CLIX: A Search Algorithm for Finding Novel Ligands Capable of Binding Proteins of Known Three-Dimensional Structure

Michael C. Lawrence and Paul C. Davis
CSIRO, Division of Biomolecular Engineering, Parkville, Victoria 3052, Australia

ABSTRACT **A computer algorithm, CLIX, capable of searching a crystallographic database of small molecules for candidates which have both steric and chemical likelihood of binding a protein of known three-dimensional structure is presented. The algorithm is a significant advance over previous strategies which consider solely steric or chemical requirements for binding. The algorithm is shown to be capable of predicting the correct binding geometry of sialic acid to a mutant influenza-virus hemagglutinin and of proposing a number of potential new ligands to this protein.**

Key words: **computer-aided drug design, database search, molecular docking, protein structure, protein–ligand interactions**

INTRODUCTION

The design of novel therapeutic agents based on the three-dimensional X-ray structure of target protein molecules is increasingly topical, given recent advances in the speed and ease with which such structures can be determined. While knowledge of the three-dimensional structure of the target molecule is recognized as highly advantageous, there is as yet no general approach to its utilization in the design strategy. Two heuristics may however be discerned: (1) the novel ligand must fit sterically into the targeted binding site of the protein; and (2) the novel ligand must achieve sufficient favorable chemical interactions with the protein and avoid unfavorable interactions.

The concept of steric fit has been employed[1,2] in a search algorithm (DOCK) which is capable of scanning a database of small molecules of known three-dimensional structure for candidates which fit geometrically into the target protein site. It is anticipated that such candidates might act as the framework for further design effort. DOCK, while providing a simple and efficient way of describing shape in terms of a set of spheres, is hindered by its lack of attention to chemical detail. Other drawbacks include its bias toward selecting large molecules and the incompleteness of its search of the orientational space available to the small molecule. Goodford[3] and Boobbyer et al.[4] have produced a computer program (GRID) which seeks to determine regions of high affinity for different chemical groups (termed *probes*) on the molecular surface of the binding site. GRID hence provides a tool for suggesting modifications to *known* ligands that might enhance binding. It may be anticipated that some of the sites discerned by GRID as regions of high affinity correspond to "pharmacophoric patterns" determined inferentially from a series of known ligands. (A pharmacophoric pattern is a geometric arrangement of features of the anticipated ligand that is believed to be important for binding.) Attempts have been made to use pharmacophoric patterns as a search screen for novel ligands[5–7]; however, the constraint of steric and "chemical" fit in the putative (and possibly unknown) receptor binding site is ignored. Goodsell and Olson[8] have used the Metropolis (simulated annealing) algorithm to dock a single known ligand into a target protein. They allow torsional flexibility in the ligand and use GRID interaction energy maps as rapid lookup tables for computing approximate interaction energies. Given the large number of degrees of freedom available to the ligand, the Metropolis algorithm is time-consuming and is unsuited to searching a candidate database of a few thousand small molecules.

This article presents a computer algorithm CLIX which searches the Cambridge Crystallographic Data Bank[9] (CCDB) for small molecules which can be oriented in the receptor binding site in a way that is *both* sterically acceptable *and* has a high likelihood of achieving favorable chemical interactions between the candidate molecule and the surrounding amino acid residues. The method is based on characterizing the receptor site in terms of an ensemble of favorable binding positions for different chemical groups and then searching for orientations of the candidate molecules that cause maximum

Received December 26, 1990; revision accepted April 15, 1991.

Address reprint requests to Dr. M.C. Lawrence, CSIRO, Division of Biomolecular Engineering, 343 Royal Parade, Parkville, Victoria, 3052, Australia.

spatial coincidence of individual candidate chemical groups with members of the ensemble. The current availability of computer power dictates that a computer-based search for novel ligands follows a *breadth-first* strategy. A breadth-first strategy aims to reduce progressively the size of the potential candidate search space by the application of increasingly stringent criteria, as opposed to a *depth-first* strategy wherein a maximally detailed analysis of one candidate is performed before proceeding to the next. CLIX conforms to this strategy in that its analysis of binding is rudimentary—it seeks to satisfy the *necessary* conditions of steric fit and of having individual groups in "correct" places for bonding, without imposing the *sufficient* condition that favorable bonding interactions actually occur. A ranked "short-list" of molecules, in their favored orientations, is produced which can then be examined on a molecule-by-molecule basis, using computer graphics and more sophisticated molecular modelling techniques. CLIX is also capable of suggesting changes to the substituent chemical groups of the candidate molecules that might enhance binding.

As an application, a search for potential influenza-virus hemagglutinin-binding molecules will be discussed.

METHODS

The CLIX algorithm can be summarized as follows. The GRID program (see Introduction) is used to determine discrete favorable interaction positions (termed *target sites*) in the binding site of the protein for a wide variety of representative chemical groups. For each candidate ligand in the CCDB an exhaustive attempt is made to make coincident, in a spatial sense in the binding site of the protein, a pair of the candidate's substituent chemical groups with a pair of corresponding favorable interaction sites proposed by GRID. All possible combinations of pairs of ligand groups with pairs of GRID sites are considered during this procedure. Upon locating such coincidence, the program rotates the candidate ligand about the two pairs of groups and checks for steric hindrance and coincidence of other candidate atomic groups with appropriate target sites. Particular candidate/orientation combinations that are good geometric fits in the binding site and show sufficient coincidence of atomic groups with GRID sites are retained.

Consistent with the breadth-first strategy, this approach involves simplifying assumptions. Rigid protein and small molecule geometry is maintained throughout. As a first approximation rigid geometry is acceptable as the high-resolution coordinates of the X-ray structure of the protein describe an energy minimum for the molecule, albeit a local one. If the surface residues of the site of interest are not involved in crystal contacts then the crystal configuration of those residues should reasonably mimic their mean solution configuration. However, in the complex both the protein and the small molecule may adopt different conformations to those found in their independent crystal structures. DesJarlais et al.[10] discuss the introduction of flexible geometry for the docking of a single trial ligand according to the distance-geometry strategy of Kuntz et al.[1] However, their method would be excessively time-consuming as a database search tool and would require the initial correct partitioning of every candidate in the database into its constituent fragments.

A further assumption implicit in CLIX is that the potential ligand, when introduced into the protein binding site, does not induce change in the protein's stereochemistry or partial charge distribution and so alter the basis on which the GRID interaction energy maps were computed. It must also be stressed that the interaction sites predicted by GRID are used in a *positional* and *type* sense only, i.e., when a candidate atomic group is placed at a site predicted as favorable by GRID, no check is made to ensure that the bond geometry, the state of protonation, or the partial charge distribution favors a strong interaction between the protein and that group. Detailed analysis should form part of more advanced modelling of candidates in the CLIX short-list.

The algorithmic details of CLIX are now described.

Identification of Target Sites

Before running CLIX it is necessary to compute GRID probe/protein interaction-energy maps for a set of probes of interest. These maps are then scanned to select favorable interaction sites in the region of interest of the protein.

Let $E_p(\mathbf{x})$ be the energy map associated with probe p, where $p \in P$, the set of all probes considered, and $\mathbf{x}$ is the map coordinate. A set of target sites $T = \{ (\mathbf{t}_j, p_j) \mid j = 1, \ldots, n_T; p_j \in P\}$ is then selected such that if $(\mathbf{t}, p) \in T$, then (1) $E_p(\mathbf{t})$ is a local energy minimum in the map E_p, (2) $E_p(\mathbf{t})$ corresponds to the binding of probe p to a residue in the protein's ligand binding site (rather than at some more remote location), and (3) $E_p(\mathbf{t}) < {E^{(m)}}_p$, where ${E^{(m)}}_p$ is a threshold energy. The set T corresponds to an ensemble of favorable interaction sites in the potential ligand binding site of the protein (note that the interaction energy maps follow the usual convention of negative energy corresponding to binding affinity). The energy thresholds ${E^{(m)}}_p$ used to select the target sites are chosen empirically to include as many potentially favorable interaction sites as is computationally manageable.

Selection of Candidate Molecules and Protein Atoms

The Cambridge Crystallographic Data Bank provides a repository of structural coordinates of small molecules that can be accessed and searched in an interactive fashion. A working subset of candidate molecules and their coordinates is readily extracted from the CCDB, for example, by excluding inorganic and organometallic compounds, and those containing known toxic elements. This subset will be denoted L. More sophisticated or customized databases can also be extracted.

The CLIX search requires a table of which substituent atoms of a trial molecule $\tau \in L$ correspond to which probes in P. The CCDB does not provide this information directly, rather it contains for each molecule two sets of atomic connectivity tables: one based on its chemical formula which includes information about the individual bond types within each molecule, and a second based on observed connectivity in the crystal, with no information about bond types. The atom labeling within these two sets of tables does not, for historical reasons, correspond. CLIX contains a precursor, graph-theoretic routine to renumber canonically the pair of connectivity tables associated with each molecule, thus establishing an isomorphic correspondence between the chemical and crystallographic connectivity tables. Details of this algorithm will be discussed elsewhere. The isomorphism achieved, CLIX proceeds further to classify as many atoms as possible in each molecule in L as corresponding to a probe in P. The classification defines groups in terms of the bonding pattern of the principal atom to its neighbors up to two nodes away in the ligand bonding tree. These groups are specified in CLIX in a way similar to that required for CCDB fragment searches. It is assumed that defining groups by the atomic and bonding detail up to second-nearest neighbor is sufficient to describe the nature of likely nonbonded interactions that the group could make with the protein and to identify the group with a particular GRID probe. Implicit also is the assumption made by GRID that all chemical groups of a given type p can be characterized by a single, uniform set of parameters when calculating E_p. More detailed modeling of substituent groups is feasible only once the candidate list has been substantially reduced.

A new database L' is therefore generated, containing for each $\tau \in L$, (1) its CCDB reference code, (2) the X-ray coordinates of its nonhydrogen atoms, (3) the connectivity and bond-type table associated with the coordinates, (4) the atomic number of each atom, (5) a list of the number of hydrogen atoms bound to each nonhydrogen atom, and (6) indices associating each atom with a probe in P. Note that the hydrogen coordinates are not retained. CLIX considers only the nonhydrogen atoms of both the protein and the candidate when assessing steric fit.

The set of atomic coordinates of a given candidate molecule $\tau \in L'$ will be denoted as $\{ \mathbf{l}_i \mid i=1, \ldots, n_\tau\}$ where n_τ is the number of nonhydrogen atoms in τ. The substituent groups or probes to which the atoms of τ correspond will be denoted as $\{ q_i \mid i = 1, \ldots, n_\tau\}$ where each q_i corresponds to an element in P. If a particular atom i in τ does not correspond to any probe in P, then $q_i = 0$. A given atom i of τ will thus also be written as $(\mathbf{l}_i, q_i)$.

CLIX must be provided with the coordinates of protein atoms in the immediate vicinity of the interaction site in order to judge the steric fit of a particular candidate orientation. These coordinates are chosen as the set $R = \{ \mathbf{r}_i \mid i = 1, \ldots, n_R\}$ of all nonhydrogen atoms on the surface of the protein which lie within a given radius of the center of the anticipated binding site. (Surface atoms are determined by the Connolly program.[11])

Fitting of Candidate Molecules Into the Binding Site

For each candidate molecule in L' a systematic search is made of all pairs of pairs $\{ (\mathbf{l}_i, q_i), (\mathbf{t}_j, p_j)\}$ for a set of indices i_1, i_2, j_1 and j_2 such that $q_{i_1} = p_{j_1}$, $q_{i_2} = p_{j_2}$, and $\mid \mid \mathbf{l}_{i_1} - \mathbf{l}_{i_2}\mid - \mid \mathbf{t}_{j_1} - \mathbf{t}_{j_2}\mid \mid < \delta_1$, i.e., for all pairs of candidate atoms that correspond in probe type and spatial separation to a pair target points. The criterion for spatial agreement, δ_1, reflects the confidence in, or "width" of, the target positions: too small a value will result in the exclusion of potentially favorable candidates, too large a value will lead to hopelessly many spurious agreements.

The candidate atomic coordinates are then oriented so that the points l_{i_1}, l_{i_2}, t_{j_1}, and t_{j_2} are colinear, with the geometric center of l_{i_1} and l_{i_2} coincident with the geometric center of t_{j_1} and t_{j_2}, and l_{i_1} lying closer to t_{j_1} than to t_{j_2} (see Fig. 1). There is one degree of freedom in orienting the candidate in this fashion: this is exploited in the next step.

With l_{i_1} and l_{i_2} kept in the position defined above, the candidate coordinates are rotated in discrete steps of θ degrees about the axis defined by l_{i_1} and l_{i_2}. At each rotational position a check is made to see whether any other candidate atom $(\mathbf{l}_{i_3}, q_{\mathrm{i}_3})$ becomes coincident with a third target point $(\mathbf{t}_{\mathrm{j}_3}, p_{j_3})$ such that $\mid \mathbf{t}_{j_3} - \mathbf{l}_{i_3}\mid < \delta_2$ and $q_{i_3} = p_{j_3}$. Similar considerations apply to the selection of δ_2 as do to δ_1. If such a match is achieved (and there may be more than one) then a check is made to see whether there are any steric clashes between the protein and candidate atoms in that orientation. This is done by computing all distances $d_{ik} = \mid \mathbf{l}_i - \mathbf{r}_k\mid$ and ensuring that $d_{ik} > \delta_3$ for all i and k. If there are no clashes, the geometry is deemed sterically acceptable.

The use of a single criterion δ_3 to assess steric fit is purely for computational speed and simplicity, an atomic-number specific constraint represents an obvious elaboration.

The search for coincidence of the third atomic group/target point pair could proceed more simply by using a distance-geometry approach,[11] wherein one would search for a target point j_3 whose distances from points j_2 and j_1 matched the distances of

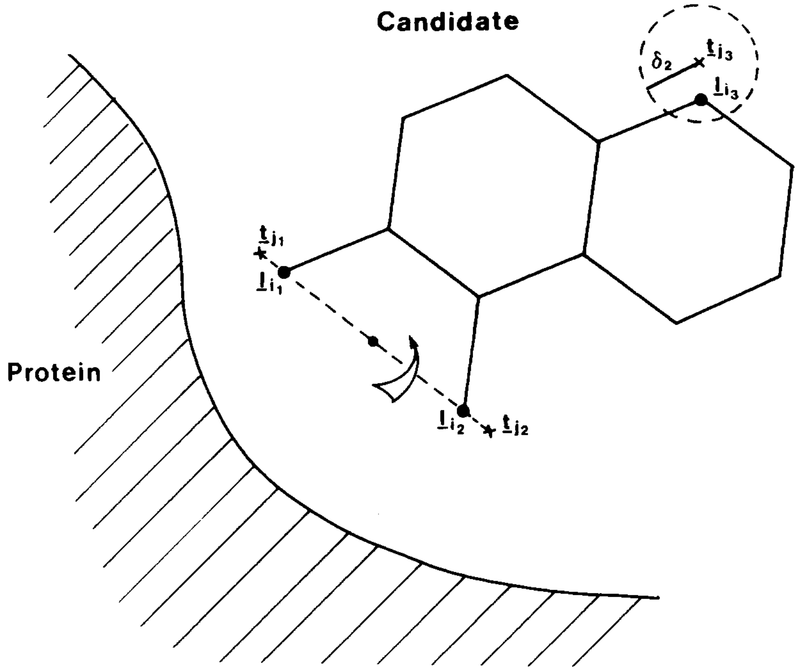

Fig. 1. Schematic diagram showing the pairwise alignment of candidate atoms with high-affinity binding positions in the protein binding site. $\mathbf{l}_{i_1}$, $\mathbf{l}_{i_2}$, and $\mathbf{l}_{i_3}$, coordinates of candidate atoms of type q_1, q_2, and q_3, respectively; $\mathbf{t}_{j_1}$, $\mathbf{t}_{j_2}$, and $\mathbf{t}_{j_3}$, target points atoms of type p_1, p_2, and p_3, respectively. The distances are such that $| |\mathbf{l}_{i_1} - \mathbf{l}_{i_2}| - |\mathbf{t}_{j_1} - \mathbf{t}_{j_2}| | < \delta_1$ and the types such that $p_1 = q_1$, $p_2 = q_2$, and $p_3 = q_3$.

a candidate atom i_3 from i_2 and i_1, respectively, and then configure the candidate so that these triplets of coordinates correspond. The method of rotating the candidate about a pair of aligned groups, however, has the advantage in that it can be extended to anticipate a more general scheme of binding, as will be mentioned in the Discussion.

Improvement of Marginally Rejected Geometries

Given (1) the discrete nature of the rotational search, (2) the way in which the atoms l_{i_1} and l_{i_2} are aligned with their corresponding target points t_{j_1} and t_{j_2}, and (3) the rigidity of the clash criterion δ_3, it can be anticipated that a number of potentially chemically favorable geometries might be rejected purely because one or more candidate atoms fail marginally to meet the clash criterion $d_{ik} > \delta_3$. A slight displacement of the candidate molecule in the protein binding site in such a case may result in a sterically acceptable fit without seriously compromising the alignment of candidate atoms and target points. However, as will become clear in the description of orientation scoring, acceptance of these geometries cannot be accommodated by simply reducing δ_3. Instead the following approach is adopted. CLIX, having determined that a particular orientation achieves steric fit subject to a value of δ_3 set slightly smaller (say by 0.5 Å) than is physically realistic, then adjusts the position of the molecule to "relax" any candidate–protein contacts that remain poor. The key to the procedure is that the motion of the candidate required to "relax" the poor contacts is such that the candidate is restrained to remain as close as possible to its original trial orientation, thus maintaining the overall correspondence of candidate atoms with target sites.

Let τ be a candidate molecule in an orientation satisfying $d_{ik} > \delta_3$ for all i and k. An attempt is then made to move τ in the protein frame so that $d_{ik} > \delta_4 > \delta_3$ for all i and k, where δ_4 represents a more realistic clash criterion. Let (i,k) be a candidate–protein atom pair such that $\delta_4 > d_{ik} > \delta_3$. For the pair (i,k) calculate the vector $\boldsymbol{\alpha}_{ik} = [\,(\delta_4 - d_{ik})/d_{ik}]\,\mathbf{d}_{ik}$ where $\mathbf{d}_{ik} = \mathbf{l}_i - \mathbf{r}_k$. The vector $\boldsymbol{\alpha}_{ik}$ represents the displacement that must be applied to atom i to move it along the line joining the centres of atoms i and k so that it is moved to a distance δ_4 away from atom k (see Fig. 2). For given atom i sum all the $\boldsymbol{\alpha}_{ik}$ resulting from protein atom clashes to obtain a displacement vector $\mathbf{s}_i = \Sigma\, \boldsymbol{\alpha}_{ik}$ (where the summation extends over all protein atoms k involved in clashes with the candidate atom i) by which to move atom i to attempt to relieve the clashes with atoms k. Define a set of displaced atom positions $\mathbf{l}'_i$ as $\mathbf{l}'_i = \mathbf{l}_i + \mathbf{s}_i$ which will act as trial new positions for the candidate atoms. If there are no candidate–protein clashes for a given candidate atom i, then $\mathbf{l}'_i = \mathbf{l}_i$. The candidate coordinates $\{\mathbf{l}_i\}$ are fitted using a least-squares method[13] onto $\{\mathbf{l}'_i\}$. The procedure is

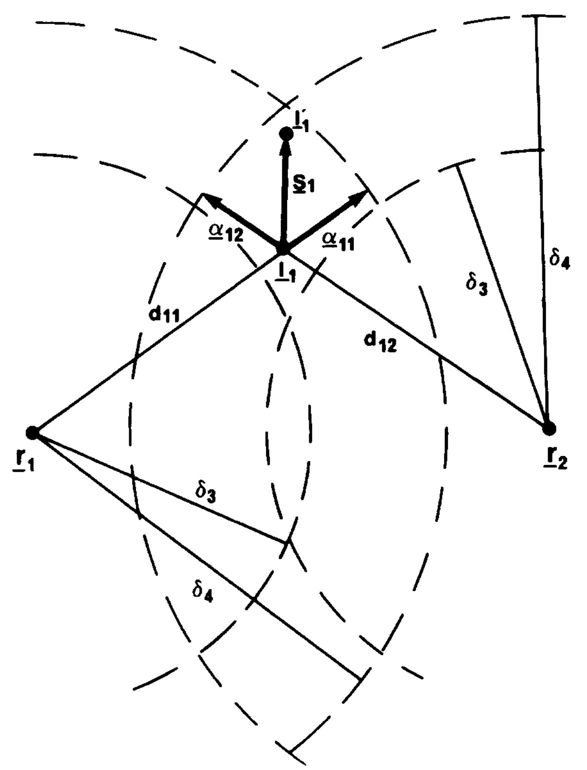

Fig. 2. Schematic diagram illustrating the vectors involved in "relaxing" marginally poor steric fits. For simplicity a candidate atom $\mathbf{l}_1$ is shown as being too close to two protein atoms. $\mathbf{r}_1$, $\mathbf{r}_2$, protein atoms in contact with candidate atom $\mathbf{l}_1$; the distances δ_3 and δ_4 and the vectors α_{11}, α_{12}, and $\mathbf{s}$ are as described in the text. If the atom $\mathbf{l}_1$ is moved to the new position $\mathbf{l}'_1$, then a better steric fit will be achieved. Repeated iteration of the algorithm described in the text should result in $d_{11} > \delta_4$ and $d_{12} > \delta_4$.

then iterated: a new set of candidate–protein clashes is detected, α_{ik} and $\mathbf{l}'_i$ recomputed, and the least-squares fitting of $\{\mathbf{l}_i\}$ onto $\{\mathbf{l}'_i\}$ repeated until no further candidate motion results. Note that the point $\mathbf{l}'_i$ does not necessarily satisfy $|\mathbf{l}_i - \mathbf{r}_k| > \delta_4$.

This procedure is stable and succeeds in relaxing the majority of marginal fits, or at least improving all the d_{ik} resulting from clashes to be significantly nearer in value to δ_4 than to δ_3. There will, however, be cases where it is not possible to reorient the candidate molecule so as to satisfy $d_{ik} > \delta_4$ for all i and k without seriously compromising the alignment of the candidate atoms and target sites.

Scoring the Candidate's Potential Binding Ability

It is not feasible to attempt a full energetic analysis of the binding of every candidate–orientation fit obtained during the search itself. Instead, the following simple scoring scheme is adopted. For each candidate atom $(\mathbf{l}_i, q_i)$ with $q_i \neq 0$ extract the energy e_i at point $\mathbf{l}_i$ in the GRID interaction energy map Eq_i, and then sum these energies over all atoms in the candidate for which $q_i \neq 0$ to obtain $e = \Sigma\, e_i$. e is then taken as a measure of the candidate's binding ability.

This is clearly a simplistic approach to the energetics of binding, as has been discussed above. Indeed no attempt is made to ensure that the orientation of the chemical bonds to a particular candidate atom i corresponds to that demanded by GRID to achieve the interaction energy e_i. However, as will be seen in the application to hemagglutinin, the value e for a known ligand (sialic acid) in its crystallographically determined binding orientation lies in the extreme tail of the distribution of energies obtained for CCDB candidates fitted by CLIX. This suggests that the score e forms a useful measure of binding probability. CLIX retains as its short-list all sterically acceptable orientations with $e < e_c$, where e_c is a user-supplied cut-off (recall that the e's of interest are numerically negative).

The geometric alignment of the candidate molecule in the protein binding site must take into account the steepness of the Lennard–Jones potential used to describe van der Waals interactions in GRID. It is for this reason that marginally poor fits cannot simply be retained in the short-list with small d_{ik} values: the repulsive energy e_i associated with such atoms would dominate the sum e and the orientation would be rejected. δ_4 should thus be chosen so that no candidate molecule, in its finally accepted orientation, is unduly penalized by high van der Waal's energies.

Chemical Substitutions

The above algorithm leads to a simple heuristic scheme for suggesting chemical changes to a given candidate that might enhance its binding ability in a particular orientation determined by CLIX.

Let $(\mathbf{l}_i, q_i)$ be a particular candidate atom corresponding to probe type q_i. Let the corresponding energy at site $\mathbf{l}_i$ in the GRID interaction energy map associated with probe q_i be e_i. Suppose further that in one of the other GRID maps (associated with probe type q_k) the energy at l_i is significantly less than e_i. Indeed energy e_i might be such that the binding of atom i at that site is unfavorable. If a molecule σ with similar stereochemistry to τ could be synthesized, but with a chemical group of type q_k substituted for atom i, then one might surmise that the ligand σ will bind more strongly to the protein. While the algorithm in no way predicts that such an altered molecule could be synthesized, (or even if it did, that it would have similar stereochemistry), it is nevertheless useful to have this information regarding atom i of τ available, and it is a trivial matter to report it during the computation of e.

In order to avoid the program generating known chemically impossible substitutions, a look-up matrix of "allowable substitutions" is used which permits only certain substitutions q_i to q_k to be suggested.

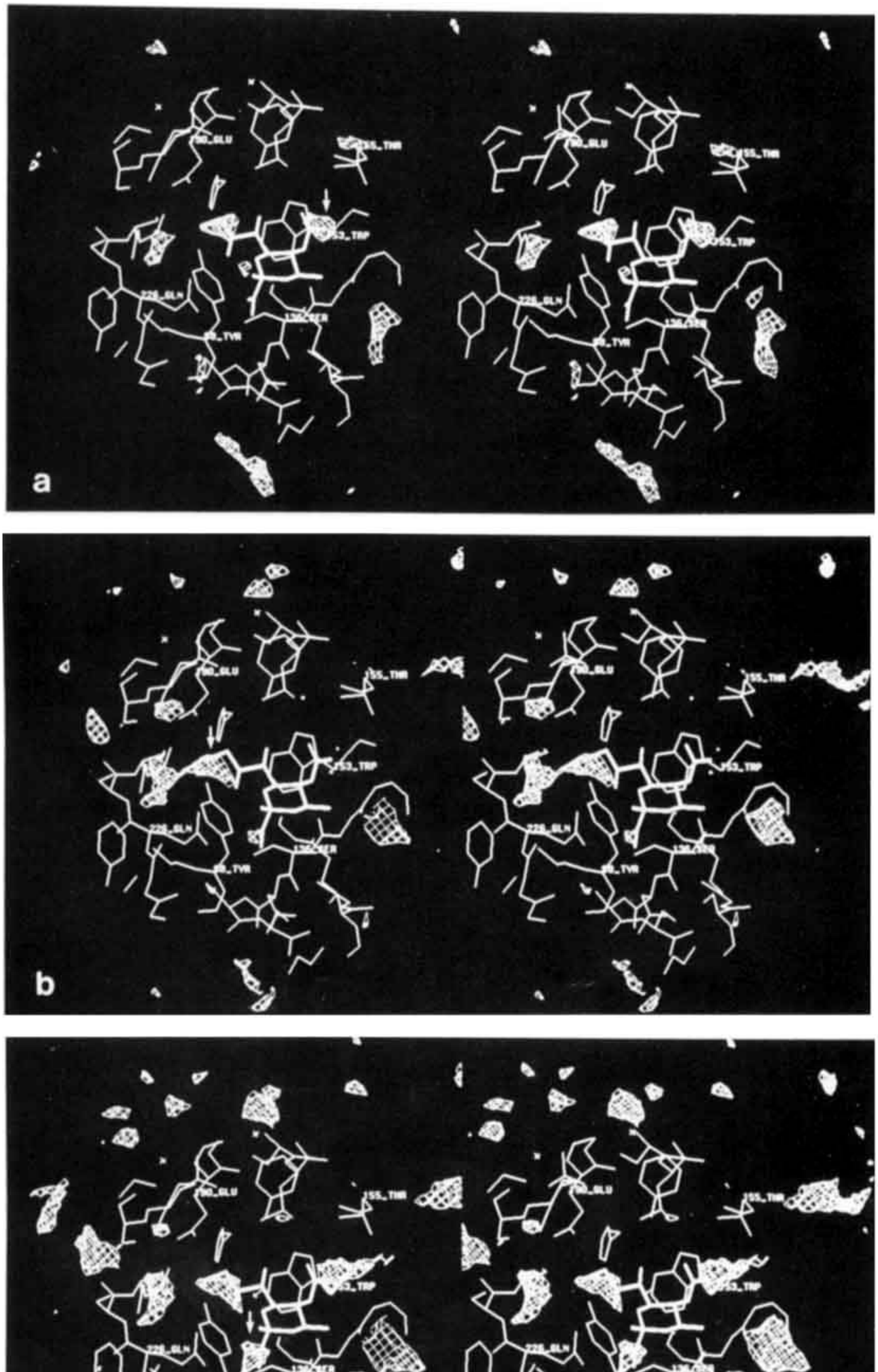

Fig. 3.

Programming Notes

CLIX is written in FORTRAN and consists of three major independent parts. The first part is concerned with the identification of chemical groups in the selected subset *L* of the CCDB. It is necessary to run this part only once in order to establish a working database *L'*. The second part concerns the orientational search and it produces the short-list as a direct-access file containing for each candidate–orientation, all the atomic information of *L'*, but with the coordinates appropriately updated to the CLIX-determined orientation. This step is the most time-consuming. The final section scores the orientations in the short-list by looking up the relevant atomic energies in the GRID maps. In order to minimize the problems of randomly accessing many large maps simultaneously, CLIX uses a double-sort strategy similar to that developed by Bricogne[14] for the noncrystallographic symmetry averaging problem in protein crystallography. Energies are obtained from the GRID maps at subgrid values by linear 8-point interpolation.

RESULTS

Potential Influenza Virus Hemagglutinin Ligands

Hemagglutinin (HA) is a surface glycoprotein of the influenza virus that binds to terminal sialic acid moieties on receptor molecules on the cell surface during infection. X-Ray structures to ~ 3 Å resolution have been determined[15] for wild-type A/Aichi/68 (H3N2) and mutant hemagglutinins, in both receptor-analogue complexed and uncomplexed forms. The electron density map of the L226Q HA mutant complexed with sialyllactose shows the location of the sialic acid binding site and a number of potential sialic acid–protein interactions are identified. Comparison with the uncomplexed structure shows that there is no appreciable conformational change in either L226Q HA or sialic acid upon binding.

It is instructive to examine the binding of sialic acid to the L226Q mutant to see whether it corresponds to the scheme anticipated by CLIX and whether CLIX is capable of recovering the correct binding orientation of this ligand. Figure 3 shows GRID probe/protein interaction–energy contour maps in the vicinity of the sialic acid binding site for probes corresponding to methyl carbon, hydroxyl oxygen, and carboxyl oxygen groups. It can be seen that the terminal hydroxyl oxygen of the triol group, the carboxyl oxygens, and the CH_3 group of sialic acid indeed lie close to a pronounced minimum in the corresponding GRID maps. Provided that these minima are included in the target set *T*, CLIX should be capable of correctly placing sialic acid in its crystallographically determined binding site.

A candidate database of 31042 organic molecules was extracted from the CCDB (Version 4.3). The chemical and crystallographic connectivity tables for each molecule were then matched using the renumbering procedure in CLIX. Successful matching was achieved for 29720 candidates and the substituent chemical groups of each molecule were identified (see Table I). The coordinates of sialic acid used were those of the refined L226Q-sialyllactose complex, as only the β-anomeric form of sialic acid is present in the CCDB and HA binds the α-anomer.

GRID maps of size 40 Å × 40 Å × 35 Å (grid spacing 0.5 Å) enclosing the sialic acid binding site were computed for a variety of probes. A set of target points *T* was then selected from the maps using the values of $E^{(m)}{}_p$ shown in Table I. Target points were restricted to be those where the primary protein atom involved in the probe–protein interaction belonged to one of the residues 98, 134–137, 153, 183, 190, 194, 226, and 228 (i.e., those residues which are potentially involved in sialic acid binding[15]). Inspection of *T* showed that it contained the three minima referred to above as being associated with sialic acid atomic groups. The set of protein atoms R was chosen as all surface atoms lying within a 20 Å radius of the sialic acid binding site.

CLIX was then used to predict the orientation of sialic acid in the HA L226Q binding site, using distance-matching criteria $\delta_1 = \delta_2 = 1.0$ Å, $\delta_3 = 2.3$ Å, $\delta_4 = 2.7$ Å, and rotational increment $\theta = 5°$. All sterically acceptable orientations were retained, i.e., no score cutoff e_c was applied. A total of seven potential orientations were predicted. The root mean square (rms) deviations of the atomic coordinates of these orientations from the crystallographic coordinates of sialic acid are shown in Table II. The orientation with the lowest *e* (−37.3 kcal/mol) has an rms coordinate deviation of 0.41 Å from the crystallographic position. The score e_x computed for sialic acid in its crystal position, using the scheme outlined in Methods, is −37.9 kcal/mol. CLIX is therefore capable of recovering the binding geometry of sialic acid in the HA mutant site. Of the remaining six orientations, the coordinates of the two with next highest score also correspond closely to the crystallographic position (see Fig. 4). These orientations might be anticipated: for every triplet of atom/target pairs, there are three distinct ways of allocating the initial pair of pairs used to determine the rotation axis in the CLIX algorithm. Their scores however vary by ~30%, depending on the precise alignment of sialic acid in the GRID maps. The remaining four orientations are spurious. Also shown in Table II are

Fig. 3. GRID interaction energy maps for (**a**) methyl carbon, (**b**) hydroxyl oxygen, and (**c**) carboxyl oxygen in the vicinity of the sialic acid binding site of the L226Q influenza-virus hemagglutinin mutant. The maps are contoured at −3.0, −7.0, and −6.0 kcal/mol, respectively. The sialic acid molecule is shown (bold) with the terminal OH, CH_3, and carboxyl O (arrowed) in their respective maps.

TABLE I. Probe Groups Used During Search for Potential L226Q HA Ligands*

Probe group	$E^{(m)}_p$ (kcal/mol)	$n(T)$	$n(L)$
Bromine atom	−7.5	2	2682
CH_2 methylene group	−3.5	2	999
CH_3 methyl group	−3.5	8	62204
Chlorine atom	−7.0	3	6412
Aromatic CH group	−3.5	5	127988
Fluorine atom	−4.0	5	4904
Iodine atom	−9.0	1	532
Amide NH_2 group	−7.0	6	454
Planar cationic NH_2 group	−9.0	3	117
Tetrahedral cationic NH_2 group	−9.0	2	429
Tetrahedral cationic NH_3 group	−9.0	5	704
Amide NH group	−6.0	4	3494
Planar cationic NH group	−9.0	1	313
Tetrahedral NH group with lone pair	−7.0	6	4020
Tetrahedral cationic NH group	−7.0	2	691
Nitrogen atom with lone pair	−4.0	3	11596
Anionic carboxy-oxygen atom	−6.0	7	1866
Carboxy hydroxyl group	−6.0	7	1866
Phenolic hydroxyl group	−8.0	5	447
Hydroxyl group	−6.0	13	11544
Carbonyl oxygen atom	−6.0	3	25847
Oxygen of nitro group	−6.0	7	5142
Unclassified atoms			362311
Totals		100	636562

*$E^{(m)}_p$, probe–protein interaction energy cut-off used for selection of target sites; $n(T)$, number of target sites associated with a given probe; $n(L)$, number of groups identified in the search database as being of particular probe type.

the DOCK scores for each predicted sialic acid orientation computed using the formula and parameters given in ref. 2. Note that the spurious orientation (5) would not be rejected solely on the basis of its DOCK score.

It should be pointed out that if α-sialic acid were in the CCDB, its coordinates determined by small-molecule crystallography would not necessarily agree with its HA-complexed coordinates. If the geometry were markedly different, then CLIX may fail to retrieve α-sialic acid's binding directly from the CCDB.

The search was then extended to the entire candidate database L' described above, with δ_1 and δ_2 set to 0.6 Å to make the coincidence criteria more stringent. The CPU time required for the search was 33 hr using a single processor on a Silicon Graphics 4D/240GTX IRIS workstation. A total of 1879 candidate–orientation combinations were retrieved, of which 115 had $e < -25.0$ kcal/mol. The latter set contained 76 unique potential ligands. The 10 highest scoring molecules on the short-list are described in Table III, together with their CLIX and DOCK scores, computed as described above. It is seen that CLIX is capable of extracting molecules with a wide variety of structure and size. Clearly some molecules in the short-list are of no interest (e.g., hexanitrobenzene): these can be swiftly dropped from the short-list, though ideally they should be eliminated by a more careful selection of the search database L. 5-Nitro-1-β-D-arabinofuranosyl-uracil (CCDB refcode GEMFEJ) is perhaps of more interest, given that it belongs to a class of molecule with known antiviral and antitumor activity (though not via any mechanism related to HA inhibition). Figure 5 shows the predicted orientation of this molecule in the sialic acid binding site of L226Q HA. The coordinates of this compound in the L226Q hemagglutinin frame are given in Table IV.

The CLIX scores of the short-list candidates all rank lower than that of the known ligand sialic acid. This should not be prematurely interpreted as implying poorer binding affinity. In the process of X-ray crystallographic refinement the sialic acid coordinates are optimized with respect to the crystallographic refinement energy function. This contrasts with CLIX which attempts no refinement of the score ("energy") of the molecule beyond aligning three substituent groups and ensuring no poor van der Waal's interactions.

DISCUSSION

The search algorithm CLIX has been shown to provide a simple method of searching a database of small molecule crystal structures for candidates that have the potential to show binding capability at a particular site of a protein of known three-dimensional structure. CLIX takes cognizance not only of

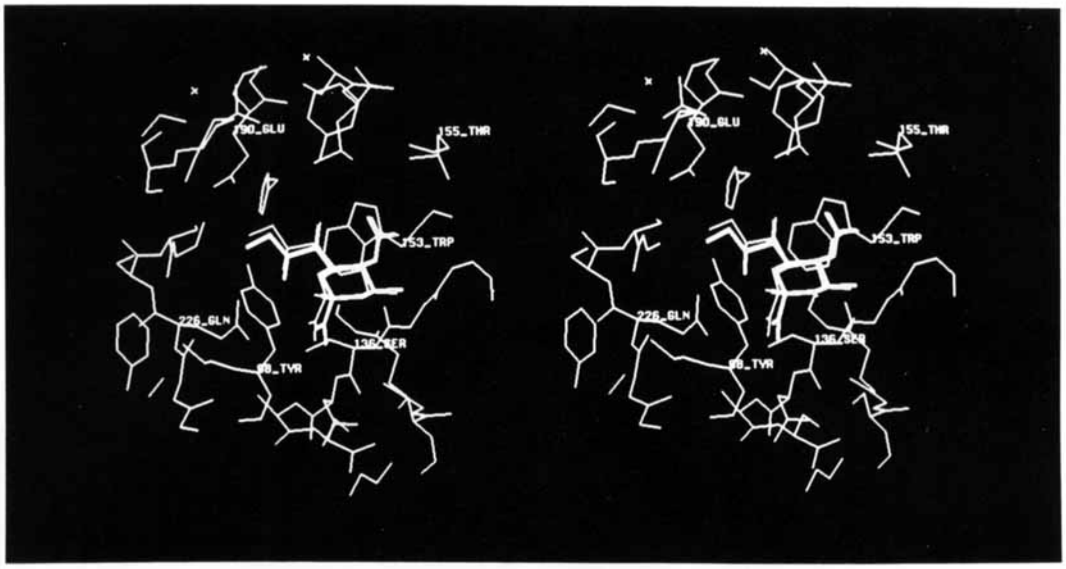

Fig. 4. CLIX-retrieved fit of sialic acid (shown in bold) in the binding site of the L226Q influenza-virus hemagglutinin mutant. The reference (crystallographic) orientation of sialic acid is shown for comparison.

TABLE II. Best CLIX Fits of Sialic Acid in L226Q HA Binding Site*

Structure	*e*	rms (Å)	Dock score
Crystal	−37.9		76
(1)	−37.3	0.41	70
(2)	−28.3	0.51	75
(3)	−27.9	0.42	78
(4)	−18.5	4.35	58
(5)	−18.0	5.35	71
(6)	−15.2	5.43	63
(7)	−12.0	6.26	50

**e*, score computed using CLIX; rms, root mean square deviation of CLIX-determined coordinates from crystal coordinates; Dock score, score computed using the DOCK formula and parameters given in Ref. 2.

the steric requirements of binding but also of the chemical requirement of achieving sufficient favorable ligand–protein interactions. Its scoring scheme, although simple, has been shown in an example to produce the correct behavior for known ligands. CLIX thus represents a considerable advance over search strategies which concentrate solely on shape (such as DOCK), and over inferential strategies (such as the pharmacophoric and quantitative structure–activity relationship methods) which concentrate on chemical similarity to known ligands without regard for steric fit in the protein receptor site.

CLIX imposes two necessary conditions for binding: steric fit and chemical coincidence of at least three candidate atoms with sites of corresponding high affinity in the protein receptor site. While sialic acid appears to bind L226Q hemagglutinin in this fashion, it is by no means obvious that the coincidence condition is not unduly restrictive in a search. Whereas affinity between individual ligand atomic groups in contact with the protein surface is anticipated, the groups do not necessarily have to lie close to pronounced local minima in the interaction energy map. It should suffice that the overall interaction energy is favorable. Preliminary studies by the authors show that sialic acid may bind influenza-virus neuraminidase according to this scheme, with only two groups (out of 10 identified with probes in Table I) being "well-placed" in the CLIX sense.

The following modification of CLIX is suggested to retrieve this more general mode of binding. The aim is to require coincidence of only two pairs (rather than three or more) of candidate atoms and target points. The algorithm is identical to the one described above except that e is computed at every rotational increment of θ that produces a sterically acceptable orientation. No check is made for coincidence of the third candidate atom/target point pair. Orientations which correspond to a local minimum e_m of $e(\theta)$ with $e_m < e_c$ are retained in the short list. This more general form of CLIX is currently being developed.

A secondary use for CLIX is to retrieve the binding geometry of a known ligand in an active site in the absence of an X-ray structure for the protein–ligand complex. As the configuration search is then limited to a single molecule, a number of trial geometries of the known ligand could be explored.

ACKNOWLEDGMENTS

We thank our colleagues Peter Colman, Jose Varghese, William Tulip, and Mark von Itzstein for

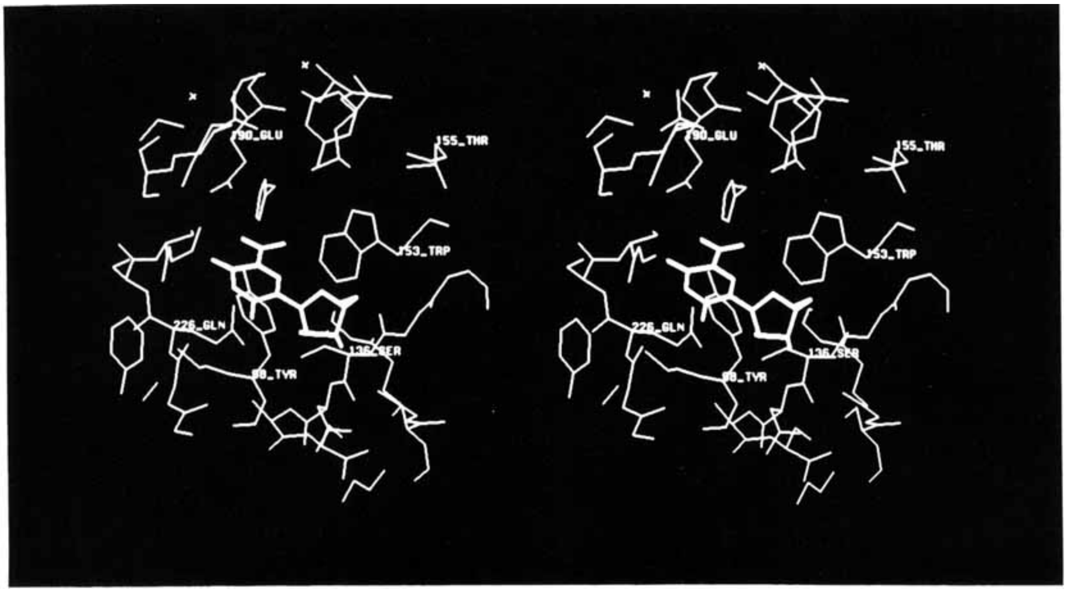

Fig. 5. CLIX-predicted fit of 5-nitro-1-β-D-arabinofuranosyl-uracil (shown in bold) in the binding site of the L226Q influenza-virus hemagglutinin mutant.

TABLE III. Highest Scoring Candidates Found During a CLIX Search for Potential L226Q Hemagglutinin Ligands*

CCDB refcode	*e*	Dock score	*n*
GAPPRHM10	−37.7	74	22
GEMFEJ	−36.7	69	20
CTCTCH	−36.6	64	27
FEPGOW	−36.0	63	18
SUCROS04	−35.3	80	23
SADDUW	−34.3	84	35
HNOBEN	−34.0	93	24
THRBIN01	−33.9	60	44
ANEUME	−33.2	64	22
OCHTET12	−33.0	59	20
(Sialic acid)	−37.9	76	21

**e*, score computed using CLIX; Dock score, score computed using the DOCK formula and parameters given in Ref. 2; *n*, number of atoms in candidate. The CCDB refcodes refer to the following molecules: GAPPRHM10, 4-*O*-β-D-galactopyranosyl-L-rhamnitol; GEMFEJ, 5-nitro-1-β-D-arabinofuranosyl-uracil; CTCTCH, cyclotricatechylene di-2-propanolate clathrate; FEPGOW, orellanine trifluoroacetic acid; SUCROS04, sucrose; SADDUW, (4*RS*,5*RS*)-2,2-dimethyl-4,5-bis(α-hydroxybenzhydryl)-1,3-dioxalane *n*-propylamine clathrate; HNOBEN, hexanitrobenzene; THRBIN01, thermorubin; ANEUME, *N*-acetylneuraminic acid methyl ester monohydrate; OCHTET12, 1,3,5,7-tetranitro-1,3,5,7-tetraazacyclooctane. Sialic acid in its experimentally determined position in L226Q hemagglutinin is included for comparison.

TABLE IV. Coordinates (in Angstrom Units) of 5-Nitro-1-β-D-arabinofuranosyl-uracil in the Orientation Predicted by CLIX as a Potential Hemagglutinin-Binding Conformation*

Atom	*x*	*y*	*z*
C1	−7.785	72.312	25.099
C2	−6.471	74.473	28.370
C3	−7.605	71.830	23.836
C4	−7.887	74.002	28.563
C5	−8.074	72.731	29.384
C6	−8.282	71.641	28.325
C7	−8.509	70.858	23.278
C8	−8.993	72.427	27.238
C9	−9.681	70.899	25.458
N10	−9.518	70.500	24.160
N11	−8.799	71.894	25.867
N12	−6.452	72.300	23.107
013	−5.661	73.412	27.879
014	−9.246	72.802	30.188
015	−7.054	71.117	27.883
016	−8.496	70.344	22.168
017	−10.530	70.451	26.184
018	−5.771	73.196	23.604
019	−6.183	71.774	22.027
020	−8.454	73.717	27.262

*The coordinates are given in the hemagglutinin mutant crystal frame.[15]

many helpful discussions, Dr. F. Allen of the Cambridge Crystallographic Data Centre for clarifying details of the CCDB connectivity tables, and Prof. D. Wiley for providing the coordinates of the L226Q HA mutant prior to its release by the Brookhaven Protein Data Bank.

REFERENCES

1. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E. A geometric approach to macromolecule-ligand interactions. J. Mol. Biol. 161:269–288, 1982.
2. DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D., Venkataraghavan, R. Using shape comple-

mentarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. J. Med. Chem. 31:722–729, 1988.

3. Goodford, P.J. A computational procedure for determining energetically favourable binding sites on biologically important macromolecules. J. Med. Chem. 28:849–857, 1985.
4. Boobbyer, D.N.A., Goodford, P.J., McWhinnie, P.M., Wade, R.C. New hydrogen-bond potentials for use in determining energetically favourable binding sites on molecules of known structure. J. Med. Chem. 32:1083–1094, 1989.
5. Jakes, S.E., Watts, N., Willett, P., Bawden, D., Fisher, J.D. Pharmacophoric pattern matching in files of 3D chemical structures: Evaluation of search performance. J. Mol. Graph. 5:41–48, 1987.
6. Brint, A.T., Willett, P. Pharmacophoric pattern matching in files of 3D chemical structures: Comparison of geometric searching algorithms. J. Mol. Graph. 5:49–56, 1987.
7. Jakes, S.E., Willett, P. Pharmacophoric pattern matching in files of 3-D chemical structures: Selection of interatomic distance screens. J. Mol. Graph. 4:12–20, 1986.
8. Goodsell, D.S., Olson, A.J. Automated docking of substrates to proteins by simulated annealing. Proteins: Struct. Funct. Genet. 8:195–202, 1990.
9. Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., Rodgers, J.R., Watson, D.G. The Cambridge Crystallographic Data Centre: Computer-based search, retrieval, analysis and display of information. Acta Crystallogr. B35:2331–2339, 1979.
10. DesJarlais, R.L., Sheridan, R.P., Dixon, J.S., Kuntz, I.D., Venkataraghavan, R. Docking flexible ligands to macromolecular receptors by molecular shape. J. Med. Chem. 29:2149–2153, 1986.
11. Connolly, M. Analytical molecular surface calculation. J. Appl. Crystallogr. 16:548–558, 1983.
12. Crippen, G.M. "Distance Geometry and Conformational Calculations, Research Studies Press." New York: John Wiley, 1981.
13. Ferro, D.R., Hermans, J. A different best rigid-body molecular fit routine. Acta Crystallogr. A33:345–347, 1977.
14. Bricogne, G. Methods and programs for direct-space exploitation of geometric redundancies. Acta Crystallogr. A32:832–847, 1976.
15. Weis, W., Brown, J.H., Cusack, S., Paulson, J.C., Skehel, J.J., Wiley, D.C. Structure of the influenza-virus haemagglutinin complexed with its receptor, sialic acid. Nature (London) 333:426–431, 1988.