Automated Flexible Ligand Docking Method and Its Application for Database Search

SHINGO MAKINO,[1,2] IRWIN D. KUNTZ[1]

[1]*Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California 94143-0446*
[2]*Central Research Lab, Ajinomoto Co., Inc., Kawasaki, Japan*

Received 13 February 1997; accepted 11 June 1997

ABSTRACT: We have developed a new docking program that explores ligand flexibility. This program can be applied to database searches. The program is similar in concept to earlier efforts, but it has been automated and improved. The algorithm begins by selecting an anchor fragment of a ligand. This fragment is protonated, as needed, and then placed in the receptor by the DOCK algorithm, followed by minimization using a simplex method. Finally, the conformations of the remaining parts of the putative ligands are searched by a limited backtrack method and minimized to get the most stable conformation. To test the efficiency of this method, the program was used to regenerate ten ligand–protein complex structures. In all cases, the docked ligands basically reproduced the crystallographic binding modes. The efficiency of this method was further tested by a database search. Ten percent of molecules from the Available Chemicals Directory (ACD) were docked to a dihydrofolate reductase structure. Most of the top-ranking molecules (7 of the top 13 hits) are dihydrofolate or methotrexate derivatives, which are known to be DHFR inhibitors, demonstrating the suitability of this program for screening molecular databases. © 1997 John Wiley & Sons, Inc. *J Comput Chem* **18**: 1812–1825, 1997

Keywords: automated docking; flexible ligand; database search; protonation; minimization

Correspondence to: I. D. Kuntz
Contract/grant sponsor: NIGMS; grant number GM-31497, GM-39552

CCC 0192-8651 / 97 / 141812-14

Introduction

The development of new computer hardware and software and the dramatic growth in the available x-ray and NMR protein structures offer many opportunities for discoveries of ligands which bind to the macromolecular structures.[1,2] The prediction of the geometric binding mode is a critical issue both for lead discovery from molecular databases[3] and for lead optimization.[4–6] Although there have been many reports dealing with rigid-body docking methods,[7–9] it is generally expected that the correct docking modes cannot be obtained unless the probe conformation is quite similar to the active conformation. Thus, consideration of the flexibility of ligands has received much recent attention. Algorithms that take the conformational flexibility of ligands into account have been reported: docking fragments and joining them together[10]; simulated annealing[11]; Monte Carlo search[12]; genetic algorithms[13,14]; and incremental searches.[1,15–17] For carrying out database searches, a program has to satisfy the following conditions: (1) all processes should be fully automated; (2) it should identify the most stable conformations; and (3) calculations should be rapid enough to search through a database efficiently. Recently, two programs, HAMMERHEAD[16] and FLEXX,[17] have been reported, which place flexible ligands into macromolecules speedily. Both methods use an heuristic search algorithm. The FLEXX program is not fully automated and manual selection of the basic fragments is necessary. The HAMMERHEAD program is fully automated. In one test, biotin was identified as the top-scoring ligand in the ACD against streptavidin. Although several other successes are reported, this search is based on the assumption that all *partial* structures should have near optimum energies. In addition, the scoring functions are not reflections of physical forces. Here, we report an automated flexible ligand docking program using a limited backtrack algorithm which searches broadly among partial structures. Computation time is reduced by skipping conformations that are similar to conformations already sampled. Other new features include: (1) Classification of flexible bonds and anchor fragment identifications are automated. (2) Multiple protonation states of a ligand are treated simultaneously, so users do not have to decide in advance which protonation model of a ligand is the best for docking. (3) Information from hydrogen bond donor and acceptor sites and ligand shape are used for the initial docking, so a ligand with few or no hydrogen bonds can also be docked. (4) Binding modes of ligands are predicted rapidly by trimming the conformational tree without loss of diversity of the conformational sampling. Although some previous studies have tried to reduce computation time by using partial structures of the ligands,[15] or by fixing the number of rotatable bonds,[13] it is difficult to apply such a special treatment to all molecules in a large database. (5) The program also only estimates the necessary partial energy in the backtrack search to reduce the computational time. In this article, we describe the theory behind this method and test the program by regenerating x-ray complex structures. Finally, the program is applied to an ACD database search.

Methods

We describe the features of the algorithm in the following order: force field scoring, flexible bond identification, anchor identification, multistate protonation, site-point generation and docking, limited backtrack search, partial energy estimation in the backtrack search, and simplex minimization. We also discuss the programming language and resource usage. All the molecular modeling and charge calculations were performed using the SYBYL program from Tripos.[18] SYBYL atom types[19] used are shown in Table I.

FORCE-FIELD SCORE

We adopted an AMBER-type potential function[20,21] for measuring the affinity of ligands. In earlier work[22] with rigid ligands, we only used the intermolecular term. For this project, both inter and *intra*molecular interactions are examined to identify the best conformation of each ligand. However, for comparison of different ligands, only the intermolecular interaction is calculated, based on the assumption that the selected best conformation of a ligand has an intramolecular energy close to the gas-phase conformational minimum. The intermolecular interaction energy is calculated using the grid-based force field[23] and the receptors are assumed to be rigid. For intramolecular interaction, we used full AMBER potential on the fly.

TABLE I.
List of SYBYL Atom Types Used in This Experiment.

C.3	carbon sp3
C.2	carbon sp2
C.1	carbon sp
C.ar	carbon aromatic
C.cat	carbocation
N.3	nitrogen sp3
N.2	nitrogen sp2
N.1	nitrogen sp
N.ar	nitrogen aromatic
N.am	nitrogen amid
N.pl3	nitrogen trigonal planer
N.4	nitrogen sp3 positively charged
O.3	oxygen sp3
O.2	oxygen sp2
O.co2	oxygen in carboxylate and phosphate
S.3	sulfur sp3
S.2	sulfur sp2
S.O	sulfoxide sulfur
S.O2	sulfone sulfur
P.3	phosphorous sp3
H	hydrogen
H.spc	special hydrogen in this work (see text)

In the conformational search before minimization, the Van der Waals (vdw) radii of the ligand atoms are scaled to 0.8. This reduction in vdw radii makes it easier to find low energy conformations through a rapid conformational search. After finding the possible conformations, the vdw scale is reset to normal values and minimizations are performed with full-sized atoms.

FLEXIBLE BOND IDENTIFICATION

Flexible bonds are automatically identified by the following method. First, all bonds not part of ring systems are selected as candidates for flexible bonds. If a selected bond has the SYBYL bond type ''1'' (a single bond) and is not prohibited from being a flexible bond (see Table II), it is identified as a flexible bond.

ANCHOR IDENTIFICATION

Anchor fragments are automatically identified by the ''ANCHOR'' program shown in Figure 1. First, ligands are divided into the largest rigid fragments. Next, fragments that contain more than eight heavy atoms are selected. If a ligand does not have any fragments with more than eight heavy atoms, then that ligand is skipped. Among the selected anchor fragment candidates, the fragment with the maximum ''donor + acceptor'' atoms is selected as the anchor fragment. The ANCHOR program automatically adds the keyword ''⟨ANCHOR⟩'' and an appropriate anchor identification flag to SYBYL multiple mol2 format files. This anchor identification process only needs to be performed once when the molecule is added to a molecular database.

MULTISTATE PROTONATION

In some cases, protonation states are critical for docking. Although previous studies selected specific protonation states in their models of ligands,[15,17] it is still very difficult to determine which state of the protonation is the best for a docking study without knowing the real docking mode. Instead of predetermining the protonation states, multistate protonation options are treated simultaneously by our program. The program

TABLE II.
Prohibited Flexible Bonds.

Number	Atom Pattern 1	Atom Pattern 2[a]	Pattern Name
1	C.3(H)(H)(H)	C.* || N.* || O.* || Si.*	Methyl
2	N.2	N.2	Imine
3	C.2	C.2 || N.am || N.2 || N.pl3	Double and amide
4	C.1	C.1	Triple
5	C.[a]	N.pl3 || N.2 (H) (H)	Planar amine
6	C.ar	C.2 || C.cat (N.pl3 | N.2) (N.pl3 | N.2)	Amidine (aromatic)
7	C.ar	S.3 (O.co2) (O.co2) (O.co2)	Sulfonic acid (aromatic)
8	C.ar	N.* (O.2)	Nitro, nitroso (aromatic)
9	C.ar	C.2 || C.cat (O.co2) (O.co2)	Carboxylic acid (aromatic)

[a] ''||'' means ''or''; ''*'' means ''any symbols'' following the UNIX convention.

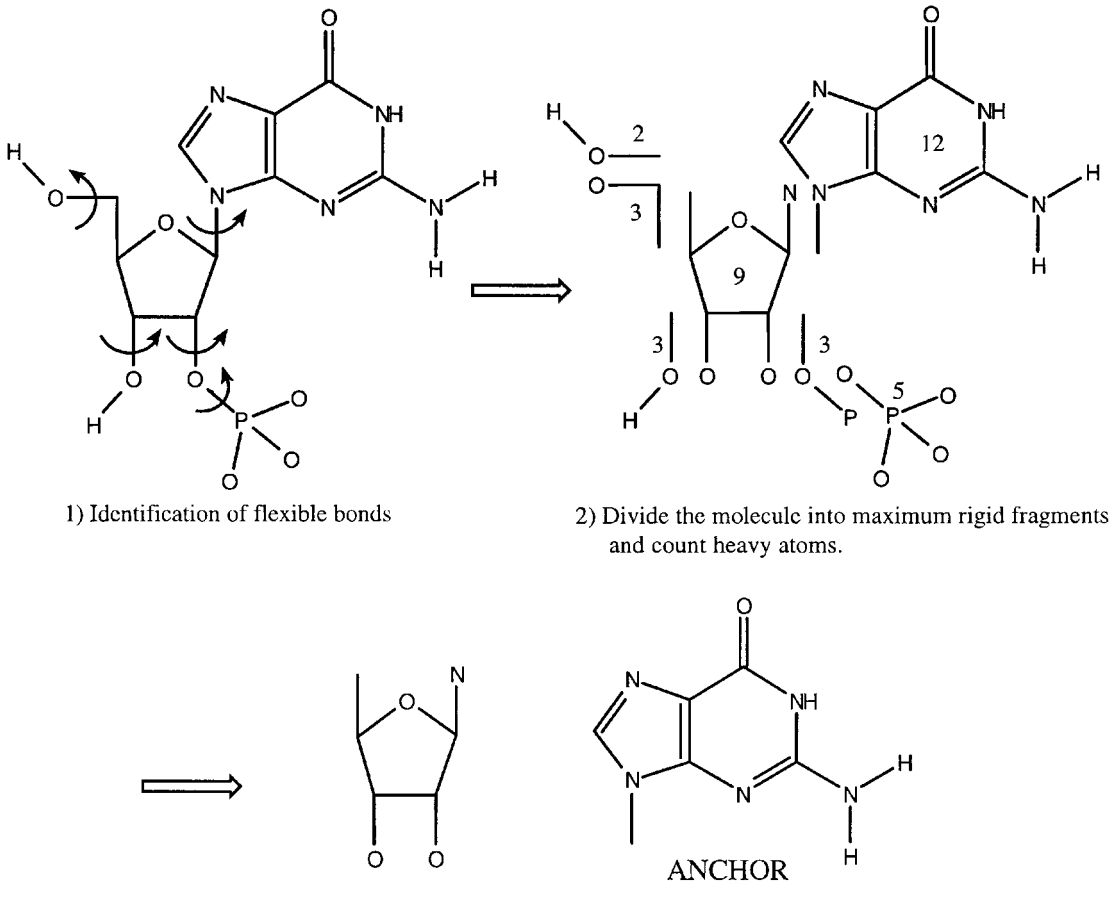

FIGURE 1. Identification of anchor fragment.

"PROTON" adds hydrogen (atom type = "H.spc") to the aromatic nitrogen (C.ar–N.ar–C.ar) in the anchor fragment (Fig. 2). The charge on the proton is 0.7 and the charges of the other atoms are not changed. The distance between N.ar–H.spc is fixed to 1.33 Å. The atom type "H.spc" is one of the SYBYL atom types, so ligands can be checked easily by using a SYBYL interface. Hindered aromatic amines that have CH_3 or NH_2 on both sides of the C.ar atom are considered hindered aromatic amines and are not protonated. After adding protons, the program evaluates the intramolecular energy between this newly added proton and atoms on the anchor fragment. If there are any atomic overlaps, this proton is removed. The force field has been modified to recognize the atom type "H.spc" and treats it in a special way. If a proton atom ("H.spc") interacts favorably with the protein, the program assumes that the proton exists. If not, the proton is ignored. Therefore, the protonation states do not need to be determined before docking, and multiple protonation states can be examined simultaneously. Protonation of ligands in databases should be done only once. Of course, it is important to examine the proposed protonation of all high-scoring ligands to establish that the protonation pattern is consistent with knowledge of ligand pK_a values.

SITE-POINT GENERATION AND DOCKING

We found that the "SPHGEN" program[24] usually provides appropriate site points for whole ligand docking; however, the site points are often

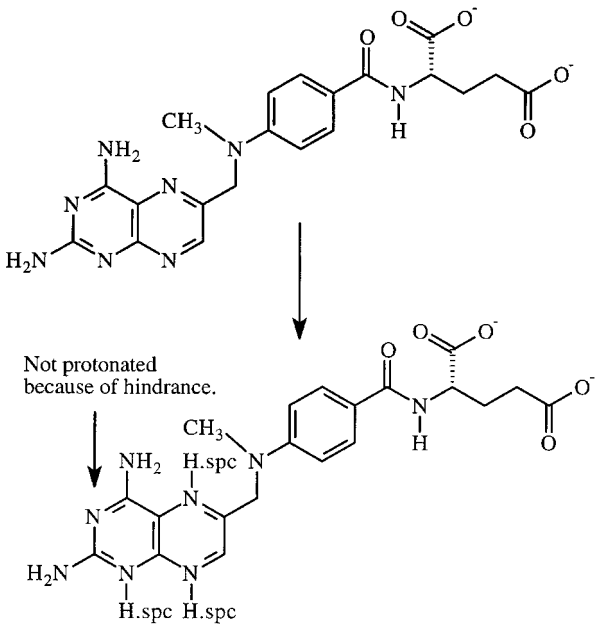

FIGURE 2. Example of multistate protonation.

not numerous enough for anchor fragment docking. We developed the program "SPHGEN++" for site-point generation of small fragment docking. First, site points are generated by the SPHGEN program (20 ~ 87 points). These site points are treated as shape descriptor site points (shape site points). Also, these site points define the receptor docking region to which the "SPHGEN++" program adds more site points. The additional site points are generated around the hydrogen acceptor and donor atoms and are called donor and acceptor site points (5259 ~ 11,546 points). The geometry for the new site points is shown in Table III. The donor and acceptor points that are near "SPHGEN"-generated site points (within the range of 2.7 Å) are selected (302 ~ 1017 points). Next, the site points are reduced by putting

TABLE III.
Geometry for Site Points.

Function Group	Nature	Geometry	Number
NH	Acceptor	1.9 Å, 5.0 ~ 35.0°	37
C=O, COO—, HO	Donor	1.9 Å, 10.0 ~ 30.0°	20
HO (Tyr)	Acceptor	2.8 Å, 5.0 ~ 65.0°	83

TABLE IV.
Matching Pattern.

Receptor Site	Ligand Site
Shape points	Heavy atoms
Donor points	Donor atoms
Acceptor points	Acceptor atoms

probe atoms on each site point. Hydrogen with a charge of 0.205 is used as a donor site probe and oxygen with a charge of −0.410 is used as an acceptor site probe (112 ~ 437 points). These points are added to shape descriptor site points and are reduced by merging the nearest site points of the same nature (personal communication from T. Ewing: 106 ~ 180 points). The matching and orientation algorithm is the same as DOCK4.[25] The site points are allowed to have several distinct types and any mismatch of the site points and the ligand atom types is rejected when the adjacent boolean matrix is built. The matching patterns of the site points are described in Table IV. For the ligand, we use the heavy atoms and also donor hydrogen atoms. DOCK4 matching has two critical parameters, minimum distance and distance tolerance, which usually depend on the system of ligand and receptor site points. We have set these parameters in the following way for database searches. The "distance tolerance" is kept constant (0.4 Å) and the "minimum distance" is increased automatically according to the maximum distance between the heavy atoms in the anchor fragment (see Fig. 3). Because anchor fragments do not have any flexible bonds, only rigid body simplex minimization is needed to dock anchor fragments. The final scores and orientation information are stored in a

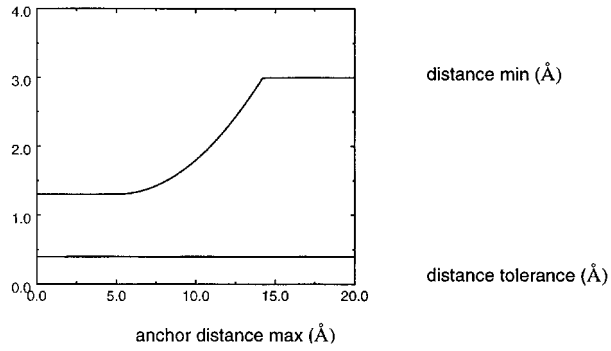

FIGURE 3. Parameters for docking.

binary tree structure. After exploring all of the possible orientations, the program clusters orientations and chooses one from each cluster. The program picks out the minimum score orientation and removes other similar orientations. Then, it chooses the second minimum score and the same cycle is repeated. Finally, the program writes the orientation information sorted by score into a file for further conformational search. Once an anchor fragment is docked in a specific location, that fragment can be displaced by the minimization procedure, but the anchor is not redocked as side chains are added.

LIMITED BACKTRACK SEARCH

Although a systematic search is an efficient way to explore the wide range of conformational space, the search space increases dramatically as the flexibility of a ligand increases. For example, if one torsion has N states ($N = 360.0/\text{search\_step\_angle}$), then a ligand with M flexible bonds has N^M conformations. However, a great amount of the conformational space can be pruned by trimming all the branches under bad nodes as shown in Figure 4a. This is called the backtrack method.[26] This method is efficient, but it will still take too long for a ligand with many flexible bonds, so we added a limitation to this backtrack method. This limitation fixes the number of conformations accepted below a certain level in the search tree as shown in Figure 4b. This version of the greedy algorithm[27] allows a reduction of sampling of similar conformations without reducing the sampling of the distinct conformations. We call this a ''limited backtrack search.'' The amount of sampling can be controlled by changing the level of the limitation. (compare numbers of sampling between Fig. 4b and 4c). This limitation should be better than taking the first N conformations, because such a sampling tends to take similar samples and skip distinct ones. For example, samples in Figure 4b are more diverse than in Figure 4d, even though the number of conformations sampled is the same. Additionally, for increased speed, at each node of the search tree, it is necessary to calculate matrices and vectors for rotation once.

PARTIAL ENERGY ESTIMATION IN THE BACKTRACK SEARCH

We tried to eliminate unnecessary calculations to the extent possible to speed up the docking calculations. On each node of the search tree, only

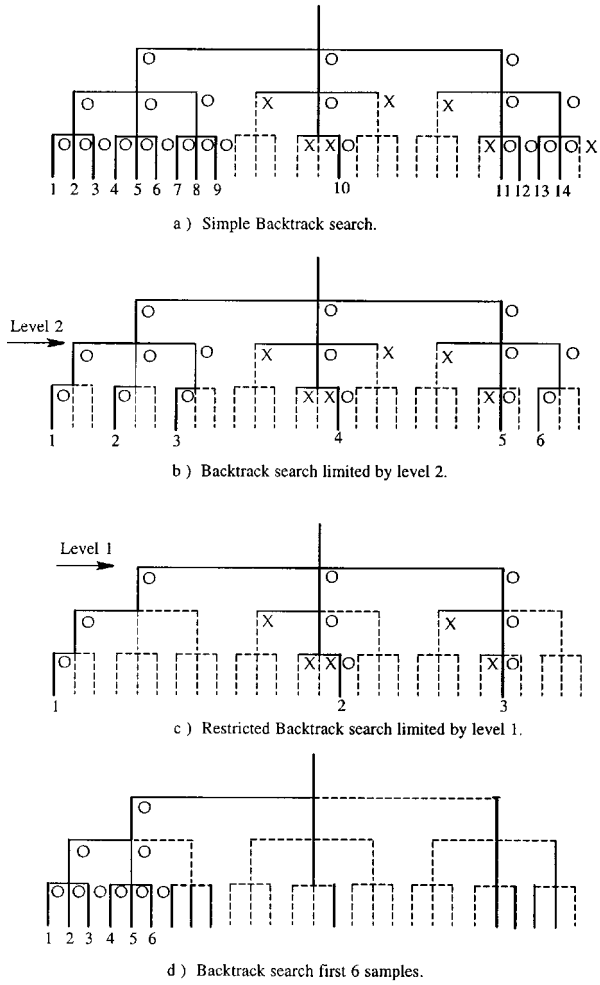

FIGURE 4. Search trees to show efficiency of the limited algorithm.

the partial energy is calculated for both inter- and intramolecular energy. Figure 5 shows how the atoms in a ligand are categorized depending on each search level. At every stage of the conformational search, the atoms designated by vertical lines and by a checkerboard pattern are not moved by changing the torsion angle, so the interaction energy between these atoms and the atoms of the receptor is constant and does not need to be recalculated. The blank atoms will be moved by the deeper flexible bond search, so their energies are not considered at this search level. Because the atoms with horizontal lines are moved by the current flexible bond and will not be moved by the deeper flexible bonds, these atoms are the only real variables at this level of the search. Thus, the energy between the atoms with horizontal lines and the atoms of the receptor are calculated as

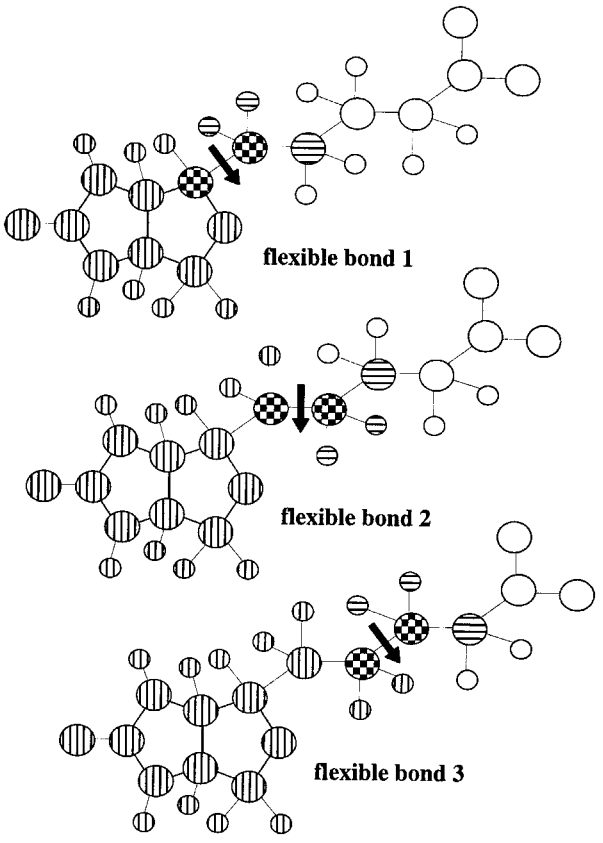

FIGURE 5. Categorization of atoms for partial energy estimation.

intermolecular energy terms and the energy between atoms with horizontal lines and the atoms with vertical lines are calculated as intramolecular energy terms on each node of the search tree. Notice that the intramolecular energy between atoms with the same pattern does not need to be calculated, because the distances between atoms with the same patern are always the same, thus their contribution to intramolecular energy remains constant. These reductions in calculation time are especially effective for intramolecular energy, which cannot be calculated with a grid-based force field. In our current version, we use a random seed to generate the initial conformation. Thus, two different runs can yield slightly different results (data not shown).

SIMPLEX MINIMIZATION

Minimization[28] is one of the most time-consuming steps in a ligand docking to a macromolecule. Thus, it is very important to accelerate a minimization. For this purpose, each branch is treated independently in the minimization process, because the dependency of the branches is considered in the search process. We will explain the efficiency of this method with the example of glucose (Fig. 6). In a backtrack search, flexible bonds are simply sorted by the ''bond weight'' (the number of atoms moved by the rotation of the bond). In a minimization, flexible bonds are categorized into branches. The simple branches from the anchor fragment are used and sub-branches or recursive branches are not used because recursive branching tends to introduce interdependence. We will compare the conformational space of independent branch minimization with the conformational space of the dependent branch minimization by this example. If each flexible bond has N status (N = minimization_range/step_degree), the conformational space for minimization will normally be N^6. If each branch is treated independently, then the minimization space will be $N^2 + 4N$ (N^2 for branch 1). For example, if $N = 12$ then $N^6 = 2{,}985{,}984$ and $N^2 + 4N = 192$. If there is only one branch, this categorization is superfluous. A simplex minimizer generates many conformations from the same conformation in each orientation, so the ''look-up matrix'' does not have to be reset for changing conformations. Each flexible bond has an independent transformation object especially for minimization, so unnecessary recalculations for matrix and vector transformations are avoided, thus speeding up the calculations.

PROGRAMMING LANGUAGE AND RESOURCE USAGE

All the programs except ''SPHGEN'' were written in C++ and compiled by GNU gcc-2.7.2. The

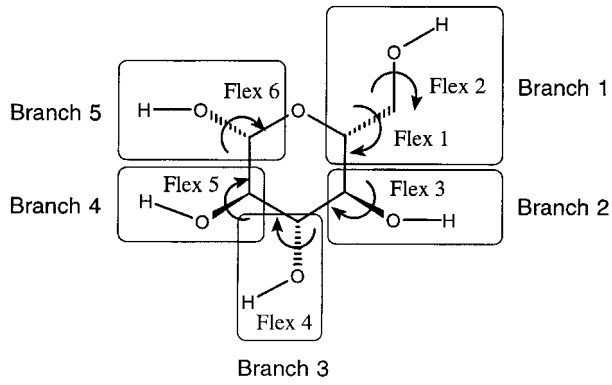

FIGURE 6. Analysis of flexible bonds.

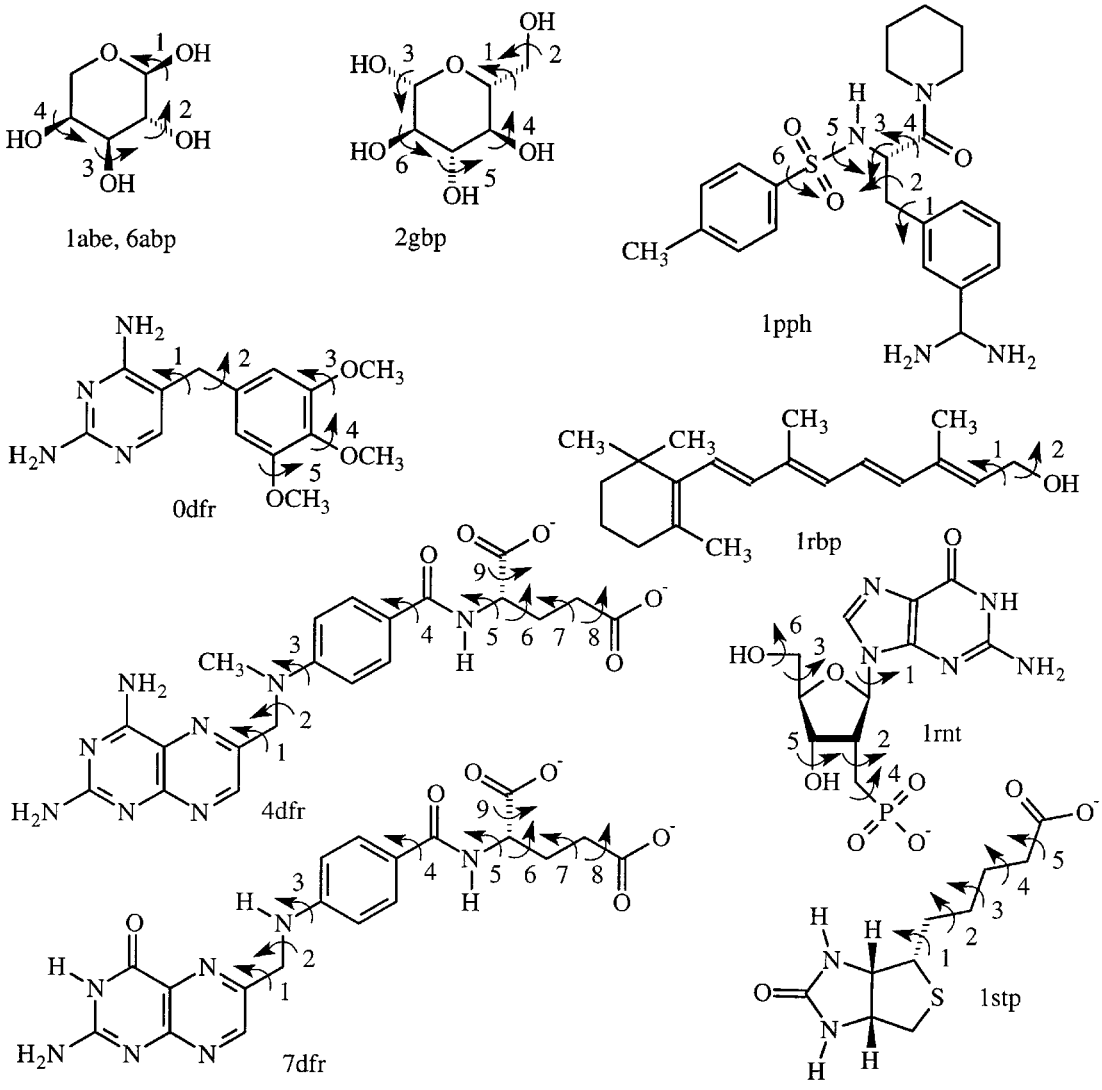

FIGURE 7. Chemical structures of the ligands for docking study.

standard template library (STL) from GNU was especially useful for hiding the complicated data structures from the users and for making the programs simpler. All calculations were performed on Silicon Graphics Indigo2 workstations with 200-MHz R4400 processors and 128 Mb of RAM. The docking anchor and the search conformation programs required approximately 7 Mb and 5 Mb of RAM, respectively, to store the scoring grids. Docking additionally required 40 Mb of RAM for matching. These additional memory requirements were so large because of the large number of the site points for the receptors used in this work (150 ~ 180).

Results and Discussion

TEST SYSTEMS

The accuracy of these methods was tested by attempting to regenerate the crystal complex structures of 10 different systems (Fig. 7, Table V). All the crystal structures were taken from the Brookhaven Protein Data Bank (PDB) except the binary complex DHFR with TMP. This structure, called ''0dfr'' in this report, was obtained from D. Matthews.[32] All crystallographic waters were removed from the x-ray structures except for dihy-

TABLE V.
X-Ray Condition and Affinities.

pdb	Resolution	Protein	Ligand	Source	pK_i[a]
1abe[29,30]	1.7	L-Arabinose-binding protein	L-Arabinose	*E. coli*	6.50
6abp[31]	1.67	L-Arabinose-binding protein M108L	L-Arabinose	*E. coli*	6.36
0dfr[32]	2.3	Dihydrofolate reductase	Trimethoprim	*E. coli*	8.89
4dfr[33]	1.7	Dihydrofolate reductase	Methotrexate	*E. coli*	8.60
7dfr[34]	2.5	Dihydrofolate reductase / $NADP^+$	Folate	*E. coli*	—
2gbp[35]	1.9	D-Galactose / D-glucose-binding protein	D-Glucose	*E. coli*	7.40
1pph[36]	1.9	Trypsin	3-TAPAP	Bovine	5.90
1rbp[37]	2.0	Retinol-binding protein	Retinol	Human serum	6.70
1rnt[38,39]	1.9	Ribonuclease T1	2′-GMP	*Aspergillus oryzae*	5.18
1stp[40]	2.6	Streptavidin	Biotin	*Streptomyces avidinii*	13.4

[a] $pK_i = -\log_{10} K_i$ (molar).

drofolate reductase. In the case of dihydrofolate reductase, two waters in the ligand-binding region, which are known to bind to the protein strongly, were preserved (WAT403, WAT405 in 4dfr and corresponding waters in 0dfr and 7dfr). After the ligands were removed from the proteins, hydrogens were added to the proteins and all the charges of the atoms were calculated by Gasteiger–Marsili method.[41,42] Then, all the hydrogens were optimized with fixed heavy atoms using MAXIMIN2. All these calculations were done using the SYBYL program. The ligand conformations were taken from the x-ray complex structures. Hydrogens were added and optimized in the same manner as the proteins, but no further minimization was done. Then, the conformations of the ligands were randomized by changing all the rotatable bonds and the orientations. The anchor fragments of the ligands were defined by the ''ANCHOR'' program and the anchor fragments were protonated by the ''PROTON'' program. The anchor fragments were docked and the top eight scoring orientations were selected for further conformational search. The search conditions were determined based on the maximum number of the flexible bonds on a branch of a ligand as shown in Table VI. As the bond depth is increased, the search step angle is also increased:

$$\begin{aligned} &\text{search\_step\_angle} \\ &\quad = \text{initial\_angle} + \text{increment\_angle} \\ &\qquad \times (\text{bond\_depth} - 1) \end{aligned} \tag{1}$$

As a result of this conformational search, the top 120 scoring conformations were stored in a binary tree structure for further minimization. After each rigid body minimization, each branch (Fig. 6) was minimized separately in its torsional degrees of freedom. These minimization steps were repeated three times. The top score conformations were then selected (Table VII). Root-mean-square

TABLE VI.
Search Conditions.

Conditions	Initial Angle	Increment Angle	Limitation Level
Max. # of flexible bonds on a branch < 2	40.0°	0.0°	2
Max. # of flexible bonds on a branch ≥ 2	30.0°	5.0°	4

TABLE VII.
The Most Stable Docking Models.

pdb Name	Score Dock[a]	Score X-ray[b]	rmsd All[c]	rmsd Anchor[d]	Dock Time[e]	Search Time[f]	No. of Flex Bonds
1abe	−27.90	−20.94	0.35	0.35	166	21	4
6abp	−25.87	−17.93	0.24	0.24	104	26	4
0dfr	−30.66	−25.44	1.03	0.79	17	94	5
4dfr	−35.56	−34.89	1.19	0.51	64	242	9
7dfr	−42.57	−26.04	1.69	0.62	11	121	9
2gbp	−29.94	−23.32	0.66	0.39	5	134	6
1pph	−16.27	−16.44	1.88	0.21	32	210	6
1rbp	−18.75	−18.65	0.24	0.23	20	45	2
1rnt	−27.84	−23.02	0.94	0.66	37	120	6
1stp	−27.30	−27.62	1.01	0.46	11	62	5

[a]Docking simulations (sum of the intra- and intermolecular force field energy).
[b]X-ray complex structures. Ligand conformations are minimized (same as ''a'').
[c]rms deviations from crystal structure.
[d]rms deviations of the anchor fragment from the crystal structure.
[e]Time (in seconds, UNIX time command) for anchor fragment docking.
[f]Time (in seconds, UNIX time command) for conformational search and minimization.

(rms) deviations are calculated by the program ''RMSD_MULTIPLE.'' This program takes symmetries of ligands into consideration and calculates all possible rms deviations, then it prints out the minimum rms deviation. We next discuss the results for the individual ligand families.

SUGAR—L-ARABINOSE AND D-GALACTOSE/D-GLUCOSE BINDING PROTEINS: 1abe, 6abp, 2gbp

Hydrogen bonds play an important role in binding sugars to their binding proteins. Although the coordinates of the hydrogens are not available from the x-ray experiments, and we do not know the actual hydrogen bond networks, hydrogen bond formation seems to have been optimized very well. The force field helped to form hydrogen bonds without special hydrogen bond scores. In the case of 2gbp, the direction of one branch of glucose is different from the x-ray structure and forms an intramolecular hydrogen bond.

FOLATE ANALOGS—DIHYDROFOLATE REDUCTASE: 0dfr, 4dfr, 7dfr

The anchor fragments were all placed very close to the x-ray structures. The orientation of the α-carboxylates of MTX and DHFR, which bind to DHFR specifically and strongly, are similar to x-ray structures.[33,34] In contrast, the orientation of the β-carboxylates of MTX and DHFR, which bind to DHFR nonspecifically and weakly, are not very similar to x-ray structures. Also, the dihedral angles that fix the orientation of the 3,4,5-trimethoxyphenyl fragment are not very close to the x-ray structure. However, the x-ray conformations and the calculated conformations have a large amount of volume overlap. A similar observation has been made for side-chain orientations during protein packing calculations when the backbone is held rigid (C. Lee, private communication). There are several reports on simulations of ligand docking with DHFR. Some reduce computation time by using partial structures of the ligands[15] or fix the number of rotatable bonds.[13] Although such approximations might be reasonable for each special case, it is difficult to apply these special conditions to all the molecules in a database.

3-TAPAP—TRYPSIN: 1pph

Although the anchor fragment and most of the other parts of the structure were placed in almost the same position as the x-ray complex structure, the tosyl group of the ligand could not be placed correctly in the narrow S3 pocket of trypsin. The

rms deviation without the tosyl group is 0.57. Nevertheless, the score is very similar to the minimized x-ray complex score, so the program might have found the alternative binding mode.

RETINOL—RETINOL BINDING PROTEIN: 1rbp

The program recognized the conjugated double bonds and treated them as nonrotatable bonds. As a result, the anchor fragment is relatively large in this case and the possibility of alternative orientations is reduced. This is the reason why the program regenerated the x-ray structure very precisely.

2′-GMP—RIBONUCLEASE T_1: 1rnt

Both the anchor fragment and the rest of the conformation were predicted correctly.

BIOTIN—STREPTAVIDIN: 1stp

The anchor fragment was oriented very close to the x-ray complex structure. There was some wobble in the methylene chain between the anchor fragment and the carboxylic acid, but because this region of the molecule lies in a hydrophobic pocket, this wobble seems acceptable.

ROBUSTNESS IN SAMPLING BY LIMITED BACKTRACK SEARCH

We employed the limited backtrack search algorithm to reduce computation time. However, this search might fail to find the conformation with the minimum energy. To test if this search method samples enough conformations, backtrack search with and without limitation were compared using ligands of DHFR. Because the approximate answer should be obtained within a reasonable computation time, it is necessary to restrict the conformational space for the search. Here, the results of the usual backtrack search with no limitations and the limited backtrack search are compared. As Table VIII shows, the full search always resulted in better energy scores, as expected. However, we note that the best scoring conformations from the limited search are actually somewhat close to the x-ray structure in two cases. Thus, we judged that this approximation reduced the quality of the results only marginally, while reducing the computation time significantly. On the whole, the program successfully predicted the binding modes of the ligands with many flexible bonds (five, nine, and nine flexible bonds) in a relatively short time.

COMPARISON OF DIHEDRAL ANGLES

Torsion angles of tested results are compared with x-ray crystal structures (see Fig. 8). The *x*-axes are the torsion angle number and the *y*-axes are the angle by degree. The solid lines represent actual torsion angles of *x*-ray and tested results of ligands, the dashed lines represent the torsion angle differences between *x*-ray and tested results. In general, the dihedral angle differences are smaller near the anchor fragment. The torsion angle differences display a ''crankshaft'' motion as reported in the previous study,[13] thus *x*-ray and tested ligands occupy very similar volume. In the case of the 1stp system, the torsion angle differences are rather large and do not show a ''crankshaft'' motion. However, as discussed in the

TABLE VIII.
Comparisons of Backtrack Search.

	No Limitation			Limitation		
pdb	Score[a]	rmsd[b]	Time[c]	Score[a]	rmsd[b]	Time[c]
0dfr	−31.71	0.89	200	−30.66	1.03	94
4dfr	−38.65	1.43	100880	−35.56	1.19	262
7dfr	−42.74	2.45	32063	−42.58	1.69	121

[a] Sum of the intra- and intermolecular force field energy.
[b] rms deviations between the tested results and the crystal structures.
[c] Time (in seconds) for conformational search and minimization.

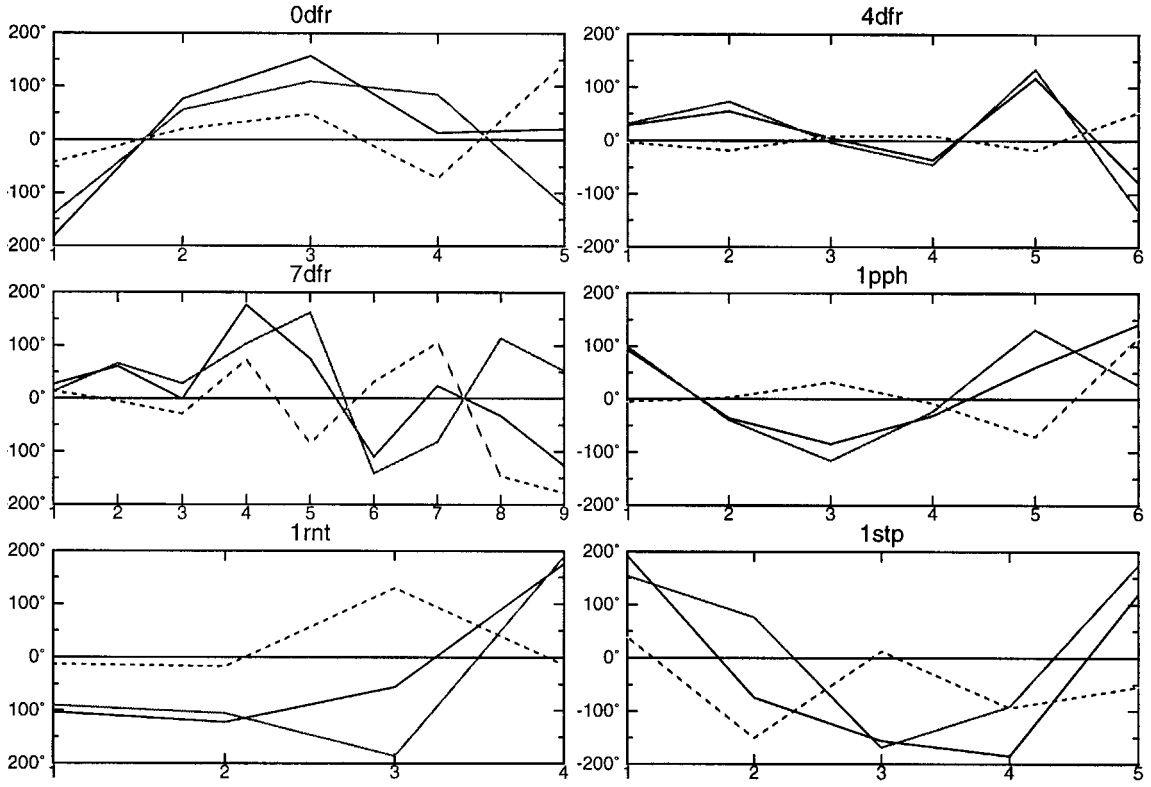

FIGURE 8. Comparison of torsion angles between x-ray and simulation results.

''Test System'' section, these large torsion angle differences are allowable since the basic binding mode is conserved.

APPLICATION FOR DATABASE SEARCH

The accuracy of this approach was demonstrated by the regeneration of these crystal structures. Our next goal was to use the new procedure in a database search. For this purpose, we selected about 20% of the molecules of the ACD database (31,891 molecules) for docking to a DHFR/NADP system. Molecules with total net charge of 1, 2, 3, and 4 were selected. The ''ANCHOR'' program eliminated about half of these candidates owing to the lack of appropriate anchor fragments or if the number of flexible bonds was more than nine (15,068 molecules). This is a reasonable screening filter for potential drug candidates. The remaining molecules were then protonated by the ''PROTON'' program. Finally, they were screened by the flexible docking program. CPU times for the anchor fragment docking and the conformation search were 2210 and 5970 minutes, respectively. The results are shown in Table IX. The program selected most of the folate (DHF) and methotrexate (MTX) derivatives, which are known to bind to dihydrofolate reductase (DHFR), as the top scoring molecules from the database. It was also interesting that the program judged that chiral S was superior to chiral R of MTX, which agrees with the experimental data. In this study, no desolvation energy of the waters or entropical contribution of the waters and ligands were included. We also note the nucleic acid derivatives were selected because the base fragment matched many features of the receptor. The fact that that this method is a valuable tool for database screening.

Conclusion

We successfully developed programs that can be applied to database searches. These programs efficiently regenerated the basic binding modes of the x-ray complex structures in ten test cases. The programs are also fully automated and rapid enough that they can be applied to large database searches. In a database search, the programs selected the well-known inhibitors for the top scores,

TABLE IX.
Results of ACD Database Search.

	Name	Compound	Score[a]
1	MFCD00150465	L-Aminopterin[b]	−73.469
2	MFCD00150769	L-Folic acid[c]	−66.955
3	MFCD00006707	L-N10-(trifluoroacetyl) pteroic acid[c]	−64.965
4	MFCD00167129	2′-Deoxyinosine 5′-monophosphate[f]	−61.371
5	MFCD00079526	3-TAPAP	−58.941
6	MFCD00064370	L-Methotrexate[b]	−57.979
7	MFCD00057319	3′,5′-dichlorofolic acid[c]	−57.970
8	MFCD00057016	Cordycepin 5′-monophosphate[d]	−57.969
9	MFCD00079147	2′-Deoxyadenosine 5′-mono phosphomorpholidate[d]	−55.112
10	MFCD00005753	2′-Deoxyadenosin 5′-monophosphate[d]	−54.838
11	MFCD00065638	Fmoc-L-Cys(Npys)[g]	−53.782
12	MFCD00057318	L-3′,5′-dibromofolic acid[c]	−53.061
13	MFCD00075823	Pteroic acid[c]	−51.454
14	MFCD00056768	1,*N*6-etheno-2′-deoxyadenosine 5′-monophosphate	−51.236
15	MFCD00210875	2′-Deoxyguanosine 5′-monophosphate[e]	−50.531
16	MFCD00079516	Npys-L-Tyr(tBu)[g]	−50.345
17	MFCD00057108	2′-*O*-monosuccinylguanosin 3′:5′-cyclic monophosphate[e]	−50.255
18	MFCD00070111	2′-Deoxyadenosine 3′-monophosphate[d]	−50.013
19	MFCD00098907	Maybridge NRB 00318	−49.854
20	MFCD00079512	Npys-L-His(tosyl)[g]	−49.831
21	MFCD00184466	Salor *S*7,530-3	−49.704
22	MFCD00057601	*N*,*O*-di (2,4-DNP)-L-tyrosine	−49.362
23	MFCD00075773	4-[*N*-(2,4-diamino-6-pteridinylmethyl)amino]benzoic acid[b]	−49.257
24	MFCD00038472	*N*,*S*-di(2,4-DNP)-L-cysteine	−48.895
25	MFCD00057369	*N*,*N*-di(2,4-DNP)-L-lysine	−48.693
26	MFCD00099689	Maybridge KM 06664	−48.502
27	MFCD00058133	2′-Deoxyadenosine 3′-monophosphate[d]	−48.400
28	MFCD00079492	Npys-L-Glu(tBu)[g]	−48.255
29	MFCD00020220	3,3′-hexamethylene diureido bis(2,4,6-triidobenzoic acid)	−48.223
30	MFCD00037966	Trp−Gly−Gly	−48.180
31	MFCD00064185	2-Naphthalene sulfonic acid	−48.022
32	MFCD00097525	Maybridge KM 06606	−47.870
33	MFCD00118936	Maybridge CD 02454	−47.783
34	MFCD00057021	2′-Deoxyadenosine 3′-monophosphate[d]	−47.601
35	MFCD00079489	Npys-L-Asp-NH_2[g]	−47.355
36	MFCD00006710	D-Amethopterin[b]	−47.249

[a]Scores are only intermolecular energy.
[b]MTX derivatives.
[c]DHF derivatives.
[d]Adenosine derivatives.
[e]Guanosine derivatives.
[f]Inosine derivatives.
[g]Protected by 3-nitro-2-pyridinesulfenyl (Npys).

showing the efficiency of the programs for use in virtual database screening.

Availability

Because this program is still under active development, interested readers should contact the authors concerning its availability.

Acknowledgments

We thank Greg Couch of the UCSF Computer Graphics Laboratory for many suggestions about programming in C++. We also thank Dr. Yaxiong Sun, Todd Ewing, Dr. Connie Oshiro, and Eric Pettersen for helpful comments and discussions. Mr. Ewing was especially helpful in making avail-

able the DOCK4 program source code in advance of its public release. Tripos Associates provided the SYBYL program and Molecular Design Ltd. provided the Available Chemicals Directory for which we are grateful.

References

1. A. R. Leach and I. D. Kuntz, *J. Comput. Chem.*, **13**, 730 (1992).
2. Distributed by Molecular Design, Ltd., San Leandro, CA.
3. E. Rutenber, E. B. Fauman, R. J. Keenan, S. Fong, P. S. Furth, P. R. Ortiz de Montellano, E. Meng, I. D. Kuntz, D. L. DeCamp, R. Salto, J. R. Rose, C. S. Craik, and R. M. Stroud, *J. Biol. Chem.*, **268**, 15343 (1993).
4. P. Y. S. Lam, P. K. Jadhav, C. J. Eyermann, C. N. Hodge, Y. Ru, L. T. Bacheler, J. L. Meek, M. J. Otto, M. M. Rayner, Y. N. Wong, C.-H. Chang, P. C. Weber, D. A. Jackson, T. R. Sharpe, and S. Erickson-Viitanen, *Science*, **263**, 380 (1994).
5. K. R. Romines, K. D. Watenpaugh, W. J. Howe, P. K. Tomich, K. D. Lavasz, J. K. Morris, M. N. Janakiraman, J. C. Lynn, M.-M. Horng, K.-T. Chong, R. R. Hinshaw, and L. A. Dolak, *J. Med. Chem.*, **38**, 4463 (1995).
6. T. L. Blundell, *Nature*, **384**, 23 (1996).
7. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. Ferrin, *J. Mol. Biol.*, **161**, 269 (1982).
8. M. Lawrence and P. C. Davis, *Proteins*, **12**, 31 (1992).
9. N. Kasinos, G. A. Lilly, N. Subbarao, and I. Haneel, *Proteins*, **5**, 69 (1992).
10. R. L. DesJarlais, R. P. Sheridan, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan, *J. Med. Chem.*, **29**, 2149 (1986).
11. D. S. Goodsell and A. J. Olson, *Proteins*, **8**, 195 (1990).
12. G. Chang, W. C. Guida, and W. C. Still, *J. Am. Chem. Soc.*, **111**, 4379 (1989).
13. C. M. Oshiro, I. D. Kuntz, and J. S. Dixon, *J. Comput.-Aided Mol. Design*, **9**, 113 (1995).
14. K. P. Clark, *J. Comput. Chem.*, **16**, 1210 (1995).
15. M. Y. Mizutani, N. Tomioka, and A. Itai, *J. Mol. Biol.*, **243**, 310 (1994).
16. W. Welch, J. Ruppert, and A. Jain, *Chem. Biology*, **3**, 449 (1996).
17. M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, *J. Mol. Biol.*, **261**, 470 (1996).
18. *SYBYL, Version 6.0.2*, Tripos Associates, St. Louis, MO, 1993.
19. *SYBYL Theory Manual 1*, Tripos Associates, St. Louis, MO, 1993.
20. S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta Jr., and P. Weiner, *J. Am. Chem. Soc.*, **106**, 765 (1984).
21. S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, *J. Comput. Chem.*, **7**, 230 (1986).
22. I. D. Kuntz, E. C. Meng, and B. K. Shoichet, *Acc. Chem. Res.*, **27**, 117 (1994).
23. E. C. Meng, B. K. Shoichet, and I. D. Kuntz, *J. Comput. Chem.*, **13**, 505 (1992).
24. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, *J. Mol. Biol.*, **161**, 269 (1982).
25. T. Ewing and I. D. Kuntz, *J. Comput. Chem.*, **18**, 1175 (1997).
26. R. M. Karp and Y. Zhang, *J. ACM*, **40**, 756 (1993).
27. M. Gondran, M. Minoux, and S. Vajda, *Graphs and Algorithms*, John Wiley & Sons, New York, 1984.
28. J. A. Nelder and R. Mead, *J. Comput.*, **7**, 308 (1965).
29. F. A. Quiocho and N. K. Vyas, *Nature*, **310**, 381 (1984).
30. M. E. Newcomer, D. M. Miller III, and F. A. Quiocho, *J. Biol. Chem.*, **254**, 7529 (1979).
31. F. A. Quiocho, D. K. Wilson, and N. K. Vyas, *Nature*, **340**, 404 (1989).
32. D. A. Matthews, J. T. Bolin, J. M. Burridge, D. J. Filman, K. W. Volz, B. T. Kaufman, C. R. Beddell, J. N. Champness, D. K. Stammers, and J. Kraut, *J. Biol. Chem.*, **260**, 381 (1985).
33. J. T. Bolin, D. J. Filman, D. A. Matthews, R. C. Hamlin, and J. Kraut, *J. Biol. Chem.*, **257**, 13650 (1982).
34. C. Bystroff, S. J. Oatley, and J. Kraut, *Biochemistry*, **29**, 3263 (1990).
35. N. K. Vyas, M. N. Vyas, and F. A. Quiocho, *Science*, **242**, 1290 (1988).
36. D. Turk, J. Sturzebecher, and W. Bode, *FEBS Lett.*, **287**, 133 (1991).
37. S. W. Cowan, M. E. Newcomer, and T. A. Jones, *Proteins*, **8**, 44 (1990).
38. R. Arni, U. Heinemann, M. Maslowska, R. Tokuoka, and W. Saenger, *Acta Cryst.*, **B43**, 548 (1987).
39. K. Takahashi, *J. Biochem.*, **72**, 1469 (1972).
40. P. C. Weber, J. J. Wendoloski, M. W. Pantoliano, and F. R. Salemme, *J. Am. Chem. Soc.*, **114**, 3197 (1992).
41. *M. Marsili and J. Gasteiger, Chim. Acta*, **52**, 601 (1980).
42. J. Gasteiger and M. Marsili, *Tetrahedron*, **36**, 3210 (1980).