Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

X-LIGAND: an application for the automated addition of flexible ligands into electron density

T. J. Oldfield

Molecular Simulations Inc., Department of Chemistry, University of York, Heslington, York YO10 5DD, England

Correspondence e-mail: tom@ysbl.york.ac.uk

Received 15 November 2000
Accepted 1 March 2001

With the advent of drug-design experiments where the interaction between a protein and a ligand is determined using X-ray crystallography, the use of automated methods for modelling the ligand into electron density represents a powerful tool. Once the protein structure has been determined by crystallography it is normal that subsequent ligand-complex structures are isomorphous, or nearly so, with the original structure and it is necessary only to determine the fit of ligand to any unsatisfied electron density. The *X-LIGAND* application was designed with this protocol in mind and provides a tool that searches for unsatisfied electron density and then fits flexible ligands to this within minutes without user intervention.

1. Introduction

The importance of defining the interaction of ligands with proteins cannot be underestimated for the understanding of many biological processes. This field has spawned a number of methods of approaching this problem using both theoretical and experimental techniques.

The best known theoretical program for defining ligand-binding sites of proteins is *GRID* (Goodford, 1985). A number of computer programs have been written subsequently that characterize ligand-binding sites and fit ligands. Some of the more common ones are *MCSS*/*HOOK* (Miranker & Karplus, 1991; Eisen *et al.*, 1994), *LUDI* (Böhm, 1992), *DOCK* (Kuntz *et al.*, 1982; DesJarlais *et al.*, 1986), *CAVEAT* (Laurie & Bartlett, 1994), *LEGEND* (Nishibata & Itai, 1993), *GROWMOL* (Bohacek & McMartin, 1994), *HIPPO*/*SPROUT*/*CAESA* (Gillet *et al.*, 1993, 1994, 1995) and *GOLD* (Jones *et al.*, 1995, 1997).

The experimental analysis of ligand interactions with proteins has spawned a number of approaches. Structure–activity relationships (SAR) by NMR studies amide chemical shifts during ligand-titration experiments (Shuker *et al.*, 1996) to provide information on the position and orientation of ligands in binding sites. X-ray crystallography can be used to determine the interaction of ligands with proteins directly by defining the atomic positions. Binding-site characterization can be carried out with multiple-solvent crystal structures (Mattos & Ringe, 1996). In this approach, protein crystals are soaked in simple organic solvents and the position of the ligands obtained from these results in an experimental multiple-copy simultaneous search (MCSS) binding-surface analysis of the protein. Another approach is to determine the crystal structure of the same protein with many differently bound ligands (Sleigh *et al.*, 1999) in an attempt to characterize less specific ligand-binding sites. All of these experimental X-ray crystallographic approaches require that either entire ligands or fragments of ligand are added to electron density.

Model-building programs used for the generation of atomic coordinate information from X-ray crystallographic data are designed to allow the placement of ligands. The most common model-building programs used for crystal structure determination include *FRODO* (Jones, 1978, 1985), *O* (Jones *et al.*, 1991; Jones & Kjeldgaard, 1997), *CHAIN* (Sack & Quiocho, 1997), *XtalView* (McRee, 1999), *MAIN* (Turk, 1992) and *QUANTA* (Oldfield, 1996; MSI). A method of real-space torsion-angle refinement has been implemented in *O* using grid summation as well as recent improvements. McRee has implemented a general real-space refinement technique within the program *XtalView*, parameterized using *xyz* and restrained by geometry. In general, these programs require that the ligand site is identified by the crystallographer and in some cases the ligand interactively placed into electron density. *X-LIGAND* was written to carry out all the steps involved in the model building of the ligand into electron density without any intervention by a user. The application is designed to determine possible binding sites using the analysis of unsatisfied electron density and then fit both rigid and flexible ligands to this with no user intervention. It is only necessary to provide the application with a macromolecule, electron-density map (usually an omit map, though any information can be used, including surfaces generated by energy calculations) and one or more ligand coordinate sets. It is even possible to provide a database of ligands to be searched and fitted to the unsatisfied map and the program returns the most likely candidate ligand.

2. Methods

The algorithms used for fitting ligands can be divided into four parts. Firstly, it is necessary to define regions within the map where ligands can be inserted. Secondly, the ligand must be parameterized to determine the rotatable bonds and geometric parameters. Thirdly, a search algorithm is used to fit flexible ligands to the electron density and finally a gradient refinement optimizes the fit.

2.1. Ligand-site searching

The first stage requires the generation of the crystallographic environment about the protein molecule as a sphere of symmetry-related atoms that extends 6 Å beyond its surface. This allows the full recognition of surface binding sites within the context of the experimental data. Next, a peak search above a threshold electron-density level within the search sphere is carried out to determine possible initial seed sites for the ligand-binding sites, though most of these are likely to be solvent if not already added. A flood fill is then performed from each of the seed sites for the map above the threshold level of electron density and truncated by non-bond interactions with the protein and by symmetry equivalence. Overlapping flood-filled sites are merged to give a unique list of possible ligand positions. Finally, the sites are sorted in descending order by volume to facilitate subsequent analysis. This process is fast, usually only a few seconds, and provides a graphical list of volumes, with the largest first.

2.2. Rotatable bond analysis

The ligand is automatically parameterized to determine the rotatable bonds and geometric restraints (Oldfield, 2001) and the search precision for each torsion angle is weighted to optimize the search efficiency. A problem encountered when searching the conformational space of most ligands is that many conformations are redundant. These conformations can be defined as those that all lie within the same energy well and would therefore converge to the same conformation using gradient-minimization methods. It is necessary only to sample a small number of these conformations during a conformation search. If a disaccharide, β-1,4-linked D-glucopyranose, is considered (Fig. 1) then there are four torsion angles to search. Two of the torsion angles are associated with the β-1,4 link between the two saccharide units and two torsion angles with the CH_2OH groups attached to the sugar ring. The two torsion angles between the monomer units are highly correlated and any change in their value will significantly affect the overall shape of the sugar. The CH_2OH groups have little impact on the overall shape of the disaccharide molecule and 50% of the torsion-angle values would refine by gradient methods to one conformation and 50% to the other conformation. The program therefore applies a precision weighting scheme for each torsion angle to define a search step size that is a function of the effective rotatable mass of the ligand affected by each torsion angle. Not only does this make the conformational search very efficient, but it also prevents the determination of many solutions where a change to a torsion angle that only affects one to two atoms fills the sample results.

To determine the rotatable bonds within the ligand molecule the application assumes that the supplied molecule has bond lengths, angles and improper torsion angles near their minima, though no assumption is made whether the rotatable bonds are at their minima. Using information on element type, connectivity and geometry, the program is able to correctly assign those bonds that would normally be rotatable at room temperature. The algorithm correctly detects the presence of

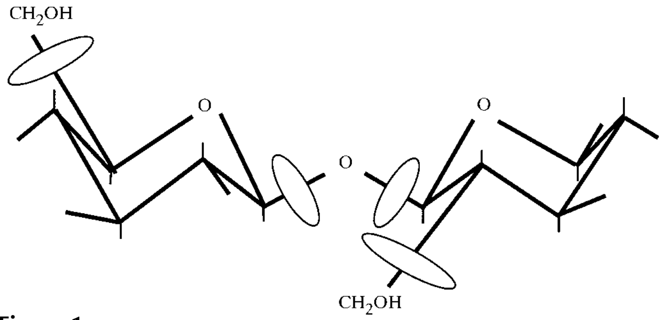

Figure 1
A disaccharide of β-1,4-linked D-glucopyranose. Thin lines mark H-atom positions and thick lines without an atom label mark the hydroxyl groups. The four ellipses show the position of rotatable bonds.

ring structures, bond orders greater than one and delocalization of electrons between atoms that would eliminate a bond from the list of rotatable bonds. The program will also use a residue-topology file (RTF) description of the ligand if one is available and can also use any previous definition entered by a user.

Each torsion angle is orientated in a 'direction' so the rotatable part of the ligand is smaller than the fixed part of the ligand and thus the maximum number of moving atoms less than or equal to half the total number of atoms. This reduces the computational overhead of rotating the atoms about the bond. The search precision of each torsion angle is defined as

$$T(i) = \tfrac{1}{4}[N_{\rm tot}/N(i)_{\rm rot}]^2,$$

where $T(i)$ is the torsion-angle precision in degrees for the ith torsion angle, $N(i)_{\rm rot}$ is the number of moving atoms for the ith torsion angle and $N_{\rm tot}$ is the total number of atoms in the ligand. For a torsion angle which rotates half the ligand, the precision is 1° and for a torsion angle which rotates 10% of the atoms, the precision is 25°. Finally, the precision of each torsion angle is rounded to the nearest angular value in degrees on the scale 1, 2, 5, 10, 20, 30, 45, 90, 120, 180°. This is to provide an integer number of sample torsion values within 360° of arc.

There are two classes of search method in *X-LIGAND*. The first considers the ligand as a free molecule that could have any position or orientation with respect to the protein atoms. The second search class uses constrained atom positions if it is known that an interaction occurs between the ligand and the protein or when it is obvious where part of the ligand is sited. The two classes of search method are further subdivided into different methods depending on a number of factors; in total there are five conformation-search methods available.

2.3. Conformation searching for flexible ligands

There are two conformation-search algorithms within *X-LIGAND* for the placement of ligands with no fixed points. The first is an exhaustive search that is used if the total calculation will not take more than 30 s; otherwise, a Monte Carlo search method is used. Generally, the Monte Carlo method is used for ligands with more than two rotatable bonds. Random conformations are generated using the random-number generator of Knuth (1981) to generate a sequence of different torsion-angle values, with a seed value defined using the time and date stamp of the computer. Efficient sampling of conformational space is a critical component of this type of calculation.

The search procedure includes optimization of non-bonding to the macromolecule, orientation, position and conformation of the ligand so as to return the best fit to the electron-density map. Typically, a ligand with nine torsion-angle degrees of freedom can be fitted in under a minute after the trial fitting of 100 000 conformations. It should be emphasized that each conformation fitted represents the optimized position and orientation of the ligand in this conformation. No searching of orientation and position is required and so this method represents a highly efficient approach to the ligand-fitting problem, allowing complex ligands of up to nine rotatable bonds to be fitted without intervention.

The following steps apply to the two algorithms for conformation searching without fixed points.

(i) Determine the centre of the electron density.

(ii) Determine the inertia tensor matrix of the density.

(iii) Generate multiple ligand conformations:

(*a*) where the total ligand conformations is 50 000 use a grid search;

(*b*) where the total ligand conformations is >50 000 use a Monte Carlo search.

(iv) For each of the ligand conformations generated:

(*a*) determine the centre of mass of the ligand;

(*b*) determine the inertia tensor matrix of the ligand.

(v) Determine a residual between the inertia tensor of the ligand and that of the density. If $I_L[V]$ is the inertia tensor matrix for the ligand and $I_D[V]$ is the inertia tensor matrix for the density, then $I_L[V] = \lambda_L[i]V_L[i]$ (for $i = 1,3$) and $I_D[V] = \lambda_D[i]V_D[i]$ (for $i = 1,3$).

To screen the conformations of the ligand against the density by shape only (so that the extent of the density is of less importance), calculate the ratios of the eigenvalues for each dimension,

$$\begin{aligned} R_L[1] &= \lambda_L[1]/\lambda_L[2], \quad & R_D[1] &= \lambda_D[1]/\lambda_D[2],\\ R_L[2] &= \lambda_L[1]/\lambda_L[3], \quad & R_D[2] &= \lambda_D[1]/\lambda_D[3],\\ R_L[3] &= \lambda_L[2]/\lambda_L[3], \quad & R_D[3] &= \lambda_D[2]/\lambda_D[3]. \end{aligned}$$

To determine the difference in shape of the ligand and density,

$$R_{\rm diff} = \left[\frac{1}{3}\sum_{i=1,3}(R_L[i] - R_D[i])^2\right]^{1/2}.$$

The best 20 $R_{\rm diff}$ values are saved during the search.

(vi) For each ligand conformation defined by the 20 $R_{\rm diff}$ values, the ligand is positioned and orientated by superposition of the tensor matrix for the ligand and the tensor matrix of the electron density. The density overlap between atoms and density is determined for the ligand as well as any non-bond interaction with the protein. Note that owing to the symmetry of the tensor matrix it is necessary to determine four fits to electron density for the ligand and thus the fit to density and orientation is taken as the best of these four fits.

The rate of the conformational search does not have a simple dependence on the number of torsion angles to change, but is linearly dependent on the total number of atoms to move to generate a new conformation. The generation of new conformations requires that each atom be moved multiple times equivalent to the number of rotatable bonds between it and the root point (Fig. 2). The total number of atoms to transform to generate a single conformation is therefore dependent on the size of the ligand, the number of rotatable bonds and the relative position of the rotatable bonds with respect to each torsion-angle root point. The generation of conformations by systematic/random changes in the rotatable bonds occurs at the rate of around 200 000 $\rm s^{-1}$ $\rm atom^{-1}$, so does not represent a rate-limiting step. A tensor matrix can be generated and compared with the electron-density tensor at around 20 000 s^{-1} atom^{-1}, while a density fit (and non-bond fit) occurs at about 1000 s^{-1} atom^{-1}. Therefore, an immediate 20-fold speed-up in the calculation is obtained by screening results using the tensor matrix before the fit to density is determined. The main calculation efficiency is obtained because the tensor match provides both a positional and orientational solution for the fit of the ligand to the electron density with just a fourfold indeterminacy. This is possible because it is only necessary to determine a ligand position, orientation and conformation that is close enough to the true final solution to be solved by real-space torsion-angle gradient refinement.

The best 20 density-fit results found during the search are displayed graphically so that a user can determine whether to stop the search when the displayed ligand solutions begin to converge to a single or small number of solutions. A time limit is also provided for the search, though this is mainly of use for database searching, where all analysis is automatic. The application allows the user to view each of the search solutions and manipulate the ligand conformation and position if necessary.

2.4. Conformation searching for ligands with fixed points

The *X-LIGAND* application provides three algorithms for the placement of flexible ligands with different numbers of fixed ligand points. For example, there may be a modified bond between a protein and its ligand or a particular atom may be known to interact with a particular region of the protein. In these cases it is possible to fix one, two or more than two atoms in the ligand and in each case a different search protocol is used. If one atom is fixed then a three-dimensional orientation search is performed about the one fixed atom and a tree search is carried out for the best 20 solutions. When two atoms are fixed in the ligand, an axis of rotation is defined linking the two fixed atoms. A one-dimensional rotation search is performed about this axis and the best 20 solutions are fitted using the tree-search algorithm. A tree search is carried out if three or more atoms are fixed. In each case, the torsion angles are both orientated and reordered so as to progressively fit the ligand away from the fixed atoms. As with the free-fit algorithms, non-bonding can be included in the analysis.

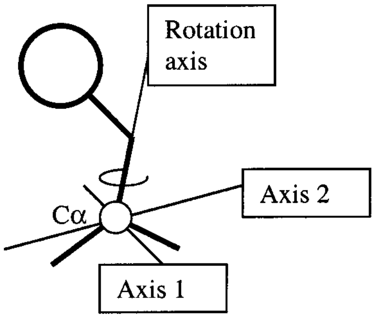

Figure 2
A simple ligand model (phenyalanine) to show the definitions of the angle flexing used within the tree search to handle error in the data and error in the seed placement of fixed atoms. Two axes are defined about the second atom of each defined torsion angle and a three-point grid search is used about each of these rotation axes. The circle indicates the root-point definition of the torsion angles.

(i) The first step requires the determination of the maximal rigid group attached to the fixed atoms. Thus, the connectivity from the fixed atoms is used to mask all those atoms 'attached' to the fixed point and not separated from the fixed point by a rotatable bond as having an occupancy of one. All remaining atoms are set to zero occupancy.

(ii) If the ligand has a single fixed point then a three-dimensional orientation search about the single fixed point with precision of 10° is performed, recording the fit to electron density (and non-bonds) in a three-dimensional grid of size $36 \times 36 \times 36$. A peak search of the grid (with surface-edge wrapping) is used to determine the best 20 orientations.

(iii) If the ligand has two fixed points then an axis between the two fixed points is generated. A one-dimensional orientation search about the axis with a precision of 10° is performed, recording the fit to the electron density (and non-bonds) in a one-dimensional array of size 36. A peak search of the array (with end wrapping) is used to determine the best 20 orientations.

(iv) A tree search is carried out for each of the 20 solutions from above or the single start point where there are more than two atoms fixed.

The tree-search algorithm progressively fits atoms starting from the fixed atoms by searching each torsion angle in turn. Since there is generally error in data and error in the seed point, flexibility in the torsion-angle tree search is allowed by providing variation in bond angle about the second atom of each torsion angle (Fig. 2). A search precision of 2° is used for the two angles defined for each torsion angle in a range of ±8°. Note that it is necessary to try all possible valence solutions from a previously fitted torsion angle. This is because it is not possible to determine which branch to progress along until the next atom set is fitted in the tree search. For example, when fitting the amino acid isoleucine it is necessary to try two possible trees from χ_1, as both are equivalent until χ_2 is determined. Since the tree search can distort the angular geometry of the ligand, it is necessary to carry out geometry refinement using the method described in Oldfield (2001) of all the ligand solutions on completion of the search.

3. Real-space torsion-angle gradient refinement

Since the solution of the conformation search is only approximate the ligand must be further refined. Real-space torsion-angle refinement (RSTR; Diamond, 1971; Chen *et al.*, 1999) using the algorithm described in Oldfield (2001) finishes off the ligand-fitting process as this provides a large radius of convergence and so represents an efficient method of fitting a ligand to electron density. The solution to the gradient represents a final fitted solution for the ligand that may or may not need subsequent refinement using standard reciprocal-space refinement programs.

Table 1
Table to show five different test calculations and the results of the ligand fitting.

No.†	Ligand/protein‡	Resolution (Å)§	Method¶	$N_{\rm atom}$††	$N_{\rm tors}$‡‡	Time (s)§§	Rate¶¶	R.m.s.d. (Å^2)†††
1	Methylparaben/insulin	1.9	Exhaustive	11	2	4	3800	0.02
2	βKDO/OppA	2.2	MC	25	7	47	990	0.02
3	βKDO/OppA	2.2	MC	25	9	63	950	0.17
4	KKK/OppA	1.4	MC + edit	28	19	382	540	0.11
5	KKK/OppA	1.4	FP(3) + edit	28	19	136	N/A	0.11
6	HDS/phospholipaseA	2.9	MC	20	15	312	1570	0.73
7	HDS/phospholipaseA	2.9	FP(0)	20	15	7	N/A	0.77

† Test calculation No. ‡ The ligand and protein used in the calculation. § Resolution of the experimental data. ¶ Search method used to fit the ligand. Exhaustive search tries all possible conformations but is only suitable for small ligands. Monte Carlo (MC) searching involves fitting a random sample of conformations either within a time limit or until convergence of the best 20 solutions occurs. Fixed point [FP(3)] is the three-dimensional orientation search plus tree fit, while the fixed point [FP(0)] is a tree search. The edit in tests 4 and 5 indicates that user intervention was required to complete the fitting. †† No. of atoms in the ligand. ‡‡ No. of torsion angles fitted. §§ Total time to fit the ligand; this includes the time for the peak search, conformation search, refinement and any user editing. ¶¶ Rate of conformation sampling per second. ††† R.m.s.d. of ligand with respect to the published coordinates.

X-LIGAND interface

The functionality of *X-LIGAND* is provided with a graphical user interface (GUI) and has the same style of interface and parameterization as the other X-ray tools of *QUANTA*98 (MSI): *X-AUTOFIT*, *X-POWERFIT*, *X-BUILD* and *X-SOLVATE*. The user has a palette containing 23 different tools, although most ligands can be fitted using three of the tools just once.

The parameter GUI allows a number of settings to be changed, but the only requirement is the setting of the threshold of the electron-density level that the application uses to flood fill each site. The aim is to define a volume approximately the size of the ligand, although the algorithm is very robust with respect to this as the ratio of principal components is screened and not absolute values. The application determines the volume of the flood-fill site and shows this volume graphically to the user. The default action is to search the whole map, but it is possible to focus on a small region. This is particularly important when model building to an incomplete macromolecule, in which case the search for unsatisfied electron density has to be localized.

Three tools allow the selection of the density site to be fitted: next site, previous sites or go to a particular site. Since the unsatisfied density potential sites are sorted by volume, in most cases the first site found by the search algorithm is the

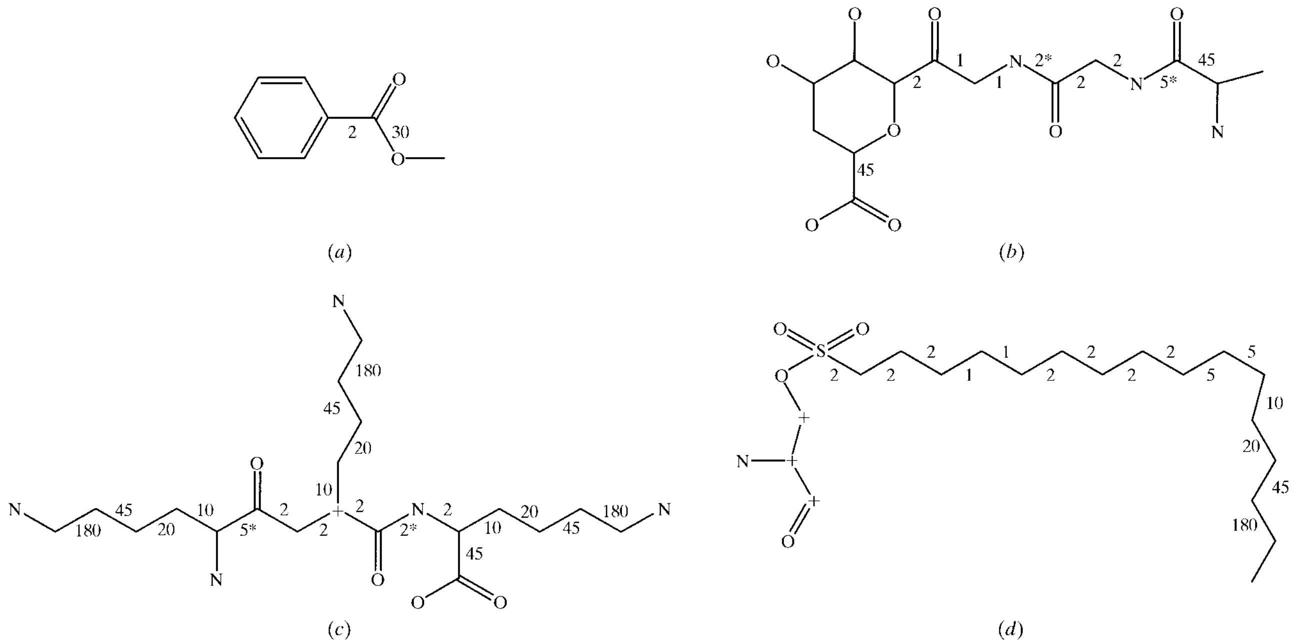

Figure 3
(*a*) Two-dimensional representation of the ligand methylparaben, with the search-angle precision marked on each bond searched during fitting. (*b*) Two-dimensional representation of the ligand 3-deoxy-β-D-*manno*-*s*-octulopyranosionic acid, with the search-angle precision marked on each bond searched during fitting. The precision numbers marked with a * are peptide bonds and were not searched in test calculation 2 but were searched in test 3. (*c*) Two-dimensional representation of the ligand Lys-Lys-Lys with the search-angle precision marked on each bond searched during fitting. The ligand is orientated in approximately the same way as the ligand coordinates shown in Figs. 6 and 7 for clarity. The precision numbers marked with a * are peptide bonds which were searched. (*d*) Two-dimensional representation of the ligand hexadecane sulfonyl (HDS) covalently attached to a serine residue with the search-angle precision marked on each bond searched. The fixed atoms within the fixed-point ligand-fitting search are shown with a +.

best site for a ligand, so the application places the view, ligand (fitted using the tensor alignment) and map at the first site after a site search.

If the ligand has no internal degrees of freedom, then placement of the view plus ligand represents the completion of the ligand-fitting process because the ligand is fitted automatically to a potential site in the current conformation whenever the first or a new site is selected. It is usually sensible to refine the ligand by real-space gradient refinement using the tool provided at this point.

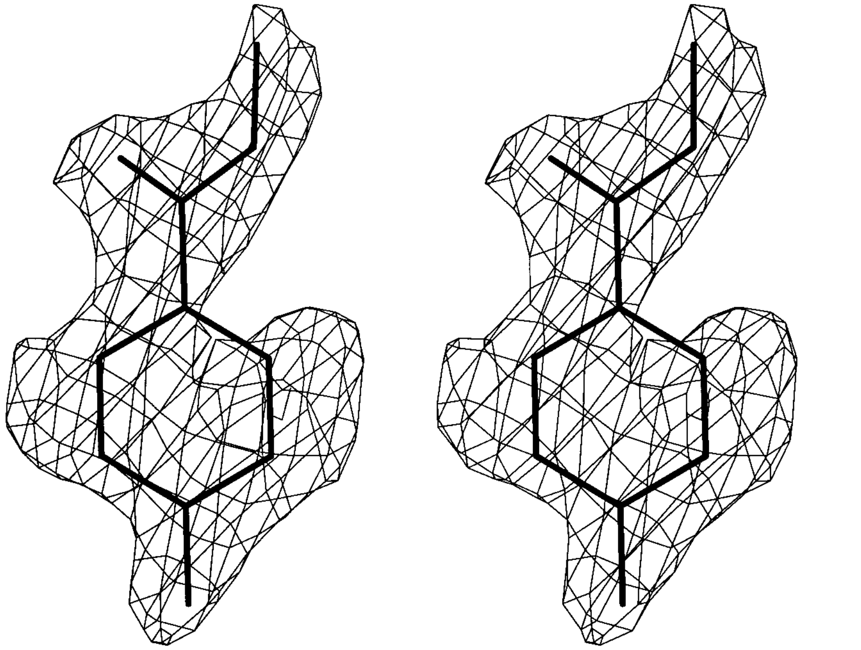

Figure 4
Stereo figure of the ligand methylparaben shown fitted to $2F_o - F_c$ electron density of resolution 1.9 Å contoured at the search level of 1.5σ.

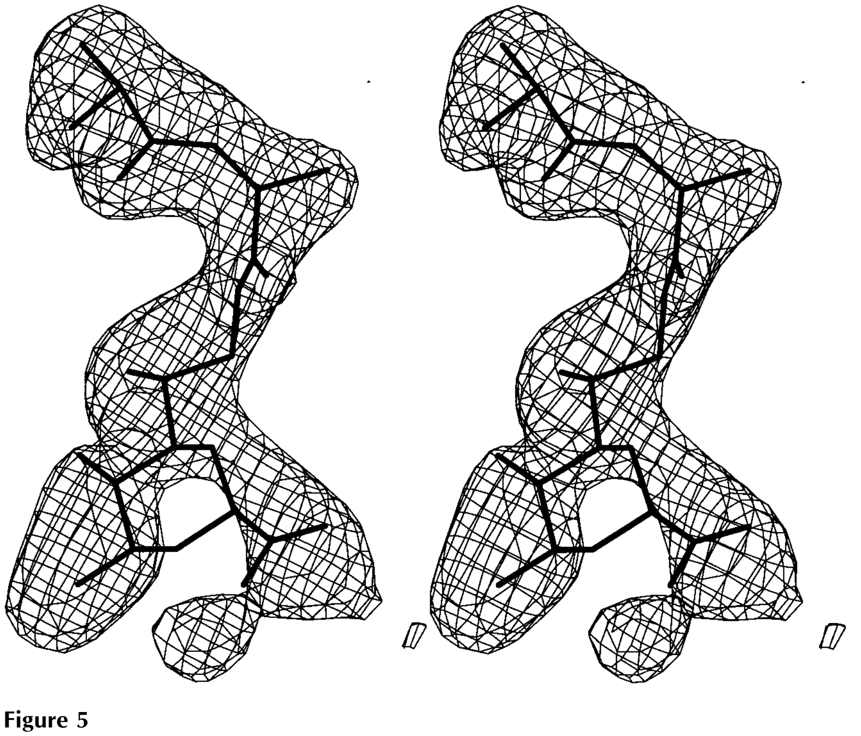

Figure 5
Stereo figure of the ligand 3-deoxy-β-D-*manno*-*s*-octulopyranosionic acid shown fitted to 2.2 Å difference electron density at 3.5σ.

If the ligand has internal degrees of freedom, then the tools for conformation searching become available. It is possible to fix points in the ligand using a tool that requests atom picking of the ligand. Each atom picked (if not already fixed) is defined as fixed; if picked a second time then this becomes free. A single tool allows all fixed points to be removed. On completion of a conformation search the user can step through the best 20 solutions or look at all 20 returned best solutions. Finally, a solution can then be refined by real-space torsion-angle refinement.

Tools are provided to edit the rotatable bonds, to define the active torsion angles to be searched, edit the position, edit the torsion angles, save the results, define a new ligand and to carry out a database search.

5. Results

Four sets of experimental data are used to demonstrate the application and to indicate the limits of the algorithm presented here. All the data were fitted using the exhaustive grid search (EG) or the Monte Carlo (MC) conformation searches, even where this would not be the natural choice for a ligand covalently bound to a protein. In addition, two ligands were fitted using the three-dimensional orientation search plus tree search [FP(3)] and a third was searched with a fixed group tree fit [FP(0)]. No example is shown using the one-dimensional axis search plus tree-search fitting. All fitting was completed using real-space torsion-angle gradient refinement (RSTR) against the electron density. Seven test calculations were carried out in all.

The first ligand (methylparaben) is small with two internal degrees of freedom and is bound to pig insulin at 1.9 Å (Whittingham *et al.*, 1995). Fig. 3(*a*) shows a two-dimensional cartoon of the molecule with the search precision marked on each torsion angle searched, while Fig. 4 shows the final result of the analysis. The second ligand example (3-deoxy-β-D-*manno*-*s*-octulopyranosionic acid; β-KDO) has seven internal degrees of freedom and two peptide bonds bound to oligo-peptide binding protein (OppA) (J.

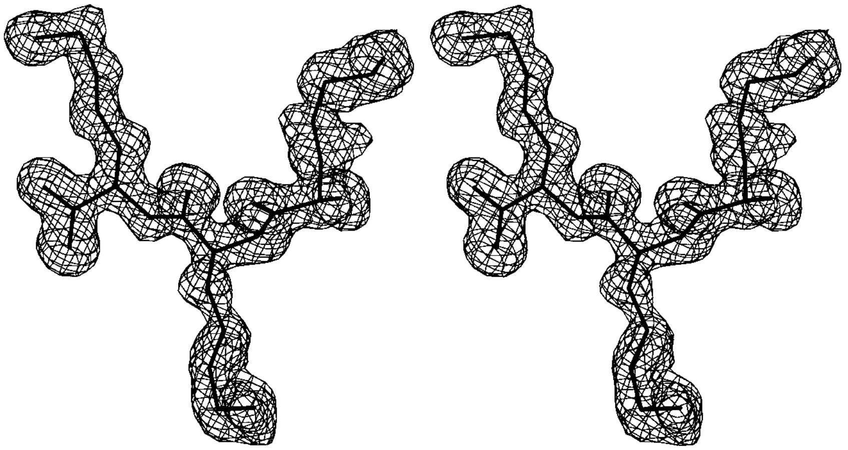

Figure 6
Stereo figure of the Lys-Lys-Lys tripeptide shown fitted to 1.4 Å difference electron density at 3.5σ. The ligand coordinates show the result from the procedure in test calculation 4, although the fit solution obtained in example 5 is within a line thickness of that shown in this figure.

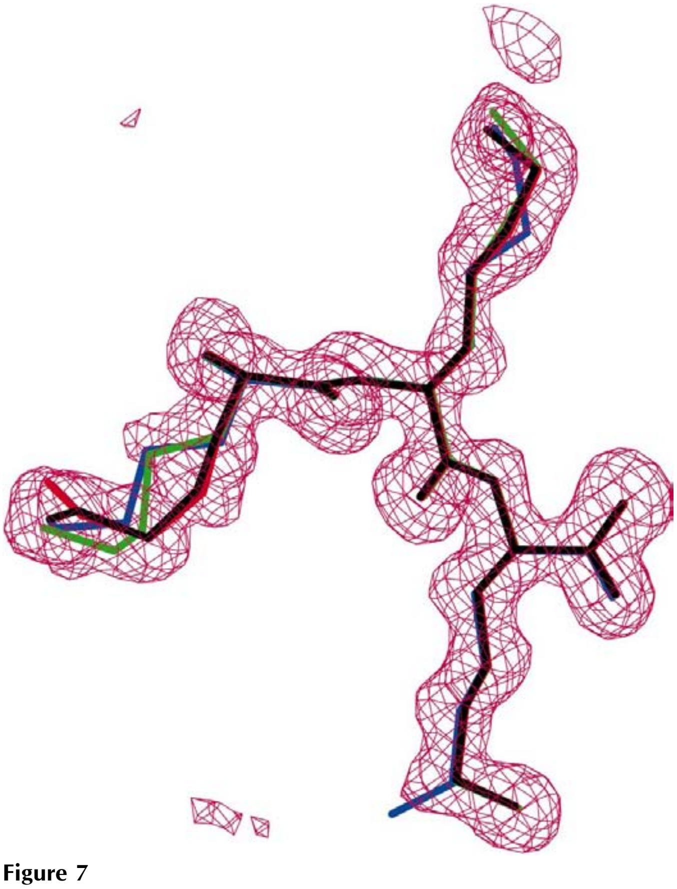

Figure 7
Figure to show the ligand Lys-Lys-Lys after each of the four stages of fitting as described in the text for test calculation 4. The first 5 min search and RSTR produce the blue coordinates, the second 1 min search and RSTR resulted in the green coordinates, the third search and RSTR resulted in the red coordinates and the user edit plus RSTR resulted in the black coordinates.

Tame, personal communication). Test calculations were carried out with the peptide bonds both fixed and searched and the results are shown in Table 1 as tests 2 and 3. Fig. 3(*b*) shows two-dimensional representation of the ligand and Fig. 5 shows the final results of fitting the ligand to experimental data when the peptide bonds were fixed. The results of the analysis with the peptide bonds searched are not shown, as they are similar to the results obtained when not searched. Test calculation 4 demonstrates the limit of a free conformation search with a large flexible Lys-Lys-Lys tripeptide (KKK) also bound to the protein OppA (Tame *et al.*, 1995). This ligand has 17 internal degrees of freedom as well as two peptide bonds; all 19 were searched. The test calculation required user intervention, as the simultaneous search of the 19 torsion angles did not converge. The ligand was fitted using the following procedure: 19 torsion angles were searched for 5 min followed by RSTR, seven torsion angles searched for 1 min followed by RSTR, five torsion angles searched for 1 min followed by RSTR, a user edit and RSTR. The best solution was taken from each MC search and used as the starting point for each subsequent search after RSTR. Fig. 3(*c*) is a two-dimensional representation of the ligand. The progressive fitting of the ligand can be seen in Fig. 7 and the final conformation is shown in stereo in Fig. 6. The fifth calculation test used the same coordinate structures as that used in test 4 but was fitted using the FP(3) search. The ligand was initially positioned so that the C$^{\alpha}$ atom of the middle lysine residue is correctly placed in the electron density (marked with a '+' in Fig. 3*c*) and orientated randomly. The ligand required an edit of the second lysine side chain before final RSTR to complete the fit. The final fitted ligand is not shown, as there is no significant difference to the result shown in Fig. 6. The fourth set of data used to demonstrate *X-LIGAND* is a hexadecane sulfonyl (HDS) covalently attached to the active-site Ser144 of the outer membrane phopholipase A from *Escherichia coli* (Snijder, 2001). Data was collected to 2.7 Å, but owing to radiation damage data is only 20.3% compete at this resolution. Therefore, it is better described as a 2.9 Å data set (completeness of 91.6%). A two-dimensional representation of the ligand and serine residue with the search precision of the searched torsion angles is shown in Fig. 3(*d*). The experimental map for the ligand was prepared by removing all solvent, ligand and Ser144 from the coordinate set and carrying out a single cycle of refinement. The ligand was fitted to the difference data at 1.5σ. The ligand was fitted using the MC search for 5 min and also with the main-chain atoms of the serine residue fixed and therefore the lipid fitted with the FP(0) search algorithm. Fig. 8 shows the result of two fitting algorithms against the ligand coordinates fitted by the crystallographer. It would be expected that the single fixed-point algorithm would normally be used, as the position of the main-chain atoms of the serine would already be known from structure tracing.

The sensitivity of the *X-LIGAND* algorithm to the extent of the ligand density was tested by fitting β-KDO to OppA with

Table 2
Calculations to indicate the sensitivity of the algorithm to density volume.

Density level (σ)	Site volume (Å^3)	Deviation (Å^2)
1.0	280	3.85
1.5	206	2.25
2.0	171	1.78
2.5	144	0.19
3.0	119	0.16
3.5	97	0.17
4.0	79	1.22
4.5	55	2.78
5.0	40	7.49

† Electron-density level as defined by the number of σ within a $F_o - F_c$ map. ‡ The volume of the electron density calculated as the number of flood-fill points determined by *X-LIGAND* to be part of the ligand electron density at a particular σ level multiplied by the volume of a map grid polyhedron. § The deviation of the ligand from the coordinates provided by J. Tame.

different electron-density σ levels used in the flood-fill calculation. The binding site of OppA is large and is able to bind many different ligands (Tame *et al.*, 1995). Changing the flood-fill search threshold resulted in significant changes of volume and shape of the target density within the binding site. The result of this analysis is shown in Table 2. The ligand was placed at coordinate (0, 0, 0) in real space at the beginning of each calculation and the torsion angles were set to random values.

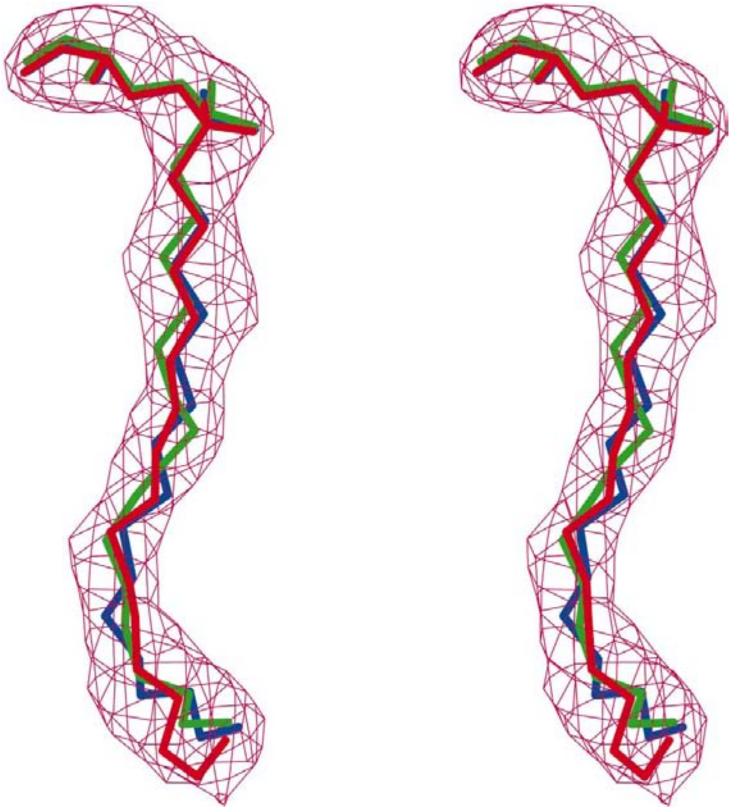

Figure 8
Stereo figure of the ligand HDS shown with difference electron density at 1.5σ. The blue coordinates were provided by A. Snijder, the red coordinates are those generated by fitting the serine atoms and using a tree search for the lipid, and the green atoms were generated using a free conformation search.

6. Discussion

The examples show that the algorithms can be used with small and large ligands, as well as different quality experimental data. Ligand molecules with two rotatable bonds can be fitted with the EG search and ligands with up to nine rotatable bonds can be fitted with no user intervention with the MC search algorithm. The ligand HDS, with 15 rotatable bonds, was fitted to the experimental data with no user intervention, but the solution obtained by Snijder was not reproduced within expected error at this resolution. More flexible ligands can be fitted with user intervention by masking search torsion angles as they are progressively fitted. The tree-search algorithm produces good results with more flexible ligands, although it does require some knowledge of the position of one or more atoms of the ligand.

The second test calculation resulted in two significant populations of conformations during the search. The second, minor, solution (results not shown) was an alternate conformation previously observed (J. Tame, personal communication) where the sugar ring is rotated approximately 180° with respect to the rest of the ligand. After approximately 20 000 conformations were searched, this alternate sugar-ring position disappeared from the top 20 solution results because the fit to density is significantly lower than the principal solution. The fact that an alternate sugar-ring conformation persisted in the 20 search results demonstrates the ability of torsion-angle precision filtering to present useful alternate conformation results to the user.

The third ligand, KKK, with 19 rotatable bonds and hence 25 degrees of freedom, represents a ligand that was too complex for the complete automation of solution determination by free fitting. It was observed that during the calculation the core of the ligand fitted first; after 5 min approximately half of the ligand was fitted in an obviously correct conformation (Fig. 7). Thus, it is possible to fit large complex ligands that would generally be considered outside the scope of this type of search algorithm, but in fact can be progressively fitted by manipulation of the searched torsion set and minor user editing. It was necessary to edit the ligand torsion angle in this example because RSTR fails to shift atoms connected by a large number of torsion angles because of correlation between these torsion angles (Diamond, 1971). The edit need only change the torsion angle a small amount to allow the refinement to proceed to completion. The definition of a fixed point shows that with a little prior thought it is possible to greatly simplify the ligand-fitting calculation. The central lysine side chain had to be edited to complete the fit after using the FP(3) fitting.

The fourth ligand was a lipid molecule covalently attached to an active-site serine residue. The experimental data used to fit the ligand was of relatively low resolution, with the result that the position of all the C atoms in the chain could not be determined with certainty. It was expected from a theoretical consideration that many of the chain torsion angles would be in a *trans* conformation as this represents an energy minimum for this type of molecule; this proved to be the case. The conformation search with no fixed points was performed with nonbonding between the ligand and protein turned off so that the serine residue would not have been displaced by a clash with residues adjacent in the protein sequence. The MC conformation fit search (Fig. 8, green) shows that the algorithm correctly orientated the molecule in the experimental data, with ligands, atoms close to the coordinates provided (Fig. 8, blue). The results of fitting HDS appeared to indicate that ligands with up to 15 rotatable bonds can be fitted, but it should be noted that this ligand has relatively few atoms and was therefore searched at the higher rate of 1570 conformations s^{-1} compared with the βKDO ligand (Table 1). The table also shows that the search rate has no simple dependence on the number of torsion angles to search, but shows a correlation to the number of atoms within the ligand. As stated before, the rate of search is directly proportional to the total number of atom moves. The FP(0) fitting produced a result with a differently fitted tail (Fig. 8, red) and this persisted even with small changes of the coordinates of the serine atoms from which the lipid chain was built. The conformation of the last two atoms of the lipid tail was unlikely and a user edit of the last torsion angle of the chain followed by RSTR produced a better fit (not shown). The ligand torsion angles were all approximately *trans*, though orientated differently, and bend at the lipid tail differently. It was not possible to define the correct conformation owing to the resolution of the data, although deviations of the coordinates fitted with *X-LIGAND* with respect to the coordinates provided were a little higher than would be expected at this resolution.

The analysis of sensitivity of the algorithm to electron-density search volume indicated that the algorithm is convergent for search volumes between 97 and 144 Å^3. The ligand volume is approximately 105 Å^3 not including H-atom volumes. The method is more sensitive to undersized target volumes of electron density. The breakdown of the algorithm was the result of the electron-density volume reducing/extending so that the ratio of the principal components of the electron density was significantly different to the principal components of the correct ligand conformation. Since the binding site of OppA is extensive and has large voids not filled by this ligand, the target volume rapidly became distorted at lower flood-fill thresholds. The algorithm is robust within a range of 0.9–1.4 times the volume of the ligand, as can be seen from Table 2.

Alternate conformations can be determined, but extensive additional electron-density volume that is not similar to a single conformation of the ligand can prevent a sensible solution from being automatically found by this type of algorithm. This algorithm is not suitable where the experimental data looks nothing like any conformation possible for the ligand coordinates.

The algorithm described provides a means to fit ligands with up to nine rotatable bonds entirely automatically to data without significant error as shown with the second ligand. The same algorithm can be used to determine a correct solution for more flexible ligands using multiple stages of analysis, as the method tends to fit a core structure first. This is shown with test calculation four with ligand three. The largest ligand so far used within *X-LIGAND* had 24 rotatable bonds and was fitted progressively as in test calculation 4. The use of the tensor to parameterize the data and ligand solves the positional and orientation problem of the ligand, resulting in a highly efficient search method. This can of course result in a limitation of the algorithm, as ligands that look very different to the electron density cannot be fitted this way. The fixed-point algorithm provides a means to fit flexible ligands where knowledge of an atom coordinate is known. Test calculations 5 and 7 demonstrate the use of this type of ligand fitting to electron density.

7. Variations on a theme

X-LIGAND can be used to place multiple occurrences of the same ligand (multiple sites or alternate conformations) or multiple ligands. It is necessary only to move to each unsatisfied electron-density site (with a single tool of the GUI) and then select the tool to fit one of the ligands at this point.

One variation of the algorithm available in *QUANTA*98 is the ability to provide a database of ligands, either as a skeleton format file (MSI) or as a list of file names. The latter would be useful where all constituents of the crystallization conditions are known and can be provided to the program. The algorithm then scans all the ligands in the list/database, carrying out a conformation search where necessary and scoring the fit quality for each ligand. The best ligand or the best 20 ligands can be returned fitted to the site.

Another variant of the algorithm has been implemented but is not available in *QUANTA*98. This is the replacement of the experimental data in the form of electron density with that of a potential energy surface determined from the protein atoms. The possible binding sites within the protein and at the surface of the protein are automatically determined and in the latter case optimized for the ligand size. The ligand can then be fitted to this using some extensions to improve the positional search variation owing to the limitation of the tensor matching. This variant of the algorithm can search approximately 500–1000 conformations per second and has successfully fitted many modelled ligands (C. S. Verma, personal communication).

8. Availability

The *X-LIGAND* functionality is part of *QUANTA*98, except for the fixed-point conformation-search algorithms, which will be available in the next release. The application can be obtained from MSI. The modelling variant that fits ligands to potential surfaces has been made available within Cerius 2 (MSI), although this is changed from the original algorithm.

I would like to thank Arjan Snijder, Jean Whittingham and Jeremy Tame for providing experimental data used within this paper. I would like to thank the members of the York Structural Biology group for both testing the algorithm and making suggestions to improve the ease of use. I would like to thank Eleanor Dodson and Garib Murshudov for help with writing this paper, and Eleanor Dodson, Garib Murshudov and Kevin Cowtan for discussions on mathematics and crystallography. This work was entirely funded by Molecular Simulations Inc.

References

Bohacek, R. S. & McMartin, C. (1994). *J. Am. Chem. Soc.* **116**, 5560–5571.

Böhm, H. J. (1992). *J. Comput. Aided Mol. Des.* **6**, 593–606.

Chen, Z., Blanc, E. & Chapman, M. S. (1999). *Acta Cryst.* D**55**, 464–468.

DesJarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D. & Venkataraghavan, R. (1986). *J. Med. Chem.* **29**, 2149–2153.

Diamond, R. (1971). *Acta Cryst.* A**27**, 436–452.

Eisen, M. B., Wiley, D. C., Karplus, M. & Hubbard, R. E. (1994). *Proteins Struct. Funct. Genet.* **19**, 199–221.

Gillet, V. J., Johnson, A. P., Mata, P., Sike, S. & Williams, P. (1993). *J. Comput. Aided Mol. Des.* **7**, 127–153.

Gillet, V. J., Myatt, G., Zsoldos, Z. & Johnson, A. P. (1995). *Perspect. Drug. Discov. Des.* **3**, 34–50.

Gillet, V. J., Newell, W., Mata, P., Myatt, G., Zsoldos, Z. & Johnson, A. P. (1994). *J. Chem. Inf. Comput. Sci.* **34**, 207–217.

Goodford, P. J. (1985). *J. Med. Chem.* **28**, 849–857.

Jones, G., Willett, P. & Glen, R. C. (1995). *J. Mol. Biol.* **254**, 43–53.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). *J. Mol. Biol.* **267**, 727–748.

Jones, T. A. (1978). *J. Appl. Cryst.* **20**, 115–157.

Jones, T. A. (1985). *Methods Enzymol.* **115**, 157.

Jones, T. A. & Kjeldgaard, M. (1997). *Methods Enzymol.* **227**, 173–230.

Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Knuth, D. E. (1981). *Seminumerical Algorithms*, 2nd edition, Vol. 2. Reading, MA, USA: Addison–Wesley.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982). *J. Mol. Biol.* **161**, 269–288.

Laurie, G. & Bartlett, P. A. (1994). *J. Comput. Aided Mol. Des.* **8**, 51–66.

McRee, D. E. (1999). *J. Struct. Biol.* **125**(2–3), 156–165.

Mattos, C. & Ringe, D. (1996). *Nature Biotechnol.* **14**, 595–559.

Miranker, A. & Karplus, M. (1991). *Proteins Struct. Funct. Genet.* **11**, 29–34.

Nishibata, Y. & Itai, A. (1993). *J. Med. Chem.* **36**(20), 2921–2928.

Oldfield, T. J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 67–74. Warrington: Daresbury Laboratory.

Oldfield, T. J. (2001). *Acta Cryst.* D**57**, 82–94.

Sack, J. S. & Quiocho, F. A. (1997). *Methods Enzymol.* **227**, 158–173.

Shuker, S. B., Hajduk, P. J., Meadows, R. P. & Fesik, S. W. (1996). *Science*, **274**, 1531–1534.

Sleigh, S. H., Seavers, P. R., Wilkinson, A. J., Ladbury, J. E. & Tame, J. R. H. (1999). *J. Mol. Biol.* **291**, 393–415.

Snijder, H. J. (2001). Submitted.

Tame, J. R. H., Dodson, E. J., Murshudov, G. & Higgins, C. F. (1995). *Structure*, **3**, 1395–1406.

Turk, D. (1992). PhD thesis. Technische Universität München, Germany.

Whittingham, J. L., Chaudhuri, S., Dodson, E. J., Moody, P. C. E. & Dodson, G. G. (1995). *Biochemistry*, **34**, 15553–15563.