

(9) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons, Inc: New York, 1990.
(10) Mezey, P. G. *Rev. Comput. Chem.* **1990**, *1*, 265.
(11) Salomaa, A. *Theor. Comput. Sci.* **1991**, *88*, 127–138.
(12) Ullman, J. R. *J. Assoc. Comput. Mach.* **1976**, *16*, 31.
(13) Sedgewick, R. *Algorithms*; Addison-Wesley: London, 1984.
(14) Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J.; Allen, F. H.; Bergerhoff, G.; Seivers, R. *Protein Data Bank*; Data Commission of the International Union of Crystallography: Bonn/Cambridge/Chester, 1987; pp 107–132.
(15) Chambers, J. L.; Stroud, R. M. *Acta Crystallogr.* **1979**, *B35*, 1861–1874.
(16) Kossiakoff, A. A.; Randal, M.; Guenot, J.; Eignebrot, C. *Proteins* **1992**, *14*, 65–74.
(17) Billeter, M. *Q. Rev. Biophys.* **1992**, *25*, 325–377.
(18) Chothia, C.; Lesk, A. M. *EMBO J.* **1986**, *5*, 823–826.
(19) Frauenfelder, H.; Hartmann, H.; Karplus, M.; Kuntz, I. D., Jr.; Kuriyan, J.; Parak, F.; Petsko, G. A.; Ringe, D.; Tilton, R. F., Jr.; Connolly, M. L.; Max, N. *Biochemistry* **1987**, *26*, 254–261.
(20) Karplus, M.; McCammon, J. A. *Annu. Rev. Biochem.* **1983**, *52*, 263–300.
(21) Weiner, P. K.; Kollman, P. A. *J. Comput. Chem.* **1981**, *2*, 287–303.
(22) Bash, P. A.; Pattabiraman, N.; Huang, C.; Ferrin, T. E.; Langridge, R. *Science* **1983**, *222*, 1325–1327.
(23) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.
(24) Connolly, M. L. *Science* **1983**, *221*, 709–713.
(25) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J. Mol. Biol.* **1992**, *161*, 269–288.
(26) Bartlett, P. A.; Shea, G. T.; Telfer, S. T.; Waterman, S. *CAVEAT: A Program to Facilitate the Structure-derived Design of Biologically Active Molecules*; Roberts, S. M., Ed.; Royal Society of Chemistry: London, 1989; pp 182–196.
(27) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. *Acta Crystallogr., Sect. B* **1979**, *B35*, 2331–2339.
(28) Fisanick, W. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145–154.
(29) Marshall, G. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. The conformational parameter in drug design: The active analog approach. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979; Vol. 112, pp 205–206.
(30) Allinger, N. L.; Yuh, Y. H.; Lii, J. J. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.
(31) Rusinko, A.; Sheridan, R. P.; Nilakatan, R.; Haraki, K. S.; Bauman, N.; Venkataghavan, R. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251–255.
(32) Dolata, D. P.; Leach, A. R.; Prout, K. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 73–85.
(33) Leach, A. R.; Prout, K. *J. Comput. Chem.* **1990**, *11*, 1193.

particular complex; and (4) proposing modifications of a lead molecule to optimize interactions.

The most systematic approach is to search through all binding orientations of all conformations of the ligand and receptor. Even with relatively crude sampling grids and a limited choice of "preferred" conformations, this brute force algorithm requires exponentially increasing resources as the molecules increase in size. It is not practical for docking two macromolecules and has severe limitations even in small molecule searches.

There are two major classes of automated searching. *Geometric* methods match ligand and receptor site descriptors. These procedures are intrinsically combinatorial and require a judicious limitation of the number of descriptors and the use of heuristic rules for pruning the search tree. Alternatively, one can align by minimizing the receptor–ligand interaction *energy*. Energy-driven searching, based on molecular dynamics (MD) and Monte Carlo (MC) simulations, is well-studied[20,34] and has been applied to a wide variety of chemical problems. The disadvantage is the enormous computational resources required for an extensive search.

Evaluation of Complementarity. The quantity of interest for ligand design is the free energy of binding (ΔG_{bind}) in aqueous solution. It can be calculated directly using free energy perturbation methods if an accurate geometric model of the complex is available.[35] Free energy calculations are demanding of computer time and cannot be carried out in conjunction with a data base search.

Many simplifications have been introduced: united atoms;[21] replacing explicit water molecules with a dielectric continuum or a distance-dependent dielectric function;[36,37] approximating an ensemble of structural configurations with a single "snapshot" structure; studying individual conformations of ligands and receptors as rigid objects; neglecting intramolecular energy contributions; and using interaction enthalpies instead of free energies.

The hydrophobic effect requires special attention. It is mainly a statistical entropic effect,[38] and its calculation requires an ensemble of configurations of explicit water molecules. Empirical models focus on solvent-exposed surface area,[37,39,40] with a recent extension to consider surface curvature.[41] Surface area contributions to the free energy have been derived from experimental free energies of transfer.[42] These empirical values include both the hydrophobic effect and enthalpic interactions of atoms with the solvent.

Methods that employ single rigid conformations and that neglect conformational energy terms depend on the assumption that complexation does not distort molecules very far from their dominant conformations in the unbound state. Many molecular complexes violate this assumption and show "induced fit".[5-7] A partial remedy is to use a variety of low-energy conformations treated as independent complexes. Alternative conformations for protein side chains can be explored,[43,44] but variants of the protein backbone are hard to predict.

Many evaluation schemes use far simpler scoring functions that bear little or no resemblance to a full force field. A common strategy is to use very simple functions during an early screening step and then more elaborate functions at later stages.

It is surprisingly difficult to determine the accuracy of calculational techniques. In the *best* cases, the free energy perturbation method has been reported to agree with experiment within ±1 kcal/mol. Energy minimization, yielding only an approximation to the enthalpy, is much less accurate. The usefulness of molecular force fields is critically sensitive to solvent representation and the description of the dielectric behavior of the system. Comparisons of a diverse series of molecules would rarely be more accurate than ±2 kcal/mol, and more often this technique is gives semiquantitative rankings with uncertainties of ±5 kcal/mol. Force field calculations are expected to be more reliable when a family of quite similar molecules are compared.

In summary, many aspects of the molecular docking problem have been studied. Algorithms have been developed that, in the best cases, yield quite accurate results for the free energy of interaction. Approximation calculations have proven useful in screening large numbers of diverse compounds (see below).

Types of Docking Programs

Finding the low-energy states of ligand–receptor complexes presents the fundamental problem that receptor sites have complicated and adjustable shapes and there are many ways of fitting a flexible ligand to them. We discuss three geometric search methods, based on descriptors, grids, and fragments. We then discuss energy-driven searches. We concentrate on programs that have been tested for ligand discovery and design.

Descriptor Matching Methods. These approaches analyze the receptor for regions of likely complementarity. Ligand atoms are placed at the "best" positions in the site, thus generating a reasonable ligand–receptor configuration that may be refined by optimization. Descriptor matching methods are rarely exhaustive, but they are fast and can usually provide satisfactory sampling of a particular region of the receptor site. Many of these algorithms use combinatorial search strategies, and small changes in parameter values can sometimes move the problem out of the feasible range of computer time.

DOCK, one of the earliest descriptor matching programs, uses spheres locally complementary to the receptor molecular surface[24] to create a space-filling negative image of the receptor site.[25] Site descriptions require 30–150 spheres. Several (four or five) ligand atoms are matched with receptor spheres to generate chiral orientations of the ligand in the site. Originally, DOCK relied on steric complementarity and used single

(34) Jorgensen, W. L.; Nguyen, T. B. *J. Comput. Chem.* **1993**, *14*, 195–205.
(35) Straatsma, T. P.; McCammon, J. A. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407–435.
(36) Daggett, V.; Kollman, P. A.; Kuntz, I. D. *Biopolymers* **1991**, *31*, 285–304.
(37) Cramer, C. J.; Truhlar, D. G. *Science* **1992**, *256*, 213–217.
(38) Dill, K. A. *Biochemistry* **1990**, *29*, 7133–7155.
(39) Chothia, C. *Nature* **1974**, *248*, 338–339.
(40) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
(41) Nichols, A.; Sharp, K. A.; Honig, B. *Proteins* **1991**, *11*, 281–296.
(42) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199–203.
(43) Ponder, J. W.; Richards, F. M. *J. Mol. Biol.* **1987**, *193*, 775–791.
(44) Wilson, C.; Mace, J. E.; Agard, D. *J. Mol. Biol.* **1991**, *220*, 495–506.

Table 1. DOCK Leads Developed at UCSF

system	affinities (μM) first lead	second generation	ref
HIV protease	100	0.8	49, 76
B-form DNA	10		77
thymidylate synthase	900	3	50
hemagglutinin	100	5	78
CD4-gp120[a]	5	1	79
malaria protease[b]	10	0.1	80

[a] Developed in collaboration with Procept, Inc., Cambridge, MA. [b] Structure obtained from homology model-building.

ligand conformations.[25] Recently, we have added molecular force fields,[45] limited conformational searches,[46] and chemical labeling of descriptors.[47] DOCK can model individual receptor–ligand complexes,[25,48] but its most common use is in the discovery of novel inhibitors.[3,49,50] Databases of small molecules (see above) are searched for candidates that complement the structure of the receptor.[51] DOCK has found novel, micromolar inhibitors for several receptors of therapeutic interest (Table 1). Its limitations are those of all descriptor matching programs: sensitivity to the quality of the negative image; nonexhaustive searches; and limited conformational exploration. As with most descriptor matching methods, it is relatively fast (Table 2).[45,52]

CAVEAT is based on directional characterization of ligands.[26] It searches for ligands with atoms located along specified vectors, typically derived from structural information from known complexes. CAVEAT rapidly searches reformatted versions of the usual ligand databases. The program has been successfully used to design α-amylase inhibitors, as well as non-peptide mimics of somatostatin. CAVEAT focuses on finding templates as starting points for chemical modification.

FOUNDATION represents an important attempt to combine models of the crucial ligand atoms (pharmacophore models) and structure-based methods.[53] The user identifies atom and bonding types that a candidate molecule must possess. Steric constraints of the receptor binding site eliminates candidates that do not complement the shape of the binding site, and candidates are oriented in the receptor site. FOUNDATION relies heavily on detailed atom-type, bond-type, chain-length, and topology constrains to restrict its search. FOUNDATION only considers the steric component of the active site, relying on its matching information to find chemically complementary ligands. The tight constraints restrict the candidates to one orientation in the site, whereas in DOCK and CLIX, many orientations are sampled.

CLIX[54] resembles DOCK by using receptor site features to define possible binding configurations. CLIX

Table 2. Docking Methods

authors	program name	algorithm	conform search	time[a] (h:min), small ligand	time[a] (h), protein ligand	accuracy	scoring	database search/ligand discovery	ligand design
Wodak and Janin[62]		grid search	no	[b]				no	no
Cherfils et al.[63]		grid search and annealing	yes	na			force field	no	no
Jiang and Kim[64]	Soft Docking	grid search	no	18:30 (v785)	120 (v785)	10?	polar/apolar contact	no	no
Hart and Read[69]		fragment	potentially	yes:no timing given	yes:no timing given	<1 Å	force field	no	potentially
Miranker and Karplus[68]	HOOK	fragment	potentially	"several"	nd	unknown	force field	no	potentially
Lewis et al.[70]	BUILDER	fragment	yes					no	yes
Goodsell and Olson[72]	AutoDock	kinetic	yes	unknown	na	na	geometry	yes	no
Caflisch et al.[74]		kinetic	yes			<1 Å	distance constraints	no	no
Kuntz et al.[45]	DOCK	descriptor	no	0:01	0:05–1	~1 Å	force field	no	no
Bartlett[26]	CAVEAT	descriptor	no	unknown	na	<1 Å	force field	yes	no
Yue[59]		descriptor	no			0.2–3.1	distance constraints	yes	no
Wang[65]		descriptor	no	nd[c]	6020	~1 Å	buried surface area	no	no
Kasinos et al.[57]		descriptor	look up	yes	~2	0.2–3.1	distance constraints	no	no
Lawrence and Davis[54]	CLIX	descriptor	no	yes:no timing given	nd	<1 Å	force field	yes	no
Smellie et al.[58]		descriptor	no	0:01	nd	unknown	matching H-bonds	no	no
Bacon and Moult[56]		descriptor	no	0:23	1–4	~1 Å	electrostatic interaction	no	no

[a] Times are approximate and depend on hardware and many parameter choices. [b] Not applicable. [c] Not done.

(45) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 505–524.
(46) Leach, A. R.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 730–748.
(47) Shoichet, B. K.; Kuntz, I. D. *Protein Eng.*, in press.
(48) Shoichet, B. K.; Kuntz, I. D. *J. Mol. Biol.* **1991**, *221*, 327–346.
(49) 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.; Craik, C. S. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 6644–6648.
(50) Schoichet, B. K.; Stroud, R. M.; Santi, D. V.; Kuntz, I. D.; Perry, K. M. *Science* **1993**, *259*, 1445–1450.
(51) DesJarlais, R.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. *J. Med. Chem.* **1988**, *31*, 722–729.
(52) Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 380–397.
(53) Ho, C. M. W.; Marshall, G. R. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 3–22.
(54) Lawrence, M. C.; Davis, P. C. *Proteins* **1992**, *12*, 31–41.

relies on an elaborate chemical description of receptor site "hot spots," using the 23 different classes of receptor environment provided by GRID.[55] It uses fewer receptor–ligand matches than does DOCK. CLIX has been used to search for ligands complementary to the sialic acid binding site of hemagglutinin, returning several interesting structures. The program is fairly fast (Table 2). It does not allow for ligand or receptor conformational flexibility. CLIX evaluates interaction energies using the GRID potential function.

Other Descriptor Methods. Bacon and Moult use surface "webs" to describe interacting proteins.[56] Evaluation includes solvation-corrected electrostatics and surface area burial. The search time is longer for small-molecule docking problems than in DOCK or CLIX, but the algorithm scales well with molecular size. The method of Kasinos et al.[57] relies heavily on graph-theoretical ideas. It requires maximal common subgraphs (the largest possible number of receptor–ligand descriptor correspondences) to generate and evaluate structures. The authors use only potential hydrogen bonding sites as their descriptors. Smellie[58] also describes an application of graph theory to the docking problem. The method is applied to the binding of Baker triazines to dihydrofolate reductase (DHFR) and might be generally useful. It suffers from an overly simple evaluation function. Yue[59] uses the structure of a known ligand–receptor complex to dock similar ligands. High accuracy is achieved, but the method is probably restricted to docking similar ligands into a receptor complex of known structure. The description of knobs and holes used by Connolly to dock α and β subunits of hemoglobin[60] includes an interesting and provocative discussion of some of the important issues in protein docking.

Grid Search Methods. Grid searches sample the six degrees of freedom of orientation space. They find the neighborhood of the correct solution(s), which cannot be guaranteed with discrete sampling methods. Accuracy is limited by the step size used in the search, which also determines the time of the search. Methods that use additional sampling in regions of high complementarity can overcome this problem.

Side Chain Spheres. Wodak and Janin explored protein–protein complexes using simplified sphere representations of side chain atoms and a grid search of four rigid degrees of freedom.[61] The approach has been extended recently to include more sophisticated surface burial evaluation algorithms,[62] full molecular force-field evaluation of complexes, and simulated annealing to refine initial docked structures.[63] An interesting observation,[63] which we have also made,[48] is the generation of structure that are dissimilar from the crystallographic result but cannot be distinguished from it by the available energy criteria.

Soft Docking[64] divides receptor and ligand surfaces into cubes to generate the translational part of the search. A pure rotational grid search samples ligand orientations in discrete angular increments. The accuracy is limited by step size, with run time scaling as the cube of the rotational step size and as the product of the number of receptor and ligand surface points. Soft Docking has been successful in docking NADPH and methotrexate into DHFR, as well as bovine pancreatic trypsin inhibitor (BPTI) into trypsin and lysozyme into an antibody. Noncrystallographic solutions are occasionally generated. The authors provide rough estimates that docking a small ligand should take ~12 h on a Vax 785.

Wang[65] provides an illustration of the perils of extensive optimization strategies in molecular docking. In his method, one ligand "knob" is fit into one receptor "hole", and the resulting complex is rotated about one angle in discrete increments, with energy optimization (based on surface area burial) of each structure. Docking BPTI to trypsin took 6020 h on a small workstation, making this method impractical in its present form.

Fragment-Joining Methods. Fragment methods identify regions of high complementarity by docking functional groups independently into receptors. They overcome most of the rigid ligand issues at the expense of adding a combinatorial search over fragment types. These approaches can suggest unsynthesized compounds, but connecting the fragments in sensible, synthetically accessible patterns is a challenging problem. These methods are attractive for chemical elaboration.

GROW is a well-tested fragment method.[66] It designs peptides complementary to proteins of known structure. A seed amino acid is placed in the receptor site followed by iterative additions of amino acids. Conformations are chosen from a library of precalculated low-energy forms. At each addition the energy of the peptide and of the peptide–receptor complex is briefly minimized and evaluated. Only the best 10–100 low-energy structures are kept at any stage. "Growing" a heptapeptide takes about 40 min of cpu time on a workstation. GROW has some very strong features. The use of peptides ensures ease of synthesis, setting it apart from most other methods, except for some of the recent work using the ACD with DOCK.[50] A heptapeptide that inhibits renin with a K_i of 30 μM was designed. A GROW program for organic molecules is under development.[67]

HOOK[68] finds "hot spots" in receptor sites by seeking low-energy locations for functional groups. HOOK differs from Goodford's GRID program by using random placement of many copies of several functional fragments followed by MD. HOOK was tested by reproducing sialic acid derivatives known to bind to hemagglutinin. The most serious drawback of HOOK is shared by all fragment methods: the need to reconnect functional groups to form complete molecules while maintaining the geometric positions of lowest energy.

(55) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849–857.
(56) Bacon, D. J.; Moult, J. *J. Mol. Biol.* **1992**, *225*, 849–858.
(57) Kasinos, N.; Lilley, G. A.; Subbarao, N.; Haneef, I. *Protein Eng.* **1992**, *5*, 69–75.
(58) Smellie, A. S.; Crippen, G. M.; Richards, W. G. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 386.
(59) Yue, S. Y. *Protein Eng.* **1990**, *4*, 177–184.
(60) Connolly, M. L. *Biopolymers* **1985**, *25*, 1229–1247.
(61) Wodak, S. J.; Janin, J. *J. Mol. Biol.* **1978**, *124*, 323–342.
(62) Wodak, S. J.; De Crombrugghe, M.; Janin, J. *Prog. Biophys. Mol. Biol.* **1987**, *49*, 29–63.
(63) Cherfils, J.; Duquerroy, S.; Janin, J. *Proteins* **1991**, *11*, 271–280.
(64) Jiang, F.; Kim, S. H. *J. Mol. Biol.* **1991**, *219*, 79–102.
(65) Wang, H. *J. Comput. Chem.* **1991**, *12*, 746.
(66) Moon, J. B.; Howe, J. W. *Proteins* **1991**, *11*, 314–328.
(67) Howe, W. J.; Moon, J. B. Preprint.
(68) Miranker, A.; Karplus, M. *Proteins* **1991**, *11*, 29–34.

Multiple-Start Monte Carlo. Hart and Read[69] use Monte Carlo searches to dock fragments of a ligand into a receptor site. MSMC found the crystallographic solutions when docking ovomucoid third domain to proteinase B and methotrexate into DHFR, though not always as the lowest energy solutions.

BUILDER. This program uses a family of docked structures to provide an irregular lattice of controllable density, which can be searched for paths that link molecular fragments. It has been shown to generate chemically reasonable compounds in the HIV protease site.[70]

LUDI. A prototypic fragment joining effort, LUDI[71] proposes inhibitors by connecting fragments that dock into microsits on the receptor. The fragments come from a list of approximately 600 molecular fragments such as benzene, adamantine, and naphthol. Microsites are defined by hydrogen bonding and hydrophobic groups using the author's own algorithm, or using the output of GRID.[55] Ligand pseudoatom (hot spot) positions are generated within microsites on the basis of the appropriate angle and distance minima for various interactions. In this respect the method resembles descriptor methods. In the last stage, the fragments are connected together using linear chains composed of one or more of 12 different functional groups, including CH_2, CO, CONH. LUDI, as all fragment methods, will have to cope with synthetic feasibility issues as *de novo* inhibitors are constructed. The program can also be used to add functionality to a known inhibitor.

Energy Search Methods. These docking techniques use MD or simulated annealing and employ full molecular mechanics force fields. They smoothly merge the configurational and conformational aspects of docking. However, the complex topography and multiple minima of molecular potential surfaces often lead to relatively long run times.

Simulated Annealing. Goodsell and Olson[72] use the Metropolis algorithm to find low-energy complexes of ligands in receptor sites, searching all configurational and several conformational degrees of freedom. The program was tested by docking phosphocholine into the antibody McPC 603, *N*-formyltryptophan into chymotrypsin, *N*-acetylglucosamine (two anomers) into lysozyme, and sulfate and citrate into aconitase. In most cases the crystallographic solution was reproduced to better than 2 Å. The program is quite efficient considering the number of degrees of freedom. It is not clear how the run time scales with conformational freedom for larger systems. Stoddard and Koshland[73] have applied this algorithm to predict the structure of the maltose binding protein–aspartate receptor complex.

Peptide Docking. Caflisch and co-workers[74] use graphics to place peptides in binding sites, followed by energy minimization and then a local search using Monte Carlo simulation. Test cases were a peptidic inhibitor of HIV-1 protease and, courageously, an undecapeptide recognized by the HLA-A2 protein where the structure was not known at the time of writing. In the HIV-1 case, Monte Carlo refinement yielded a configuration closely resembling the crystallographic complex. In the HLA-A2 case, the peptide was docked in a helical conformation and then subjected to local refinement by Monte Carlo searching. The helical complex, while consistent with the mutagenesis work then available, is in conflict with recent crystallographic results in which peptides bind in an extended geometry.[75]

Table 3. Examples of Structure-Based Drug Design Leading to Clinical Trials

system	company	ref
thymidylate synthase	Agouron	4
purine nucleoside phosphorylase	Biocryst, Ciba-Geigy	81
HIV-1 protease	Merck	82
	DuPont Merck	83

Conclusions

We conclude this Account with a summary of our experience with the DOCK program and the challenges that lie ahead for structure-based design. Certain aspects of the docking problem have been solved. (1) The negative imaging procedure used with DOCK automatically locates concave features that are potential binding pockets. (2) We can reassemble the components of a complex into a geometry within 1 Å rms of the experimental structure. That is, the problem of constructing a "three-dimensional jigsaw puzzle" from *rigid pieces of proper conformation* has been solved. (3) Multiple binding geometries, plausible on steric and chemical grounds, are routinely seen. The number of alternatives increases when conformational freedom is introduced. To sort among these states requires quite accurate determinations of free energy (e.g., ±1 kcal/mol).

Table 1 shows the status of projects in which DOCK was used to identify inhibitors. The procedure finds novel lead compounds in the micromolar range for a very wide range of macromolecular systems. Clearly the "lock-and-key" model has some utility at this level. Table 3 lists a few examples from the literature where structue-based efforts have led to compounds considered for clinical trials.

Looking across all the methods presented in Table 2 and the relative paucity of results to data, it is premature to make definitive statements. Instead, we summarize the progress toward four uses of structure-driven design: screening for new leads; rank-ordering similar and diverse compounds; proposing preferred ligand–receptor geometries; and rapid, semiautomatic optimization of a lead compound.

Present programs are relatively successful at lead identification. This is, of course, the easiest task since false positives are acceptable and false negatives are not recognized. Typical hit rates of 1–10% are competitive with high-throughout experimental screens, and computer screening is much less expensive. An important issue is to determine the false negative rate in computer screening. An appropriate test would be to

(69) Hart, T. N.; Read, R. J. *Proteins* **1992**, *13*, 206–222.
(70) Lewis, R. A.; Roe, D. C.; Huang, C.; Ferrin, T. E.; Langridge, R.; Kuntz, I. D. *J. Mol. Graphics* **1992**, *10*, 66–78.
(71) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 593–606.
(72) Goodsell, D. S.; Olson, A. J. *Proteins* **1990**, *8*, 195–202.
(73) Stoddard, B. L.; Koshland, D. E. *Nature* **1992**, *358*, 774–776.
(74) Caflisch, A.; Niederer, P.; Anlinker, M. *Proteins* **1992**, *13*, 223–230.
(75) Silver, M.; Go, H.; Strominger, J.; Wiley, D. *Nature* **1992**, *360*, 367–369.

run a substantial, diverse database of compounds through both computer and experimental screenings.

The task of rank-ordering the binding energies for a diverse set of compounds is more difficult. Our experience is that force fields and empirical energy functions can rarely achieve better than ±2 kcal/mol accuracy except within a family of compounds that have little conformational flexibility and that all bind in a very similar manner. A comparison of scores with K_i or IC_{50} data spanning several orders of magnitude for a set of known inhibitors would provide a stringent test of current scoring functions.

The most difficult task is to propose accurate (±1 Å rms) geometric models. With DOCK, the best cases have shown displacements of 1–2 Å from the predicted geometry. In the worst cases, the displacements are about 5 Å.[50,76] Complications include the conformational freedom of the ligand and the receptor, the possibility of alternative binding modes (configurational freedom), and the inclusion of water molecules and ions as part of the binding complex. We have had problems with all of these phenomena. The obvious design/test protocol is to combine structural experiments with computations to provide rapid assessment of the degrees of freedom of a particular system of interest.[50] The most successful efforts at structure-based design have used one X-ray structure per one to two compounds synthesized![4]

De novo strategies have the potential to assist in the optimization process, especially if coupled to a rule-based approach to what modifications are synthetically feasible. An appropriate challenge for such methods is to re-create known high-affinity inhibitors in the proper conformations, given the relevant binding site.

The fundamental limitations of computational methods are in sampling conformational and configurational space (the "induced-fit" problem) and evaluating the free energy of interaction. Hardware advances, reparametrization of force fields, and improved heuristics should all contribute to significant improvements in the near future. Opportunities for directed ligand design will increase dramatically as the number of solved structures grows and the molecular mechanics of disease are clarified. It will be an exciting challenge to make maximal use of this new information to design new molecules.

Many of our colleagues contributed to the software described here; Drs. Renée DesJarlais, J. Scott Dixon, George Seibel, Robert Sheridan, and Dale Bodian especially should be thanked. We are grateful for the interactions provided by the UCSF Computer Graphics Laboratory. Support for the work was provided by the National Institutes of Health, the Advanced Research Project Agency, Molecular Design Limited Information Systems, Tripos Associates, Parke-Davis, Smithkline Beecham, and Glaxo.

(76) 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.; Stroud, R. M. *J. Biol. Chem.* **1993**, *268*, 15343–15346.

(77) Kerwin, S. M.; Kuntz, I. D.; Kenyon, G. L. *Med. Chem. Res.* **1991**, *1*, 361–368.

(78) Bodian, D. L.; Yamasaki, R. B.; Buswell, R. L.; Stearns, J. F.; White, J. M.; Kuntz, I. D. *Biochemistry* **1993**, *32*, 2967–2978.

(79) McGregor, M.; Cohen, F. E.; Kuntz, I. D. Unpublished results.

(80) Ring, C. S.; Sun, E.; McKerrow, J. H.; Lee, G. K.; Rosenthal, P. J.; Kuntz, I. D.; Cohen, F. E. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 3583–3587.

(81) Montgomery, J. A.; Niwas, S.; Rose, J. D.; Secrist, J. A., 3rd; Babu, Y. S.; Bugg, C. E.; Erion, M. D.; Guida, W. C.; Ealick, S. E. *J. Med. Chem.* **1993**, *36*, 55–69.

(82) Ghosh, A. K.; Thompson, W. J.; Lee, H. Y.; McKee, S. P.; Munson, P. M.; Duong, T. T.; Darke, P. L.; Zugay, J. A.; Emini, E. A.; Schleif, W. A.; Huff, J. R.; Anderson, P. S. *J. Med. Chem.* **1993**, *36*, 924–927.

(83) Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C.-H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. *Science* **1994**, *263*, 380–384.