Journal of Computer-Aided Molecular Design, 8 (1994) 153–174
ESCOM

J-CAMD 226

FLOG: A system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure

Michael D. Miller[a], Simon K. Kearsley[b], Dennis J. Underwood[a] and Robert P. Sheridan[b]

[a]*Molecular Systems Department, WP42-3, Merck Research Laboratories, West Point, PA 19486, U.S.A.*
[b]*Molecular Systems Department, RY50S-100, Merck Research Laboratories, Rahway, NJ 07065, U.S.A.*

Received 20 April 1993
Accepted 22 July 1993

Key words: DOCK; Dihydrofolate reductase; Merck Index; 3D databases

SUMMARY

We present a system, FLOG (Flexible Ligands Oriented on Grid), that searches a database of 3D coordinates to find molecules complementary to a macromolecular receptor of known 3D structure. The philosophy of FLOG is similar to that reported for DOCK [Shoichet, B.K. et al., J. Comput. Chem., 13 (1992) 380]. In common with that system, we use a match center representation of the volume of the binding cavity and we use a clique-finding algorithm to generate trial orientations of each candidate ligand in the binding site. Also we use a grid representation of the receptor to assess the fit of each orientation. We have introduced a number of novel features within this paradigm. First, we address ligand flexibility by including up to 25 explicit conformations of each structure in our databases. Nonhydrogen atoms in each database entry are assigned one of seven atom types (anion, cation, donor, acceptor, polar, hydrophobic and other) based on their local bonded chemical environments. Second, we have devised a new grid-based scoring function compatible with this 'heavy atom' representation of the ligands. This includes several potentials (electrostatic, hydrogen bonding, hydrophobic and van der Waals) calculated from the location of the receptor atoms. Third, we have improved the fitting stage of the search. Initial dockings are generated with a more efficient clique-finding algorithm. This new algorithm includes the concept of 'essential points', match centers that must be paired with a ligand atom. Also, we introduce the use of a rapid simplex-based rigid-body optimizer to refine the orientations. We demonstrate, using dihydrofolate reductase as a sample receptor, that the FLOG system can select known inhibitors from a large database of drug-like compounds.

INTRODUCTION

Molecular modeling methods that use the 3D atomic coordinates of receptors are becoming more commonly used now that more structures of pharmaceutically interesting macromolecules are being characterized. These methods can be divided into those which design new molecules to

complement the receptor and those which select existing molecules. For the latter purpose, there has recently been a rapid expansion in techniques [1–6] to search databases of 3D coordinates to find small molecules that fit the receptor.

One of the first publications in this field [1] showed how the DOCK algorithm [7] could be used for that purpose. DOCK addressed three important steps: representing the binding cavity, placing candidate ligands into the cavity, and scoring the quality of fit. Originally, DOCK was used to search the Cambridge Crystallographic Database [8] to find molecules that fit a receptor on the basis of shape only. The intent was to use the molecules with the best shape fit as starting places to design novel ligands for the receptor. Later, DesJarlais et al. [9] reported the results of a DOCK search in which the crystal structure of bromoperidol was seen to fit the active site of HIV protease. Fortuitously (since only steric fit was considered), bromoperidol in its putative binding mode showed reasonable electrostatic complementarity as well. The parent compound haloperidol was tested and shown to be a weak inhibitor of the enzyme, and binding could be increased by a small chemical modification. (The crystallographically observed binding mode of a thioketal derivative of haloperidol to a mutant of HIV protease was later seen to be much like the one suggested by DOCK, although the binding mode to the native enzyme was quite different [10].) The use of DOCK searches as a selection rather than a design method is attractive, since it is usually much easier to test an existing compound than to synthesize and test a new one. The new approach requires (a) a way to score molecules on their full physical complementarity to the receptor (including H-bonding, hydrophobicity, etc.), not just on steric fit; and (b) a 3D database of compounds for which samples are available. Kuntz [11] has reported encouraging results in selecting compounds from the Cambridge Database or the CONCORD-generated [12] 3D version of the Fine Chemicals Directory, using a full intermolecular force field to generate the score. For seven receptors, 2–20% of the best scoring compounds show inhibition of the appropriate receptor in the micromolar range. The details of these searches have been published only for thymidylate synthase [13] and for cercarial elastase and trophozite cysteine protease [14].

A number of groups have published search techniques, most of them based on DOCK, but varying in detail in how the binding cavity is represented and in how ligands are placed and/or scored. A common feature is that they treat the receptor and ligands as being rigid. Techniques are available to address ligand flexibility [15–17]. Generally, though, these methods are too computationally expensive to permit timely searches of large databases. One way to allow for both ligand flexibility and rapid searches is to prepare databases that explicitly store multiple conformations of each structure. We call these databases 'Flexibases'. In this paper we describe a system, FLOG (Flexible Ligands Oriented on Grid), that performs DOCK-like searches over Flexibases. We describe a number of changes to the DOCK algorithms that allow searches over databases of 1 million conformations in a reasonable time. We show that FLOG can select known substrates and inhibitors of dihydrofolate reductase from a database of drug-like compounds.

METHODS

The methods section will cover:

(1) How ligands are represented in the database and how their atoms are assigned to a type.

(2) How the receptor is represented as interaction energy field and how this is used for grid-based scoring.

(3) How match centers are generated to fill the binding cavity.

(4) How searches are done. This has two important aspects: how candidate ligands are docked in the binding cavity and how each docking is evaluated.

(5) Implementation of the FLOG strategy.

The database

Flexibases are constructed by an automatic procedure involving the application of distance geometry, followed by steric and geometric filters to produce conformations which are uniformly 'dissimilar'. The procedure generates an average of 8 and a maximum of 25 conformations per compound, the minimum root-mean-squared distances between any two conformations being at least 1.2 Å. Each conformation is treated as one entry in the database. Details are given elsewhere [18]. The coordinates and connection table of only nonhydrogen atoms are stored. We devised a way to represent the physical properties of individual nonhydrogen atoms by assigning them to one of seven types:

(1) Cations;

(2) Anions;

(3) Neutral hydrogen-bond donors;

(4) Neutral hydrogen-bond acceptors;

(5) Polar. This includes nonhydrogen atoms which are simultaneously donors *and* acceptors, e.g., hydroxyl oxygen, and atoms which can be donors *or* acceptors via tautomerization but cannot be both simultaneously, e.g., nitrogens of an imidazole ring;

(6) Hydrophobic. This includes carbon, sulfur, halogens, etc., in nonpolar environments;

(7) None of the above. Polar atoms which do not have a proton or lone pair (e.g., nitrogens in nitro groups), nonpolar atoms which have a polar environment (carbons in carbonyl groups), or the central atom of sulfones, phosphates, etc., are assigned to type 7.

Except for the addition of type 7, these are similar to the six types used by Jiang and Kim [19] for assigning protein atoms. Atoms are assigned to a type based on their occurrence in specific functional groups which are stored in a user-defined library. Charge states for ionizable groups are appropriate for near-neutral pH. For instance, oxygens in carboxylate groups are assigned as type 2. Details of atom typing are given elsewhere [20].

Representation of the receptor

It is now common to represent the receptor as an interaction energy field instead of a set of explicit atoms [3,4,21,22]. The energy field is sampled by a 'probe' atom at equally spaced points, arranged in a 3D rectangular array (or grid) with sides parallel to the coordinate axes. Since many energy features are short-range, the grid spacing must be ≤ 0.5 Å to provide a proper description [3]. We need to have a separate grid corresponding to each ligand atom type, so we store a 4D array **GRID** where element GRID(ix,iy,iz,k) represents the interaction energy that a probe of type k, located at the grid point with indices ix,iy,iz, feels from the receptor. Grid indices are calculated as follows:

$$ix = (x - x0)/dg + 1$$

where x is the coordinate of the grid point, x0 is the x coordinate of the origin of the grid and dg is the grid spacing. Once **GRID** is precalculated as described below, scoring a ligand on the

receptor becomes simply a series of table lookups and additions. The total score of a ligand with Natom atoms would be:

$$\text{Score} = \sum_{j=1}^{\text{Natom}} \text{GRID}(jx, jy, jz, kj)$$

where jx,jy,jz are the indices of the nearest grid point to the location of atom j and kj is the type of atom j. An atom that falls outside the grid has a contribution of zero.

Our receptors include crystallographically observed nonhydrogen atoms plus polar hydrogens. Typing for receptor atoms is different from that for ligand atoms. Receptor atoms are assigned as donors, acceptors or neither. There are also atoms, designated as 'polar', that can act as donors or acceptors. Types are assigned for atoms in specific amino acid residues (see Appendix). To prepare **GRID**, the total sum of six potentials (van der Waals, electrostatic, hydrogen bonding from donors, hydrogen bonding from acceptors, hydrogen bonding from polar atoms, and hydrophobic) is calculated at each grid point ix,iy,iz for each probe of type k:

$$\begin{aligned}\text{GRID}(ix,iy,iz,k) = {} & C(k,\text{vdW}) \times V_{\text{vdW}} + C(k,\text{ESPOT}) \times V_{\text{ESPOT}} + C(k,\text{H-BondD}) \times V_{\text{H-bondD}} + \\ & C(k,\text{H-bondA}) \times V_{\text{H-bondA}} + C(k,\text{H-bondP}) \times V_{\text{H-bondP}} + \\ & C(k,\text{Hydrophobic}) \times V_{\text{Hydrophobic}}\end{aligned}$$

Each potential is, in turn, a sum over receptor atoms. Whether various probes feel individual potentials as attractive or repulsive is controlled by the matrix **C**, the values of which are shown in Table 1. For instance, a type-4 probe (acceptor) will be attracted to a region near a receptor N-H (a donor), since C(4,H-BondD) = 1. On the other hand, it would be repelled from a receptor C=O (an acceptor) or a receptor aliphatic group, since C(4,H-bondA) and C(4,Hydrophobic) = −1. A type-7 probe feels only the vdW potential, since C(7,vdW) = 1, but all other values in that row are zero.

The user has a choice on whether cation and anion probes will feel the electrostatic potential or the hydrogen-bond potentials (thus the two rows each for cations and anions in Table 1). In the latter case, they will be treated as equivalent to neutral donors and acceptors. Our default is to treat them as equivalent, since the electrostatic potential is sometimes too dominated by the overall charge on the receptor. Moreover, a recent paper [23] showed that individual atomic interactions

TABLE 1
VALUES OF MATRIX C, WHICH CONTROLS HOW PROBES OF VARIOUS TYPES INTERACT WITH SIX POTENTIALS

		Potentials					
k	Type	vdW	ESPOT	H-bondD	H-bondA	H-bondP	Hydrophobic
1	Cation	1	−1	0	0	0	0
		1	0	−1	1	1	−1
2	Anion	1	1	0	0	0	0
		1	0	1	−1	1	−1
3	Donor	1	0	−1	1	1	−1
4	Acceptor	1	0	1	−1	1	−1
5	Polar	1	0	1	1	1	−1
6	Hydrophobic	1	0	−1	−1	−1	1
7	None of the above	1	0	0	0	0	0

in proteins are well predicted by hydrogen bonding but poorly by the electrostatic potential.

In formulating the individual potentials, we keep the traditional DOCK convention [1] that more positive scores mean a more attractive interaction. (This is opposite to the usual convention for many of the functional forms for nonbonded interactions.) The magnitudes of the potentials are set by means of a parameter file, which is listed in the Appendix. It should be emphasized that, although the individual potentials have the form and units of energy, the total score cannot be interpreted as a true binding energy, only as a qualitative measure of complementarity.

The van der Waals potential is a standard 6–9 London–Jones function:

$$V_{vdW} = -\sum_{i}^{\text{all atoms}} \text{depth(probe,i)} \left[\frac{2\times\text{dist0(probe,i)}^9}{\text{dist(gp,i)}^9} - \frac{3\times\text{dist0(probe,i)}^6}{\text{dist(gp,i)}^6} \right]$$

where dist(gp,i) is the distance between the grid point and the receptor atom i. The distance of maximum attraction dist0(probe,i) is the sum of the radii of the probe and the atom i; the well depth depth(probe,i) is the geometric mean of the depth of the probe and the depth of atom i.

The electrostatic potential is:

$$V_{ESPOT} = -\sum_{i}^{\text{all atoms}} q(i)/(e(i)\,\text{dist(gp,i)})$$

where q(i) is the partial charge on receptor atom i and e(i), the local dielectric as defined by Boobbyer et al. [22], is a function of the accessibility of the grid point and atom i.

The hydrogen bonding potential, for example from donors, is:

$$V_{H\text{-}bondD} = \sum_{i}^{\text{donors}} \text{radial term} \times \text{angle term} \times \text{dihedral term}$$

The radial, angular and dihedral components are given in Fig. 1. The radial component, new to this work, is unusual in that it is Gaussian in form and attractive only, as opposed to the standard London–Jones-like form [22] which has attractive and repulsive terms. An H-bond potential with a repulsive term can give rise to an artifactual short-range attraction when it is subtracted. Our radial H-bond term, when added to the vdW term, gives a London–Jones-like potential surface for H-bonds with energies of about 3 kcal/mol and maxima at the correct probe–receptor atom distance (roughly 2 Å for donors, i.e., polar hydrogens, and 3.0 Å for acceptors and polar atoms). Metals in the receptor are given pseudo-H-bond potentials (radial term only) with depths of about 5 kcal/mol and maxima at about 2 Å.

The form of the hydrophobic potential is derived from the idea [24] that hydrophobic regions are attractive van der Waals regions where there is little H-bonding:

$$V_{Hydrophobic} = \begin{cases} 0.0 & \text{when } V_{vdW} < 0 \\ 0.0 & \text{when } V_{H\text{-}bondD} + V_{H\text{-}bondA} + V_{H\text{-}bondP} + V_{vdW} > F_{min} \\ F \times V_{vdW} & \text{otherwise} \end{cases}$$

The hydrophobic potential is zero wherever the van der Waals term is repulsive or where the total H-bond attraction is greater than a threshold F_{min}. Otherwise, it is set to a multiple F of the van der Waals potential. F_{min} and F are adjustable by the user. By default, we set $F_{min} = 1.5$, i.e., half

the standard H-bond well depth. By experimenting with how well the total scoring function predicts the location of hydrophilic and hydrophobic atoms in a number of crystallographic receptor–ligand complexes, we settled on F = 1.0 as a reasonable default value. This means that the 'bond' between hydrophobic atoms has a total well depth of 0.3 kcal, twice the depth of the van der Waals potential alone and about 1/10 the depth of an H-bond. This is consistent with semiempirical solvent-accessible surface area terms for hydrophilic and hydrophobic protein atoms [25]. The energy value for burying a square Angstrom of the surface of a hydrophobic atom is 1/10 to 1/30 the magnitude for a hydrogen-bonding atom, depending on what experimental data is used.

Generation of match centers

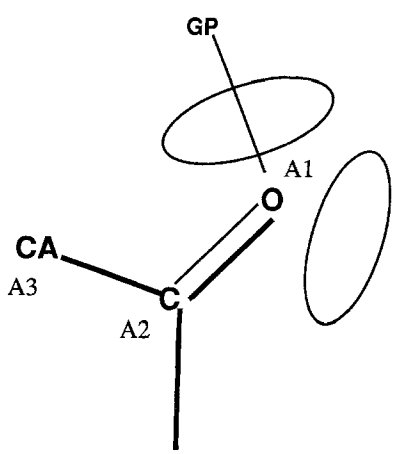
Many search methods place a ligand into a binding cavity by pairing ligand atoms to 'match centers' that fill the volume of the cavity. The centers can be generated in a number of ways. The original method [7] involves growing of spheres along the normals of the Connolly surface of the receptor, so that the spheres would fill concavities in the surface; these 'receptor spheres' then serve as match centers. Alternatively, a web of evenly spaced match centers may be generated on the solvent-accessible surface [6]. One can place match centers on the atomic positions of co-crystallized ligands if the coordinates are available. It is also common to place match centers at favorable sites of interaction in the grid [4]. We prefer this last method. For all atom types k, we

$V_{H\text{-}BOND}$ = RADIAL TERM * ANGLE TERM * DIHEDRAL TERM

RADIAL TERM = D * EXP(–(dist0 – distance(GP,A1)) ** 2

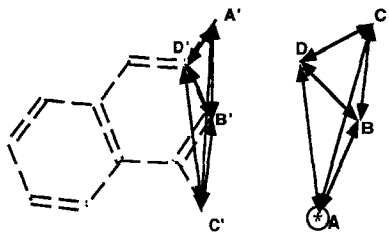
ANGLE TERM = COS(ABS(ang0 – angle(GP,A1,A2))) ** n1 or 1.0 if no A2 defined

DIHEDRAL TERM = 1/2 [COS(n2 * dihedral(GP,A1,A2,A3) – dihed0)] or 1.0 if no A3 defined



A1	type	A2	ang0	n1	A2	dihed0	n2	dist0	D
O	–1	C	145.0	2	CA	0.0	2	2.5	3.5

Fig. 1. The radial, angle, and dihedral components of H-bond potentials. A backbone carbonyl oxygen from the receptor, a hydrogen-bond acceptor (type = –1), is shown as an example. The energy of the interaction depends on the location of the grid point (GP) and the atoms adjacent to the acceptor. In this case maximum hydrogen bonding occurs near the lone-pair directions in the plane of the carbonyl group.



Match center distances		Ligand atom distances	
AB	2.73	A'B'	2.36
AC	5.10	A'C'	4.69
AD	3.98	A'D'	1.33
BC	2.49	B'C'	2.32
BD	1.97	B'D'	1.39
CD	2.12	C'D'	3.63

subGraph 1

Assigned Pairs	A A'	A B'	A C'	A D'	B A'	B B'	B C'	B D'	C A'	C B'	C C'	C D'	D A'	D B'	D C'	D D'
1) A/A'	--	--	--	--	--	(0.37)	1.96	1.40	--	2.74	0.41	3.77	--	1.62	0.71	2.65
2) B/B'	--	--	--	--	--	--	--	--	--	--	(0.17)	1.1	--	--	0.35	0.58
3) C/C'	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	1.51
4) none																

subGraph 3

Assigned Pairs	A A'	A B'	A C'	A D'	B A'	B B'	B C'	B D'	C A'	C B'	C C'	C D'	D A'	D B'	D C'	D D'
1) A/C'	--	--	--	--	1.96	0.41	--	0.90	0.41	2.78	--	1.47	0.71	1.66	--	(0.35)
2) D/D'	--	--	--	--	0.64	(0.58)	--	--	0.79	0.73	--	--	--	--	--	--
3) B/B'	--	--	--	--	--	--	--	--	(0.13)	--	--	--	--	--	--	--
4) C/A'	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
5) mask full																

Fig. 2. An example of the operation of our 'minimum residual' algorithm. Four match centers (A–D) and four ligand atoms (A'–D') are considered. In the text possible center-atom pairings are treated as a 2D array. For clarity in this figure, they are written out in a line. In this example searches are being performed with *nodlim* = 4 and *dislim* = 1.5. Match center A is an essential point, so all cliques start with an initial pairing of A with one of the atoms A'–D'. Subgraph 1 shows the situation where center A is initially paired with atom A' (denoted A/A'). In Step 1, all other possible pairings involving A and A' are masked out. Of the remaining possibilities, B/B' has the smallest residual ($|\text{Dist}(AB) - \text{Dist}(A'B')| = 0.37$). Since 0.37 < *dislim*, the clique building can continue. In Step 2, all other possible pairings involving B and B' are masked out. Of the remaining possibilities, C/C' has the smallest residual ($|\text{Dist}(BC) - \text{Dist}(B'C')| = 0.17$). Since 0.17 < *dislim* and all the distances in the A-B-C triangle are within *dislim* to the corresponding distances in the A'-B'-C' triangle, the clique building can continue. In Step 3, all the other possible pairings involving C and C' are masked out. The only possibility left is D/D'. Since the residual 1.5 > *dislim*, no further pairing is possible and the clique building stops. There are only three pairs (A/A', B/B', C/C'). Since 3 < *nodlim*, this clique is not recorded. Subgraph 3 shows the situation where center A is initially paired with atom C'. In Step 1, all other possible pairings involving A and C' are masked out. Of the remaining possibilities, D/D' has the smallest residual. Since 0.35 < *dislim*, the clique building can continue. In Step 2, all other possible pairings of D and D' are masked out. Of the remaining possibilities, B/B' has the smallest residual. Since 0.58 < *dislim* and the distances in the A-D-B triangle are the same as the corresponding distances in the C'-D'-B' triangle within *dislim*, the clique building can continue. In Step 3, all the other possible pairings involving B and B' are masked out. Of the remaining possibilities, C/A' has the smallest residual. Since 0.13 < *dislim* and the distances in the A-D-B-C tetrahedron are the same as the corresponding distances in the C'-D'-B'-A' tetrahedron within *dislim*, the clique building can continue. In Step 4, all the other possibilities involving C and A' are masked out. There are no possibilities left, and the clique building stops. Since there are four pairs (A/C', D/D', B/B', C/A') and 4 ≥ *nodlim*, this clique is recorded.

look in **GRID**(ix,iy,iz,k) for grid points which are local energy maxima. We check all pairs of maxima, and if two maxima are closer than a distance cutoff (typically 1.5 Å), we eliminate the one with the lower energy. Remaining maxima with energies greater than some threshold (typically 2.0 kcal) are saved as match centers. We find that match centers fall closer to the atoms of cocrystallized ligands when relatively fine grids (≤ 0.3 Å spacing) are used.

Search strategy

The strategy to test each candidate conformation for complementarity to the receptor is as follows:

(1) Read in the grid and match centers. Calculate the center–center distances.

(2) For each conformation in the database:

- 2.1 Read the coordinates of the atoms. Calculate the atom–atom distances.
- 2.2 Find all sets of atom-center pairs that are allowed as matches (see below).
- 2.3 For each match:
  - 2.3.1 Produce an initial orientation for the candidate in the active site.
  - 2.3.2 Optimize the orientation corresponding to the match (see below) until the score is maximal.

  End 2.3
- 2.4 Write the score of the best orientation of this conformation.
- 2.5 Repeat 2.3–2.4 for the mirror-reflection of the conformation.

End 2.

As in the DOCK algorithm, initial dockings of each conformation in the binding site (step 2.2) are generated by a method of systematic distance matching, where a clique-finding algorithm looks for sets of match center–ligand atom pairs such that all the center–center distances are the same as the corresponding atom–atom distances within a certain tolerance. The minimum number of pairs *nodlim* and the tolerance *dislim* are under user control. Exhaustive clique searches are not feasible for all but the smallest cases, because the number of possible permutations of pairs grows exponentially with the number of atoms and centers. Additional constraints need to be imposed if the searches are to be completed in a reasonable time.

We designed a new clique-finding algorithm to be highly vectorizable, so as to take advantage of our computer resources. Our heuristic is that at each step in the search, the algorithm selects a new center-atom pair such that the absolute distance difference relative to the previously selected pair is a minimum. In other words, the algorithm takes the path of 'minimum residuals', where the residual is |dist(previous center, new center) – dist(previous atom, new atom)|. A 'data mask' in the form of a two-dimensional array **M** is useful in keeping track, in a vectorizable fashion, of which centers (rows) or atoms (columns) are available for matching (1) or had been previously selected or rejected (0). In the following discussion the notation **M**(i,∗) means all the elements in row i and **M**(∗,j) means all the elements in column j. The algorithm is:

2.2.1 For match center i=1 to Nmatch:

2.2.2 For ligand atom j=1 to Natom:

- 2.2.2.1 Initialize $\mathbf{M}(*,*)=1$. Set $\mathbf{M}(i,*)=0$ and $\mathbf{M}(*,j)=0$
  npair = 0
  i' = i, j' = j
- 2.2.2.2 For all centers k and atoms m where $\mathbf{M}(k,m)=1$ calculate the residuals
  $R(k,m)=w(\{i',k\}\{j',m\})\times|\mathrm{Dist}(i',k)-\mathrm{Dist}(j',m)|$

2.2.2.3 k'=k and m'=m where R(k,m) is a minimum and **M**(k,m)=1

2.2.2.4 If R(k',m') ≤ *dislim* and all the center–center distances match the corresponding atom–atom distances within dislim then

npair=npair+1

add the k'/m' pair to the clique

i'=k', j'=m'

endif

Set **M**(k',∗)=0 and **M**(∗,m')=0

2.2.2.5 If all elements of **M**=0, go to 2.2.2.6. Otherwise go to 2.2.2.2.

2.2.2.6 The clique is finished. If npair≥*nodlim*, record the clique.

End 2.2.2

End 2.2.1

This algorithm is comprehensive, in that all center-atom combinations are tried as an initial pair, and deterministic, in that the order of the centers and atoms is irrelevant in the absence of rare ties in the distance differences. We tried a number of weighting schemes for computing R(k,m), including a distance-dependent weighting that biased the search toward longer distances. However, we found none better than a simple difference, i.e., w({i',k},{j',m})=1. The user may assign certain match centers to be 'essential points'. Given Ness essential points, the user may specify that at least Nreq essential points must appear in every clique, where 1 ≤ Nreq ≤ Ness. This substantially constrains the clique search. Using essential points is equivalent to insisting that

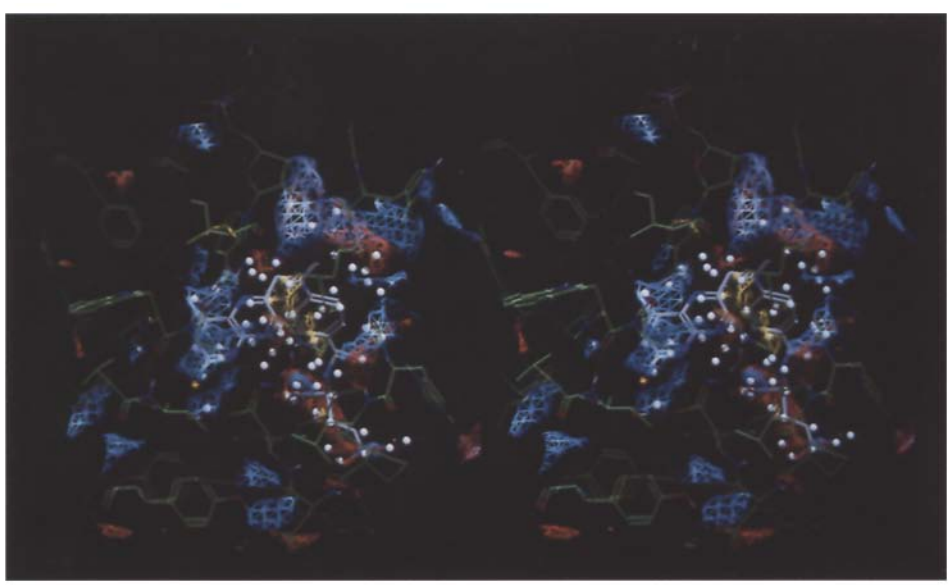


Fig. 3. A contour map of the grid at the active site of dihydrofolate reductase. Contoured regions represent the interaction energy for a ligand atom of a particular type. The interaction energy shown here is the sum over five types of potentials (van der Waals, hydrogen bonds from donors, hydrogen bonds from acceptors, hydrogen bonds from polar atoms, and hydrophobic). Blue: the interaction score for a hydrogen-bond donor or cation (contoured at 2.5); red: the interaction score for a hydrogen-bond acceptor or anion (at 2.5); yellow: the interaction score for a hydrophobic atom (at 2.5). Match centers are shown as spheres, the orange sphere being the essential point. The crystallographically observed binding mode of methotrexate is shown in atom colors.

certain regions of the active site cavity be filled with a ligand atom. Step 2.2.1 loops over match centers i to be used in the initial pairings. Constraining i to be one of Ness essential points guarantees that each clique contains at least one essential point and results in a substantial speedup if Ness $\ll$ Nmatch. If Nreq > 1, cliques with fewer than Nreq essential points are discarded. An example of the operation of the algorithm is presented in Fig. 2.

One generates an 'orientation' from each match (Step 2.3.1) by rotating/translating the candidate conformation such that the subset of atoms listed in the match are best fit onto their paired match centers. The orientation must then be scored (including all its atoms) and optimized on the grid (Step 2.3.2). To reduce the dependence on the initial orientation, we allow the candidate to move as a rigid body in the active site to increase its score. This is done by performing a simplex optimization [26] on seven degrees of freedom (the four elements of the quaternion [27] that describes the rotation and three components of translation), allowing up to 12 degrees of rotation and 1.0 Å of translation per step. Optimization is considered complete when the rotations differ by no more than 1.2 degrees from the previous step and the translations differ by no more than 0.1 Å or until a total of 500 steps have been taken.

In computing the score at each stage of the optimization, we use a process called 'shiver' to partly compensate for the rigidity of the receptor, the discreteness of the grid representation and the discreteness of the conformations. The energy contribution of each atom is taken from the larger of two values: the energy at the nearest grid point or the average of the 27 grid points surrounding the nearest grid point. The effect is that each atom is allowed to independently move around (i.e., shiver) to find the best local environment.

A note on chirality: within a search it is usually desirable to try both the original coordinates and their mirror-reflection (Step 2.5) of each conformation. The absolute chirality of compounds in the database is often arbitrary, having been built from connection tables that in many instances lack stereochemical information. When compounds are achiral, the mirror-reflection may represent an alternative conformation.

FLOG outputs a summary record for the highest scoring orientation of each conformation it has examined. The record includes the structure identifier, the conformation number, the (optimized) score, the final quaternion and translation vector and a mirror-reflection flag. Original coordinates and mirror-reflection are listed in separate records.

Implementation

We have implemented FLOG on several platforms, including a Cray YMP with four processors under the UNICOS operating system and a cluster of four RS/600 Model 580 workstations operating under AIX. To take advantage of the multiple processors, we divide databases into pieces or 'stripes', containing a few tens of thousands of conformations each. Up to four stripes can be searched concurrently. After all the stripes have been examined, the summary records are combined and postprocessed.

RESULTS

As an example we will show the results of a FLOG run on dihydrofolate reductase (DHFR) from *L. casei*. The complex with the cofactor NADP and the inhibitor methotrexate (MTX) has been crystallized at 1.7-Å resolution [28] and is available from the Brookhaven Protein Data

Bank [29] as dataset 3DFR. The pteridine ring of the MTX is seen to bind in a deep cylindrical cavity with the benzoic acid and glutamate portions at a surface groove. The NADP and two active-site waters (HOH^{201} and HOH^{253}) were retained as part of the receptor. MTX and other water molecules were removed. Polar hydrogens, including those of the waters, were placed so that they could best form intrareceptor hydrogen bonds. We calculated a 0.3-Å grid around the active site, the grid extending 5 Å beyond MTX in each direction. This was the best overall compromise between sampling the receptor field finely enough, enclosing most of the active site, and yet not having to store so many grid points that program memory requirements would become a problem. Figure 3 shows the grid contoured for selected atom types. We generated 60 match centers from local maxima in the grid around the active site. We chose as the single essential point the center that falls nearest to the 2-amino group of MTX. This constrains the search to binding modes where the ligand penetrates to the bottom of the active site, as is seen with crystallographically observed inhibitors.

Searches were done with *nodlim* = 5 and *dislim* = 1.5. We used the Flexibase MINDEX, which was generated from connection tables in the 11th edition of the Merck Index [30]. MINDEX contains a total of 57 531 conformations from 7636 structures, divided into four stripes. The run took ~12 h of CPU time on our Cray YMP for all stripes, which were searched concurrently.

The best score per compound was taken as the best score of all orientations of both enantiomers of all conformers of that compound. Figure 4 shows the best score as a function of the number of atoms in the compound (Natom). The median score for compounds appears to be linear with log10 (Natom) for Natom > 10. Individual molecules may fall below this line if they cannot sterically fit into the active site; this is especially common as Natom increases. Visual inspection of the compounds docked in the active site shows that the compounds with the highest scores are not necessarily the most interesting ones. As expected from the graph, the list of high-scoring compounds is dominated by the largest ones. Most of these do not seem to fit the receptor very well; many protrude from the active site. On the other hand, small compounds with relatively low scores

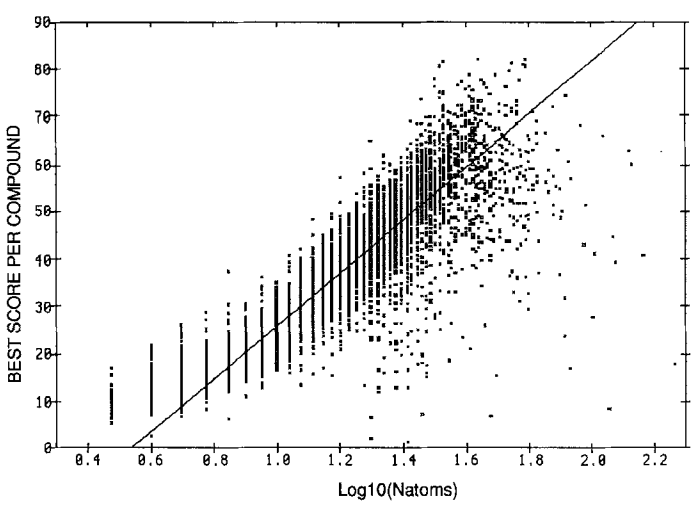


Fig. 4. The best score per MINDEX structure as a function of the number of atoms. The line Score = 55.9 × log10 (Natom) − 30.0 was derived from a robust linear fit over structures with Natom > 10 nonhydrogen atoms.

TABLE 2
MINDEX COMPOUNDS WITH THE HIGHEST NORMALIZED SCORES

MINDEX compound	Normalized score	Score	MINDEX ident.	Name	No. nonhydrogen atoms
1	26	82	4141	folinic acid[a]	34
2	26	81	5908	methotrexate[a]	33
3	24	78	485	aminopterin[a]	32
4	22	65	4187	fructose-1,6-diphosphate	20
5	22	76	6135	mitoxantrone	32
6	22	73	5466	lucifer yellow	29
7	20	75	6471	ninopterin[a]	33
8	20	82	8356	scilliroside	44
9	20	57	4188	fructose-6-phosphate	16
10	19	75	142	A-denopterin[a]	34
11	19	69	8883	sulfaloxic acid	27
12	19	75	1360	brassinolide	34
13	19	74	5907	methopterin[a]	33
14	19	79	9201	theaflavin	33
15	18	77	7889	proscillaridin	38
16	18	77	6276	nalorphine	39
17	18	69	7542	polydatin	28
18	17	70	1641	calcitriol	30
19	17	62	7711	pratensein	22
20	17	72	9817	validamycins	34
21	17	58	6261	nadoxolol	19
22	17	68	1734	comostat	29
23	17	72	9960	withaferin A	34
24	17	70	8208	ribostamycin	31
25	16	72	2879	denopterin[a]	34
26	16	59	6215	mureride	20
27	16	76	238	alizarin cyanine green F	40
28	16	76	6345	naringin	41
29	16	48	6131	mitoquazone	13
30	16	70	5851	methocycline	32
31	16	71	4522	hamamelitannin	34
32	16	61	6610	nordihydroguaiaretic acid	22
33	15	77	762	androst-16-en-3-ol	45
34	15	71	1916	ceanothic acid	35
35	15	68	28888	febantel	31
36	15	67	5175	ketanserin	29
37	15	52	386	ambazore	16
38	15	66	6366	ergosterol	28
39	15	65	790	D-araboflavin	27
40	15	67	4628	hexaprenaline	30
41	15	70	7404	pimozide	34
42	15	64	796	arbaprostil	26
43	15	75	7052	penimepicycline	42
44	14	70	5336	leucoglycodrin	34
45	14	69	691	A-ninopterin[a]	33
46	14	66	4239	galactoflavin	29
47	14	59	5694	melarsoprol	22
48	14	69	8045	quercimeritrin	33
49	14	60	5699	melibiose	23
50	14	77	2035	chartreusin	46

[a] Folate analog.

can fit to the receptor quite snugly. Therefore, we need a way to normalize the scores for size, so that we can find compounds which fit well on a per atom basis. We used a robust linear fit to generate the best line through these data. Since the scatter around the line appears uniform with log (Natom), we can define a normalized score simply as the displacement above the line. Table 2 lists the 50 compounds (Natom > 10) with the highest normalized scores. Of these, a number are analogs of folate, including the familiar inhibitors MTX and aminopterin. Melarsoprol (5694, 47th in the list) is a diaminotriazole inhibitor of Trypanosome DHFR. It is an example of a smaller inhibitor that might have been missed if the sorting had been done by raw score.

Given the discrete sampling of conformational space and the simplicity of the scoring function, we do not expect the highest scoring conformation of a compound to exactly mimic the crystallographically observed binding mode. However, in some cases we come close. In the highest scoring conformation of MTX, the α-carbon is inverted, and the benzoate portion is flipped relative to the crystallographic mode, but the major sites of interaction are maintained (see Fig. 5). Three other conformations of MTX that score in the range 75–80 show similar small variations. (The crystallographic conformation itself has a score of 93 after rigid-body optimization.) Folinic acid, an intermediate substrate with a 4-carbonyl group, appears with the pterin ring in the 'inverted' mode expected for folate on DHFR. The purpose of FLOG is to identify new classes of compounds which present similar groups to DHFR. One example is sulfaloxic acid, which is shown in Fig. 5. The carbonyldiamino nitrogens donate two hydrogen bonds, which replace the two

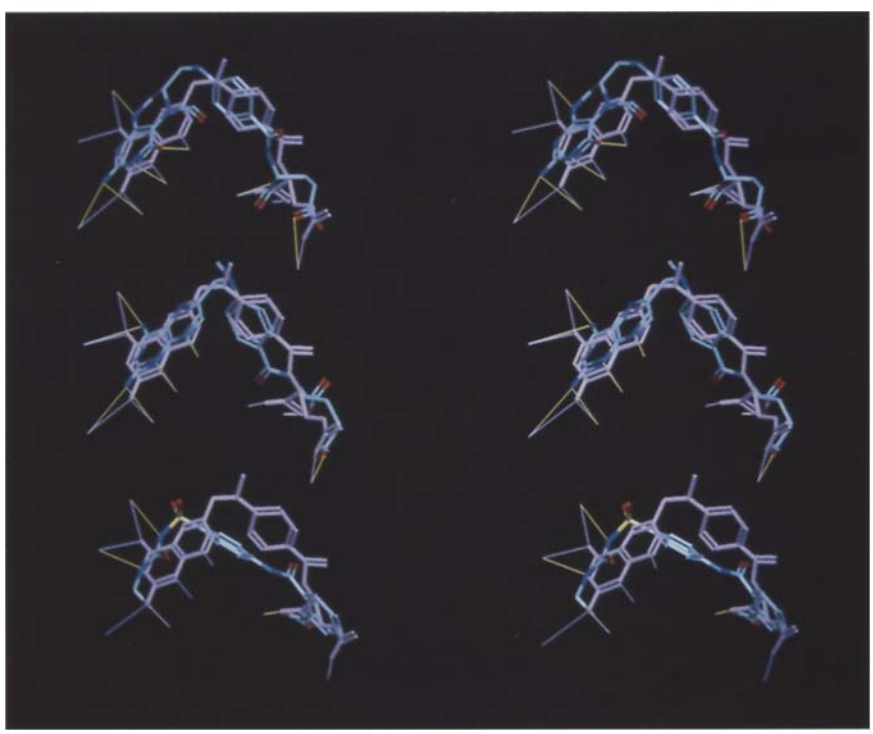


Fig. 5. The dockings of highest-scoring conformations of selected MINDEX structures from Table 1. The crystallographically observed binding mode of methotrexate is shown in purple. Top: folinic acid (4141); center: methotrexate (5908); Bottom: sulfaloxic acid (8883). Putative hydrogen bonds to the receptor are shown as thin lines.

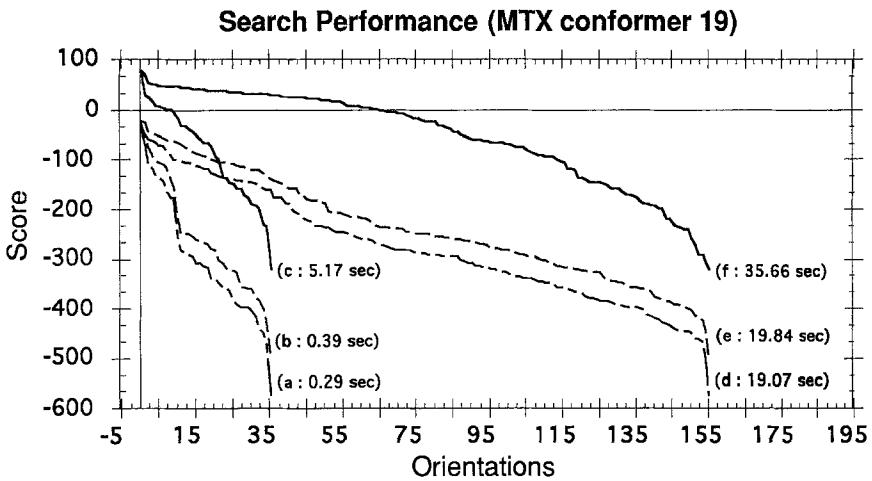


Fig. 6. Graph showing the search performance using essential points and simplex optimization on the MINDEX entry for the conformation (No. 19) of methotrexate (5908) that is shown in Fig. 5. The score of each orientation is plotted vs. its rank order. Orientations from both the original coordinates and their mirror-reflection are included. Computation times are indicated. Most orientations have at least one atom of MTX in collision with the receptor. Thus, the score of most orientations < 0. (a) using a single essential point; (b) using a single essential point, plus the shiver procedure; (c) using a single essential point, plus the shiver procedure and simplex optimization; (d) using all match centers; (e) using all match centers, plus the shiver procedure; (f) using all match centers, plus the shiver procedure and simplex optimization.

hydrogen bonds donated by the 4-amino group of MTX. A hydroxyl group replaces the 2-amino group as a donor, and an *ortho*-carboxylate on a phenyl ring replaces the α-carboxylate as a hydrogen-bond acceptor.

Using the highest scoring conformation of MTX (conformer 19 out of 25) as an example, we can demonstrate the utility of essential points, the shiver procedure and rigid-body optimization. The results are presented in Fig. 6. The score of each orientation, both from the original coordinates and mirror-reflection, is plotted as a function of rank order of the orientation. Using only the single essential point to form the initial pairs, the clique search produced 35 orientations (curves a, b and c). Using all the match centers, it produced 155 orientations (curves d, e and f). The highest scores for curves c and f are about the same, as are b and e, and a and d. This illustrates that using the essential point saves computation time without losing the highest scoring orientations. Comparing b with a and e with d shows the effect of the shiver procedure, which is mainly to rescue atoms which are very close to the repulsive receptor walls. Comparing b with c and e with f shows the effect of optimization, which increases the scores substantially with a modest increase in computation time. In this case, optimization was necessary for this interesting conformation to have a score > 0, i.e., to be recognized as complementary to the enzyme.

DISCUSSION AND CONCLUSIONS

The goal of FLOG, like many systems using the DOCK paradigm, is to select from a large database a small set of compounds suitable for assaying for the appropriate biological activity. These searches can never find *all* the potentially interesting compounds in the database, for a

number of reasons. First, the receptor is assumed to be rigid and the conformational space of each ligand is not fully explored. Second, the ligand orientations depend on a matching between atoms and match centers, and matching algorithms cannot be exhaustive. Finally, the true binding energy cannot be calculated by simple scoring functions which, apart from other deficiencies, commonly ignore solvation. However, one must inevitably make approximations and sacrifice completeness, so that large databases can be searched in a reasonable time. Fortunately, for drug discovery purposes it does not matter that some compounds are missed, as long as the compounds selected are significantly enriched in novel actives relative to a randomly chosen set.

There are several new features in FLOG relative to DOCK. The first is the use of multiple conformations for each compound in the database. It is obvious that multiple conformations cover conformational space better than a single conformation (as is generated, for instance, by CONCORD [12]). By no means, however, is conformational space fully covered for many compounds. One must inevitably trade coverage of the conformational space for search speed.

One scoring function reported in the literature for DOCK [3] involves the nonbonded potentials from the AMBER force field [31], which includes the electrostatic and van der Waals potentials. This can make fairly fine distinctions between chemical groups. To be compatible with this type of scoring function, the candidate ligands must have explicit hydrogens and a charge must be assigned to every atom. In contrast, our representation of ligands as a set of nonhydrogen atoms, each assigned to one of seven atom types, is very simplified. One advantage of such a simple representation is that the atoms can be more easily defined opportunistically. We need not worry about the direction in which hydrogens on terminal rotamers (e.g., OH groups) are pointed. Also, we need not specify tautomers or charge states in advance when there is more than one possibility. For example, an imidazole in a ligand might be neutral with either nitrogen being a donor and the other an acceptor, or a cation with both nitrogens being donors, depending on the environment of the receptor. Assigning the nitrogens as type 5 allows for all possibilities. (Similarly, certain atoms in the receptor can be defined as polar, allowing for the same opportunities.) The drawbacks are, of course, that hydrogen bond angular information around the ligand atoms is lost and the hydrogen bond count, and therefore the interaction score, may be overestimated. Also, one might find chemically nonsensical binding modes, for instance with both nitrogens of the imidazole accepting hydrogen bonds from the receptor. However, these drawbacks are not critical. The binding modes of the selected compounds can be inspected visually and can be further screened by more detailed energy calculations.

A commonly observed feature of DOCK-like searches is that the highest scoring compounds are often not the most complementary to the receptor by visual inspection. This is because for many scoring functions, the score of a molecule depends on its size. We are the first to present a well-defined method for normalizing the score by size.

Our clique-finding routine, which pairs ligand atoms to match centers, is different from those described previously. It achieves speed by strictly limiting the number of possible cliques that are examined in a 'minimum residual' heuristic. This differs from the 'longest distance' [7] and 'binned distance' [3] heuristics of the two versions of the distance matching routine from DOCK and also from the algorithm in CLIX [4]. Our routine works well on computers with vector or RISC architecture. An important feature of the routine is that specifying some match centers as essential points leads to a substantial additional saving of time. Using essential points often makes sense biochemically; in many cases one can specify certain regions of active-site cavities,

e.g., near catalytic residues of enzymes, that must be occupied by inhibitors or substrates.

There is a certain arbitrariness in using discrete match centers to represent an active-site cavity and there may be no one best way to generate them. The orientations generated by any clique search routine are very sensitive to the number and exact location of the centers. We introduced an inexpensive rigid-body optimization step to reduce this sensitivity by allowing the candidate ligands to adjust to the receptor independently of the centers. In many cases, optimization finds many potentially high-scoring conformations that would otherwise be missed. The simplex procedure does not require derivative information and it is robust, normally reaching convergence in a few tens of cycles. The average time spent on clique searching and optimization is less than five seconds per conformation, fast enough for us to search databases as large as one million conformations in a reasonable time, using coarse parallelization.

NOTE ADDED IN PROOF

Meng et al. [32] have described the use of a rigid-body optimizer within DOCK that uses numerical derivatives. How the behavior of our simplex-based optimizer differs from that of Meng et al. will be published at a later date.

REFERENCES

1 DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 31 (1988) 722.
2 Shoichet, B.K., Bodian, D.L. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 380.
3 Meng, E.C., Shoichet, B.K. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 505.
4 Lawrence, M.C. and Davis, P.C., Protein Struct. Funct. Genet., 12 (1992) 31.
5 Sudarsanam, S., Virca, G.D., March, C.J. and Srinivasan, S., J. Comput.-Aided Mol. Design, 6 (1992) 223.
6 Bacon, D.J. and Moult, J., J. Mol. Biol., 225 (1992) 849.
7 Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
8 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. and Watson, D.G., Acta Crystallogr., B35 (1979) 2331.
9 DesJarlais, R.L., Seibel, G.L., Kuntz, I.D., Furth, P.S., Alvarez, J.C., Ortiz de Montellano, P.R., DeCamp, D.L., Babe, L.M. and Craik, C.S., Proc. Natl. Acad. Sci. USA, 87 (1990) 6644.
10 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., Rose, J.R., Craik, C.S. and Stroud, R.M., J. Biol. Chem., 268 (1993) 15343.
11 Kuntz, I.D., Science, 257 (1992) 1078.
12 Pearlman, R.S., Rusinko III, A., Skell, J.M., Balducci, R. and McGarity, C.M., CONCORD, distributed by Tripos Associates Inc., St. Louis, MO.
13 Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D. and Perry, K.M., Science, 259 (1993) 1445.
14 Ring, C.S., Sun, E., McKerrow, J.H., Lee, G.K., Rosenthal, P.J., Kuntz, I.D. and Cohen, F.E., Proc. Natl. Acad. Sci. USA, 90 (1993) 3583.
15 DesJarlais, R.L., Sheridan, R.P., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 26 (1986) 2149.
16 Leach, A.R. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 730.
17 Goodsell, D.S. and Olson, A.J., Protein Struct. Funct. Genet., 8 (1990) 195.
18 Kearsley, S., Underwood, D.J., Sheridan, R.P. and Miller, M.D., J. Comput.-Aided Mol. Design, (1994) submitted for publication.
19 Jiang, F. and Kim, S.-H., J. Mol. Biol., 219 (1991) 79.
20 Bush, B.L. and Sheridan, R.P., J. Chem. Inf. Comput. Sci., 33 (1993) 756.
21 Goodford, P.J., J. Med. Chem., 28 (1985) 849.
22 Boobbyer, D.N.A., Goodford, P.J., McWhinnie, P.M. and Wade, R.C., J. Med. Chem., 32 (1989) 1083.

23 Bohacek, R.S. and McMartin, C., J. Med. Chem., 35 (1992) 1671.
24 Weber, A.E., Halgren, T.A., Doyle, J.J., Lynch, R.L., Siegel, P.K.S., Parsons, W.H., Greenlee, W.J. and Patchett, A.A., J. Med. Chem., 34 (1991) 2692.
25 Wesson, L. and Eisenberg, D., Protein Sci., 1 (1992) 227.
26 Shavers, C.L., Parsons, M.L. and Deming, S.N., J. Chem. Educ., 56 (1979) 307.
27 Kearsley, S.K., Acta Crystallogr., A45 (1989) 208.
28 Filman, D.J., Bolin, J.T., Matthews, D.A. and Kraut, J., J. Biol. Chem., 257 (1982) 13663.
29 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
30 The Merck Index, 11th ed., Merck & Co., Rahway, NJ, 1989.
31 Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta Jr., S. and Weiner, P., J. Am. Chem. Soc., 106 (1984) 765.
32 Meng, E.C., Geschwend, D.A., Blaney, J.M. and Kuntz, I.D., Protein Struct. Funct. Genet., 17 (1993) 266.

APPENDIX

PARAMETERS FOR GENERATING GRID

HYDROGEN-BONDING PARAMETERS

Res[a]	A1	Charge	Type[b]	A2	ang0	n1	A3	dihed0	n2	dist0	D
***	O	−0.500	−1	C	145.000	2	CA	0.000	2	2.500	3.500
***	OT	−1.0	−1	C	145.000	2	CA	0.000	2	2.500	3.500
***	OCT	−1.0	−1	C	145.000	2	CA	0.000	2	2.500	3.500
***	N	−0.520	0								
***	HN	0.248	1	N	180.000	1				1.500	3.500
***	C	0.526	0								
!											
! NT is the backbone N which is not bonded (H position unspecified)											
***	NT	−0.272	1	CA	110.000	2				2.500	3.500
!											
! NTER is the true N-terminus, a cation											
***	NTER	0.728	1	CA	110.000	2				2.500	5.000
!											
ALA	CA	0.215	0								
ALA	CB	0.031	0								
!											
ARG	CA	0.237	0								
ARG	CB	0.049	0								
ARG	CG	0.058	0								
ARG	CD	0.111	0								
ARG	NE	−0.493	0								
ARG	HE	0.294	1	NE	180.000	1				1.300	5.000
ARG	CZ	0.812	0								
ARG	NH1	−0.635	0								
ARG	HH11	0.362	1	NH1	180.000	1				1.300	5.000
ARG	HH12	0.362	1	NH1	180.000	1				1.300	5.000
ARG	NH2	−0.635	0								
ARG	HH21	0.362	1	NH2	180.000	1				1.300	5.000
ARG	HH22	0.362	1	NH2	180.000	1				1.300	5.000
!											
ASP	CA	0.246	0								
ASP	CB	−0.208	0								
ASP	CG	0.620	0								
ASP	OD1	−0.706	−1	CG	145.000	2	OD2	0.000	2	2.300	5.000
ASP	OD2	−0.706	−1	CG	145.000	2	OD1	0.000	2	2.300	5.000

HYDROGEN-BONDING PARAMETERS (continued)

Res[a]	A1	Charge	Type[b]	A2	ang0	n1	A3	dihed0	n2	dist0	D
ASN	CA	0.217	0								
ASN	CB	0.003	0								
ASN	CG	0.675	0								
ASN	OD1	–0.470	–1	CG	145.000	2	ND2	0.000	2	2.500	3.500
ASN	ND2	–0.867	0								
ASN	HD1	0.344	1	ND2	180.000	1				1.500	3.500
ASN	HD2	0.344	1	ND2	180.000	1				1.500	3.500
!											
CYS	CA	0.146	0								
CYS	CB	0.100	0								
CYS	SG	0.000	0								
!											
GLN	CA	0.210	0								
GLN	CB	0.053	0								
GLN	CG	–0.043	0								
GLN	CD	0.675	0								
GLN	OE1	–0.470	–1	CD	145.000	2	NE2	0.000	2	2.500	3.500
GLN	NE2	–0.867	0								
GLN	HE1	0.344	1	NE2	180.000	1				1.500	3.500
GLN	HE2	0.344	1	NE2	180.000	1				1.500	3.500
!											
GLU	CA	0.246	0								
GLU	CB	0.000	0								
GLU	CG	–0.208	0								
GLU	CD	0.620	0								
GLU	OE1	–0.706	–1	CD	145.000	2	OE2	0.000	2	2.300	5.000
GLU	OE2	–0.706	–1	CD	145.000	2	OE1	0.000	2	2.300	5.000
!											
GLY	CA	0.246	0								
!											
! Histidine protonated at ND1											
HID	CA	0.219	0								
HID	CB	0.060	0								
HID	CG	0.089	0								
HID	ND1	–0.444	0								
HID	HD1	0.320	1	ND1	180.000	1				1.500	3.500
HID	CD2	0.145	0								
HID	NE2	–0.527	–1	LE2	0.000	2				2.500	3.500
HID	LE2	0.0	0								
HID	CE1	0.384	0								
!											
! Histidine protonated at NE2											
HIE	CA	0.219	0								
HIE	CB	0.060	0								
HIE	CG	0.112	0								
HIE	ND1	–0.527	–1	LD1	0.000	2				2.500	3.500
HIE	LD1	0.0	0								
HIE	CD2	0.122	0								
HIE	NE2	–0.444	0								
HIE	HE2	0.320	1	NE2	180.000	1				1.500	3.500
HIE	CE1	0.384	0								
!											
! Histidine cation											
HIP	CA	0.195	0								
HIP	CB	0.211	0								
HIP	CG	0.103	0								
HIP	ND1	–0.613	0								
HIP	HD1	0.478	1	ND1	180.000	1				1.300	5.000
HIP	CD2	0.353	0								
HIP	NE2	–0.686	0								
HIP	HE2	0.486	1	NE2	180.000	1				1.300	5.000
HIP	CE1	0.719	0								

HYDROGEN-BONDING PARAMETERS (continued)

Res[a]	A1	Charge	Type[b]	A2	ang0	n1	A3	dihed0	n2	dist0	D
!											
! Histidine in ambiguous protonation state											
HIS	CA	0.219	0								
HIS	CB	0.060	0								
HIS	CG	0.100	0								
HIS	CD2	0.134	0								
HIS	ND1	−0.326	9	LD1	0.0	2				2.5	3.5
HIS	LD1	0.0	0								
HIS	NE2	−0.326	9	LE2	0.0	2				2.5	3.5
HIS	LE2	0.0	0								
HIS	CE1	0.385	0								
!											
! Histidine with proton at ND1 and metal at NE2											
HMD	CA	0.219	0								
HMD	CB	0.060	0								
HMD	CG	0.089	0								
HMD	ND1	−0.444	0								
HMD	HD1	0.320	1	ND1	180.000	1				1.500	3.500
HMD	CD2	0.145	0								
HMD	NE2	−0.527	0								
HMD	CE1	0.384	0								
!											
! Histidine with proton at NE2 and metal at ND1											
HME	CA	0.219	0								
HME	CB	0.060	0								
HME	CG	0.112	0								
HME	ND1	−0.527	0								
HME	LD1	0.0	0								
HME	CD2	0.122	0								
HME	NE2	−0.444	0								
HME	HE2	0.320	1	NE2	180.000	1				1.500	3.500
HME	CE1	0.384	0								
!											
ILE	CA	0.199	0								
ILE	CB	0.030	0								
ILE	CG1	0.017	0								
ILE	CD1	−0.001	0								
ILE	CG2	0.001	0								
!											
LEU	CA	0.204	0								
LEU	CB	0.016	0								
LEU	CG	0.054	0								
LEU	CD1	−0.014	0								
LEU	CD2	−0.014	0								
!											
LYS	CA	0.227	0								
LYS	CB	0.039	0								
LYS	CG	0.053	0								
LYS	CD	0.048	0								
LYS	CE	0.218	0								
LYS	NZ	−0.272	0								
LYS	HZ1	0.311	1	NZ	180.000	1				1.300	5.000
LYS	HZ2	0.311	1	NZ	180.000	1				1.300	5.000
LYS	HZ3	0.311	1	NZ	180.000	1				1.300	5.000
!											
MET	CA	0.137	0								
MET	CB	0.037	0								
MET	CG	0.090	0								
MET	SD	−0.025	0								
MET	CE	0.007	0								
!											
PHE	CA	0.214	0								

HYDROGEN-BONDING PARAMETERS (continued)

Res[a]	A1	Charge	Type[b]	A2	ang0	n1	A3	dihed0	n2	dist0	D
PHE	CB	0.038	0								
PHE	CG	0.011	0								
PHE	CD1	−0.011	0								
PHE	CD2	−0.011	0								
PHE	CE1	0.004	0								
PHE	CE2	0.004	0								
PHE	CZ	−0.003	0								
!											
PRO	N	−0.257	0								
PRO	CA	0.112	0								
PRO	CB	−0.001	0								
PRO	CG	0.036	0								
PRO	CD	0.084	0								
!											
SER	CA	0.292	0								
SER	CB	0.194	0								
SER	OG	−0.550	−1	CB	110.000	2	HO	180.000	1	2.500	3.500
SER	HO	0.310	1	OG	180.000	1				1.500	3.500
!											
! Serine where the position of hydroxyl proton is unspecified											
SER$	CA	0.292	0								
SER$	CB	0.194	0								
SER$	OG	−0.240	9	CB	110.000	2				2.500	3.500
!											
THR	CA	0.268	0								
THR	CB	0.211	0								
THR	OG1	−0.550	−1	CB	110.000	2	HO	180.000	1	2.500	3.500
THR	HO	0.310	1	OG1	180.000	1				1.500	3.500
THR	CG2	0.007	0								
!											
! Threonine where the position of hydroxyl proton is unspecified											
THR$	CA	0.268	0								
THR$	CB	0.211	0								
THR$	OG1	−0.240	9	CB	110.000	2				2.5	3.50
THR$	CG2	0.007	0								
!											
TRP	CA	0.248	0								
TRP	CB	0.020	0								
TRP	CG	0.046	0								
TRP	CD1	0.117	0								
TRP	CD2	−0.275	0								
TRP	NE1	−0.330	0								
TRP	HE1	0.294	1	NE1	180.000	1				1.500	3.500
TRP	CE2	0.000	0								
TRP	CE3	0.145	0								
TRP	CZ2	0.029	0								
TRP	CZ3	−0.082	0								
TRP	CH2	0.034	0								
!											
TYR	CA	0.245	0								
TYR	CB	0.022	0								
TYR	CG	−0.001	0								
TYR	CD1	−0.035	0								
TYR	CD2	−0.035	0								
TYR	CE1	0.100	0								
TYR	CE2	0.100	0								
TYR	CZ	−0.121	0								
TYR	OH	−0.368	−1	CZ	120.000	2	HO	180.000	1	2.500	3.500
TYR	HO	0.339	1	OH	180.000	1				1.500	3.500
!											
! Tyrosine where position of hydroxyl proton is unspecified											
TYR$	CA	0.245	0								

HYDROGEN-BONDING PARAMETERS (continued)

Res[a]	A1	Charge	Type[b]	A2	ang0	n1	A3	dihed0	n2	dist0	D
TYR$	CB	0.022	0								
TYR$	CG	−0.001	0								
TYR$	CD1	−0.035	0								
TYR$	CD2	−0.035	0								
TYR$	CE1	0.100	0								
TYR$	CE2	0.100	0								
TYR$	CZ	−0.121	0								
TYR$	OH	−0.029	9	CZ	120.000	2	CE1	0.0	2	2.500	3.500
!											
VAL	CA	0.201	0								
VAL	CB	0.033	0								
VAL	CG1	0.006	0								
VAL	CG2	0.006	0								
!											
NDP	NN1	0.0	0								
NDP	NC2	0.0	0								
NDP	NC3	0.0	0								
NDP	NC4	0.0	0								
NDP	NC5	0.0	0								
NDP	NC6	0.0	0								
NDP	NC7	0.0	0								
NDP	NO7	0.0	−1	NC7	145.000	2	NN7	0.000	2	2.5	3.5
NDP	NN7	0.0	0								
NDP	H71	0.0	1	NN7	180.000	1				1.5	3.5
NDP	H72	0.0	1	NN7	180.000	1				1.5	3.5
NDP	NC1$	0.0	0								
NDP	NC2$	0.0	0								
NDP	NO2$	0.0	9	NC2$	110.000	2				2.5	3.5
NDP	NC3$	0.0	0								
NDP	NO3$	0.0	9	NC3$	110.000	2				2.5	3.5
NDP	NC4$	0.0	0								
NDP	NO4$	0.0	−1	NC1$	110.0	2	NC4$	180.000	1	2.5	3.5
NDP	NC5$	0.0	0								
NDP	NO5$	0.0	−1	NC5$	110.0	2	NP	180.000	1	2.5	3.5
NDP	NP	0.0	0								
NDP	NO1	−0.5	−1	NP	180.0	0				2.3	5.0
NDP	NO2	−0.5	−1	NP	180.0	0				2.3	5.0
NDP	O3	0.0	−1	NP	110.0	2	AP	180.000	1	2.5	3.5
NDP	AP	0.0	0								
NDP	AO1	−0.5	−1	AP	180.0	0				2.5	3.5
NDP	AO2	−0.5	−1	AP	180.0	0				2.5	3.5
NDP	AC5$	0.0	0								
NDP	AO5$	0.0	−1	AC5$	110.0	2	AP	180.000	1	2.5	3.5
NDP	AC4$	0.0	0								
NDP	AO4$	0.0	−1	AC1$	110.0	2	AC4$	180.000	1	2.5	3.5
NDP	AC3$	0.0	0								
NDP	AO3$	0.0	9	AC3$	110.0	2				2.5	3.5
NDP	AC2$	0.0	0								
NDP	AO2$	0.0	−1	AC2$	110.0	2	AP2$	180.000	1	2.5	3.5
NDP	AP2$	0.0	0								
NDP	AOP1	−0.667	−1	AP2$	180.000	0				2.3	5.0
NDP	AOP2	−0.667	−1	AP2$	180.000	0				2.3	5.0
NDP	AOP3	−0.666	−1	AP2$	180.000	0				2.3	5.0
NDP	AC1$	0.0	0								
NDP	AN1	0.0	−1	LN1	0.0	2				2.5	3.5
NDP	LN1	0.0	0								
NDP	AC2	0.0	0								
NDP	AN3	0.0	−1	LN3	0.0	2				2.5	3.5
NDP	LN3	0.0	0								
NDP	AC4	0.0	0								
NDP	AC5	0.0	0								
NDP	AC6	0.0	0								

HYDROGEN-BONDING PARAMETERS (continued)

Res[a]	A1	Charge	Type[b]	A2	ang0	n1	A3	dihed0	n2	dist0	D
NDP	AN6	0.0	0								
NDP	H61	0.0	1	AN6	180.0	1				1.5	3.5
NDP	H62	0.0	1	AN6	180.0	1				1.5	3.5
NDP	AN7	0.0	−1	LN7	0.0	2				2.5	3.5
NDP	LN7	0.0	0								
NDP	AC8	0.0	0								
NDP	AN9	0.0	0								
!											
HOH	O	−0.6	−1	H1	126.0	2	H2	180.000	1	2.5	3.5
HOH	H1	0.3	1	O	180.0	1				1.5	3.5
HOH	H2	0.3	1	O	180.0	1				1.5	3.5
!											
! Water where positions of the proton unspecified											
HOH$	O	0.0	9							2.5	3.5
!											
ZN	ZN	2.0	1							1.8	6.0
FE	FE	2.0	1							1.8	6.0
CA	Ca	2.0	1							1.8	6.0

Charges are from the AMBER force field, Ref. 31.

[a] *** means 'any residue'.

[b] Type: 1 = donor; −1 = acceptor; 9 = donor and/or acceptor; 0 = neither.

VAN DER WAALS PARAMETERS

Element	Radius	Depth
! Hydrogens are ignored		
H	1.00	0.00
C	1.85	0.12
N	1.75	0.16
O	1.65	0.15
S	2.00	0.20
P	2.10	0.20
!		
! Dummy atom		
L	0.0	0.00
!		
! Special PDB atom names for NDP		
NN	1.75	0.16
NO	1.65	0.15
NC	1.85	0.12
NP	2.10	0.20
AN	1.75	0.16
AO	1.65	0.15
AC	1.85	0.12
AP	2.10	0.20
!		
! Metals		
ZN	0.65	0.10
FE	0.65	0.10
!		
! Calcium distinguished from alpha carbon		
Ca	0.65	0.10
!		
! Probe		
PR	1.65	0.15
!		
! Hydrophobic threshold and multiplication factor F		
FO	1.50	1.0

Most parameters are from the AMBER force field, Ref. 31.