Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Modelling bound ligands in protein crystal structures

**P. H. Zwart,‡ G. G. Langer and V. S. Lamzin***

European Molecular Biology Laboratory, c/o DESY, Notkestrasse 85, Building 25A, 22603 Hamburg, Germany

‡ Present address: SAIC-Frederick/Department of Biology, Argonne National Laboratory, Argonne, IL 60439, USA.

Correspondence e-mail: victor@embl-hamburg.de

Received 26 February 2004
Accepted 28 May 2004

Methods for automated identification and building of protein-bound ligands in electron-density maps are described. An error model of the geometrical features of the molecular structure of a ligand based on a lattice distribution of positional parameters is obtained *via* simulation and is used for the construction of an approximate likelihood scoring function. This scoring function combined with a graph-based search technique provides a flexible model-building scheme and its application shows promising initial results. Several ligands with sizes ranging from 9 to 44 non-H atoms have been identified in various X-ray structures and built in an automatic way using a minimal amount of prior stereochemical knowledge.

## 1. Introduction

Automated model-building techniques in protein crystallography form an essential component of any hardware and software pipeline that is aimed at delivering protein crystal structures with minimum user intervention (*e.g.* Brunzelle *et al.*, 2003). Model-building routines such as *ARP*/*wARP* (Perrakis *et al.*, 1999), *RESOLVE* (Terwilliger, 2003) and *MAID* (Levitt, 2001) are able to construct almost complete protein structures in a fully automated manner (Badger, 2003) given a set of reasonable phase estimates and X-ray data of sufficient resolution. Although the protein part of a structure is recognized, other compounds, such as DNA, RNA and ligands, cannot be built fully automatically. The problem of ligand building is of particular interest both from a theoretical and a practical point of view. The chemical variety of ligands bound to proteins is enormous: at the time of writing, more than 4000 entries are present in the Hetero-compound Information Centre (HIC-Up; http://xray.bmc.uu.se/hicup) and over 2000 ligand dictionary entries are contained in the *REFMAC5*/*CCP*4 monomer library (Vagin *et al.*, 2003; Collaborative Computational Project, Number 4, 1994). Finding a means of handling the basic chemical knowledge of ligands in the interpretation of electron densities at a resolution lower than atomic resolution and with phase error present is particularly challenging. The practical interest stems largely from pharmaceutical companies and large-scale X-ray crystallography facilities that desire to automate drug-discovery efforts or to build up a general infrastructure for structure solution. Ligand-building procedures play a central role in the automation and practical feasibility of high-throughput X-ray crystallographic screening for lead identification and optimization, as carried out by, for example, Abbott (Nienaber *et al.*, 2000) and Astex Technology (Sharff & Jhoti, 2003). Existing methods for construction of non-protein models are either based on the use of torsion angles, interatomic distance matrices or on topological analysis of the electron density. The

methods implemented in *XLIGAND* (Oldfield, 2001*b*) or *BLOB* (Diller *et al.*, 1999) fit ligands to the electron density by varying the torsion angles. *XLIGAND* performs a shape matching and requires initial guesses of the location of the ligand obtained *via* segmentation of the difference density. A ligand molecule is placed into the density in several trial conformations and a local optimization to maximize the fit to the electron density is carried out (Oldfield, 2001*a*). *BLOB* utilizes global optimization techniques to find the orientation, location and conformation of the ligand. An example of a distance matrix-based interpretation technique is the pioneering work of Koch (1974) and extensions thereof (Main & Hull, 1978; Cascarano *et al.*, 1991; Altomare *et al.*, 2002). These distance matrix-based map-interpretation methods use iterative procedures for the construction of molecular models in maps on the basis of known geometrical features and approximate atomic positions obtained by peak-picking methods. Recently, distance matrix-based methods have also been applied to the interpretation of high-resolution protein electron-density maps (Oldfield, 2002). The interpretation of electron-density maps *via* a topological analysis of electron density additionally invokes other topological features in the interpretation process such as pits and saddle points (Leherte *et al.*, 1997; Menéndez-Velázquez & García-Granda, 2003). Although all three methods have their specific advantages, we chose to investigate ligand-building techniques on the basis of distance matrices because of their close link to the model-building techniques implemented in *ARP*/*wARP*. Furthermore, distance-matrix approaches may allow the construction of algorithms for building of partially disordered ligands in a more straightforward way than using torsion angle-based approach.

Although building of ligand structures in electron density may seem to be a different problem to building a protein on the basis of repetitive peptide motifs (Lamzin & Wilson, 1997), it can be shown that the underlying principles are based on the same concepts (Bart & Busetti, 1976).

In a crystallographic restrained refinement, the following function is optimized by varying the atomic positions $\{\mathbf{x}\}$:

$$\mathrm{LL}(\{\mathbf{x}\}) = \ln[f(\text{chemical sense}|\{\mathbf{x}\})] + \sum_{\mathbf{h}} \ln[f(F_{\mathbf{h}}^{\mathrm{obs}}|\{\mathbf{x}\})]. \quad (1)$$

The $f(F_{\mathbf{h}}^{\mathrm{obs}}|\{\mathbf{x}\})$ term in (1) models the probability distribution of the X-ray data given the estimated set of atomic positions $\{\mathbf{x}\}$. $f(\text{chemical sense}|\{\mathbf{x}\})$ expresses the prior knowledge of the stereochemistry of the system. In protein crystallography, this expression is usually modelled by the product of a set of Gaussian distributions centred on the 'ideal' values of geometrical features such as distances and angles. When $f(F_{\mathbf{h}}^{\mathrm{obs}}|\{\mathbf{x}\})$ is also modelled by a Gaussian, (1) results in a standard least-squares refinement. Modelling the X-ray part of (1) by a Rice distribution results in the so-called maximum-likelihood refinement (Pannu & Read, 1996; Bricogne, 1997*a*; Murshudov *et al.*, 1997).

The approach we adopted for ligand building is related to the described refinement example. However, instead of varying the positional parameters for optimization of the total log likelihood (LL), we keep them fixed and the interpretation in the form of a set of atomic labels is modified to optimize (1). Furthermore, although we model the prior distribution by a (weighted) sum of independent log-probabilities, the individual probability density functions do not have a Gaussian form. As is the case for the amplitude part of (1), the prior of our chemical sense can be derived on the basis of a suitable error model of the positional parameters. The stereochemical quality of an interpretation is gauged by the modelled distribution of the geometric features, but the correspondence to the X-ray data is accounted for in a simpler way. The likelihood of an atom has been modelled by a monotonically increasing function of the density height in order to drive the interpretation towards high electron density. Owing to the approximations in the developed function, we will use the generic term scoring function rather than log-likelihood.

From a wide spectrum of various topological and geometrical descriptors, we only use the information from bonded atoms (1–2 distances), bonding angles (1–3 distances), the chirality of the atoms and van der Waals repulsions. Although a large number of other sources of information, such as planarity restraints, *cis–trans* specifications, possible intramolecular hydrogen-bonding patterns and unfavourable combinations of specific torsion angles are ignored, the number of geometrical descriptors in combination with the electron density allows one to obtain a suitable estimate of the position, orientation and conformation of the ligand.

## 2. Methods

The following prototype procedure has been developed for the automated building of ligands in residual electron density. Firstly, the protein part of a macromolecular model is subjected to manual or automatic remodelling of side-chain and or main-chain conformations to overcome ligand-induced non-isomorphism. This structure is then used to obtain phases and figures of merit and a corresponding difference electron-density map. An orthogonal grid is then constructed from which points are selected that are likely to belong to the

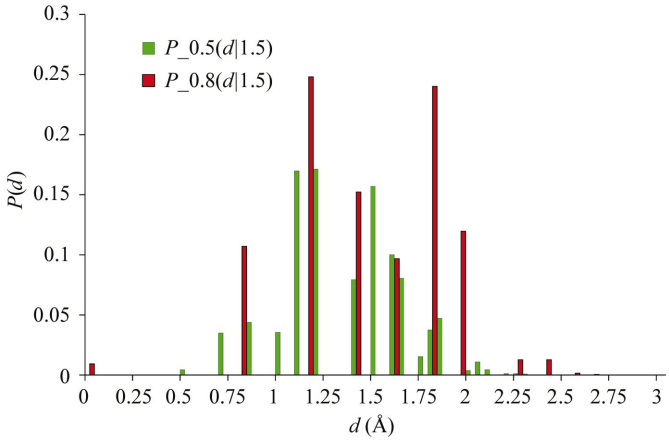


**Figure 1**
Distribution of distances on a grid given an ideal distance of 1.5 Å and a grid spacing of 0.5 and 0.8 Å, respectively.

ligand. The geometrical features from the ligand molecule are used to construct an error model for the positional parameters of the ligand atoms. A search algorithm designed to optimize the constructed scoring function results in ligand-atom names being assigned to grid points. A geometrized ligand is obtained and is subsequently subjected to a restraint refinement of the ligand–protein complex using *REFMAC*5 (Murshudov *et al.*, 1997). The procedure can be iterated to locate other ligands if they are present.

### 2.2. Trial atom generation

A difference electron-density map that is supposed to contain a ligand is parameterized by an orthogonal grid with a minimum spacing $d_{\rm grid}$ between two grid points. $d_{\rm grid}$ is set to 0.5 Å, which is linked to the error model used, as will be explained in §2.3. The orthogonal grid is constructed in such a way that it covers the complete macromolecule, with an added border of appropriate size. Crystallographic symmetry is ignored at this stage. Each grid point is associated with three parameters: density height, occupancy and cluster number. The density height is the value of the electron density at the location of the grid point in the unit cell. The occupancy is either 0 or 1 and determines whether the grid point is used in trial-atom generation. The cluster numbers divide the set of grid points into clusters in which the elements are path-connected.

The grid points are selected if their electron-density value is above a certain threshold $\rho_{\rm thres}$. The selected grid points are clustered using an approach that is related to the well known skeletonization procedures (Greer, 1974; Swanson, 1994).

(i) Set the occupancy of all grid points with an associated density height larger than $\rho_{\rm thres}$ equal to 1; set the cluster number of all grid points to 0.

(ii*a*) Move to the next grid point with an occupancy of 1 that has a neighbouring grid point with an occupancy equal to zero.

(ii*b*) Flag this grid point indicating 'to be removed' unless it only has neighbours with occupancy 0 or neighbours flagged to be removed or if a removal of this grid point disconnects the neighbouring grid points.

(iii) Go to (ii*a*) until all grid points have been visited.

(iv) Set occupancies of the 'to be removed' grid points to 0; go to (ii*a*) until no further changes occur.

(v) Assign different cluster numbers to each grid point with a non-zero occupancy.

This algorithm, known as constrained erosion (Heijmans, 1992), delivers a number of isolated grid points. It can be shown that the remaining grid points have not been path-connected given the definition of the neighbourhood in step (ii). In the present implementation, two grid points are defined as neighbours when their distance is smaller than or equal to $3^{1/2}d_{\rm grid}$.

The inverse of this algorithm, geodesic reconstruction (Heijmans, 1992), is applied.

(i) Initialize $C$ to 0.

(ii) $C = C + 1$

(iii) Move to the next grid point with occupancy equal to 1 and cluster number equal to $C$.

(iv) Select all neighbours of this grid point with an associated density height larger than $\rho_{\rm thres}$; set the cluster numbers and occupancies to $C$ and 1, respectively.

(v) Go to (iii) until no further changes occur.

(vi) Go to (ii) until all clusters are constructed.

Other algorithms can be constructed that would perform the clustering in a similar fashion.

The number of grid points grouped in a connected cluster is related to the volume of the cluster of the difference density which is used as a possible signal-to-noise classifier. In a practical implementation, the largest cluster is assigned to the ligand to be built, or in the case of multiple ligands, the volume-ordered list of clusters is matched to the list of ligands ordered by their size.

The density threshold $\rho_{\rm thres}$ used in the clustering algorithm is selected on the basis of the sizes of the obtained clusters.

To reduce the amount of grid points even further, another selection procedure is carried out that resembles constrained erosion.

(i) Move to the grid point with the highest density and occupancy 1.

(ii) Select all grid points within a distance of $d_\rho$ and set their occupancy to 0.

(iii) Go to (i) until convergence.

Since the height of the electron density is correlated with the proximity of atoms, this procedure is more likely to preserve the grid points that are close to the position of ligand atoms. Choice of the selection radius $d_\rho$ should reflect the bonding distances present in the ligand that is sought and the choice of the grid spacing $d_{\rm grid}$. Setting $d_\rho$ to 1.3 Å for a grid spacing of 0.5 Å gives satisfactory results. The grid-based selection procedure is insensitive to the shape or topological properties of the electron density around an atom, but has the disadvantage of generating a large surplus of initial trial atoms.

### 2.3. The distribution of distances

An error model of the geometric features of the ligand is needed in the design of a scoring function, as mentioned above. The positional parameters of the trial atoms are not continuously distributed as assumed in a free-atom model (Zwart & Lamzin, 2003, 2004), but follow a discrete so-called lattice distribution (Abramovicz & Stegun, 1974; Bricogne, 1974). We assume that the best possible interpretation is that which maps the ligand atoms to their closest neighbours on the grid. The proposed error model of the positional parameters thus consists of a rounding-off operation of the positional parameters of the 'true' ligand atoms to the positional parameters of the grid. The distribution of interatomic distances after the rounding-off operation can be obtained *via* simulation. The sampling of a point distributed on a sphere is carried out using rejection sampling; the algorithm is outlined in Appendix *A*. Inclusion of any uncertainty or 'natural spread' of a given interatomic distance can also be taken into account.

Empirical distributions, which are clearly non-Gaussian, for an interatomic distance of 1.5 Å and an orthogonal grid spacing of 0.5 and 0.8 Å are shown in Fig. 1. The choice of a grid spacing of 0.5 Å is made to prevent the positional parameters of bonded (non-H) atoms being rounded off to the same grid point. Furthermore, the 0.5 Å grid spacing ensures that the distance of a 'true' ligand atom from the nearest grid point is smaller than or equal to $3^{1/2} \times 0.5/2 = 0.43$ Å. This is a positional error that should lie well within the radius of convergence of restrained refinement procedures for the ligand.

### 2.4. The distribution of chirality

The chirality of an atom is defined by the sign of the scalar triple product of the interatomic vectors between the chiral atom $j$ and three bonded neighbours $k$, $l$ and $m$,

$$C_j = \text{sign}[\mathbf{d}_{jk} \cdot (\mathbf{d}_{jl} \times \mathbf{d}_{jm})]. \quad (2)$$

$\mathbf{d}_{jx}$ denotes the vector between the chiral atom $j$ and a neighbouring atom $x$. The order of the bonded atoms is determined on the basis of the order of appearance in the input ligand structure, rather than by the standard priority rules, since it is only required to have an internal standard.

The distribution of chirality of an atom is constructed in a way similar to the construction of the distance distributions. Chiral atoms and their bonded neighbours are randomly oriented and placed on a grid. Generation of random orientations is performed by sampling from a uniform distribution of points on a four-dimensional unit sphere. These four numbers can be considered to form a quaternion and are used to reorient the fragment under consideration (Appendix *A*). After rounding off the positional parameters to the nearest grid points, the chirality is recomputed. An example distribution of the sign of the chiral volume is shown in Fig. 2.

### 2.5. Repulsion

Another source of information on the internal geometry of a molecular fragment is van der Waals repulsions. A repulsion term models our prior knowledge that a $1{-}n$ distance, with $n$ larger than 3, is on average larger than an average $1{-}3$ distance. Repulsion terms prevent crumpled trial assignments being recognized as possible molecular fragments. The repulsion term used has the following form:

$$W(d|a, b) = \tfrac{1}{2}\{1 + \tanh[(d - a)b]\}. \quad (3)$$

By varying $a$ and $b$, the location of the inflection point and shape of the repulsion function can be modified, as shown in Fig. 3. From a probabilistic viewpoint, this function could be seen as an improper prior (Bernardo & Smith, 2000) on the $1{-}n$ ($n > 3$) distances, although its role should be seen more as an activation function (Bishop, 1995) whose logarithmic form only gives penalties for interatomic distances involved in short non-bonded interactions.

### 2.6. Electron density

One of the most essential sources of experimental information is the height of the difference electron density. The ligand molecule is constructed from the selected grid points that not only satisfy the above described stereochemical criteria but also lie in the highest possible density. A monotonic scoring function is used that is similar to that describing the van der Waals repulsions,

$$W(\rho|s) = \frac{1}{2}\left[1 + \tanh\left(\frac{2}{s}\rho - 2\right)\right], \quad (4)$$

where $s$ is the mean electron density for the selected cluster of the grid points and $\rho$ is the value of the electron density for each of the grid points.

### 2.7. Searching and scoring

A graph of the known ligand is constructed by assuming that $1{-}2$ distances lie between 1.1 and 1.9 Å. A graph of the grid representation can be constructed in a similar way. The distance limits for the putative $1{-}2$ distances are obtained by transforming the $1{-}2$ distance boundaries to a grid by using the Monte Carlo procedure described in §2.3.

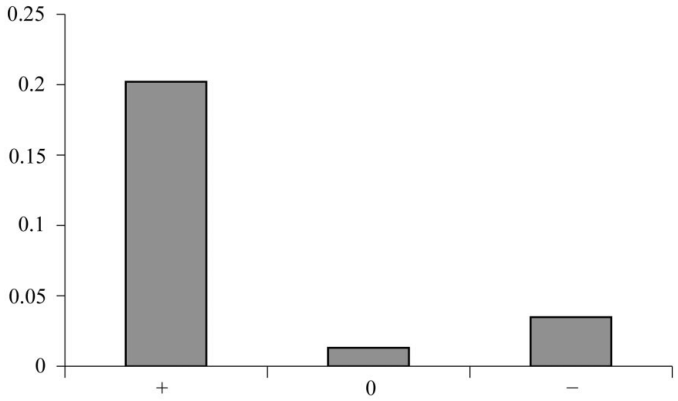


**Figure 2**
Conditional distributions of the sign of a chiral volume of 2.77 Å$^3$ and a grid spacing of 0.5 Å.

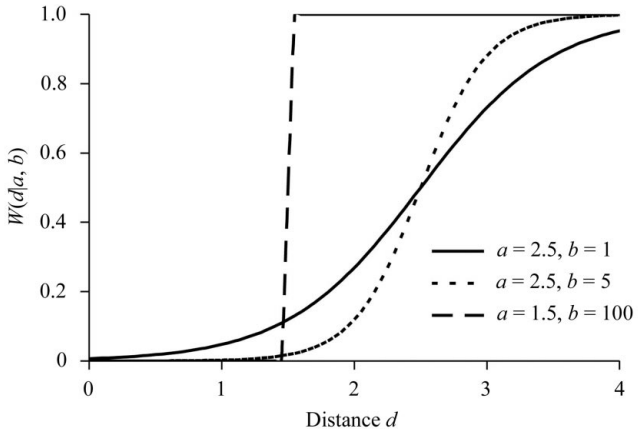


**Figure 3**
Repulsion function $W(d|a, b)$ with various choices of location parameter $a$ and shape parameter $b$.

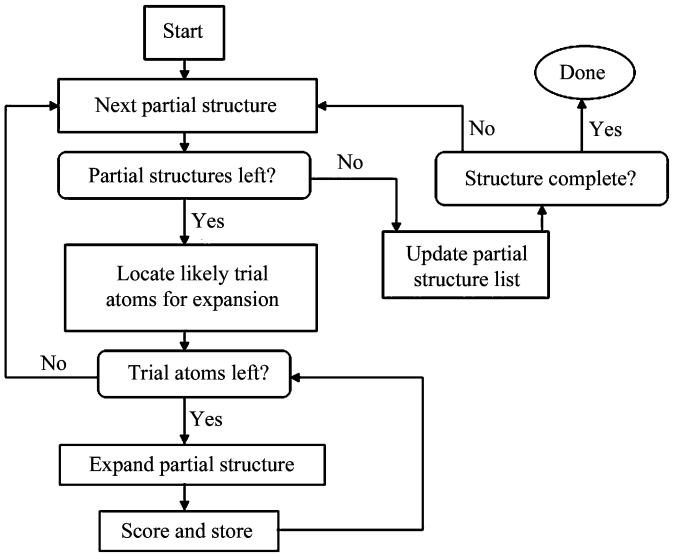


**Figure 4**
Flowchart of the search procedure. See text for details.

The search procedure starts with the generation of a set of partial interpretations by assigning the label of a given ligand atom to each grid point within the available cluster. These partial interpretations are then expanded by addition of one fixed ligand label. Expansions are generated on the basis of the constructed graph of the trial atoms, taking into account constraints dictated by the graph of the ideal ligand, the graph on the trial atoms and the available partial interpretation. Each expanded interpretation is scored, but only $N_{\rm store}$ partial interpretations with the best scores are stored. When all possible single-atom expansions have been tried, the stored partial interpretations are used for further expansions until completion of the ligand. The search procedure is illustrated in Fig. 4. We denote the order in which specific atoms of the ligand are assigned to the grid points as the expansion order. By default, the first atom to be assigned is that with the largest number of bonded neighbouring atoms. The order in which other atoms are 'attached' to the partial interpretation depends on the amount of geometrical information to be gained by the addition of this atom. The larger the amount of geometrical information available on a partial structure, the easier it is to recognize it as a correct fragment. For this reason, atoms are added to a partial structure in the order that provides the maximum expected amount of information in the subsequently generated structure. Conceptually, this procedure should minimize the chances of a correct interpretation falling outside the $N_{\rm store}$ best partial interpretation.

The partial interpretations are scored as follows:

$$\begin{aligned} Q(\text{grid}|\text{ligand}) = {} & w_{\rm prior} \sum_m \ln[P_{\rm prior}(d_m^{\rm obs}|d_m^{\rm tar})] \\ & + w_{\rm c} \sum_n \ln[P_C(C_n^{\rm obs}|C_n^{\rm tar})] \\ & + w_{\rm rep} \sum_o \ln[W(d_o^{\rm obs}|a,b)] \\ & + w_{\rm dens} \sum_j \ln[W(\rho_j|s)]. \end{aligned} \quad (5)$$

**Table 1**
Data-set characteristics.

| PDB code | $d_{\rm min}$ (Å) | $B_{\rm Wilson}$ (Å$^2$) | Ligand | $B_{\rm ligand}$† (Å$^2$) | Non-H atoms | R.m.s.d.‡ (Å) | CPU for G4 Mac OSX 1 GHz (min) |
|---|---|---|---|---|---|---|---|
| 1ee2 | 1.5 | 15 | NADH | 11 | 44 | 0.09 | 27 |
| | | | Cholic acid | 16 | 29 | 0.09 | 5 |
| 1obd | 1.4 | 14 | ATP | 23 | 31 | 0.07 | 7 |
| | | | AMP | 32 | 23 | § | § |
| 1o2d | 2.2 | 38 | Propamidine | 34 | 23 | § | 7 |
| 1a28 | 1.8 | 24 | Progesterone | 25 | 23 | 0.17 | 2 |
| 1cbs | 1.8 | 13 | Retinoic acid | 13 | 22 | 0.22 | 2 |
| 1ld8 | 1.8 | 18 | FDP¶ | 17 | 24 | § | 12 |
| | | | IC49†† | 19 | 33 | 0.28 | 10 |
| | | | Sucrose | 31 | 23 | § | § |
| 1ok4 | 2.1 | 19 | DHAP‡‡ | 17 | 9 | 0.30 | 5 |

† Average $B$ value of the ligand atoms. ‡ Root-mean-square displacement from the deposited structure after restrained refinement of the protein–ligand complex. § See text. ¶ Farnesyldiphosphate. †† Inhibitor compound 49. ‡‡ Dihydroxyacetone phosphate

$P_{\rm prior}(d_m^{\rm obs}|d_m^{\rm tar})$ denotes the probability of the observed distance given the assigned target distance. $P_C(C_n^{\rm obs}|C_n^{\rm tar})$ gives the probability of the observed chirality given the target chirality. These distributions are obtained as described in §2.3 and §2.4. $W(d_o^{\rm obs}|a,b)$ denotes the repulsion terms discussed in §2.5. The $W(\rho_j|s)$ term accounts for the density values. The multipliers $w_{\rm prior}$, $w_{\rm c}$, $w_{\rm rep}$ and $w_{\rm dens}$ are relative weights for the contributions of the four features.

Global optimization algorithms such as simulated annealing (Kirkpatrick *et al.*, 1983) and the cross-entropy method (Rubinstein, 1999) have been tried as an alternative to the outlined optimization procedure, but seemed to lack the ease of incorporating geometrical constraints dictated by the connectivity matrix of the search and target graphs during random-search procedures. However, preliminary implementations of these algorithms did show successes, but required a considerably longer time and fine-tuning of parameters in order to converge to the correct solution.

### 2.8. Geometrization

Once the grid points have been assigned to the ligand atoms, the ligand model is fit to the density and is geometrized using the 1−2 and 1−3 distance restraints. The target deviations from the bonded and angle-bonded distances are set to 0.02 and 0.04 Å, respectively. Least-squares minimization is carried out using first-order derivatives and the diagonal approximation of the normal matrix, with a formulation similar to that described by Agarwal (1978).

## 3. Results

A number of tests have been carried out on moderate-size ligands using data obtained from the PDB (Bernstein *et al.*, 1977; Berman *et al.*, 2000). The parameters $a$ and $b$ in (3) were set to 2.5 and 2.0, respectively. The weights $w_{\rm prior}$, $w_{\rm c}$, $w_{\rm rep}$ and $w_{\rm dens}$ were set to 0.7, 10, 12 and 6, respectively. The number of putative 1−2 distances within the selected set of grid atoms is obtained by constructing a graph on the selected grid points

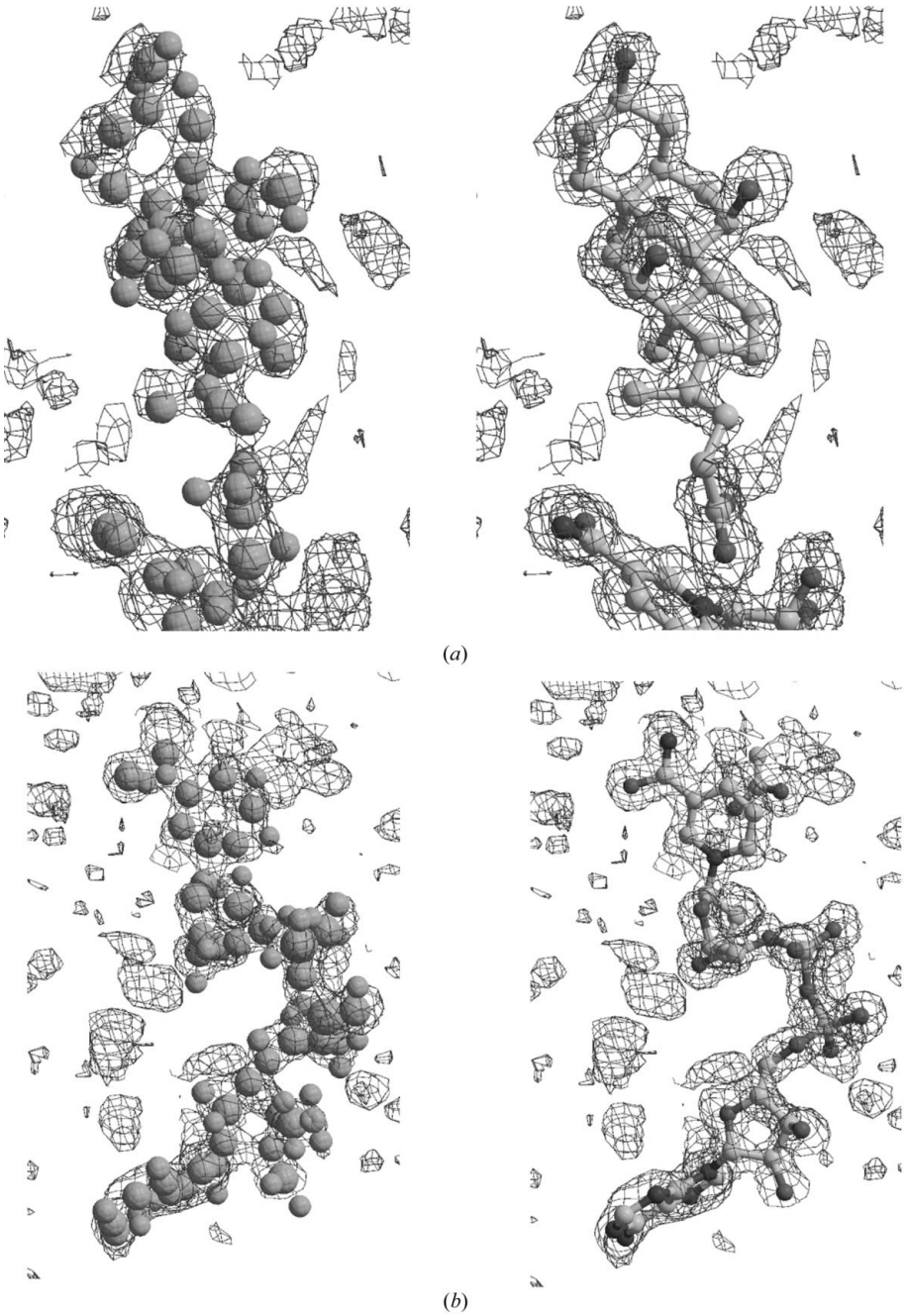

**Figure 5**
Trial atoms (left) and refined interpretations (right) of cholic acid (*a*) and NADH (*b*) in the original difference density.

with the computed distance limits (§2.7). The maximum number of partial structures stored during each expansion cycle was five times the number of putative 1—2 neighbours observed in the set of trial atoms. The characteristics of the structures used and the X-ray data sets are summarized in Table 1. The procedure has been run with the specified parameters unless stated otherwise. Electron-density thresholds during the building were determined by the procedure outlined in §2.1. By default, the interpretation with the highest score has been used to validate the procedure. The results for all the test structures are also summarized in Table 1. Detailed descriptions of the building for each case are given in the following subsections.

### 3.1. Cholic acid and NADH

The X-ray data and dimeric atomic model of SS-LADH (PDB code 1ee2; Adolph *et al.*, 2000) contains two cholic acid molecules, two NADH molecules, 2 × 374 residues and approximately 1000 water molecules. Phases obtained from a rigid-body refinement of the protein part of the structure have been used as a starting point for the building of cholic acid and NADH. Cluster construction reveals four clusters of connected density with a volume larger than 80 Å$^3$. The clusters with the approximate volumes of 150 Å$^3$ were interpreted as NADH and the clusters with volumes around 85 Å$^3$ as possible cholic acids. Fig. 5 shows the initial difference density with the placed grid atoms and the model after refinement with *REFMAC*5 for one of the the cholic acid clusters. The r.m.s.d. (root-mean-square displacement) of the built model to the deposited model is 0.30 Å after geometrization and 0.09 Å after restrained refinement of the protein–ligand complex with *REFMAC*5.

The building of NADH resulted in a structure with an r.m.s.d. of 0.11 Å from the deposited coordinates (Fig. 5). In order to prevent the algorithm discarding correct partial interpretations during the early stages of the building, the number of partial expansion stored during the iterative extension had to be enlarged by a factor of four from the default value.

### 3.2. ATP and AMP

The atomic model of saicar synthetase (PDB code 1obd; Levdikov *et al.*, 1998) contains AMP and ATP. Because of the relatively large amount of noise in the difference electron density, the described clustering procedure was unable to determine the locations of the ligands within a reasonable amount of time. For this reason, the positions of the ATP and AMP were used in the cluster selection and assignment. The building and subsequent refinement of ATP resulted in a structure matching the deposited coordinates (r.m.s.d. = 0.09 Å; Fig. 6). Building of AMP was unsuccessful owing to the ill-defined/absent difference density for the phosphate and sugar moiety. The deposited AMP structure has an occupancy of 0.5. A

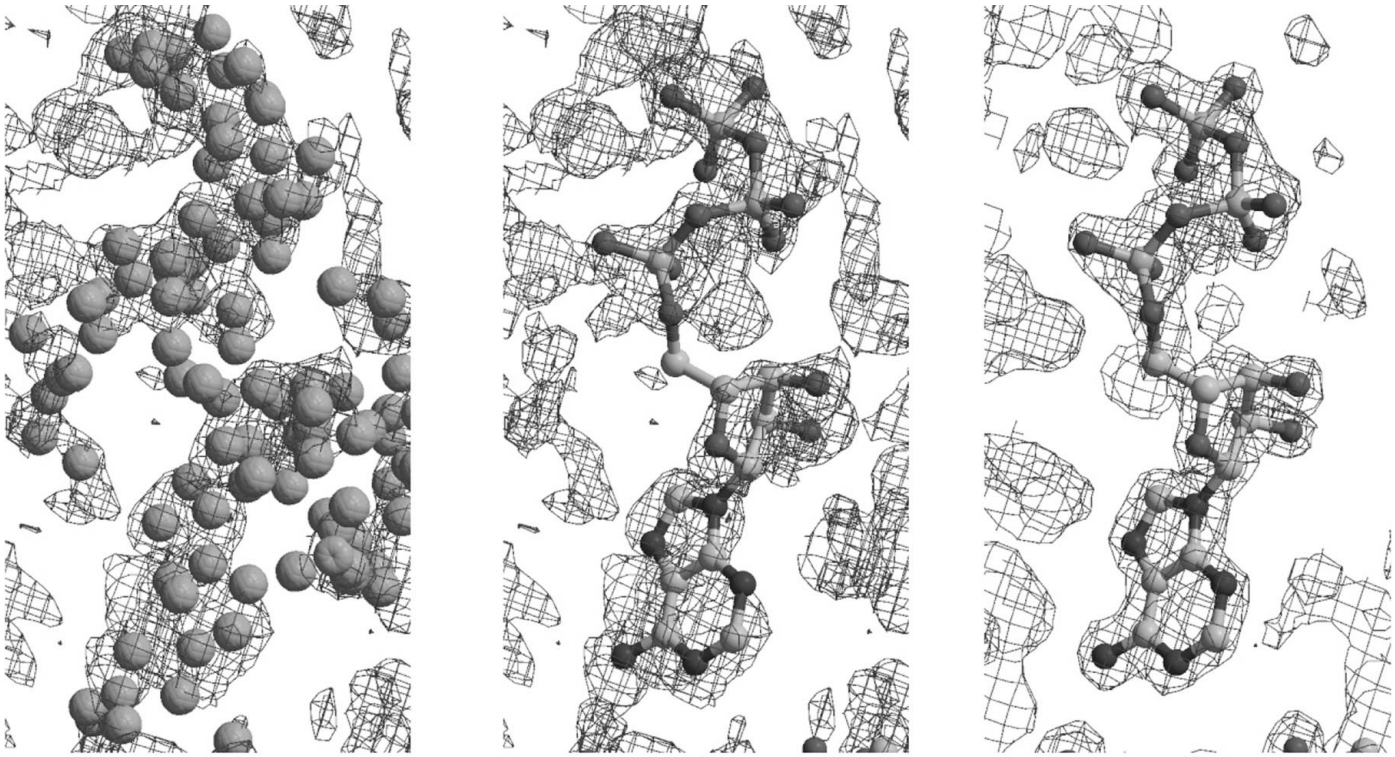

**Figure 6**
Difference density with trial atoms for ATP (*a*), refined interpretation in the original difference density (*b*) and the density after refinement (*c*).

Wilson plot of the deposited structure factors and a completeness analysis of the X-ray data indicates that about 15% of the strongest reflections around 3.0 Å resolution are missing. This could be a reason for the relatively noisy difference map and the subsequent unsuccessful building of AMP.

### 3.3. Propamidine

The location of a propamidine molecule in a double-stranded DNA structure (PDB code 1o2d; Schwarzenbacher *et al.*, 2004) has been determined with the default parameters of the described clustering algorithm using phases from rigid-body refinement of the non-ligand part of the atomic model. Interpretation of the difference density and subsequent refinement resulted in the placement of the ligand with a different conformation compared with the deposited structure (Fig. 7). In the same figure, the best six geometrized interpretations are shown. The relatively weak density of part of the propamidine molecule possibly explains the difference between the deposited and automatically built models.

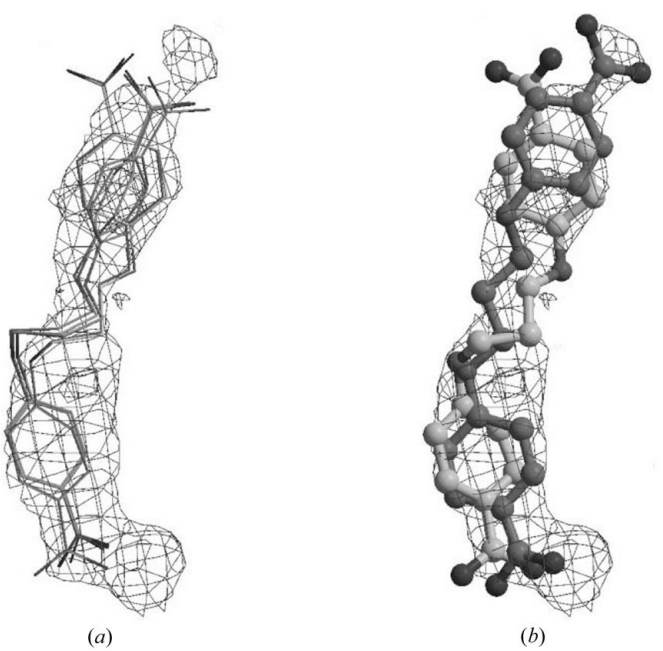

**Figure 7**
Top 6 interpretations of propamidine density (*a*). The best interpretation (light grey) and the deposited structure (dark grey) are shown in (*b*).

### 3.4. Progesterone

The position of the steroid in a human progesterone receptor (PDB code 1a28; Williams & Sigler, 1998) was located using default parameters. The built and deposited model differ in the orientation of the keto group (Fig. 8). The interpretation that is consistent with the deposited crystal structure has a slightly lower score but shows more favourable protein contacts than the interpretation with the flipped keto group. These interactions are currently not taken into account in our scoring function. The r.m.s.d. of the built and refined

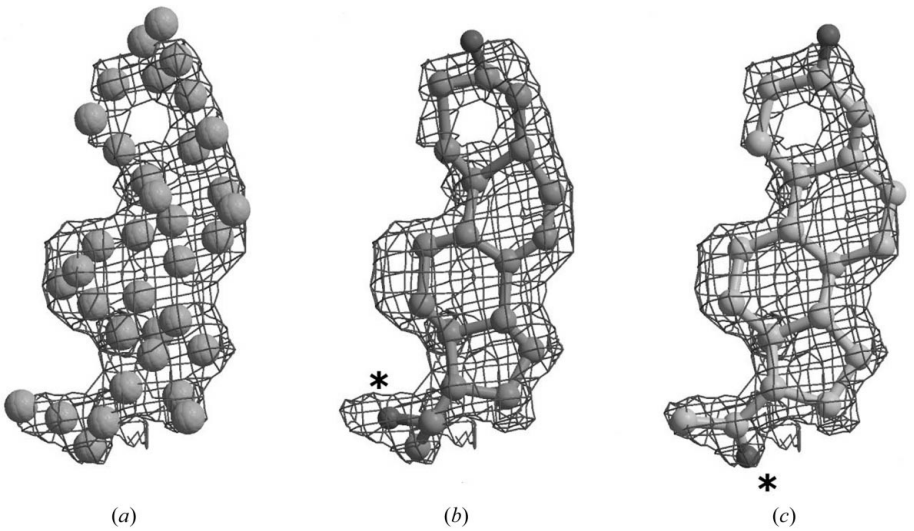


**Figure 8**
Difference density with trial positions (*a*), non-geometrized interpretation (*b*) and deposited structure of progesterone (*c*). * denotes the position of the keto oxygen that is different in the interpretation and deposited structures.

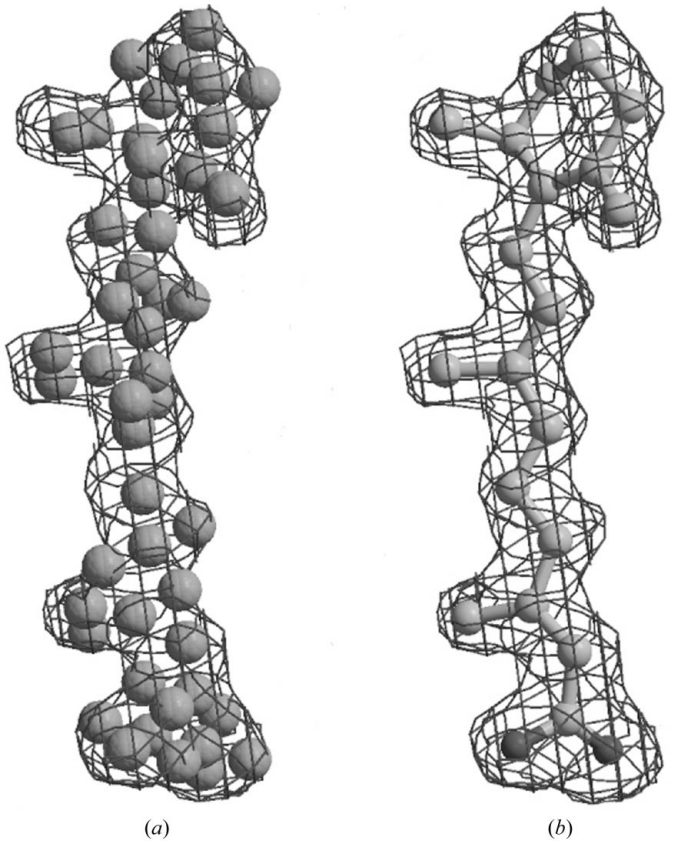


**Figure 9**
Difference density with trial positions (*a*) and refined interpretation of retinoic acid (*b*).

model from the deposited coordinates is 0.17 Å. Inclusion of the flipped keto group increases the r.m.s.d. to 0.29 Å.

### 3.5. Retinoic acid

The retinoic acid in the difference electron density of a retinoic acid transport protein (PDB code 1cbs; Kleywegt *et al.*, 1994) was located and built using default parameters (Fig. 9). The r.m.s.d. of the build model to the deposited model was 0.22 Å.

### 3.6. Farnesyldiphosphate (FDP), inhibitor compound 49 (IC49) and sucrose

Location of the ligands FDP, IC49 and sucrose in the difference density of human farnesyltransferase (PDB code 1ld8; Leahy *et al.*, 1992) was carried out as follows. The largest three difference density clusters have been assigned to the individual ligands on the basis of the cluster volumes. Once one ligand had been built, the protein–ligand complex was re-refined and the new density map was subsequently used to build the remaining ligands. Owing to the size of the ligands, the number of intermediate partial interpretations was increased by a factor of two. Whereas IC49 was built and refined to an r.m.s.d. of 0.28 Å, FDP was built in a *cis* rather than *trans* conformation compared with the deposited structure (Fig. 10). Attempts to build sucrose failed under various settings. This is attributed to the fact that the ligand has a high apparent symmetry, resulting in a high probability that the interpretation process does not retain the correct partial structure after each iteration and converges to false minima.

### 3.7. Dihydroxyacetone phosphate (DHAP)

The location of the DHAP molecule in the difference map of an aldolase structure (PDB code 1ok4; Lorentzen *et al.*, 2003) was determined using the clustering algorithm around the residues where the ligand was known to bind *a priori*. The refined interpretation is shown in Fig. 11. Inclusion of protein–ligand interactions would have made the interpretation easier, as DHAP is covalently bound to the protein.

## 4. Discussion and conclusions

The modelling of the distribution of distances *via* a marginalization of a lattice distribution proved to be an adequate tool in modelling prior geometrical knowledge in grid-based model-building routines. The resulting approximate distributions can be fairly quickly obtained *via* Monte Carlo simulations. Approximate distributions of relative complex quantities, such as the distribution of the sign of the chiral volume of an atom, can also be obtained using simulations in a straightforward way. It must be noted that the constructed error model on the positional parameters is an approximation. The interatomic distances are not independent and correlations should in principle be taken into account. If efficient ways of storing and handling multidimensional distributions of geometrical features can be implemented, one could attempt

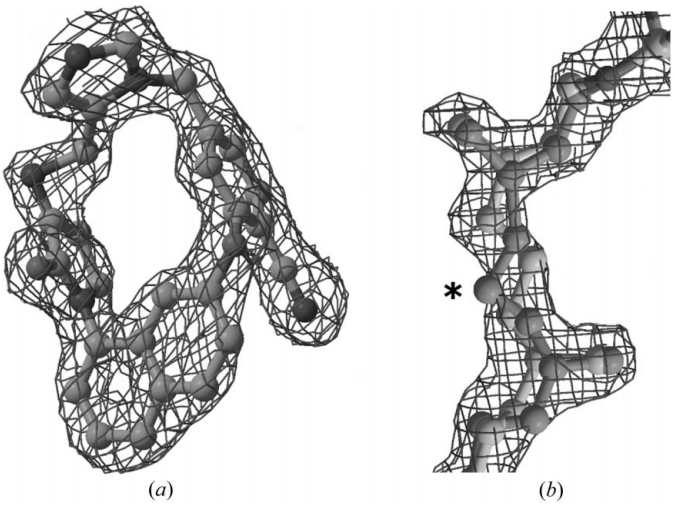


**Figure 10**
Difference densities and refined interpretations for IC49 (*a*) and an overlay of the interpretation and deposited FDP model (*b*). * marks the incorrectly built part of FDP.

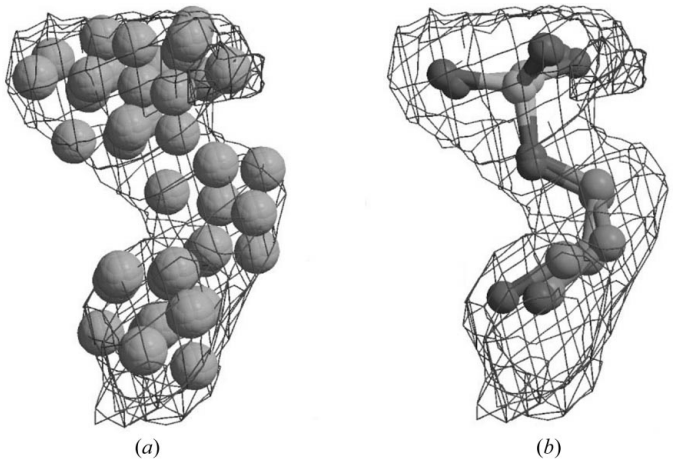


**Figure 11**
Difference density with trial positions (*a*) and refined interpretation overlaid on the correct dihydroxyacetone phosphate model (*b*).

to obtain the joint probability distribution of all of them for the whole search molecule. Furthermore, the grid-point selection algorithm designed to eliminate atoms with low density values affects the possible set of distances between grid points. This set of distances may have a distribution that is different from that constructed in the simulations and probably depends on the spatial distribution of density heights within the cluster. However, the designed classifier proved to be good enough to recognize the correct solution. A similar situation is present for handling the available prior knowledge which is limited to 1−2 and 1−3 distances, chiral signs and repulsions. Even with this limited amount of information one is able to recognize complex models in difference density. Additional information, such as planarity restraints, prior 1–4 distance distributions and ligand–protein interactions, will most likely enhance the performance of the recognition process.

The models constructed on the grid are close enough to their correct positions that *REFMAC*5 was able to straightforwardly refine the protein–ligand complex. Combining the geometrization with a real-space fit to the electron density has further enhanced the interpretation process.

As seen from the progesterone example and, to a certain extent, the DHAP example, internal ligand geometric information alone is not always sufficient to interpret the difference density. Inclusion of protein–ligand contacts in the decision-making process would help to resolve possible ambiguities, prevent chemically unreasonable contacts between protein and ligand atoms and could possibly limit the search space. A similar approach would also be useful for building structure with internal repeats, such as glycosylation sites. If in the initial stage the sugar backbone can be fitted, subsequent placement of the (carbon) oxygen groups can be carried out using restraints on the parts that are already present. Ideally, the building procedure should be able to identify these modularities automatically and use them to enhance the speed and performance of the recognition process.

The search algorithm is able to build ligands in a difference electron density, based on the proposed scoring function. A present limitation of the software may be its speed: most ligands were build in approximately 10 min, whereas ATP took 15 min and NADH about half an hour. In future implementations, the building algorithm will be optimized for CPU efficiency. An essential part of future development will be the implementation of efficient mechanisms for the decision whether an addition of an atom or set of atoms to the available partial structure results in a better description of the observed difference electron density. This will facilitate the process further, also enabling the construction of partially disordered ligands, such the AMP example in §3.2, to be carried out automatically.

The ligand building routine described has been incorporated into version 6.1 of the *ARP/wARP* suite, which was introduced in July 2004.

## APPENDIX *A*
## Sphere and hypersphere point picking

Uniform sampling of points on a sphere with a unit radius is carried out using a method developed by Marsagalia (1972) that consists of sampling two random numbers, $A$ and $B$, distributed independently and uniformly on $(-1, 1)$. Pairs of $(A, B)$ for which $A^2 + B^2 < 1$ can be used to construct a vector $(x, y, z)$ that is distributed uniformly on a sphere,

$$x = 2A(1 - A^2 - B^2)^{1/2}, \quad (6)$$

$$y = 2B(1 - A^2 - B^2)^{1/2}, \quad (7)$$

$$z = 1 - 2(A^2 + B^2). \quad (8)$$

Sampling points $(a_0, a_1, a_2, a_3)$ on a four-dimensional sphere with unit radius can be carried out in a similar way. Four random numbers $(A, B, C, D)$ are drawn independently from a uniform distribution on $(-1, 1)$. Random numbers for which

the pairs $(A, B)$ and $(C, D)$ satisfy $A^2 + B^2 < 1$ and $C^2 + D^2 < 1$ are used in the following transformation

$$a_0 = A, \tag{9}$$

$$a_1 = B, \tag{10}$$

$$a_2 = C\left(\frac{1 - A^2 - B^2}{C^2 + D^2}\right), \tag{11}$$

$$a_3 = D\left(\frac{1 - A^2 - B^2}{C^2 + D^2}\right). \tag{12}$$

The vector $(a_0, a_1, a_2, a_3)$ is then uniformly distributed on a four-dimensional sphere with radius 1. This vector can be considered as a quaternion,

$$\mathbf{q} = a_0 + a_1 i + a_2 j + a_3 k \tag{13}$$

and can be used to reorient a molecular fragment (Weisstein, 1999). More efficient sampling methods based on the correspondence of the rotation group $SO(3)$ and a four-dimensional sphere are described elsewhere, *e.g.* by Bricogne (1997*b*).

The authors would like to thank R. J. Morris and A. Perrakis for stimulating discussions. PHZ thanks K. Cowtan for his help with the use of the Clipper libraries and the EMBL for a PhD fellowship.

## References


Abramovicz, M. & Stegun, I. A. (1974). *Handbook of Mathematical Functions.* New York: Dover Publications Inc.

Adolph, H.-W., Zwart, P., Meijers, R., Hubatsch, I., Kiefer, M., Lamzin, V. S. & Cedergren-Zeppezauer, E. (2000). *Biochemistry*, **39**, 12885–12897.

Agarwal, R. (1978). *Acta Cryst.* A**34**, 791–809.

Altomare, A., Giacovazzo, C., Ianigro, M., Moliterni, A. G. G. & Rizzi, R. (2002). *J. Appl. Cryst.* **35**, 21–27.

Badger, J. (2003). *Acta Cryst.* D**59**, 823–827.

Bart, J. C. J. & Busetti, A. (1976). *Acta Cryst.* A**32**, 927–933.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bernardo, J. M. & Smith, A. F. M. (2000). *Bayesian Theory.* New York: Wiley.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

Bricogne, G. (1974). *Acta Cryst.* A**30**, 395–405.

Bricogne, G. (1997*a*). *Methods Enzymol.* **276**, 361–423.

Bricogne, G. (1997*b*). *Methods Enzymol.* **276**, 424–449.

Brunzelle, J. S., Shafaee, P., Yang, X., Weigand, S., Ren, Z. & Anderson, W. F. (2003). *Acta Cryst.* D**59**, 1138–1144.

Cascarano, G., Giacovazzo, C., Camalli, M., Spagna, R. & Watkin, D. J. (1991). *Acta Cryst.* A**47**, 373–381.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Diller, D., Pohl, E., Redinbo, M., Hovey, B. & Hol, W. (1999). *Proteins*, **36**, 512–525.

Greer, J. (1974). *J. Mol. Biol.* **82**, 279–301.

Heijmans, H. J. A. M. (1992). *Nieuw Arch. Wisk. (4)*, **10**, 237–276.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). *Science*, **220**, 671–680.

Kleywegt, G. J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K. & Jones, T. A. (1994). *Structure*, **2**, 1241–1258.

Koch, M. H. J. (1974). *Acta Cryst.* A**30**, 67–70.

Lamzin, V. & Wilson, K. (1997). *Methods Enzymol.* **277**, 269–305.

Leahy, D. J., Axel, R. & Hendrickson, W. A. (1992). *Cell*, **68**, 1145–1162.

Leherte, L., Glasgow, J. I., Baxter, K., Steeg, E. & Fortier, S. (1997). *J. Artif. Intell. Res.* **7**, 125–159.

Levdikov, V. M., Barynin, V. V., Grebenko, A. I., Melik-Adamyan, W. R., Lamzin, V. S. & Wilson, K. S. (1998). *Structure*, **6**, 363–376.

Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.

Lorentzen, E., Pohl, E., Zwart, P., Stark, A., Russell, R., Knura, T., Hensel, R. & Siebers, B. (2003). *J. Biol. Chem.* **278**, 47253–47260.

Main, P. & Hull, S. E. (1978). *Acta Cryst.* A**34**, 353–361.

Marsaglia, G. (1972). *Ann. Math. Stat.* **43**, 645–646.

Menéndez-Velázquez, A. & García-Granda, S. (2003). *J. Appl. Cryst.* **36**, 193–205.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Nienaber, V. L., Richardson, P. L., Klighofer, V., Bouska, J. J., Giranda, V. L. & Greer, J. (2000). *Nature Biotechnol.* **18**, 1105–1108.

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

Oldfield, T. J. (2001*b*). *Acta Cryst.* D**57**, 696–705.

Oldfield, T. J. (2002). *Acta Cryst.* D**58**, 963–967.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Rubinstein, R. (1999). *Methods Comput. Appl. Prob.* **1**, 127–190.

Schwarzenbacher, R. *et al.* (2004). *Proteins*, **54**, 174–177.

Sharff, A. & Jhoti, H. (2003). *Curr. Opin. Chem. Biol.* **7**, 340–345.

Swanson, S. M. (1994). *Acta Cryst.* D**50**, 695–708.

Terwilliger, T. C. (2003). *Acta Cryst.* D**59**, 38–44.

Vagin, A., Murshudov, G., Dodson, E., Henrick, K., Richelle, J. & Wodak, S. (2003). *MONLIB, a Multi-purpose Dictionary for Macromolecules.* Unpublished results.

Weisstein, E. (1999). *CRC Concise Encyclopedia of Mathematics.* New York: Chapman & Hall/CRC Press.

Williams, S. P. & Sigler, P. B. (1998). *Nature (London)*, **393**, 392–396.

Zwart, P. H. & Lamzin, V. S. (2003). *Acta Cryst.* D**59**, 2104–2113.

Zwart, P. H. & Lamzin, V. S. (2004). *Acta Cryst.* D**60**, 220–226.