Conformational Analysis of Flexible Ligands in Macromolecular Receptor Sites

Andrew R. Leach[1]* and Irwin D. Kuntz[2]

[1]Computer Graphics Laboratory and [2]Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California 94143-0446

Received 26 September 1991; accepted 13 January 1992

A computational method for exploring the orientational and conformational space of a flexible ligand within a macromolecular receptor site is presented. The approach uses a variant of the DOCK algorithm [Kuntz et al., *J. Mol. Biol.*, **161**, 288 (1982)] to determine orientations of a fragment of the ligand within the site. These positions then form the basis for exploring the conformational space of the rest of the ligand, using a systematic search algorithm. The search incorporates a method by which the ligand conformation can be modified in response to interactions with the receptor. The approach is applied to two test cases, in both of which the crystallographically determined structures are obtained. However, alternative models can also be obtained that differ significantly from those observed experimentally. The ability of a variety of measures of the intermolecular interaction to discriminate among these structures is discussed.

INTRODUCTION

The growing number of therapeutically important macromolecules whose 3-dimensional structures are known to atomic resolution has led to considerable interest in "site-directed molecular design." This can be conceptually regarded as the design of a suitable small molecule "key" to fit the macromolecular "lock." Computer molecular modeling plays an important role in this process, not only by the provision of hardware and software for interactive display but also through methods to calculate interaction energies, simulate dynamic properties, and determine relative free energies of binding. In this article, we describe a computational method for exploring the orientational and conformational space of a flexible molecule such as a putative ligand within a macromolecular active site. The method has been developed to address three issues in this aspect of molecular modeling and drug design. The first issue is to provide a means of rapidly determining chemically plausible structures for the intermolecular complex. These structures may then serve as starting models for more detailed computational analysis. Elucidating all the possible binding modes is difficult to perform manually and so theoretical methods will clearly be of some value. Our second objective was to develop a strategy for performing the conformational analysis of flexible molecules within the confines of an external "environment" (i.e., the receptor site) and to determine the extent to which current conformational search methods for isolated molecules can be applied to this problem. Finally, the method will be used to provide structures for an ongoing investigation to determine which evaluation functions (e.g., steric fit, electrostatic energy, etc.) are most useful in discriminating between different models of an intermolecular complex.

Two other methods that take account of conformational flexibility when docking small molecule ligands have previously been reported. These are the distance geometry method of Ghose and Crippen[1] and the simulated annealing approach of Goodsell and Olsen.[2] Both use a rigid receptor structure. In distance geometry, a bounds matrix giving upper and lower bounds on the interatomic distances is first calculated. For a ligand in a receptor site, this bounds matrix contains two blocks, one appropriate to the ligand atoms and the other to the site atoms. Elements within the ligand block refer to intraligand distances whereas those in the site block refer to intrasite distances (in the case of a rigid receptor the upper and lower bounds for the site atoms are equal). Elements (**i**, **j**) where **i** and **j** are in different blocks refer to ligand-site interatomic distances. The lower bound of the distances that involve at least one ligand atom (both intra and intermolecular) are typically set equal to the sum of the appropriate van der Waals (VDW) radii. The upper distance constraints are initially set to a

*Author to whom all correspondence should be addressed at Department of Chemistry, University of Southampton, Southampton, Hampshire SO9 5NH, England, UK.

CCC 0192-8651/92/060730-19$04.00

large value. The upper and lower bounds are then refined using triangle inequality rules. Conformations of the ligand are generated by randomly assigning distances between the upper and lower bounds of the smoothed distance matrix, embedding, and then optimizing the resulting coordinates against the initial distance bounds. Ligand atoms can be positioned in close proximity to specific site atoms (e.g., to form hydrogen bonds) by assigning appropriate bounds. Clique-finding algorithms have been used to automatically determine such sets of hydrogen-bonding ligand-site atom pairs.[3] However, the computational resources required for the embedding and optimization steps can be quite heavy and the full procedure of clique finding and structure generation may require a significant amount of computer time.

The simulated annealing approach of Goodsell and Olsen uses the Metropolis algorithm at a decreasing sequence of temperatures to derive a structure for the interaction complex. The ligand is permitted six degrees of translational/rotational freedom and conformational flexibility is accommodated by rotating about single bonds. Currently, there is no way to deal with conformationally variable rings. At each step, a new configuration is generated by making random changes to each variable (the rotatable bonds, $x/y/z$ translations, and $x/y/z$ rotations). The energy of the configuration is calculated and it is accepted or rejected according to the Metropolis criterion. This is performed for decreasing temperatures and eventually a final configuration is obtained. The configuration energies are calculated using a grid-based approach with parameters from the AMBER force field. It was found necessary to significantly increase the relative contribution of the hydrogen bond term to produce the desired results. Simulated annealing should (in theory) always converge to the same minimum. To guarantee this, however, would require an infinite number of temperature decrements at each of which the system comes to thermal equilibrium, and for practical purposes it is necessary to use a relatively small number of temperatures with a finite number of equilibration steps. It is consequently more generally used in molecular modeling to traverse a large part of the energy hypersurface to identify a variety of reasonable minima. For flexible molecules, convergence could require a great increase in computational effort (the cases studied, which had a maximum of four rotatable bonds, required up to 1 h of Convex CPU time for each run, each of which only produces a single structure). Nevertheless, simulated annealing should be an efficient method for local sampling and refinement, perhaps using as a starting point a structure from an alternative approach.

The algorithm we report here represents a synthesis of our previous investigations into the docking of ligands to macromolecular receptor sites (viz. the DOCK program[4,5]) and in conformational analysis and search (the WIZARD[6] and COBRA[7] programs). In their respective fields, these methods have proved capable of performing rapid searches of the relevant space to give good quality models. An underlying principle of the combined approach is the incorporation of "chemical intuition" to enhance efficiency and improve the quality of the resulting models. Two frequently observed features of macromolecule–ligand complexes are the high degree of shape complementarity and the presence of hydrogen bonds, elements of which we have therefore tried to include in the algorithm. In addition to directing the search toward structures exhibiting these two characteristics, we have assumed that the ligand prefers to attain a reasonably low-energy conformation (i.e., one that is related to, but not necessarily identical to, a minimum energy conformation of the isolated molecule). As will be described in detail below, a two-stage process is used. First, the relationship between the coordinate frames of the ligand and the receptor is established by determining a number of orientations of a part of the ligand (conceptually the "most rigid" part of the molecule), herein termed the *anchor fragment* within the site, using a variant of the DOCK algorithm termed *Directed DOCK*. These orientations are pruned to eliminate those of less interest and then clustered. One orientation is selected from each family and the resulting set then forms the basis for the exploration of the conformational degrees of freedom of the remainder of the ligand, within the confines of the receptor site. As in previous work, a rigid receptor structure is assumed throughout.

The DOCK Algorithm

The DOCK algorithm provides a method for rapidly determining possible binding modes of a ligand (in a fixed conformation) within a receptor site. A number of publications describe the method in some detail[4,5,8] so only a brief outline will be given here. The molecular surface of the macromolecule is first calculated using the MS program.[9-11] A "negative image" of the site, consisting of a set of overlapping spheres, is then calculated by a program called SPHGEN from the points that define the molecular surface and the surface normals of these points. Orientations of the ligand conformation within the site are generated by matching the ligand atoms with the sphere centers to find sets containing at least four atom/sphere center pairs in which all the interatomic distances are equal to the corresponding sphere center–sphere center dis-

tances, within some tolerance. The translation-rotation matrix that best superimposes the ligand atoms on the sphere centers is then determined and the ligand oriented within the site. The orientation is checked to ensure there are no unfavorable steric interactions between the atoms of the ligand and the receptor. If it is acceptable, an interaction energy is computed and the orientation stored. Initially, the algorithm was used to dock single ligands into a variety of protein receptor sites and was found capable of producing orientations that are close to those determined experimentally. More recently, it has been used to tackle the problem of docking two macromolecules[12] and to search "3-dimensional" databases for the purpose of discovering novel lead compounds.[5,13] Two particular advantages of the approach that are relevant to the current discussion are its speed and the high degree of shape complementarity of the resulting structures. This second characteristic is in part a consequence of the use of a site description derived from the molecular surface. A limited amount of conformational flexibility was previously incorporated into the DOCK approach in two ways. In the first, a number of conformations of each ligand were docked.[14] This approach requires the structure of the ligand to be predetermined (typically conformations at local energy minima are selected), with no possibility for the ligand to adapt to the receptor environment. The second method was to dock fragments of the ligand, which were then reformed using energy minimization.[8] This is an attractive approach for ligands that are composed of a small number of relatively large fragments, but it might be difficult to implement for very flexible ligands. Moreover, the intrinsic reliance on minimization to reform the ligand adds to the computational cost of such an approach.

INCORPORATING INFORMATION ABOUT THE RECEPTOR SITE INTO THE DOCK ALGORITHM: DIRECTED DOCK

The Directed DOCK algorithm aims to actively incorporate chemical "knowledge" when orienting a ligand or a fragment of a ligand within the site. Such knowledge was previously incorporated in the DOCK approach only in a passive sense by the use of a variety of scoring functions to evaluate the orientations (which, as indicated above, are determined solely on the basis of shape). The originally employed scoring function (the *contact score*, Fig. 1) was designed to measure the degree of shape complementarity between the ligand and the receptor site using a distance-based pairwise interatomic function. Electrostatic scoring functions have subsequently been incorporated[13,15] and others are under development.[16] The Directed DOCK algorithm provides the means by which other types of interactions can be taken into account during the matching procedure, with the aim of providing higher-quality structures for subsequent modeling experiments.

Here, we report how information about the hydrogen-bonding features of the receptor can be incorporated in the form of *hydrogen-bonding site points*. Each of these is associated with a hydrogen-bonding donor or acceptor receptor atom and defines a point within the site where an acceptor or donor ligand atom (as appropriate) could be placed to form a hydrogen bond with the receptor atom. The Directed DOCK algorithm thus uses an extended set of site points comprising both the original sphere centers and the additional hydrogen-bonding site points. These site points are then matched with the ligand atoms in a scheme that takes into consideration the "character" of the ligand atom and the site point. For example, a hydrogen-bonding site point associated with a donor site atom can only match a ligand acceptor atom and vice versa. First, we describe the method by which the coordinates of the hydrogen-bonding site points are determined.

Determination of Hydrogen-Bonding Site Points

In many cases, the receptor structures come from X-ray crystallography and hydrogen atom positions are not determined. First, therefore, the positions of *donor hydrogen atoms* (i.e., hydrogen atoms bonded to an appropriate electronegative atom) are calculated from the heavy atom coordinates for all groups except $-OH$ (in serine and threonine), $-SH$ (in cysteine), and $-NH_3^+$ (in lysine), where it is not possible to unambiguously determine the hydrogen positions. Each potential hydrogen bond donor or acceptor atom in the site is then considered in turn. A sphere with points evenly distributed upon its surface is positioned so that the center of the sphere coincides with the atom. The radius of the sphere is chosen to be such that an appropriate atom when placed at the sphere's surface would be at a distance corresponding to the "ideal" hydrogen bond distance between the two atoms. Where an explicit hydrogen atom is involved the radius of the sphere is 1.8 Å; where no hydrogen is involved, it is 2.8 Å. Each point on the sphere's surface is then examined. Any point that does not lie within at least one of the spheres that form the negative image of the site is rejected (Fig. 2). This eliminates points that are either buried in the receptor or on the outside of the macromolecule. In addition, when the receptor atom is either a donor hydrogen atom or an

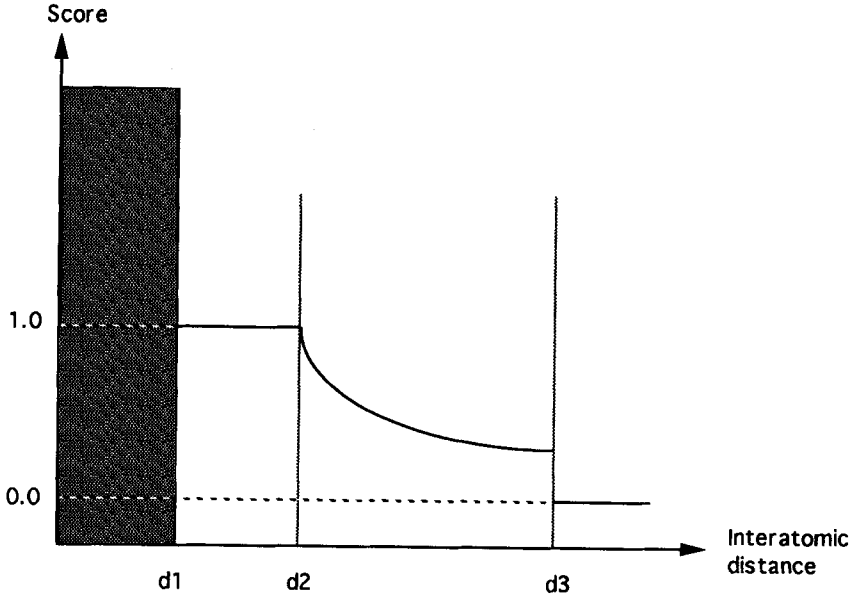

Figure 1. DOCK contact scoring function. Orientations with an atom closer than d1 are rejected. Otherwise, the score for the orientation is given by

$$\sum_{i=1}^{N}\sum_{i=1}^{M} S_{ij}(r)$$

where N is the number of atoms in the receptor and M is the number in the ligand. $S_{ij}(r)$ is the interatomic scoring function, which has a value of 1 for $d1 \leq r \leq d2$, a value of $\exp[-(r - d2)^2]$ for $d2 \leq r \leq d3$, and a value of 0 for $r > d3$. Typical values are d1 = 2.5 Å; d2 = 3.5 Å, d3 = 5.0 Å. In this article, this function is only applied to interactions between heavy (i.e., nonhydrogen) atoms.

acceptor atom the angle subtended at the atom is also evaluated. If the angle is less than 90°, then the point is rejected (Fig. 2). This recognizes that hydrogen bonds have an angular preference.[17] The cutoff value of 90° corresponds to the criteria expounded by Baker and Hubbard.[18] Each of the points that satisfies these two criteria is added to the original set of sphere centers.

The number and radii of the spheres used for each hydrogen-bonding atom varies according to its nature (donor or acceptor; hydrogen or "heavy" atom). A single 1.8 Å sphere is used for a donor hydrogen atom, the points on its surface indicating site positions where an acceptor atom could be placed. For a heavy donor atom, where no hydrogens are defined, such as Nζ of lysine, a 2.8 Å sphere is used. Two concentric spheres are used for an acceptor atom such as a carbonyl oxygen. The larger, with radius 2.8 Å, corresponds to points where a heavy donor atom from the ligand could be placed. The smaller, 1.8 Å sphere, identifies positions where a donor hydrogen could be positioned. For an atom such as the oxygen of a hydroxyl group that can act as both donor and acceptor, three spheres are used: one appropriate to an acceptor ligand atom, one to a heavy donor atom, and one to a donor hydrogen.

Determination of Orientations Within the Site

Having defined the set of site points, a *matching algorithm* is used to generate *matching sets.* A matching set contains at least four ligand atom/site point pairs from which the unique translation-rotation matrix to orient the anchor fragment of the ligand in the site is calculated. The use of more than one type of site point has prompted the development of a matching algorithm that differs from those previously employed.[4,13] One key objective was that it should not be inherently biased either toward the formation of hydrogen bonds or toward the attainment of high shape complementarity. Rather, we desired that it should be capable of providing a range of structures, some with a high degree of hydrogen-bonding complementarity, some with a high degree of shape complementarity, and some with both hydrogen-bonding and shape characteristics. Consequently, the matching algorithm is designed to provide matching sets in which a high proportion of the atom/site point pairs are be-

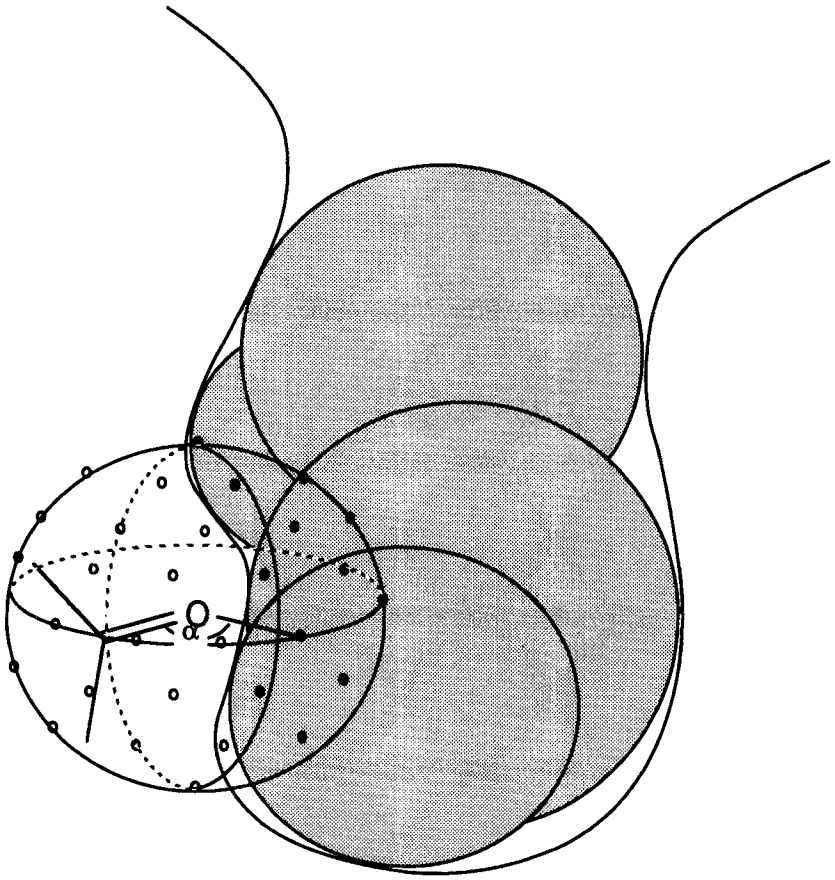

Figure 2. Schematic illustration of the derivation of hydrogen-bonding site points. A sphere with points uniformly distributed on its surface is positioned at each potential hydrogen bond donor or acceptor atom in the receptor site. For a point to be accepted, it must lie within at least one of the spheres that form the negative image of the site (as derived from the SPHGEN program), and must also form an acceptable angle (where applicable). In this diagram, acceptable points are indicated by solid circles and unacceptable points by open circles.

tween hydrogen-bonding atoms/site points (hydrogen-bonding complementarity), sets in which a high proportion of the matching pairs are between ligand atoms and the original sphere centers (shape complementarity), and sets that contain both types of matching pairs. A second distinguishing feature of the matching algorithm is that it is more exhaustive than those previously reported. As a consequence, it usually requires more computer time, but the additional computational effort is compensated by a more complete coverage of the site and thus results of higher quality.

The matching algorithm (outlined in Fig. 3) first determines the numbers of donor and acceptor atoms in the ligand. Their sum gives an upper limit on the number of possible hydrogen-bonding pairings that any matching set may contain (**max_h_pairs**). The user defines a minimum number of hydrogen-bonding matches that must be present in each matching set (**min_h_pairs**). The algorithm then generates all possible sets of complementary hydrogen-bonding atom/site point pairs containing **min_h_pairs**, **min_h_pairs** + 1, . . . **max_h_pairs** subject to the requirement that all the site point distances must be equal to the corresponding interatomic distances within the user-specified tolerance. Thus, for example, if there are two donor hydrogen ligand atoms (α, β), three hydrogen-bonding site points (A, B, C), each associated with an acceptor atom, and at least one hydrogen-bonding match is required (i.e., **min_h_pairs** = 1) then the following sets will be produced: [αA], [αB], [αC], [βA], [βB], [βC], [αA, βB], [αA, βC], [αB, βA], [αB, βC], [αC, βA], [αC, βB], [αC, βC] (assuming of course that the interatomic distance α-β is equal to the distances A-B, A-C, C-B within the tolerance).

Each of the hydrogen-bonding sets (many of which will, of course, contain fewer than the minimum four pairs required to define the translation-

```
max_h_pairs := number of h-bonding ligand atoms
number_of_h_bond_matches := min_h_pairs

For each conformation of the anchor fragment do

  While number_of_h_bond_matches ≤max_h_pairs do

    Generate h-bonding sets containing number_of_h_bond_matches:

      Take the next h-bonding set:

        Extend the set by matching ligand atoms with sphere centres:
          Determine next ligand atom to be matched
          Find matching sphere centre
        Until the set cannot be extended further

        Calculate translation-rotation matrix
        Orient, check for problems, evaluate hydrogen bonds and score

        Backtrack and find a new sphere centre to match the ligand atom last
          considered

      Until no more sets based on the current h-bonding set can be found

    Until all h-bonding sets containing number_of_h_bond_matches have been
      considered

    Increase number_of_h_bond_matches by 1

  enddo

enddo
```

Figure 3. Outline the matching algorithm.

rotation matrix) is extended by matching additional ligand atoms with the sphere centers. To examine all possible combinations would require too much computer time and so a variant on the original DOCK matching algorithm is used to determine the "next" ligand atom to be matched at each stage. This atom is chosen from those not yet in the set and is the one that is "furthest" from the atoms already paired with site points. The furthest atom is defined to be the one with the largest sum of interatomic distances from the paired ligand atoms. The sphere centers are then examined to find a match for this ligand atom such that all the interatomic distances are equal to the corresponding distances between site points, within the specified tolerance. If such a sphere center can be found, the next ligand atom is determined and a matching sphere center assigned. The process continues until no more matches between atoms and sphere centers can be found. If at this point the set contains four or more pairs the translation-rotation matrix is determined using the algorithm of Ferro and Hermans[19] and the fragment is oriented within the site. The orientation is checked for steric problems, and if satisfactory all possible hydrogen-bonding interactions are evaluated (not just those in the matching set) using criteria similar to those of Baker and Hubbard[18] to determine which are acceptable. These criteria depend on whether or not an explicit hydrogen atom is involved. Where an explicit hydrogen is involved, the distance from acceptor to hydrogen must be less than 2.8 Å, the distance from the heavy donor atom to acceptor must be less than 3.8 Å, the angle subtended at the hydrogen by the heavy donor atom and the acceptor atom must be greater than 90°, and the angle

subtended at the acceptor atom by the donor and the atom bonded to the acceptor (e.g., the carbon in a carbonyl group) must also be greater than 90°. When there is no explicit hydrogen atom, the distance from donor to acceptor must be less than 3.8 Å and the angle subtended at the acceptor atom by the donor and the atom bonded to the acceptor must be greater than 90°. The algorithm then backtracks to find a new sphere center for the last (non-hydrogen-bonding) ligand atom to be considered, and proceeds until all possible matching sets based on the initial hydrogen-bonding set have been generated. The next hydrogen-bonding set is then generated, is extended by the addition of atom/sphere center pairs, and so on until all the possible hydrogen-bonding sets have been considered.

The space covered by the receptor is divided into "cells" of length equal to the VDW cutoff (usually around 2.5 Å), which enables the nearest neighbors of a ligand atom when it is oriented within the site (as must be done to check for steric problems) to be determined in a time independent of the number of atoms in the site. This is of particular value when an all-atom model of the receptor is used in which all hydrogen atoms are defined, not just the donor hydrogens. Second, due to the greater concentration of site points and the desire to produce structures of higher quality, a tighter tolerance is used to accept or reject a particular pairing (~0.7 Å rather than the 1.5 Å typically used in DOCK). Two consequences of this are that a much smaller proportion of the matching sets contain more than the minimum four pairs necessary to determine a unique translation-rotation matrix than in the original DOCK algorithm, and that fifth, sixth, etc. pairs do not change the orientation much in a rms sense. The matching algorithm thus generates sets containing the minimum of four matching pairs; this has little effect on the number of orientations obtained and the degree of coverage but does reduce the time required.

Pruning and Clustering the Orientations

The number of orientations of each conformation of the anchor fragment produced by the directed dock algorithm can be quite large—sometimes several hundred. This initial set of orientations will inevitably contain some of less interest than others because they make fewer hydrogen bonds with the site, have a lower score, or are deemed inferior according to some other measure of the interaction. In the first instance, these would be considered less viable starting points for the conformational search and so should be eliminated. In addition, many of the orientations can be very similar and would lead to much repetition were they all to be used. The orientations are thus subjected to a combined pruning and clustering process, the object of which is to first eliminate the "less interesting" orientations and then to cluster those that remain based upon their position and orientation in the receptor site. One representative orientation is then taken from each of the clusters for the conformational search. In this way, a number of spatially disparate orientations, all of which meet some minimum "standard," obtained. These serve to define the orientational relationship between the ligand and the receptor site from which to explore the conformational space of the remainder of the ligand.

A number of measures can be used to assess and then prune a set of orientations of a ligand within a receptor, be they from DOCK, Directed DOCK, or some other approach such as distance geometry. In the extreme, one could minimize each orientation using molecular mechanics and then use the intermolecular energy as the basis for pruning. However, the computational effort required for this would be prohibitive and so alternative measures need to be devised. In this work, three measures to assess the intermolecular interaction are used: the number of hydrogen bonds made with the site, the DOCK score, and the electrostatic interaction energy. The latter is calculated using a grid-based approach[15] in which the electrostatic potential at each point in a regular grid superimposed on the receptor site is determined using charges from the AMBER[20,21] force field and a distance-dependent dielectric function ($\varepsilon\alpha 1/r$). The idea of mapping the potential field of a receptor onto a regular grid was first introduced by Goodford[22] and has subsequently been used in a variety of applications including comparative molecular field analysis (CoMFA)[23] and Goodsell and Olsen's simulated annealing method.[2] The contribution of each atom to the electrostatic energy of a particular orientation can then be obtained by calculating the electrostatic potential at the atom using linear interpolation from the surrounding grid points and multiplying by the atomic charge. The individual atom contributions when added together give the electrostatic interaction energy of the orientation.

The improved leader algorithm[24] is used to cluster the orientations that remain after pruning. In this algorithm, a "central" object is first determined; in our case, this is taken to be the orientation that is most similar to the "average" coordinates of the whole set. The similarity/dissimilarity measure is defined as the rms distance separation for all M atoms in the fragment or ligand:

$$\text{Similarity} = \sqrt{\left[\left(\sum_{i=1}^{M} |r_i(a) - r_i(b)|^2\right)\Big/ M\right]}$$

for two orientations a and b

This gives an indication of how different any two orientations are, in both a translational and rotational sense. The orientation most dissimilar to this central orientation is then found; these then define two "leaders" and the entire set of orientations is then sorted into two clusters according to which of the two they are closest to. During sorting, the orientation most dissimilar from its leader is found; this then becomes the third leader. In the next step, the orientations are sorted into three clusters, and so on. After $N - 1$ passes, the orientations are thus divided into N clusters. Should all clusters fall below a given size, as measured by the similarity between the leader and the most dissimilar member (e.g., 1 Å), then all the clusters are regarded as sufficiently small and any further division is unwarranted. This algorithm has the advantage of speed over other methods that require a similarity matrix of order N to be precomputed, without some of the drawbacks associated with other quick-partition algorithms (although it is still dependent on the ordering of the orientations). It is worth noting that the commonly used single-linkage method is generally unsuitable because of its tendency to produce elongated clusters. When the orientations have been partitioned into the desired number of clusters, a representative orientation for each is chosen by determining the one with the best score, according to the currently operable evaluation measures.

THE CONFORMATIONAL SEARCH

Every orientation that remains after pruning and clustering defines a position of the anchor fragment relative to the site. For each of these, the conformational space of the remainder of the ligand is then searched, keeping the coordinates of the anchor fragment fixed. The conformational part of the procedure is similar in some ways to some other programs (e.g. SYBYL[25]) in which the coordinates of part of a molecule remain fixed while searching the conformational space of the remainder. However, there are some significant differences in our algorithm, as will be described below. The conformational search must find conformations for the ligand that are not only internally acceptable but that also have no unfavorable intermolecular interactions with the environment. We have chosen to use a systematic search algorithm to perform the search. A number of novel features are incorporated to try and minimize the effect of the "combinatorial explosion" to which such algorithms are subject. This is magnified in our case as the search is to be performed for a number of different orientations. Rings and ring systems can be particularly problematic for systematic dihedral searches and so for these we use preformed fragments. A relatively small number of dihedral values are used for the remaining acyclic rotatable bonds, chosen to cover the minimum energy regions of the bond as it rotates. By implication, we thus assume that the ligand prefers to adopt a low-energy structure. However, as interactions with the receptor may cause the ligand to be deviate from a gas-phase minimum-energy structure we have incorporated a method (an *adjusting algorithm*) that "modifies" the ligand in response to the environment. The negative image of the site is used to prune the appropriate parts of the search tree when the ligand conformation moves into "uninteresting" regions of the conformational space.

The acyclic single bonds and preconstructed ring fragments constitute the *variables* which with a given conformation of the anchor fragment define an internal conformation of the ligand. In this respect, the approach differs from the WIZARD[6] and COBRA[7] programs, in which the entire molecule (both cyclic and acyclic portions) is constructed from preformed templates. Here, only the rings and ring systems in the molecule are constructed from prestored fragments. The conformations of rings and ring systems are computed before the search proper commences by joining preformed templates using a modification of the COBRA program.[7] Each rotatable bond, conformationally variable ring, and ring system is then ordered according to the number of bonds between it and the anchor fragment. This gives the order in which they will be changed during the search. The variables thus define a search tree that is searched using a depth-first algorithm. As each variable is modified (by rotating a dihedral or inserting a new conformation for a ring system), the *group* of atoms whose positions will not be further affected until this variable is changed again are examined to check for intramolecular and intermolecular problems. If there is any intramolecular strain (ligand atoms too close together), backtracking occurs: The variable is set to its next value or the search returns to the previous variable as appropriate. If there is intermolecular strain, the adjusting algorithm is used to modify the conformation to try and eliminate the problem as described below. When all the atoms in the group have been examined, the next variable is assigned its value and so on until a conformation has been constructed. The interaction energy is then computed and the algorithm backtracks to the last variable to be changed. The search continues until the entire search tree has been explored. At this point, the next orientation is considered and the search procedure repeated. The process continues until all the orientations have been considered.

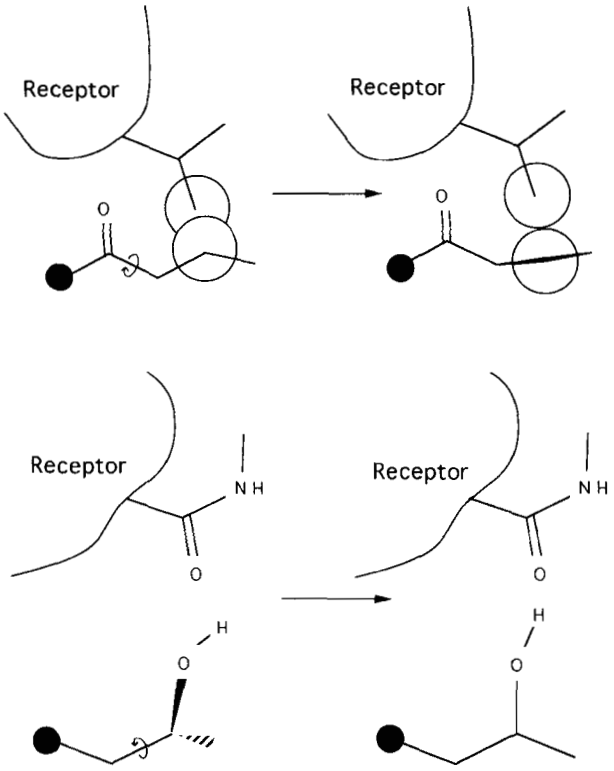

Figure 4. Schematic illustration of the adjustment of the ligand in response to steric and hydrogen-bonding interactions with the receptor site.

Adjusting the Ligand Conformation to Account for Intermolecular Interactions

Two types of interactions are recognized when a ligand atom is examined for intermolecular strain with the site: too-close steric repulsions and hydrogen-bonding attractions. The adjusting algorithm attempts to resolve such *strain* in a logical fashion by making changes to the dihedral values of the ligand's rotatable bonds. A steric problem is successfully resolved by moving the ligand atom away from the offending receptor atom(s) to an appropriate distance, given by the currently operable intermolecular VDW cutoff value; hydrogen-bonding strain is resolved by moving the ligand atom closer to the appropriate site atom so that the interatomic distance corresponds to that of an ideal hydrogen bond (Fig. 4). In each case, the objective is thus to position the ligand atom an appropriate distance from one or more site atoms. The adjusting algorithm we currently use is related to one previously developed for resolving intramolecular strain in molecules.[26] In that work, the algorithms were used to resolve strain in a complete conformation of an isolated molecule. Here, however, the conformation of the ligand is adjusted as it "grows" within the receptor site.

In the case of a steric problem, all the receptor atoms to which the ligand atom is too close are identified and collected in a list. Each rotatable bond that lies between the ligand atom and the anchor fragment of the molecule is then examined to determine whether a change in its dihedral value could move the ligand atom out of conflict with the receptor atom(s) and, if so, the exact amount of rotation required. The *deformations* so obtained are ranked according to the amount of energy required to perform them to take into consideration the fact that (for example) to rotate about an amide bond requires more energy per degree of dihedral change than a bond between two sp^3 carbon atoms. Any deformations that require an energy greater than some threshold (e.g., 3 kcal/mol) are discarded. The lowest-energy deformation is then tested: the appropriate bond rotation is performed and the resulting ligand conformation is examined to check that the deformation has introduced no intraligand strain, that other ligand atoms have not been moved into steric conflict with the site, and that no hydrogen bonds previously formed have been broken. If the deformation is acceptable, the next atom in the group is examined. However, should this lowest-energy deformation not prove satisfactory, the one next lowest in energy is tried, and so on until all the possibilities have been considered. If none of the deformation mechanisms is successful, the algorithm backtracks to the previous atom within the group for which a deformation was made, and the next possibility attempted for that atom. In this way, the space of deformation mechanisms for the group of atoms is explored to try and find a sequence of adjustments that will resolve all the strain in the group. If no satisfactory resolution mechanisms can be found that resolve all the strain in the group, the conformational search continues by backtracking to the appropriate point in the search tree.

Resolution of hydrogen-bonding strain between the ligand and the site is performed concurrently with the resolution of steric strain. All complementary receptor atoms within a given distance of a hydrogen-bonding donor or acceptor ligand atom that is not in steric conflict with the site are found. This distance is typically chosen to be 2 Å larger than the ideal hydrogen bond distance. A similar analysis to that used to resolve steric strain is then performed to determine which (if any) bond rotations will adequately position the ligand atom at the appropriate hydrogen-bonding distance from each of the site atoms. These deformation mechanisms are ranked and then examined in turn to determine if any will indeed position the ligand atom so that it makes a reasonable hydrogen bond with the appropriate receptor atom but without introducing any intra- or intermolecular strain. If no satisfactory resolution mechanism can be found, the atom is simply left in its original posi-

tion. Should there be more than one complementary site atom "within range" of the ligand atom, the algorithm tries each possibility in turn. The same set of variable values may thus give rise to more than one conformation, each making a different set of hydrogen bonds with the site. It is this feature that gives the approach an advantage over an alternative method of resolving strain, which would be to use a few steps of energy minimization. Minimization algorithms produce only a single structure from a given starting point, and so significant structures could be missed if there is more than one complementary site atom within range.

Three advantages accrue from this ability to adjust the conformation in response to interactions with the receptor. The first is that higher-quality structures can be produced because hydrogen-bonding atoms can be positioned so as to attain a more favorable orientation than might otherwise be the case. Second, by permitting the ligand to respond to the external influence of its environment fewer dihedral values are required for the rotatable bonds, as there is less chance that a structure might be missed because too few values were used. The dihedral values permitted to a given rotatable bond are typically chosen to cover the low-energy regions of the potential energy surface as the bond is rotated. For example, three values would be used for a bond between two sp^3 carbon atoms, corresponding to the anti and two gauche conformations. In cases where the potential energy surface has a flat "valley," a number of values would be chosen to give a reasonable coverage in these regions. The third advantage is that fewer structures are produced. Conformational search algorithms can generate many conformations that are very similar to each other and that on refinement produce the same minimum energy structure. A smaller family of structures requires less time to refine and is more easily analyzed. Nevertheless, the entire search procedure can generate a large number of structures, which must then be pruned and clustered as above.

Pruning the Search Tree

The search tree is pruned when there are either unfavorable intraligand interactions or unacceptable steric interactions with the receptor site that cannot be resolved. Additional improvements in search efficiency can be achieved by eliminating other parts of the search tree corresponding to structures that, although problem-free, will be of no subsequent interest. In many cases, it is a reasonable assumption to discard structures where the ligand extends far out of the site in favor of those where the ligand remains in close proximity to the macromolecule. This could, of course, be performed when the search is complete, but can result in considerable savings in computer time if done during the search. The original set of spheres that provide the negative image of the site can be used to eliminate such structures, for should the ligand conformation cease to lie completely within at least one of the spheres then it can be regarded as no longer being inside the site and the appropriate part of the search tree pruned from further consideration.

EXAMPLES

Here, we describe the application of the approach described above to the binding of methotrexate (MTX) to dihydrofolate reductase (DHFR) and of netropsin to the DNA duplex d(CGCGATATCGCG). In both cases, the objective was to not only test the effectiveness of the approach in finding structures similar to those obtained by X-ray crystallography but also to determine how wide a range of possibilities could be found and to examine how well various measures of the intermolecular interaction were able to discriminate between them.

MTX–DHFR

The coordinates of *E. coli* DHFR[27] were taken from the file 4DFR in the protein databank.[28] This contains the crystal structure of the enzyme without the NADPH cofactor, the larger site providing a more complete test of the algorithm. Of the two protein molecules in the asymmetric unit molecule, "B" was chosen as it has a more complete chain and is less perturbed by intermolecular contacts.[29] All waters were removed with the exception of HOH639, which has a very low temperature factor and appears to play an important role in the stabilization of certain ligands. Analogous computations without this water present produced very similar results. The receptor site was defined using the residues in Table I. The molecular surface of the site was calculated using the MS program, keeping the rest of the molecule for surface checks. The SPHGEN program produced 8 clusters of which the two largest, containing 90 and 12 spheres respectively, when merged into a single cluster gave a good representation of the entire site (covering both the MTX and NADPH binding regions). Hydrogen-bonding site points were added as described above, giving a grand total of 1222 site points. Of the 1120 hydrogen-bonding site points, 459 were points that could match a heavy donor atom, 202 were points that could match a heavy acceptor atom, 71 were points associated with an atom able to act as either donor or acceptor, and 388 were points that could match a donor hydro-

Table I. Residues used to define the active site of DHFR.

ILE 5	ALA 6	ALA 7	ILE 14	GLY 15	MET 16	GLU 17	ASN 18
ALA 19	MET 20	PRO 21	TRP 22	ASN 23	LEU 24	PRO 25	ALA 26
ASP 27	LEU 28	ALA 29	TRP 30	PHE 31	LYS 32	GLY 43	ARG 44
HIS 45	THR 46	TRP 47	GLU 48	SER 49	ILE 50	GLY 51	ARG 52
PRO 53	LEU 54	PRO 55	GLY 56	ARG 57	GLN 65	ILE 94	GLY 95
GLY 96	GLY 97	ARG 98	VAL 99	TYR 100	GLU 101	GLN 102	ILE 115
ASP 122	THR 123	HIS 124	HOH 639				

gen atom. Hydrogen atoms were added to all protein atoms with the exception of —OH and —SH groups, for which it is difficult to determine a unique position. The pteridine ring system of MTX (including all hydrogen atoms and protonated at N1[30]) was chosen as the "rigid" portion of the ligand; this exists in a single conformation and has five potential hydrogen-bonding (donor) atoms (Fig. 5). Three different intermolecular VDW cutoffs were used: 2.5 Å for interactions between two heavy (i.e., nonhydrogen) atoms, 1.75 Å for interactions between one heavy and one hydrogen atom, and 1.5 Å for interactions between two hydrogen atoms.

A total of 1530 orientations of the pteridine ring system were produced by the Directed DOCK algorithm (CPU time ~4 min on the Computer Graphics Laboratory's MIPS RC6280). Each matching set was required to contain at least one specific hydrogen-bonding match. The rms deviations of the orientations to that of the pteridine ring in the crystallographically determined structure ranged from 0.5–19.3 Å. The orientations were first pruned by eliminating those making fewer than three hydrogen bonds with the site. The 593 that remained had contact scores between 10–120 and electrostatic energies between −38 and +22 kcal/mol. The electrostatic interaction energies were calculated using the grid-based approach[15] with atomic charges derived by electrostatic potential fitting[31] to an *ab initio* quantum mechanical calculation (UCSF/Gaussian-80[32]) at the STO-3G level on the pteridine ring system. The orientations were pruned further by requiring each to have a contact score greater than 60 and an electrostatic interaction energy less than −15 kcal/mol to be acceptable. These values were chosen so that approximately 50% of the orientations had better values. The 239 orientations that satisfied all three criteria were then clustered. A total of 16 families were obtained, all of which were smaller than 1 Å. The members of each family were ranked according to both contact score and electrostatic energy and the orientation with the best combined rank chosen as the representative of that family. The two scores cannot be simply added together as they are on different scales. The rms differences of these 16 orientations from the crystal orientation of the pteridine ring system are given in Table II. A significant proportion of the orientations are "close" to the crystal structure (within 1.5 Å or so) but there are, however, orientations that are significantly different.

MTX has 11 rotatable acyclic bonds; the values that each bond was permitted to adopt in the con-

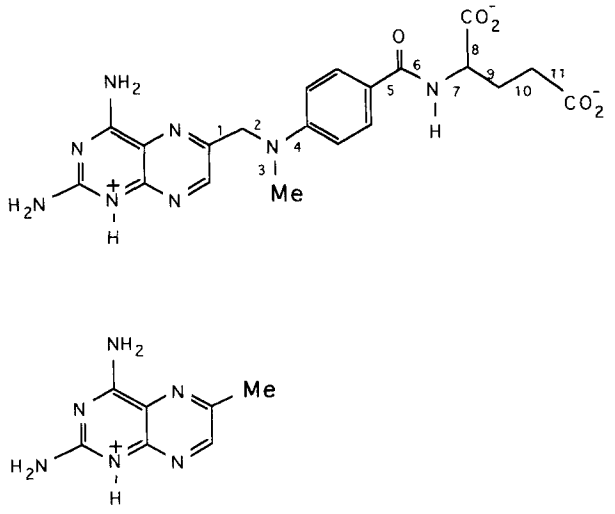

Figure 5. Methotrexate. The pteridine ring system is chosen as the anchor fragment. Table III gives the dihedral values permitted to each of the rotatable bonds during the conformational search.

Table II. Summary of orientations of pteridine ring system obtained from the Directed DOCK program.

Number	rms with crystal orientation (Å)	Number of MTX conformations
1	1.0	716
2	6.8	18
3	14.3	516
4	0.5	858
5	0.8	453
6	1.1	737
7	1.8	123
8	4.0	0
9	10.3	98
10	1.6	288
11	1.4	0
12	1.6	331
13	2.1	713
14	2.2	111
15	1.4	1078
16	1.9	755

Table III. Dihedral values used for the rotatable bonds of MTX.

Torsion number (see Fig. 5)	Permitted values (°)
1	45, 90, 135, 225, 270, 315
2	45, 90, 135, 225, 270, 315
3	120
4	0
5	0, 180
6	0, 180
7	60, 170, 215, 260, 305
8	0, 90
9	60, 180, 300
10	60, 180, 300
11	0, 90

formational search are shown in Table III. The total number of configurations to be explored by the search was nearly 26,000 for each of the 14 orientations of the pteridine ring system. The "flexible" part of the ligand has six potential hydrogen-bonding atoms (five acceptor oxygens and one donor hydrogen). The conformational search yielded 6795 structures of the MTX/DHFR complex (requiring approximately 20 min CPU time). The ligand orientations had rms values between 0.5–14.0 Å with the crystallographically determined structure of MTX. The initial set was then pruned and clustered to provide a maximum of 20 structures for molecular mechanics refinement. First, structures making fewer than five hydrogen bonds with the site were eliminated. The contact scores of the remaining conformations ranged from 78–257 while the electrostatic energies varied between −142 and −26 kcal/mol. A contact score greater than 160 and an electrostatic energy less than −55 kcal/mol were the criteria for a conformation to be accepted for clustering. Approximately 15% of the orientations had scores better than each of these criteria. The representative structure for each family was chosen to be the one with the best combined ranking for both the contact score and the electrostatic energy. The 20 structures obtained after clustering had the rms fits with the crystal orientation shown in Table IV. The pteridine orientation from which each of these MTX conformations is derived is also shown in Table IV. Each of the 20 structures was minimized using AMBER[33] (keeping the protein rigid while permitting the ligand complete conformational, translational, and rotational flexibility with an 8 Å cutoff and a distance-dependent dielectric); the final AMBER energies of each complex structure are shown in Table IV.

The conformations of MTX fall into two broad families. In all of these, the pteridine ring system occupies approximately the same part of the site as in the crystallographically determined structure. There are then two "grooves" where the remainder of the ligand can be placed; one of these corresponds to the crystal orientation, giving structures with rms differences less than 3 Å. The remaining structures have rms deviations around 10 Å with the flexible portion of MTX filling the region usually occupied by the cofactor. Of the final family of 20 structures, conformation 4 (Table IV) is most similar to the crystal orientation of MTX and is drawn using the MIDAS molecular graphics display system[34] in Figure 6 with the hydrogen-bonding interactions it makes with the receptor. This orientation has a contact score of 177 and an electrostatic interaction energy of −138 kcal/mol. It has the best AMBER energy (−272.1 kcal/mol) of the 20 after minimization. (For comparison, the crystal orientation of MTX from the 4DFR file has a contact score of 157 and an electrostatic interaction energy of −104 kcal/mol and when minimized the final interaction energy was −270 kcal/mol.) An alternative structure (15 in Table IV) is shown in Figure 7; this differs considerably (10.3 Å rms) from the crystal orientation. As can be seen, this structure makes a number of hydrogen bonds with the receptor; it has a contact score of 183 and an electrostatic interaction energy of −97 kcal/mol. The AMBER energy after minimization (−200 kcal/mol) is, however, significantly higher than that of the crystal structure (−270 kcal/mol). Indeed, all of the structures that are "close" to the crystal give the lowest-energy structures on minimization. For this system, both the grid-based electrostatic interaction function and the AMBER energy appear to be good discriminators between the "correct" (i.e., crystallographically determined) and "incorrect" structures. However, the picture may be somewhat less clear-cut than it would appear at first sight. An energy component analysis of the minimized crystal structure indicated that interactions between the terminal carboxylate and ARG57 and between the glutamate carboxylate group and ARG52 make very substantial contributions to the interaction energy of the complex. These two carboxylate groups lie close to the surface of the protein. As such, solvent effects may be important and a molecular mechanics energy only provides a partial picture.

Netropsin–d(CGCGATATCGCG)

A slightly different approach was used to investigate the binding of netropsin (Fig. 8) to the DNA duplex d(CGCGATATCGCG) through the use of a model structure of the receptor (generated with the NUCGEN program in AMBER). In this way, we hoped to minimize the effects of structural bias present in the crystal structure[35] and determine whether it would be possible to both reproduce the experimental result in an *ab initio* modeling exper-

Table IV. Summary of the MTX/DHFR structures.

Number	rms with crystal orientation (Å)	Complex energy of minimized complex (kcal/mol)	rms of minimized structures with crystal (Å)	rms of initial and minimized structures (Å)	Contact score	Grid-based electrostatic interaction energy (kcal/mol)	Number of pteridine orientation from which derived
1	1.2	−252.7	0.8	1.4	173	−99	1
2	9.4	−156.8	9.1	0.7	176	−60	4
3	1.5	−216.1	1.4	0.5	168	−64	4
4	1.1	−272.1	0.9	0.6	177	−138	5
5	9.3	−190.1	9.7	1.2	205	−77	6
6	9.6	−190.8	9.7	1.1	180	−89	6
7	1.3	−202.9	1.2	0.7	208	−105	6
8	1.5	−218.4	1.5	0.8	167	−123	6
9	9.3	−162.0	9.2	1.1	257	−78	7
10	10.9	−191.6	10.7	0.7	162	−77	12
11	10.6	−181.6	10.6	0.9	209	−78	13
12	2.7	−235.1	2.5	0.9	173	−123	13
13	2.6	−224.0	2.7	0.9	173	−110	13
14	2.9	−167.8	2.9	0.9	163	−57	13
15	10.3	−199.7	10.8	1.2	183	−97	15
16	10.8	−176.7	11.0	1.1	176	−69	16
17	10.7	−196.8	10.5	1.0	192	−86	16
18	10.5	−186.7	11.0	1.3	201	−87	16
19	10.8	−168.2	11.2	1.3	191	−80	16
20	10.5	−147.4	10.5	1.0	194	−68	16

The complex energy is the sum of the internal energy of the ligand and the interaction energy of the ligand with the protein.

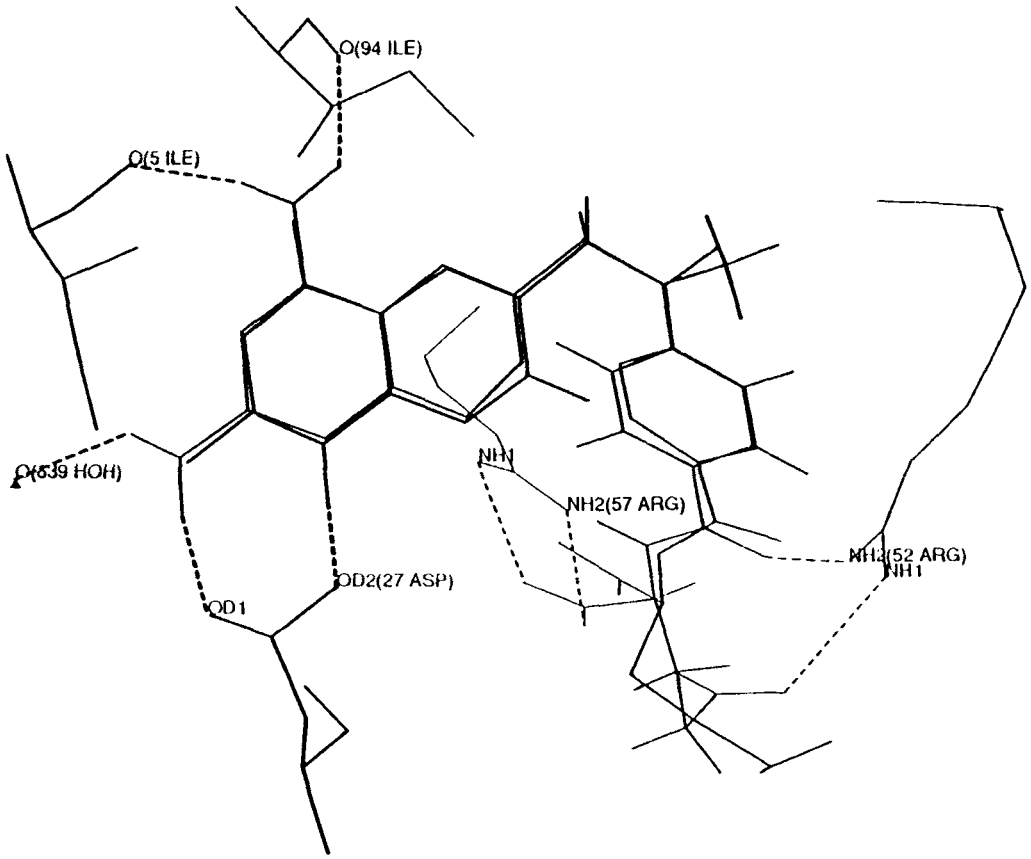

Figure 6. Structure of MTX (from the final family of 20; 4 in Table IV) that is closest to the crystal orientation. The crystal orientation is also shown and can be identified as the one without hydrogen atoms.

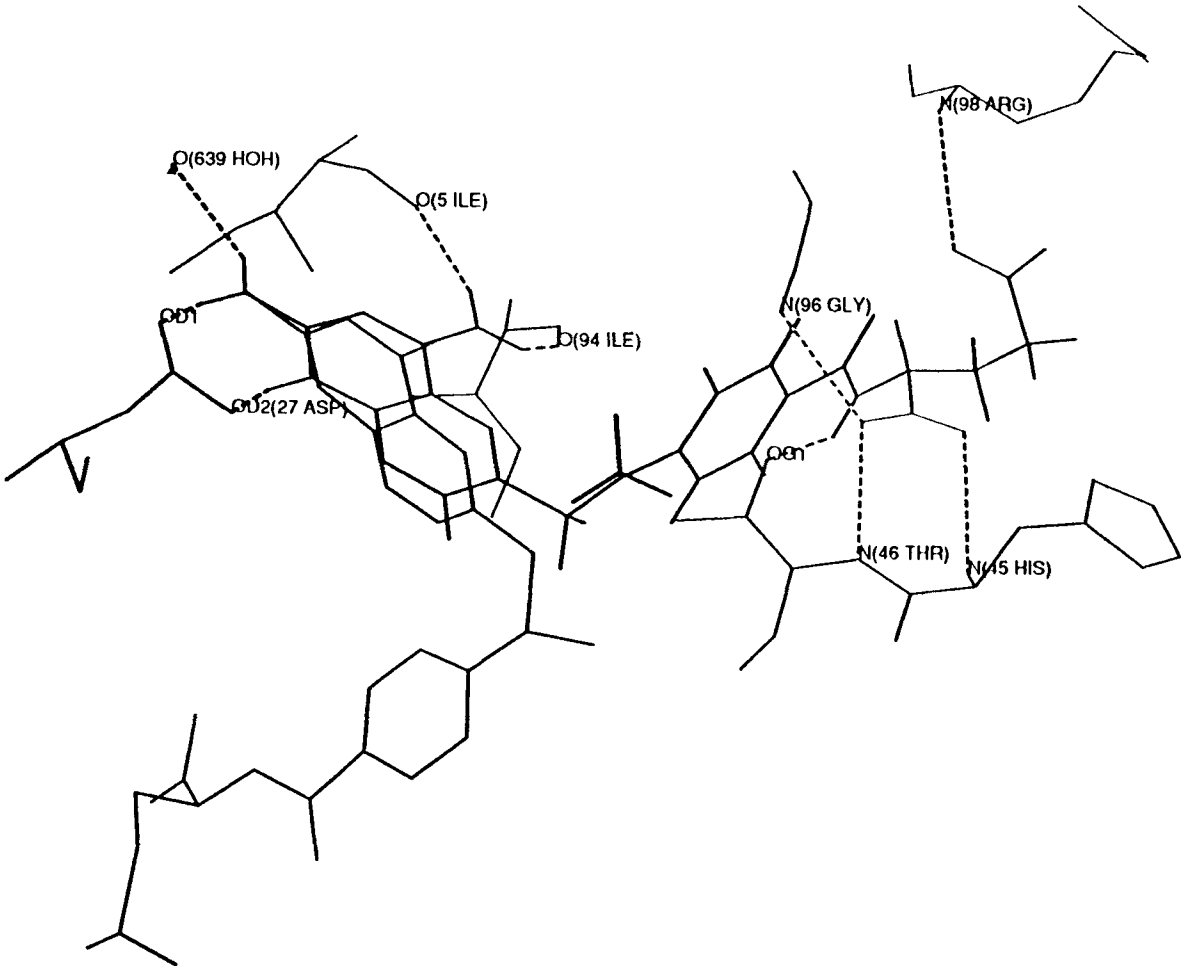

Figure 7. Alternative MTX conformation (15 in Table IV), shown with the crystal orientation.

iment and investigate the possibility of alternatives. The largest sphere cluster produced by SPHGEN (containing 86 spheres) filled the minor groove of the DNA, to which were added 1402 hydrogen-bonding site points (52 points where a ligand acceptor could be placed and 1350 points to which a donor atom could be matched). The anchor fragment of netropsin was defined to include the two pyrrole rings and the adjoining amide groups (Fig. 8). This forms an extensive delocalized system which to a first approximation can be treated as planar. More than one low-energy conformation should be considered for this fragment; a conformational search followed by molecular mechanics minimization using SYBYL[25] produced the eight low-energy conformations shown in Figure 9 (hereafter referred to as fragments 1–8). Each fragment has six hydrogen-bonding atoms: three amide hydrogens and three carbonyl oxygens. Orientations of these fragments in the minor groove of the duplex were generated using the Directed DOCK procedure; as for MTX, at least one hydrogen-bonding pair was required for each matching set. The numbers of orientations obtained for each fragment are shown in Table V. Each set of orientations was then pruned and clustered. In all cases, an orientation was required to form at least one hydrogen bond with the receptor

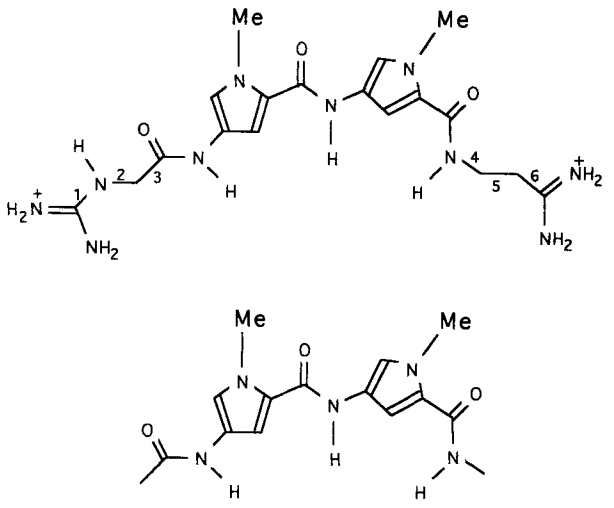

Figure 8. Netropsin and its anchor fragment. The dihedral values permitted to each of the bonds are indicated in Table VI.

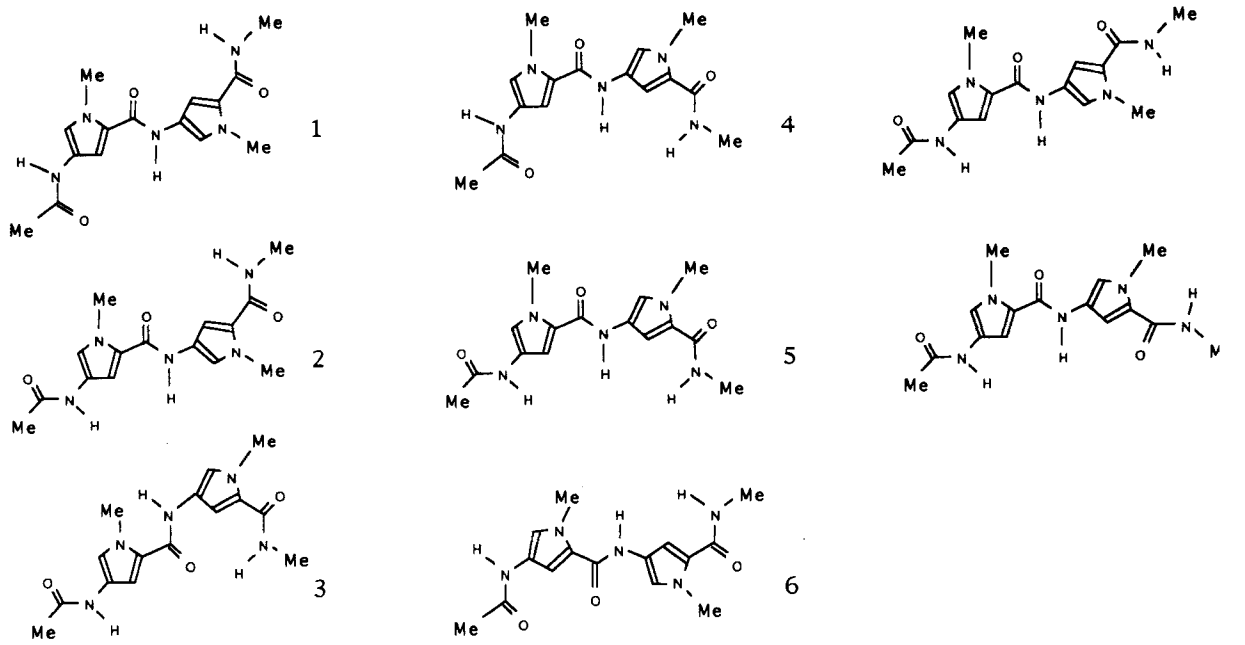

Figure 9. Eight conformations of the anchor fragment of netropsin used in this study.

Table V. Summary of orientations of the netropsin fragment obtained from the Directed DOCK algorithm.

Fragment number	Number of initial orientations	Final number of structures after pruning and clustering	Range of contact scores for final set of orientations	Range of electrostatic energies (kcal/mol)
1	4	0	—	—
2	32	3	134-157	−0.9--0.3
3	61	10	127-169	0.0-7.1
4	37	1	131	3.2
5	102	11	123-200	−2.4-3.3
6	41	0	—	—
7	3	0	—	—
8	83	2	122-135	1.5-5.9

and to have a contact score greater than 120. The electrostatic interaction energy was not used in this case due to the narrow range of values. The numbers of orientations remaining for each fragment after pruning and clustering are shown in Table V. These orientations were used as the basis for the conformational search. The dihedral values permitted to each rotatable bond were as shown in Table VI. The total numbers of structures of the netropsin-duplex complex for each of the fragments are given in Table VII.

The netropsin structures were finally pruned and clustered. A maximum of 20 clusters were determined for each of the eight sets with the criteria for acceptance being four hydrogen bonds, a contact score greater than 200, and an electrostatic energy better than −110 kcal/mol. As before, the clustering was terminated when all the clusters were smaller than 1 Å and the representative for each family was chosen to be the one with the best combined rank for both contact score and electrostatic energy. The numbers of structures obtained for each fragment are shown in Table VII. These structures were minimized with AMBER using the same protocol as for MTX (i.e., rigid receptor; distance-dependent dielectric, 8 Å nonbonded cutoff). "Hydrated" counterions (R^* = 5.0 Å, ϵ^* = 0.1) were placed near each phosphate group as

Table VI. Dihedral values used for netropsin.

Torsion number (see Fig. 8)	Permitted values (°)
1	0
2	0, 90, 180, 270
3	75, 180, 285
4	75, 180, 285
5	60, 180, 300
6	0, 90, 180, 270

Table VII. Summary of the netropsin/d(CGCGATATCGCG) structures.

Number	Number of netropsin conformations	Number of conformations after pruning and clustering	Range of contact scores of pruned set	Range of electrostatic energies	Range of complex energies after minimization (AMBER; kcal/mol)
1	0	0	—	—	—
2	1185	7	200–211	−117–−134	−172.4–−142.3
3	834	20	202–243	−113–−142	−170.8–−136.5
4	0	0	—	—	—
5	786	20	201–285	−118–−153	−171.4–−140.4
6	0	0	—	—	—
7	0	0	—	—	—
8	912	2	204–210	−128–−135	−144.3–−142.6

these have been shown to give better results for *in vacuo* studies of DNA.[36] The results of the minimizations are summarized in Table VII.

The crystallographically determined structure of netropsin bound to this particular duplex has been determined to 2.4 Å resolution (R = 20.0%).[35] Of some interest and relevance to this discussion is the fact that the drug binds in two orientations in the minor groove. A number of observations can be made concerning the results of our modeling experiment. First, structures corresponding to those experimentally determined are generated. Of these two structures, one has the second highest contact score (277) and the third lowest electrostatic interaction energy (−144 kcal/mol) of all the complex structures generated. The other has a somewhat lower contact score (246) but does have the best electrostatic interaction energy (−153 kcal/mol). It is, however, interesting to note that quite reasonable structures (according to hydrogen bonding, contact score, grid-based electrostatic interaction energy, and AMBER energy) can be obtained for conformations that differ considerably from the "correct" ones, particularly those derived from fragments 2 and 3.

Fragment 2 is able to fit into the groove via hydrogen-bonding interactions involving the two unidirectional amide hydrogens. The third amide group points away from the DNA with the pyrrole methyl group fitting snugly into the groove. A large number of conformations based on the three orientations of this fragment were produced, in all of which the positively charged amidinium moiety protrudes out of the groove. Although many of these structures are rejected, the seven complex structures obtained after clustering from this initial fragment do contain the ones with the lowest AMBER energies after minimization. This is largely due to electrostatic interactions between the amidinium group and the phosphate/sugar backbone. The use of a rigid receptor structure and the lack of solvent may all contribute to this seemingly anomalous result, but they do serve to emphasize the danger in the use of purely enthalpic measures. It is noteworthy that the contact scores of all the conformations of fragment 2 are only slightly higher than the threshold value and much lower than those for the other fragments. This is because a significant portion of the ligand lies outside the groove and is thus unable to make as many nonbonded contacts.

Surprisingly, a number of seemingly reasonable conformations (on the basis of their AMBER energies) were obtained from fragment 3, which can only be accommodated if the central carbonyl oxygen is positioned deep inside the groove. Although the electrostatic interaction energies of the orientations of this fragment are slightly higher than those of fragments 2 and 5 (Table V), the values are not sufficiently worse for a clear distinction to be made at this stage. This may to be due to the fact that the electrostatic potential within the groove is almost uniformly negative, which means that there is little difference between structures in which the oxygen atom points "into" and "out of" the groove. The subsequent conformational search yields a number of structures. Figure 10 shows one of these, which has an AMBER energy of −159 kcal/mol. The contact score (234) and the electrostatic energy (−135 kcal/mol) are quite respectable, as are the hydrogen-bonding interactions made with the receptor. In this orientation, the netropsin binds to the interface between the GC and AT base pairs. Most of the structures of the netropsin–duplex complex did, however, bind to the central ATAT region of the duplex; netropsin has a well-documented preference for AT-rich regions of DNA.[37] This has been ascribed to a number of factors, not least of which is the narrowing of the minor groove in such regions, which enhances the VDW interaction between the drug and the duplex. In addition, the 2-amino functionality of guanine that protrudes into the groove may impose steric restrictions on the binding of such molecules to GC regions. However, with a very different netropsin conformation it does appear plausible that binding could occur in these regions.

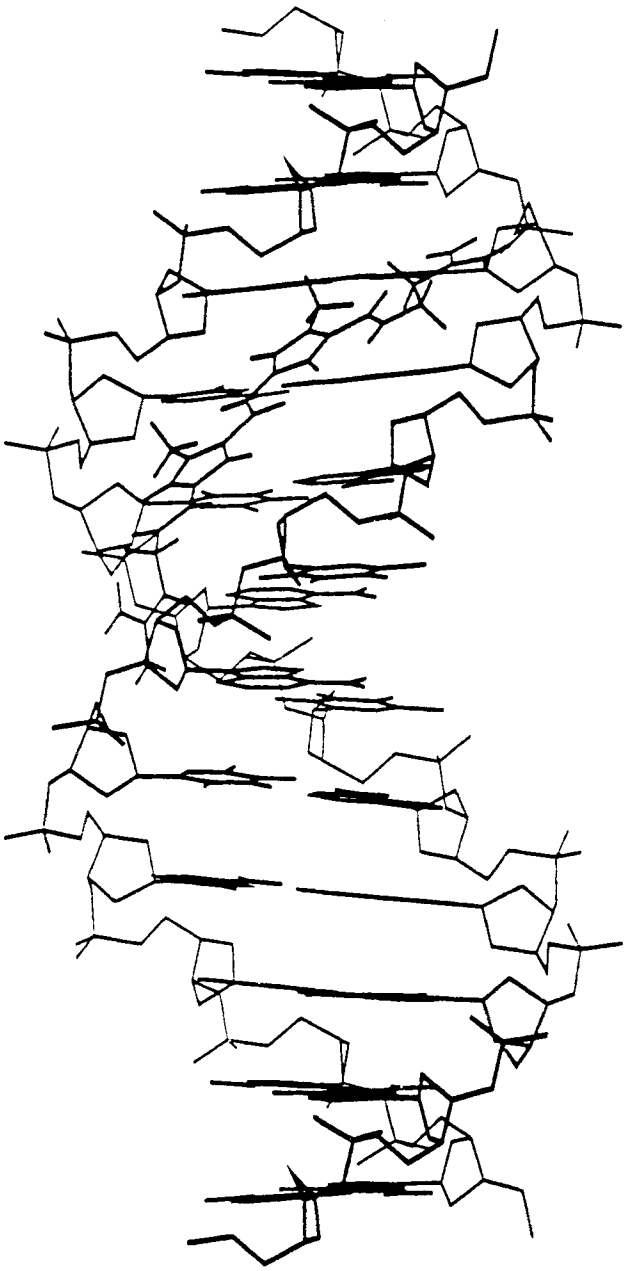

Figure 10. Netropsin structure derived from fragment 3. The ligand binds to the interface between the GC and AT regions with the central carbonyl oxygen pointing into the minor groove.

CONCLUSIONS

As the 3-dimensional structures of more biologically important receptors become available, there will be an increasing need to predict how ligands will bind to them. Even in the most favorable case when it is known how a ligand does bind (e.g., the substrate or a substrate analog) it does not necessarily follow that other ligands will bind in the same way.[38] Elucidating all the possible binding modes is difficult to perform manually and so theoretical methods will clearly be of some value. However, there remain many issues of concern to those who would develop such methods and those who would use them. These include the importance of giving full conformational flexibility to the ligand, the relevance of the conformational energy surface of the isolated molecule, how critical is the "rigid receptor" assumption, the role of solvent, and how best to assess any models that are obtained. At present, there is no clear evidence to indicate which of these represents the most serious limitation, partly because there are at present only a few structures of such complexes available.

The Directed DOCK procedure described in this article provides a method by which information

about the receptor site other than its shape characteristics can be incorporated into the DOCK method. Our initial efforts have been focused on the inclusion of information about hydrogen bonding because such interactions are frequently observed in ligand–receptor complexes and because they possess a directional quality that makes them more easily accommodated than other types of interactions. The Directed DOCK approach could be easily extended to cover similar types of interaction such as the formation of salt bridges or binding at a metal site. An alternative could be to incorporate regions of high potential from a program such as GRID[22] in the form of a cluster of points where a suitable ligand atom could be placed. The increased number of site points in the Directed DOCK algorithm does, however, make the method less attractive for applications involving large numbers of molecules, but the use of a suitable matching scheme might partly ameliorate this.

In contrast to the distance geometry and simulated annealing approaches, the method described here separates the orientational and conformational search problems. This requires the molecule to be divided into an "anchor fragment" and a "conformationally flexible portion." Ideally, the anchor fragment should be a part of the molecule that exists in a single rigid conformation or for which there are high-energy barriers between conformations. In some respects, the approach mirrors the common practice of drug development, where different functional groups are added to a single core (for example, a set of MTX variants all based on the pteridine ring system). The same set of DOCKed orientations of the anchor fragment could then be used for each molecule. Alternatively, should there be unambiguous evidence to indicate where a part of the ligand binds (e.g., at a metal ion) then the appropriate part of the ligand could be positioned in the correct geometry and the conformational space of the remainder of the molecule searched. By this means, a point is provided within the receptor from which the ligand can "grow," adapting its conformation to take account of the environment. It is this capability that gives the approach an advantage over docking multiple conformations of the ligand in which the ligand is unable to respond to the surrounding receptor site.[14]

In common with other approaches, our algorithm relies upon measures of the intermolecular interaction to select the correct answer from a large number of possibilities. At the present time, it is impractical to perform energy minimization (let alone molecular dynamics and then minimization, which would be preferable) on all the configurations that might be produced. Alternative approaches for measuring the interaction that are amenable to rapid calculation must therefore be devised. Such approaches are by definition approximate, and thus it is important that the criteria by which a structure is deemed acceptable or not are sufficiently lax, but not to the extent that they have no discriminatory power. Another potential difficulty when using such measures that is exemplified by the two examples discussed above is that the optimal measure may comprise a subset of those available. In both the MTX/DHFR and netropsin/duplex systems, the contact and the electrostatic measures were given equal weight. However, the results in Tables IV and VII imply that the electrostatic energy alone would be a better way to identify the correct answer for MTX/DHFR while the contact score would be preferred for netropsin/d(CGCGATATCGCG). Accurate evaluation of many of the factors that contribute to the binding of a ligand to a receptor are beyond the scope of even the most sophisticated calculations, but some of these can be addressed using approximate treatments as will be discussed elsewhere.[15,16]

The authors thank Robert Langridge, Elaine Meng, and Brian Shoichet for helpful discussions. A.R.L. thanks the Science and Engineering Research Council (UK) and NATO for a postdoctoral fellowship and the Computer Graphics Laboratory for their kind hospitality. The Computer Graphics Laboratory is supported by the National Institutes of Health (grant R1081; R. Langridge PI). Additional support was provided by the Defence Advanced Research Projects Agency under contract N00014-86-K-0757 administered by the Office of Naval Research and grants GM-39552 (G.L. Kenyon, PI) and GM-31497 (IDK). Molecular graphics images were produced using the MidasPlus software system from the Computer Graphics Laboratory, University of California, San Francisco. I.D.K, acknowledges the provision of the SYBYL software from Tripos Associates.

References

1. A.K. Ghose and G.M. Crippen, *J. Comp. Chem.*, **6,** 350 (1985).
2. D.S. Goodsell and A.J. Olson, *Proteins*, **8,** 195 (1990).
3. A.S. Smellie, G.M. Crippen, and W.G. Richards, *J. Chem. Inf. Comp. Sci.*, **31,** 386 (1991).
4. I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, and T.E. Ferrin, *J. Mol. Biol.*, **161,** 288 (1982).
5. R.L. DesJarlais, R.P. Sheridan, G.L. Seibel, J.S. Dixon, and I.D. Kuntz, *J. Med. Chem.*, **31,** 722 (1988).
6. D.P. Dolata, A.R. Leach, and K. Prout, *J. Comp-Aided Mol. Design*, **1,** 73 (1987).
7. A.R. Leach and K. Prout, *J. Comp. Chem.*, **11,** 1193 (1990).
8. R.L. DesJarlais, R.P. Sheridan, J.S. Dixon, I.D. Kuntz, and R. Venkatarghavan, *J. Med. Chem.*, **29,** 2149 (1986).
9. R. Langridge, T.E. Ferrin, I.D. Kuntz, and M.L. Connolly, *Science*, **211,** 661 (1981).
10. M.L. Connolly, *J. Appl. Crystallogr.*, **16,** 548 (1983).
11. M.L. Connolly, *Science (Washington, DC)*, **221,** 709 (1983).

12. B.K. Shoichet and I.D. Kuntz, *J. Mol. Biol.*, **221,** 327 (1991).
13. B.K. Shiochet, D. Bodian, and I.D. Kuntz, *J. Comp. Chem.*, in press.
14. R.L. DesJarlais, PhD thesis, University of California, San Francisco, CA, 1987.
15. E.C. Meng, B.K. Shoichet, and I.D. Kuntz, *J. Comp. Chem.*, in press.
16. B.K. Shoichet, A.R. Leach, and I.D. Kuntz, in preparation.
17. P. Murray-Rust and J.P. Glusker, *J. Am. Chem. Soc.*, **106,** 1018 (1984).
18. E.N. Baker and R.E. Hubbart, *Prog. Biophys. Mol. Biol.*, **44,** 97 (1984).
19. D.R. Ferro and J. Hermans, *Acta Cryst.*, **33,** 345 (1977).
20. S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, 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. Comp. Chem.*, **7,** 230 (1986).
22. P.J. Goodford, *J. Med. Chem.*, **28,** 849 (1985).
23. R.D. Cramer, *J. Am. Chem. Soc.*, **110,** 5959 (1988).
24. J.A. Hartigan, *Clustering Algorithms*, chap. 3, Wiley, New York, 1975.
25. Sybyl, v.5.3, Tripos Associates Inc., St. Louis, MO, 1991.
26. A.R. Leach, K. Prout, and D.P. Dolata, *J. Comp. Chem.*, **11,** 680 (1990).
27. J.T. Bolin, D.J. Filman, D.A. Matthews, R.C. Hamlin, and J. Kraut, *J. Biol. Chem.*, **257,** 13650 (1982).
28. F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112,** 535 (1977).
29. Comment in the PDB file 4DFR.
30. K. Hood and G.C.K. Roberts, *Biochem.*, **171,** 357 (1978).
31. U.C. Singh and P.A. Kollman, *J. Comp. Chem.*, **5,** 129 (1984).
32. U.C. Singh and P.A. Kollman, UCSF Gaussian-80, *QCPE Bull.*, 2, 17 (1982); program 446.
33. D.A. Pearlman, D.A. Case, J.C. Caldwell, G.L. Seibel, U.C. Singh, P. Weiner, and P.A. Kollman, Amber 4.0 (UCSF), University of California, San Francisco, CA, 1990.
34. T.E. Ferrin, C.C. Huang, L.E. Jarvis, and R. Langridge, *J. Mol. Graph.*, **6,** 13 (1988).
35. M. Coll, J. Aymami, G.A. can der Marel, J.H. van Boom, A. Rich, and A. H.-J. Wang, *Biochemistry*, **28,** 310 (1989).
36. U.C. Singh, S.J. Weiner, and P.A. Kollman, *Proc. Natl. Acad. Sci. USA*, **82,** 755 (1985).
37. C. Zimmer and U. Wahnert, *Prog. Biophys. Mol. Biol.*, **47,** 31 (9186).
38. D. Ringe, *Nature*, **351,** 185 (1991).