J. Mol. Biol. (1982) **161**, 269–288

A Geometric Approach to Macromolecule–Ligand Interactions

IRWIN D. KUNTZ, JEFFREY M. BLANEY, STUART J. OATLEY
ROBERT LANGRIDGE AND THOMAS E. FERRIN

Department of Pharmaceutical Chemistry
School of Pharmacy, University of California
San Francisco, Calif. 94143, U.S.A.

(Received 4 January 1982, and in revised form 18 May 1982)

We describe a method to explore geometrically feasible alignments of ligands and receptors of known structure. Algorithms are presented that examine many binding geometries and evaluate them in terms of steric overlap. The procedure uses specific molecular conformations. A method is included for finding putative binding sites on a macromolecular surface.

Results are reported for two systems: the heme–myoglobin interaction and the binding of thyroid hormone analogs to prealbumin. In each case the program finds structures within 1 Å of the X-ray results and also finds distinctly different geometries that provide good steric fits. The approach seems well-suited for generating starting conformations for energy refinement programs and interactive computer graphics routines.

1. Introduction

To position two molecules so that they interact favorably with one another is a problem of general interest to chemists and biochemists. It is a problem of considerable difficulty for molecules of any complexity, because of the large number of internal degrees of freedom and the attendant local minima in the molecular conformation space.

Our approach is to reduce the number of degrees of freedom using simplifying assumptions that still retain some correspondence to a situation of biochemical interest. Specifically, we treat the geometric (hard sphere) interactions of two rigid bodies, where one body (the "receptor") contains "pockets" or "grooves" that form binding sites for the second object, which we will call the "ligand". Our goal is to fix the six degrees of freedom (3 translations and 3 orientations) that determine the best relative positions of the two objects.

Earlier studies of molecular docking (Wodak & Janin, 1978; Greer & Bush, 1978; Levinthal *et al.*, 1975; Salemme, 1976) made use of approximate potential functions for the intermolecular interactions and grid searches to fix the degrees of freedom. Here we use a very simple interaction function containing only two terms: hard sphere repulsions and "hydrogen bonding". A zero value for this function will correspond to a docking geometry having: (1) no hard sphere overlaps between

0022-2836/82/300269-20 $03.00/0 © 1982 Academic Press Inc. (London) Ltd.

receptor and ligand atoms; (2) all hydrogen bonding atoms of the ligand with a nitrogen or oxygen atom of the receptor within 3·5 Å; and (3) all ligand atoms within the receptor binding site.

We expect, and find, that more than one receptor–ligand geometry can provide low values of the interaction function. Thus, we extend the goals of the calculation to produce representatives of *all* geometrically feasible solutions.

The degrees of freedom are fixed using distance-comparison techniques (Lesk, 1979) derived from distance geometry methods (Crippen, 1981; Kuntz *et al.*, 1979) and optimization procedures. The method also uses molecular surface calculations (Richards, 1977; Connolly, 1981).

We should emphasize that the assumption of rigid molecules is a severe restriction, since it requires prior knowledge of the two structures to atomic detail. Hence, these procedures should not be construed as a general solution to the docking problem. The methods described may be extended to include some internal degrees of freedom at a later date. We do not view the calculated structures as ends in themselves; rather, we see them as reasonable starting points for molecular mechanics (Weiner & Kollman, 1981; Potenzone *et al.*, 1977; Hagler *et al.*, 1974; Momany *et al.*, 1974) and for interactive computer graphics studies (Langridge *et al.*, 1981).

We also note that some of the algorithms presented here have interesting independent applications such as: location and characterization of macromolecular binding sites, conformational comparison among a series of compounds and design of new ligands.

2. Methods

(a) *Approach*

We divide the problem into three parts.

(1) Representation of the receptor and ligand structures, a process that includes identification of the possible binding sites on the receptor molecule.
(2) Matching of the receptor and ligand representations.
(3) Optimization of ligand position within the binding site.

We outline the overall procedure and then describe the details.

The representation program generates a set of spheres that fill all pockets and grooves on the surface of the receptor molecule. These spheres are collected into a number of presumptive "binding" sites; each site can then be examined independently for geometric matching with the ligand. The ligand molecule is also represented by a set of spheres that approximately fill the space occupied by the ligand. If the ligand provides a good match to the receptor site, the set of ligand spheres should, in some sense, fit within the set of receptor spheres. It is helpful to think of the "lock and key" analogy often used to describe enzyme–substrate interactions. The program produces a representation of the key (the ligand spheres) and a representation of the key-*hole* (the receptor spheres).

The pairing rule is based on a comparison of internal distances in both ligand and receptor. A ligand sphere can be paired with a receptor sphere if *each* sphere belongs to a set of spheres with the following property: the internal distances of *all* the spheres in the ligand set must match all the internal distances within the receptor set, within some error limit on each distance. This rule allows the identification of geometrically similar clusters of spheres in the receptor site and in the ligand, without requiring explicit rotations of one structure onto the other.

The final stage in the program explores the suggested pairings. It carries out the rotation of the ligand spheres onto the corresponding receptor spheres using the least-squares algorithm of Ferro & Hermans (1977). The rotation/translation matrix generated from the spheres is then applied to all ligand *atoms*. The optimization procedure manipulates the ligand atom co-ordinates to reduce atom overlaps and ensure hydrogen-bonding partners. It also makes an effort to place ligand *atoms* within receptor *spheres*. The output of this part of the program is a set of ligand structures expressed in the receptor molecule co-ordinate system. Each structure has a score based on the degree of overlap. Geometrically similar structures are grouped together for display on a computer graphics system or for further processing. The entire procedure takes hours of computer time on a minicomputer such as a PDP 11/70 (Digital Equipment Corporation).

(b) *Detailed algorithms*

(i) *Representation*

The guiding principle is to represent the ligand and the receptor site so that the two representations are *identical* in the limit of a "perfect" geometric fit. The usual atomic description of molecules can never achieve this goal since the ligand and receptor atoms are never coincident in their optimum geometry. The definition of molecular surface by Richards (1977) and a program to calculate this surface developed by Connolly (1981,1982) provide a useful starting point. In the limit of perfect fit, the ligand and receptor surfaces will be in exact correspondence within the receptor binding site.

A brief description of the molecular surface is required. The surface, as implemented by Connolly's program, is a collection of points and the vectors normal to the surface at each point. The points fall into two classes, following Richards: contact points that lie on the van der Waals' surface of the solvent-accessible atoms and re-entrant points that lie on the inward-facing surface of a "probe" sphere. As Richards has emphasized, the surface and volume of a collection of atoms can only be defined completely with reference to a probe object of some form. A spherical probe is most commonly used. A probe of zero radius yields the van der Waals' surface of a molecule, whose volume is the sum of the atomic volumes and whose appearance is very similar to the normal "space-filling" atomic models in common use. A probe of infinite radius generates a solid polyhedron called the "convex hull" of the collection of atoms. It is the smallest object with no concavities that contains all the atoms. For the work presented here, we used a probe sphere of radius 1·4 Å to approximate a water molecule. The results are not sensitive to the precise probe radius if the probe is small compared to the ligand molecule.

The surface calculation also requires specification of all atomic radii (Table 1). Since hydrogen co-ordinates are rarely available from X-ray diffraction experiments on macromolecules, we use "united atom" radii that are somewhat larger than the usual van der Waals' radii. A straightforward refinement would use explicit hydrogen atoms and smaller heavy-atom radii. Such a refinement would not alter the major results of this paper.

In principal, docking algorithms could be designed that use the molecular surfaces directly. The large number of points per surface cause several difficulties, so we developed a more compact representation.

For a surface composed of n surface points, we construct a set of spheres with the following properties.

(1) Each sphere touches the molecular surface at two points (i, j) and has its center on the surface normal from point i (Fig. 1).
(2) Each receptor sphere lies on the *outside* of the receptor surface.
(3) Each ligand sphere lies on the *inside* of the ligand surface.

The co-ordinates of the centers of the spheres are found analytically. Special cases involving surface normals lying along the principal axes or symmetry-related points must be treated separately.

TABLE 1

Atomic radii and overlap distances

A. *van der Waals' radii* (Å) (united atoms)

	Surface calculations	Refinement calculations
N	1·80	1·50
O	1·50	1·50
C	1·80	1·75
S	1·85	1·75
I	2·15	2·15
Br	1·95	1·95
Fe	1·50	1·50

B. *Overlap distances* (Å)

	N	O	C	S	I	Br
N	2·5	2·5	3·5	3·5	3·0	2·8
O		2·5	3·5	3·5	3·0	2·8
C			3·5	3·5	4·0	3·8
S				3·5	4·0	3·8
I					4·5	4·35
Br						4·25

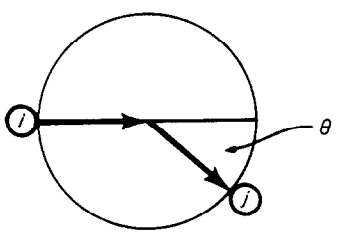


FIG. 1. Sphere generated tangent to surface points i, j, with center on surface normal at point i.

We reduce the number of spheres in several ways. First, while $n-1$ spheres can be constructed at each surface point, we are only concerned with the sphere of smallest radius, since all larger spheres at that point must intersect the molecular surface. Second, for each smallest sphere, we calculate the angle formed by the 2 surface points, i and j, and the sphere center (Fig. 1). We retain spheres with angles less than 90° since such spheres are more likely to span across an invagination than spheres with large angles, which tend to lie in shallow grooves. The exact choice of angular cut-off is not crucial. Two additional restrictions are applied to the receptor only. Receptor spheres are accepted only if points i and j belong to atoms in amino acid residues that are more than 4 apart in sequence. This eliminates the grooves formed by every alpha helix. We also eliminate receptor spheres with radii greater than 5 Å. Inspection showed that such spheres extend out of the "top" of the binding pockets. Finally, we reduced the data storage and computation requirement by retaining only one sphere per atom. Specifically, we choose the largest sphere formed from the *contact* surface points of each *receptor* atom and the largest sphere formed from the *re-entrant* surface points of each *ligand* atom. While more spheres could be retained, this set appears sufficient for our purposes.

The result of these manipulations is a set of spheres for each molecule of interest (Fig. 2). We finish with no more than one sphere per atom. Each is characterized by its center coordinates, radius and internal angle (Fig. 1).

At this point, the list of receptor spheres includes *all* the invaginations of the receptor surface. We separate the list into several possible binding sites with the rule: spheres belong to the same site if the spheres overlap. This procedure identifies a small number of sites scattered about the surface of a typical globular protein (Figs 3 and 4). In the proteins we have examined, the binding site is always the largest such feature. This approach should be useful in characterizing the binding sites of any macromolecule whose structure is available.

The entire computation for generating these spherical representations is relatively rapid: a few minutes of PDP 11/70 time for ligands of 40 to 50 atoms and approx. 30 min for small

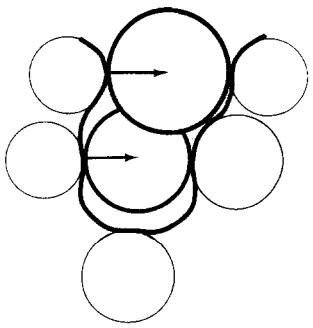


FIG. 2. Schematic representation of a small binding site formed from five atoms (thin circles). The molecular surface is shown as the thick line. The two receptor spheres (thick circles) are constructed as described in the text, with their centers lying along the surface normals (arrows).

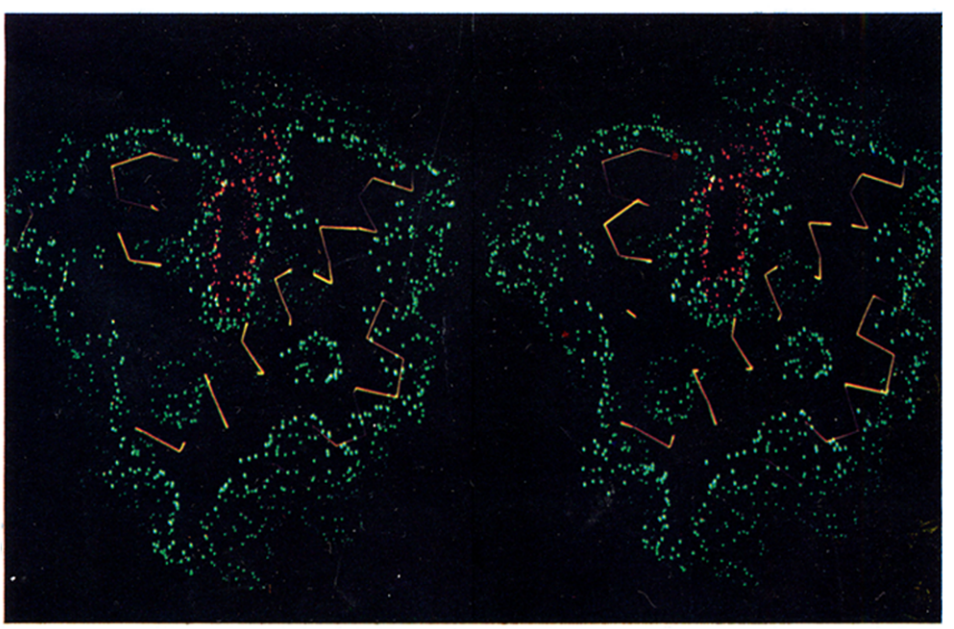


FIG. 3. Cross-section of the molecular surfaces for myoglobin (green) and heme (red). Part of the α-carbon chain is also shown. The most prominent feature is the E helix.

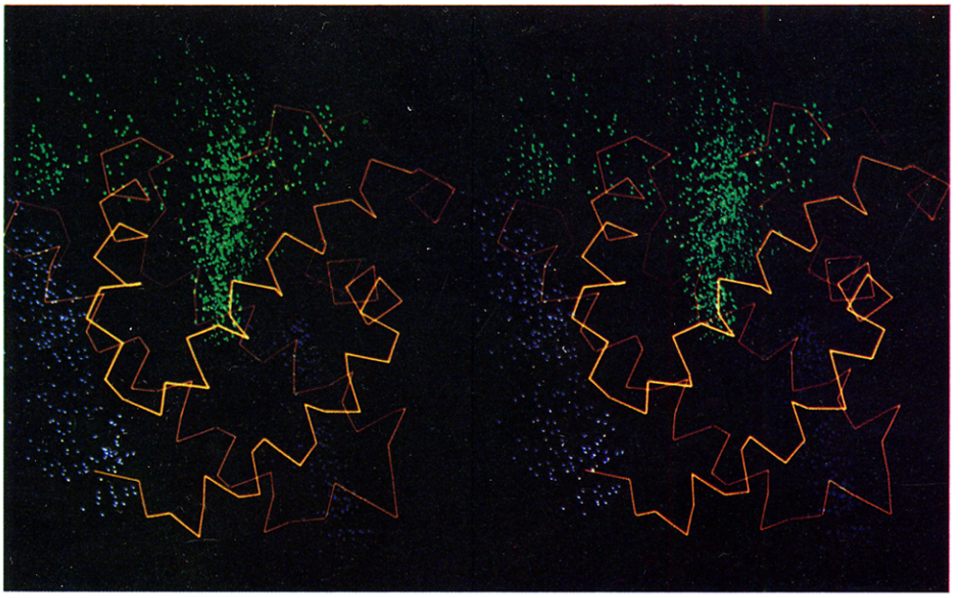


FIG. 4. Major pockets or invaginations in the myoglobin surface. The heme pocket (green) and 2 regions formed by helical surfaces (violet, blue) are discussed in the text.

globular proteins. The surface calculation, which must precede these efforts, takes 5 s/atom on the 11/70.

(ii) *Matching*

Given the set of spheres for a particular ligand and the set of spheres for one of the binding sites of the receptor, we must next find some method of alignment of the two representations. The major hurdle is that one cannot rely on having information that pairs specific ligand and receptor spheres. The two general approaches to the problem are grid searches or combinatorial matching.

One cannot try all possible combinations. For m ligand spheres and n receptor spheres there are $n!/(n-m)!$ possibilities. For $n = 50$ and $m = 25$ there are choices of the order of 10^{39}. While the vast majority of such arrangements would be geometrically impossible, the number is too large for even the simplest test strategy. There is a large redundancy here, since fixing 4 specific pairs is sufficient to determine the rigid docking. Even so, for moderate-sized ligands the number of possibilities is large. We feel it is desirable to retain as thorough a search as is computationally feasible. Our procedure is as follows: systematically pair each ligand sphere, i, with each receptor sphere, k. Consider the set of distances from i to all other *ligand* spheres j, d_{ij}, and from k to all other receptor spheres l, d_{kl}. Assign a second pair ($j = l$) of spheres so that a maximum number of spheres obey the condition:

$$abs(d_{ij} - d_{kl}) < \varepsilon,$$

where ε is a parameter that specifies the allowed deviation between the ligand and receptor internal distances. We find that a value for ε between 1 and 2 Å works well for the molecules in this study.

Once the best second pair of ligand–receptor spheres has been identified, proceed to a third pair of spheres subject to the additional constraint that the distances from the new spheres to the previously assigned pairs must also obey the error check. This process continues until no further pairs can be assigned. If the number of assignable pairs is less than 4, the

orientation problem is underdetermined and the match is rejected. Otherwise, the match is retained for the refinement procedure. Lesk (1979) has described a somewhat similar approach that also makes use of internal distance comparisons.

Some caveats should be clearly stated. The matching procedure outlined here does not guarantee a globally optimum solution. The process can require of the order of n^4 steps to run to completion. For the work here we have reduced the computation time by only accepting ligand and receptor spheres whose centers are approximately the same distance from the centroid of each respective representation. The allowed deviations were typically 1 to 2 Å. This restriction is equivalent to the assumption that the ligand fills the receptor pocket but does not extend much outside it. This turns out to be a good approximation for the systems studied in this paper but could not be used for a small ligand in a large pocket or a large ligand that extends far outside the receptor site.

With the assumptions and parameters given above, the matching algorithm ran to completion in a few hours of computer time†. Typically, several hundred assignment lists were generated that met the criteria listed above.

The output of this stage of the program consists of short lists of pairs of ligand and receptor sphere numbers. These collections of spheres have been generated to have *all* internal distances matched within ε. The lists require further processing for 3 reasons.

(1) Internal distances retain an ambiguity in handedness that must be resolved.
(2) The *unmatched* ligand spheres (those *not* on a particular list) may be forced to lie in totally unacceptable locations to achieve the geometric requirements of the list assignments.
(3) We need to establish a co-ordinate transformation to locate the ligand with respect to the receptor.

These questions are resolved in the final stage of the program.

(iii) *Optimization routines*

The first task is to convert the list assignments into Cartesian co-ordinates. Then the ligand is moved somewhat to improve its "fit" according to the criteria mentioned earlier: minimization of overlap, hydrogen-bond pairing, and positioning of the ligand into the receptor binding site.

The conversion into co-ordinates is straightforward. The ligand spheres of each list are translated and rotated onto their receptor sphere partners using the least-squares algorithm ORIENT of Ferro & Hermans (1977), which has proved robust and efficient for our purposes. The ORIENT routine returns a measure of how closely the 2 lists of spheres correspond to each other. If the correspondence is poor (that is, a least-squares error greater than 3 Å) the program reverses the handedness of the set of ligand spheres and ORIENT is called again. If this error is also large, the particular assignment set is rejected and the next one is tried.

If the least-squares error from ORIENT is acceptable, the translation/rotation matrix from the procedure is applied to all ligand atoms. This process yields Cartesian co-ordinates for the ligand atoms referenced to the centroid of the receptor sphere cluster.

The next step is to optimize the placement of the ligand. There are many options one could explore, including energy minimization and interactive computer graphics manipulation. We devised a few simple procedures to sort rapidly among the large number of possible structures generated in the matching routine. Our goal was not a full-scale optimization but rather a weeding out of implausible structures.

We first compute an overlap error between all ligand *atoms* and all receptor *atoms*.

$$E_{\text{overlap}} = \sum r_i + r_k - d_{ik}, \tag{1}$$

† The equivalent grid search must scan 3 orientation angles with approx. 100 steps per angle and 3 translational motions that can be restricted, by the centroid assumption, to approx. 10 steps per axis. This grid contains 10^9 points, each point requiring calculations of approx. 10^3 distances to check for overlaps, too large a calculation for the minicomputer.

where r_i, r_k are the van der Waals' radii of the ligand and receptor atoms (Table 1) and d_{ik} is the interatomic distance. The sum is only taken for positive values. The overlap error is roughly proportional to a van der Waals' repulsion with a proportionality constant of 0·1 to convert into kcal/mol for non-bonded methyl groups.

Next, we attempt to improve E_{overlap} by moving the ligand further into the receptor site as follows. The receptor *sphere* closest to each ligand *atom* is found. If the sphere is within 5 Å, we use the receptor sphere center as a target for the ligand atom. This is repeated for all ligand atoms. Then the ORIENT routine is called and a new set of ligand co-ordinates is calculated by rotating the ligand atoms onto their target positions. If the overlap error is reduced, the process continues.

Finally, we use a displacement procedure that further reduces the overlap error and establishes potential hydrogen-bonding patterns. A displacement is calculated for any ligand atom that violates the overlap constraint or for any polar ligand atom that is further than 3·5 Å from a receptor polar atom. The displacement is along the line of atom centers for the 2 atoms involved in each violation. The magnitude of the displacement is just sufficient to remove the error. If these displacements were applied to the atoms directly, or as pseudo forces, they would result in distortion of the rigid ligand-molecule geometry. To avoid this problem, the displacements are treated as targets and the ORIENT routine is used to find the best rigid-body translation/rotation onto the desired locations. The process is reasonably convergent. It terminates inelegantly if the number of displacements becomes less than 4. An alternative method that uses functional optimization for rigid-body displacements has been described by Cox (1967).

The result of the refinement procedure is a set of ligand co-ordinates that have been adjusted to a locally good fit to the restraints. In keeping with the general purpose of this study, no effort has been made to carry the process to complete convergence since we anticipate further refinement using molecular mechanics. Approximately 100 structures are produced in the two test cases described below. Each is scored using the overlap error. We also group the structures into classes whose members have atom to atom r.m.s.† displacements of less than 1 Å. This classification greatly aids inspection of the various docking arrangements. The refinement and classification programs require approximately 5 min per structure on the PDP 11/70.

3. Results

Our primary purpose in this paper is to test the techniques described above. Specifically, we ask: does the program reproduce known ligand–receptor geometries? If so, does it also provide alternative structures that are geometrically reasonable? To these ends, we have examined two systems for which the ligand–receptor geometry has been established by crystallographic means. The first is the heme–myoglobin interaction in metmyoglobin. The second is the binding of thyroxine to prealbumin. We also discuss some preliminary results for the docking of modified thyroxines in prealbumin for which no direct structural data are currently available.

(a) *Myoglobin–heme binding*

We select this system for several reasons. The crystallographic results are unambiguous so that it affords an excellent control calculation. The heme group has very little internal flexibility. The high symmetry and the "flatness" of the heme provide a demanding test of the representation and matching parts of the algorithms.

† Abbreviation used: r.m.s., root-mean-square.

We review the complete procedure and provide quantitative details. Two surfaces are calculated at a density of five points/Å^2. The heme surface is obtained from the heme co-ordinates taken directly from the myoglobin structure (Takano, 1977). The receptor surface is simply that of the myoglobin with the heme atoms removed. No internal water molecules are included. The ligand and receptor surfaces are closely complementary in the heme pocket (Fig. 3), as one expects for a tight complex. These surfaces are the input data for the representation program. The heme group contains 43 non-hydrogen atoms, for which 408 re-entrant surface points were used to construct 42 spheres meeting the various constraints above. (The iron atom has no re-entrant surface when the heme surface is constructed using a 1·4 Å probe.) Three minutes of computer time were required. The radii of the 42 spheres ranged from 1·75 to 2·15 Å slightly larger, on average, than the van der Waals' radii of the various atoms with which they were associated. The sphere centers are displaced "inward" of the atom centers so that the ligand surface is well-filled (Fig. 5). The visual correspondence between heme surface and the envelope of the spheres is quite good. We should note that a perfect correspondence is *not* essential for the subsequent stages of the program. The important point is to have a sufficiently accurate set of internal distances so that the matching algorithm will function properly†. The fit between the ligand surface and the surface of the spherical representation can be improved to almost any desired limit by increasing the number of surface points in the original surface and by increasing the number of spheres retained in the representation. The choice of one sphere/atom was for convenience, and proved to be of sufficient accuracy for this study.

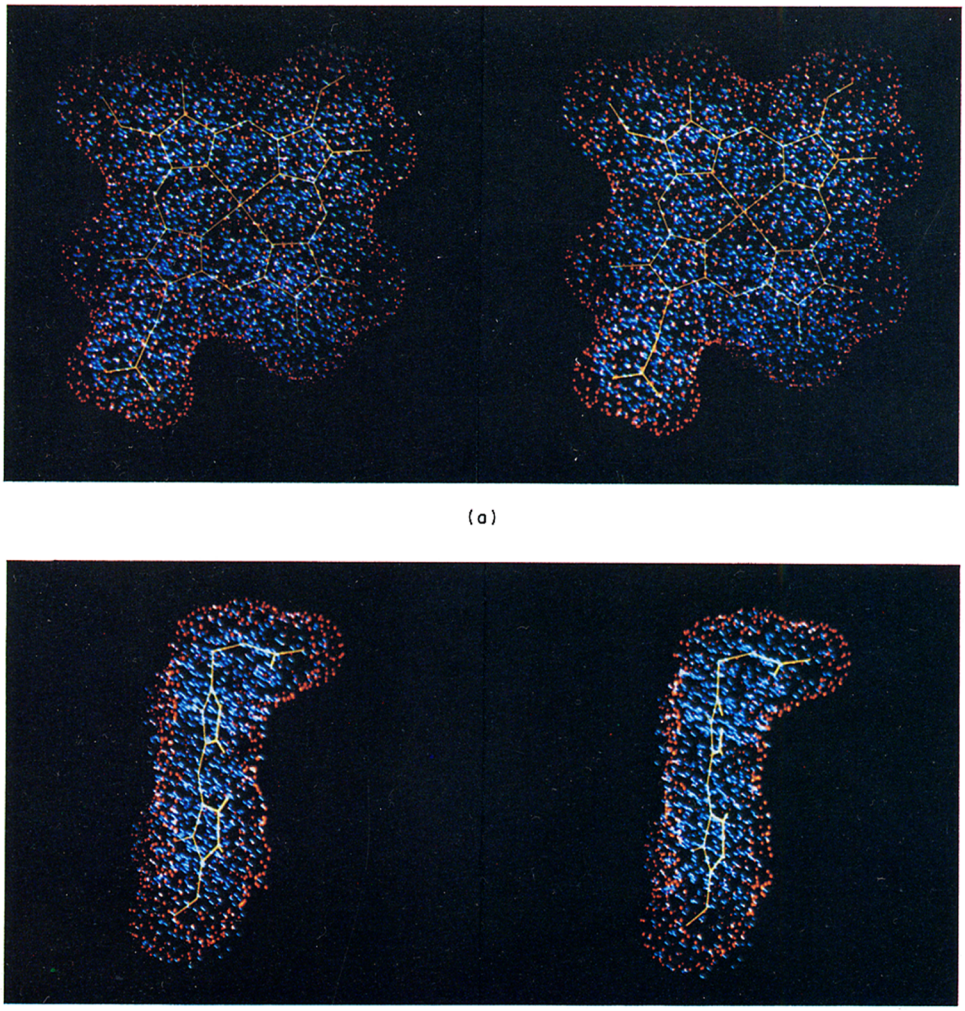
Metmyoglobin, without the heme, contains 1217 non-hydrogen atoms. We used 2289 contact surface points to construct 204 spheres. The small ratio of spheres/atoms reflects the lack of solvent accessibility of many myoglobin atoms and the more restrictive conditions for receptor sphere selection. This calculation took 40 minutes of 11/70 time.

The metmyoglobin spheres fall into three major clusters of 54, 23 and 17 spheres and a number of smaller clusters (Fig. 4 and Table 2). The largest cluster is clearly the heme pocket. It is the only one large enough to accommodate a ligand of that size. The spherical radii range from 1·5 to 4·8 Å, with the larger values associated with the top and "front" (E–F) face of the pocket. Closer inspection confirms that the spherical representation provides a good approximation to the pocket (Fig. 6). The only features that do not agree well are two substantial extensions along the outer surface that come from the large spheres for the side-chain oxygens of Gln91 and Ser92. These could be removed without altering the results below.

The second largest cluster is formed by the N terminal of the protein, the EF corner, and the side of the H helix. The third largest is a shallow invagination on the GH corner that continues some distance along the B helix/G helix interface (Fig. 4). Some of the smaller "sites" will be discussed in more detail in a later paper.

Returning to our main theme, the identification of the largest site with the heme

† The more conventional choice of using interatomic distances for the ligand has the advantage of greater accuracy and numerical stability, but the interatomic distances for the receptor are not a suitable match set.



(a)

FIG. 5. Front (a) and side (b) views of the heme surface (red) filled with spheres (blue) calculated using eqn (1).

pocket allowed us to use the matching algorithm with just the spheres for that site and those for the heme. The distance matching parameter, ε, was 1·5 Å. There are 54×43 or 2322 choices for the initial pairing. Of these 573 were retained after screening the distances from the ligand and receptor centroids. All of these starting points provided at least six ligand–receptor pairs. The median number of entries was eight per list, with about 25% of the arrangements having nine or ten entries. Remember that eight entries imply that all $8 \times 7/2$ or 28 internal distances among the eight ligand spheres and the 28 distances for the corresponding eight receptor spheres agree within the 1·5 Å limit.

The matching algorithm required about 30 seconds/structure for the heme

TABLE 2

Myoglobin surface sites

Site	Spheres	Helices	Atoms
1	54	B, C, CD, E, F, FG, G, H	Leu32 CD2; Glu41 O; Lys42 O CG NE; Phe43 CD1 CE1 CE2 CZ; Asp44 N; Asp44 OD1; Arg45 CD NH1 NH2; Lys63 CE; His64 CE1; Thr67 OG1 CG2; Val68 O CG1 CG2; Ala71 CB; Leu72 CA CD1; Leu89 CG CD1; Gln91 OE1; Ser92 OG; His93 CE1 NE2; Thr95 CG2; Lys96 O NZ; His97 CB CG ND1 CE1 NE2; Ile99 CG1 CG2 CD1; Tyr103 O CD1 OH; Leu104 CD2; Ile107 CB CG2 CD1; Ser108 OG; Ile111 CD1; Phe138 CE1 CZ; Leu149 CG CD2
2	23	NA, S, EF, F, H	Val1 N CG2; Leu2 O; Glu4 DE1; Lys79 O; Gly80 O; His81 CE1 NE2; Glu83 CA O CB OE2; Lys87 CE NZ; Gln91 NE2; Leu137 CD2; Asp141 OD2; Lys145 CG CD CE NZ; Glu148 CD OE1
3	17	B, C, G, H	Arg31 NE NH1; His36 CE1 NE2; Glu109 C CB CG OE1; Ala110 N; His116 CG CD2 CE1 NE2; Gly124 CA; Ala125 N; Gln128 OE1 NE2

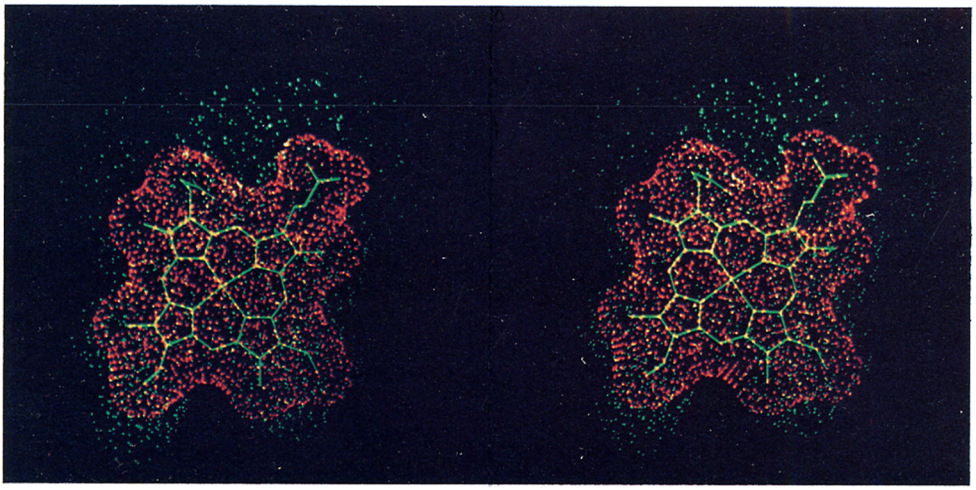
problem. We found no way of deciding by simple inspection which of the assignment lists would yield geometrically reasonable structures. For example, the longest lists generally led to excessive overlap (see below). Thus we decided to retain all the assignment lists with six or more entries for processing by the refinement section of the program.

Refinement

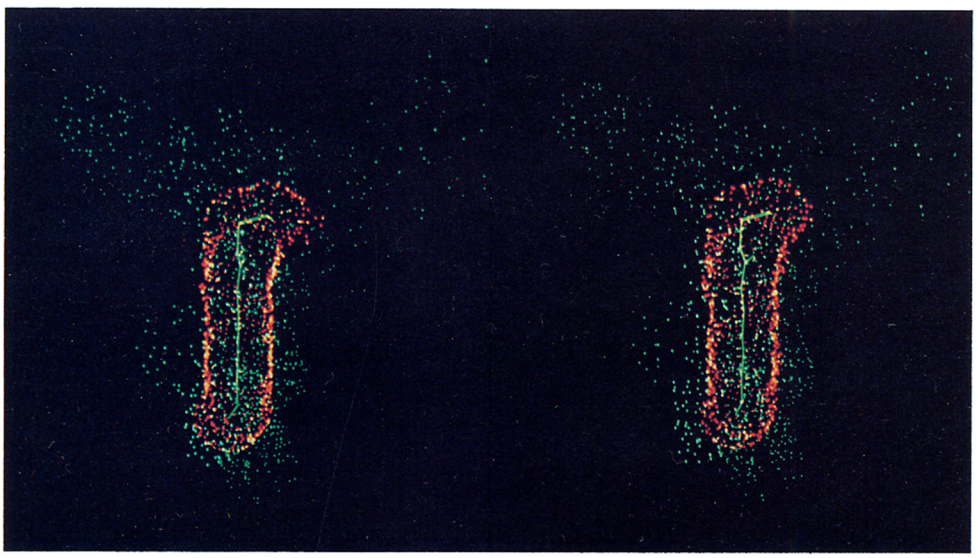
Two parameters were evaluated for the matching lists.

(1) The overlap error (eqn (1)).
(2) The root-mean-square deviation from the heme co-ordinates in the original X-ray structure.

Of course, the deviations from the crystal structure would not be available for an "unknown" but were calculated here to test the docking routines. The refinement program required approximately four minutes per structure for the myoglobin–heme system. The initial orientation phase reduced the overlap errors by approximately fourfold; the displacement phase of the routine made a roughly equivalent improvement. In Table 3 some of the final structures are listed in order of increasing overlap error. The structures fall into classes; the class with the smallest overlap error being the one closest to the original co-ordinates. Four structures were found that were within 1 Å of the X-ray heme co-ordinates (class 1). All of these had small overlap errors (typically 0·2 Å) for the van der



(a)



(b)

FIG. 6. Front (a) and side (b) views of the spherical representation of the heme pocket (green) and the heme surface (orange) (see the text).

Waals' radii we used. No attempt was made to refine these structures beyond the displacements described above. The overlap function (eqn (1)) had values of 16 to 25 for these four structures compared with 20 for the X-ray co-ordinates. As noted earlier, these overlap values are equivalent to approximately 2 kcal/mol repulsive contribution. At somewhat higher values of the overlap function ($\sim$30) there is an interesting group of docking geometries (class 2), in which the heme is rotated 180 degrees about an axis passing through CHA and CHC so that the propionic acid

TABLE 3

Myoglobin–heme dockings arranged in order of overlap function

Overlap†	r.m.s.‡	Class§
16·8	0·50	1
21·6	0·80	1
23·8	0·80	1
24·3	1·02	1
27·5	6·57	2
27·5	6·56	2
29·9	6·54	2
29·9	6·54	2
30·0	6·54	2
30·9	6·54	2
32·5	6·72	2
34·3	1·36	1
34·9	1·37	1
36·3	5·55	3
37·0	5·83	3
39·0	4·67	3
39·2	4·62	3
39·2	4·67	3
40·0	6·27	–
40·0	4·71	3
41·4	5·38	3
41·5	4·84	3
41·9	4·38	3
43·0	5·34	3
46·2	7·01	–
48·3	4·77	3
51·9	1·65	–
53·1	5·15	3
61·9	6·89	4
63·4	6·33	–
64·0	6·96	4
65·2	6·54	–
65·8	6·97	4
67·2	7·47	–
67·9	6·90	4
70·4	6·97	4
76·1	5·37	3
80·7	6·99	4

† As defined for eqn (1). The X-ray structure had an overlap value of 20. These values can be converted into conventional units of kcal/mol by multiplying by 0·1 (see the text).

‡ r.m.s. co-ordinate error compared with heme co-ordinates from X-ray data.

§ All structures in a class are within 1 Å r.m.s. co-ordinate error of another class member. Class 1, X-ray; class 2, inverted; class 3 and 4, 90° rotation (see the text).

side-chain positions are interchanged. These structures are really quite similar to the X-ray structure, although there are more close contacts between ligand and the heme pocket. There have been reports of such heme "inversion" in some insect hemoglobins (La Mar *et al.*, 1981). The large r.m.s. co-ordinate errors arise from the large displacements of particular atoms during the heme rotation. Structurally, they are replaced with nearly identical atoms so that the net change in shape is quite small. At still larger values of the overlap parameter ($E_{\text{overlap}} > 35$) are

structures that correspond to various 90 degree rotations of the heme, with or without inversion (classes 3 and 4). These arise because of the high symmetry of the heme core but are of much less biochemical interest because of the awkward placement of the various heme side-chains. The large overlap values mark these structures as poorer docking candidates.

In summary, the myoglobin–heme calculation confirms that the algorithms produce a sampling of the reasonable geometries for this system. Some of the structural classes are quite far apart in conformation space. While we cannot claim that the program has produced all structural classes of interest, it has certainly found the most obvious geometric variants, given the basic symmetry of the heme group. Of course, we make no claim that this approach could start with a proper apomyoglobin structure and proceed to the (presumably) significantly different myoglobin geometry.

(b) *Thyroxine–prealbumin*

The human serum thyroxine-transport protein prealbumin was the first hormone-binding protein to be characterized fully. Its atomic structure has been determined, and extensively refined, at 1·8 Å resolution by X-ray analysis (Blake *et al.*, 1978; S. J. Oatley, unpublished results). Prealbumin is a tetramer of identical subunits, each containing 127 residues; the subunits associate in an ellipsoidal shape forming a long central channel containing two hormone binding sites (Blake & Oatley, 1977). The symmetry of the molecule requires not only that the two sites are identical but also that each site itself has twofold symmetry.

The binding of thyroxine to prealbumin has also been investigated at 1·8 Å resolution (Blake *et al.*, 1982; Blake, 1981). The interpretation and refinement of these data are still in progress (S. J. Oatley, J. M. Burridge & C. C. F. Blake, unpublished results). The initial difference electron-density map was dominated by the features corresponding to the electron-dense iodine atoms, although extensive small conformational changes in the protein were also evident. A preliminary interpretation of this map was obtained by adjusting the torsion angles of a thyroxine molecule (Cody, 1974) so that its iodine atoms best fitted their corresponding electron-density features (Blake *et al.*, 1981).

Thyroxine is oriented with its phenolic hydroxyl group buried deep within the binding channel and its carboxyl and amino groups ion-paired with the Lys15 and Glu54 residues at the mouth of the binding channel. The positions of these charged residues in the native protein structure are such as to be in apparent close contact with thyroxine; they must be displaced on hormone binding, but their new positions were not clearly resolved in the electron-density maps. Accordingly, these close contacts were relieved by energy refinement of the complex (Blaney *et al.*, 1982*b*), allowing only the side-chains of Lys15 and Glu54 and the amino acid moiety of thyroxine to move from their X-ray co-ordinates. This provided the description of the binding site used as the starting point in our calculations.

Thyroxine (T4) has 24 non-hydrogen atoms, each of which yields a sphere in the ligand representation. The prealbumin "site" we used contained only one of the two binding sites and required 43 spheres for its representation. As was seen for

myoglobin, the spherical representations, even at this level of approximation, are quite similar to the surfaces from which they are derived. The matching program examined 360 potential docking arrangements.

The refinement program provided three major groupings of structures (Table 4). The set with the least overlap was, curiously enough, an inverted form in which the amino acid moiety was innermost in the cavity and the phenolic ring was directed outward toward the solvent (class 1). Although this arrangement had been considered at an early stage of the crystallographic investigation (Blake & Oatley, 1977), the present findings and the results of energy calculations indicate that this

TABLE 4

Thyroxine–prealbumin structures

Overlap†	r.m.s.‡	Class§
0·0	8·89	1
1·4	7·16	1
1·4	8·19	1
2·7	9·19	1
3·6	6·71	1
5·5	8·32	1
9·1	7·55	1
11·7	9·26	1
13·8	7·29	1
16·7	7·05	1
19·3	5·84	1
22·7	0·51	2
22·7	9·40	1
23·2	0·36	2
23·6	0·37	2
23·7	0·62	2
23·8	0·37	2
24·4	0·61	2
25·0	0·70	2
25·2	8·22	1
26·6	9·24	1
28·5	0·45	2
30·3	0·97	2
30·5	9·12	1
30·9	1·59	3
31·1	1·45	2
31·1	1·61	3
31·1	7·16	1
31·3	0·94	2
31·3	1·60	3
31·4	1·60	3
31·5	1·60	3
32·2	1·58	3
32·2	1·67	2
32·6	1·58	2

† As defined for eqn (1). The X-ray structure has an overlap value of 40 (see footnotes to Table 3 and text).

‡ r.m.s. co-ordinate error compared with heme co-ordinates from X-ray data.

§ All structures in a class are within 1 Å r.m.s. co-ordinate error of another class member. Class 1, inside-out; class 2, X-ray; class 3, C-2 (see the text).

is not a favored conformation. The simple hydrogen-bond procedure we used did not rule out this class of structures because of the central Thr hydroxyl groups; a more realistic test that forced ion-pairing would have done so. The next group of structures, ranked by overlap, are quite close to the conformation we chose for a reference point (class 2). The overlap values range from 23 to 45 (not shown) with the reference structure having a value of 40, and are equivalent to 2 to 5 kcal/mol repulsive energy. The overlaps are primarily with the I-3 and I-5 iodines. These structures show a small (0·3 Å) displacement inward compared to the starting structure, possibly because we omitted an internal water molecule that hydrogen bonds to the phenolic oxygen (Fig. 7). The third group of structures place the thyroxine at the equivalent C-2 position (class 3). The equivalence is not exact because the energy refinement procedures broke the crystallographic symmetry. The rest of the structures produced by the program had relatively poor overlap scores and were wedged in various ways into the mouth of the cavity. These are not listed in Table 4 and were not explored further. Thus, as with the myoglobin–heme example, the thyroxine calculation yields groups of structures that span geometrically reasonable possibilities.

Visual inspection, using computer graphics, of some of the docking geometries suggested significant differences. Specifically, we noticed that the precise matching of the ligand and receptor surfaces was much better for thyroxine oriented with the phenolic group inward than with the amino acid moiety inward. In the latter case, it was not possible to position the iodines into well-defined pockets.

We examined the docking into the prealbumin site of four thyroxine derivatives in which the phenolic ring was replaced by naphthol. The results were quite similar to those for thyroxine: three classes of structures were found for each derivative. These three classes were closely related to the "inverted", "normal" and "C-2" structures for thyroxine. Our calculations suggest that all four isomers can fit within the binding pocket.

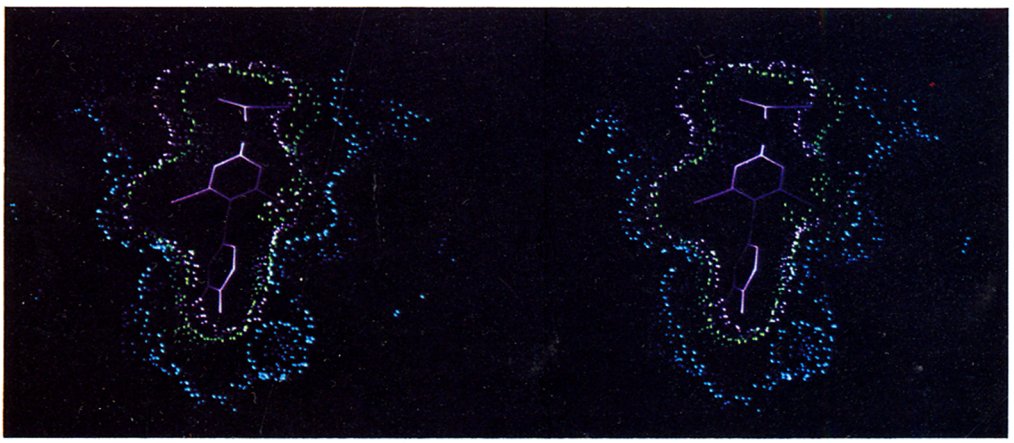


FIG. 7. Surface for hormone binding pocket in prealbumin (blue); the surface and molecular framework for thyroxine (purple) in its reference position; the surface for thyroxine (green) in its best "docked" position. The blue sphere in the lower right corner of the pocket is a water molecule.

The binding constants of these compounds to prealbumin have been measured (Blaney *et al.*, 1982*a*; and Table 5). Visual inspection showed the 3′-Br-, naphthol derivative (W) made a better fit than any of the other naphthol derivatives while the 7-OH group of the Z analog caused particularly unfavorable displacements (Fig. 8), forcing the I-3 and I-5 iodines out of their receptor pockets.

To approach the question of "goodness of fit" more quantitatively, we developed the following scheme. Imagine a thin spherical shell (0·25 Å thick) extending beyond the van der Waals' sphere for each ligand atom. Each *surface point* of the receptor must lie in one of the following classes: (1) inside the van der Waals' sphere of a ligand atom, (2) inside the spherical shell of a ligand atom, or (3) outside the spherical shells of all ligand atoms. We define a "fit" parameter as the number of the surface points in category 2 (shell points) minus the surface points in category 1 (overlap points). The fit value provides two useful discriminations. It ranks the three thyroxine structural classes in the order: "phenolic in" > "C-2" > "amino acid in", an ordering that is intuitively reasonable and in agreement with the X-ray results. The ordering for the thyroxine analogs is T4 > W > Z > Y > X. While the 7-OH naphthol analog is placed too high on this list, the overall agreement is encouraging for so simple an approach. We recognize that serious comparisons of this type require more detailed calculations, but it is not difficult to construct algorithms that approximate the same quantity that the eye detects as "goodness of fit".

TABLE 5

Docking thyroxine and derivatives to prealbumin

A. *Docking geometries for thyroxine*

Class†	Geometry	Overlap‡	r.m.s.§ (Å)	Fit‖
1	Inside out	2	8·03	62
2	X-ray	23	0·45	172
3	C-2 related to class 2	31	1·60	158

B. *Comparison of thyroxine and derivatives in class 2 docking geometry*

Compound	Relative binding affinity ¶	Overlap‡	r.m.s.§ (Å)	Fit‖
T4	1·00	23	0·45	172
W	0·29	17	1·11	166
Y	0·016	8	0·90	111
X	0·063	7	1·14	107
Z	0·007	12	1·29	133

† See the text.

‡ Defined for eqn (1), averages for best 5 structures (in arbitrary units).

§ r.m.s. deviation per atom compared to starting conformation, best 5 structures.

‖ See the text, average for best 5 structures (in arbitrary units).

¶ Fraction of the apparent association constant of the derivative to that of T4. Data from Blaney *et al.* (1982*a*).

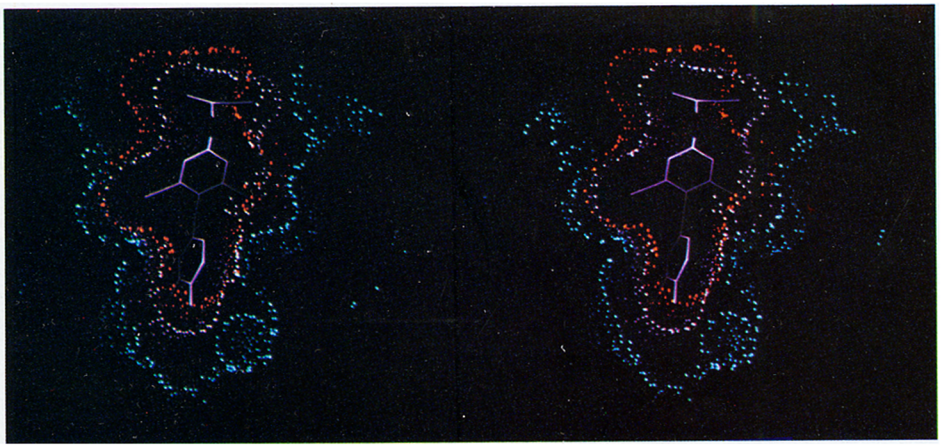


FIG. 8. Surface for thyroxine binding pocket and thyroxine as described for Fig. 7. The red surface is the best docked position for the 7-OH naphthyl analog (Z) (see the text).

4. Discussion and Conclusions

The main objective of this work was to design and test a set of algorithms that can explore in a reasonably complete manner the geometrically feasible alignments of a ligand and receptor of known rigid structure. This limited goal appears to be met by the program we have described above. The results for both test systems are:

(1) Structures quite near the "correct" structures are readily recovered and identified as feasible solutions.

(2) Other families of structures are found that are geometrically reasonable and that can be tested by simple scoring schemes, chemical intuition, or visual inspection with computer graphics.

The important underlying assumptions that made these calculations feasible within a few hours of minicomputer time should be restated explicitly.

(1) Both ligand and receptor structures were known in advance and were assumed to be rigid. The binding site of the ligand need not be known in advance, but was taken as given for prealbumin and readily selected by the clustering algorithm for myoglobin.

(2) The center of the ligand and the center of the receptor pocket were assumed to be approximately coincident.

(3) The numerical parameters used to select the spheres, to match distances, and to discard matches and structures certainly influence the total time. The parameters were chosen to be the same for both test cases, but further experiment is needed to establish their range of application.

(a) *Limitations*

We summarize the limitations of the program in its present form.

(1) There is no simple way to overcome the need for known ligand and receptor structures. Molecular mechanics can provide a sampling of the favorable

conformations of ligands of reasonable size, but the need for a detailed receptor geometry will remain a severe limit. There are techniques in various stages of development that build up a picture of the receptor geometry from ligand binding data (Crippen, 1980; Marshall *et al.*, 1978; Cramer, 1980) and these may offer useful starting points for our procedures.

(2) Without allowing molecular flexibility, many aspects of ligand–receptor interactions are not properly described. While the present program could explore the docking of several specific conformers of the ligand, of more interest would be changes in the matching and/or refinement packages to permit distortions of the rigid geometry. As noted earlier, one way to do this would be to use the structures generated here as the starting point for energy minimization or molecular dynamics calculations.

(3) This paper assumes that the ligand and the receptor pocket are of roughly the same size. Two other situations are commonly found. First, there are examples of large binding sites that are incompletely filled by the ligand, such as that for trimethoprim in dihydrofolate reductase (Baker *et al.*, 1981). We anticipate that this can be handled by the existing program, although the time needed may increase dramatically. The more difficult situation is the docking of a portion of a ligand into a site. A good example would be the docking of trypsin inhibitor with trypsin or, more generally, the matching of any two macromolecular surfaces. The program as it now stands might well be swamped by the large number of extraneous internal distances from portions of the ligand that were not involved in the active interface. What is needed is a way to break a large ligand up into a number of "projections" in analogy to the separation of the receptor surface into a number of binding "concavities". There are a number of ways this could be done.

(4) The relationship between the manipulations and simple scoring schemes used here, and energy optimization techniques needs to be established. A major concern is the use of united atoms instead of explicit hydrogen atoms. The surfaces will certainly be modified somewhat when hydrogen atoms are introduced. Further, the local energy terms will be altered significantly.

(b) *Other applications*

Some of the procedures have applications beyond their use here. For example, the various pockets, and packing defects for protein surfaces can be identified and examined in a systematic manner as suggested for myoglobin. The matching algorithm may prove useful for identification of common geometric features in a set of compounds. Finally, the spherical representation of the receptor pocket provides a tool with which to explore ligand modification.

Discussions with Y. Martin stimulated this project. M. Connolly, F. Cohen, P. Kollman and P. Weiner provided useful suggestions. S. J. Oatley is a Mr and Mrs John Jaffé Donation Research Fellow of the Royal Society. Funding from the National Institute of Health (GM-19267, I. D. Kuntz, RR-1081, R. Langridge) is gratefully acknowledged. J.M.B. is supported in part by the American Foundation for Pharmaceutical Education.

REFERENCES

Baker, D. J., Beddell, C. R., Champness, J. N., Goodford, P. J., Norrington, F. E. A., Smith, D. R. & Stammers, D. K. (1981). *FEBS Letters*, **126**, 49.

Blake, C. C. F. (1981). *Proc. Roy. Soc. London, sec. B*, **211**, 413–431.

Blake, C. C. F. & Oatley, S. J. (1977). *Nature (London)*, **268**, 115–120.

Blake, C. C. F., Geisow, M. J., Oatley, S. J., Rérat, B. & Rérat, C. (1978). *J. Mol. Biol.* **121**, 339–356.

Blake, C. C. F., Burridge, J. M. & Oatley, S. J. (1982). *Proc. VI Internat. Endocrin. Congr.*, in the press,

Blaney, J. M., Weiner, P. K., Dearing, A., Kollman, P. A., Jorgensen, E. C., Oatley, S. J., Burridge, J. M. & Blake, C. C. F. (1982*a*). *J. Amer. Chem. Soc.* in the press.

Blaney, J. M., Jorgensen, E. C., Connolly, M. L., Ferrin, T. E., Langridge, R., Oatley, S. J., Burridge, J. M. & Blake C. C. F. (1982*b*). *J. Med. Chem.* **25**, 785–796.

Cody, V. (1974). *J. Amer. Chem. Soc.* **96**, 6770–6725.

Connolly, M. L. (1981). Thesis, University of California.

Connolly, M. L. (1982). *Acta Crystallogr.* in the press.

Cox, J. M. (1967). *J. Mol. Biol.* **28**, 151–156.

Cramer, R. D. (1980). *J. Amer. Chem. Soc.* **102**, 1849–1859.

Crippen, G. M. (1980). *J. Med. Chem.* **23**, 599–606.

Crippen, G. M. (1981). *Distance Geometry and Conformational Calculations*, Research Studies Press, John Wiley & Sons, Chichester & New York.

Ferro, D. R. & Hermans, J. (1977). *Acta Crystallogr.* **33**, 345–347.

Greer, J. & Bush, B. L. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 303–307.

Hagler, A. T., Huler, E. & Lifson, S. (1974). *J. Amer. Chem. Soc.* **96**, 5319–5327.

Kuntz, I. D., Crippen, G. M. & Kollman, P. A. (1979). *Biopolymers*, **18**, 939–957.

La Mar, G. N., Anderson, R. R., Budd, D. L., Smith, K. M., Langry, K. C., Gersonde, K. & Sich, H. (1981). *Biochemistry*, **20**, 4429–4436.

Langridge, R., Ferrin, T. E., Kuntz, I. D. & Connolly, M. L. (1981). *Science*, **211**, 661–666.

Lesk, A. M. (1979). *Comm. ACM*, **22**, 219–224.

Levinthal, C., Wodak, S. J., Kahn, P. & Dadivanian, A. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 1330–1334.

Marshall, G. R., Barry, C. D., Bosshard, H. E., Dammkoehler, R. A. & Dunn, D. A. (1978). In *Computer Assisted Drug Design*, *ACS Symposia Series*, vol. 112, pp. 205–226, (Olson, E. C., ed.), American Chemical Society, Washington D.C.

Momany, F. A., Carruthers, L. M., McGuire, R. F. & Scheraga, H. A. (1974). *J. Phys. Chem.* **78**, 1595–1620.

Potenzone, R., Cavicchi, E., Weintraub, H. J. R. & Hopfinger, A. J. (1977). *Computer Chem.* **1**, 187–194.

Richards, F. M. (1977). *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.

Salemme, F. R. (1976). *J. Mol. Biol.* **102**, 563–568.

Takano, T. (1977). *J. Mol. Biol.* **110**, 569–584.

Weiner, P. K. & Kollman, P. A. (1981). *J. Comput. Chem.* **2**, 287–303.

Wodak, S. J. & Janin, J. (1978). *J. Mol. Biol.* **124**, 323–342.

Edited by R. Huber