Error Estimation and Pattern Recognition Techniques in Protein Crystallography
Zwart, P.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 4

Towards Automated Ligand Building

Abstract

Methods for the automated identification and building of protein-bound ligands in electron density maps are described. An error model of observed geometrical features of a ligand based on a lattice distribution 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 results in a flexible model building scheme and shows promising initial results. Several ligands, with a size between 9 and 44 non-hydrogen atoms, have been identified and build in an automatic way using a minimal amount of prior knowledge.

4.1 Introduction

Automated model building techniques in protein crystallography are an essential component for any hard- and software pipeline that is aimed to deliver protein crystal structures with minimum amount of user intervention (Brunzelle et al., 2003). Model building routines such as ARP/wARP (Perrakis et al., 1999), Resolve (Terwilliger, 2003a), 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 recognised, other compounds, such as DNA, RNA and ligands, cannot be build automatically at present. The problem of ligand building is of particular interest, both from a theoretical and practical point of view. The chemical variety of ligands bound to proteins is enormous: up to date, more than 4,000 entries are present in the Hetero-compound Information Centre (HIC-Up, http://xray.bmc.uu.se/hicup) and over 2,000 ligand dictionary entries are in the REFMAC5/CCP4 monomer library (Vagin et al., 2003). Means of handling basic chemical knowledge of ligands in the interpretation of electron densities at resolutions lower than atomic and with possible phase errors, is challenging. The practical interest
stems largely from pharmaceutical companies and large scale X-ray crystallography facilities that desire to automate drug discovery efforts or 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 optimisation, as carried out by pharmaceutical companies such as Abbott (Nienaber et al., 2000) and Astex-Technologies (Sharff & Jhoti, 2003).

Existing methods for automated building of non-protein models are either based on the use of torsion angles, interatomic distance matrices or a topological analysis of the electron density. The ligand building methods implemented in XLIGAND (Oldfield, 2001b) or BLOB (Diller et al., 1999a) fit ligands to the electron density by varying the torsion angles. XLIGAND performs a shape matching and needs initial guesses of the location of the ligand obtained via a segmentation of the difference density. A ligand molecule is placed in several trial conformations into the density and a local optimisation to maximise the fit to the electron density is carried out (Oldfield, 2001a). BLOB utilises global optimisation 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 E-maps on the basis of known geometrical features and approximate atomic positions obtained by peak-picking E-maps. Recently, distance matrix based methods have also been applied to the interpretation of high resolution protein electron densities (Oldfield, 2002b). The interpretation of electron density maps via a topological analysis of electron density maps is closely related to the distance matrix based approach, with the difference that other topological features such as pits and saddle points are considered in the interpretation process (Leherte et al., 1997; Menendez-Valazquez & Garcia-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 a 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 seems to be a different problem than building a protein on the basis of free atoms (Isaacs & Agarwal, 1985; Lanzin & Wilson, 1997), it can be shown that the underlying principles are based on the same concepts (Bart & Busetti, 1976). The main difference between the automatic building of protein structures and ligands stems from the repetitive nature of the protein backbone, allowing search strategies where putative peptides, with in advance known stereochemistry, are located and subsequently linked into larger fragments. Ligands usually lack these repetitive motifs and thus other strategies need to be adopted.
4.1. Introduction

The methods for building protein structures implemented in ARP/wARP are based on pattern recognition techniques. A trial polypeptide is constructed on the basis of available free atoms and is accepted for further analyses if some stereochemical and density criteria are met. Decision boundaries for acceptance of a fragment are constructed from empirical estimates of the distributions of noise and signal. These estimates can be obtained from a specific database (Terwilliger. 2003a,b; Kleywegt & Jones. 1996. 1998). The precomputed distributions can be either hard-coded in the implementation, or read in during the course of the program. For ligands the situation is somewhat different. Since the variety of possible ligands is large, precomputing and subsequently hard-coding distributions of geometrical and density features on the basis of a database does not seem a realistic option. Thus one needs to resort to other methods of characterising the distribution of features. One way of generating possible discriminators is using force field techniques and/or approximating binding free-energy functions, as done by protein-ligand docking programs such as Dock (Oshiro & Kuntz, 1995) and AutoDock (Morris et al., 1998). An interesting notion is that these molecular modelling techniques can, from a Bayesian point of view, be seen as a source of prior information for the recognition of a ligand in an electron density map, whereas the correspondence of the molecular model to the X-ray data can be seen as the likelihood term. This viewpoint is closely linked to crystallographic restrained refinement, where the following function is optimised by varying the atomic positions \( \{x\} \):

\[
LL(\{x\}) = \ln[f(\text{chemical sense}|\{x\})] + \sum_h \ln[f(F_h^{\text{obs}}|\{x\})]
\]  

(4.1)

The \( f(F_h^{\text{obs}}|\{x\}) \) term in expression (4.1) models the probability distribution of the X-ray data given the estimated set of atomic positions \( \{x\} \). \( f(\text{chemical sense}|\{x\}) \) expresses the prior knowledge of the stereochemistry of the system. In (protein) crystallography this expression is usually modelled by a product of a set Gaussian distributions centred on the ‘ideal’ values of geometrical features such as distances and angles. When \( f(F_h^{\text{obs}}|\{x\}) \) is also modelled by a Gaussian, expression (4.1) results in a standard least-squares refinement. Modelling the X-ray part of expression (4.1) by a Rice distribution results in the so-called maximum likelihood refinement (Pannu & Read, 1996; Bricogne, 1997b; Murshudov et al., 1997).

The approach we adopted for ligand building is related to the described refinement example. There are two important differences though. First of all, rather than varying the positional parameters in order to optimise the total log likelihood (LL), the positional parameters and available phases are kept fixed and the interpretation in the form of a set of atomic labels is modified to optimise expression (4.1). The second point is the form of the prior probability of our chemical sense. The crystallographic practice of modelling the prior distribution by a (weighted) sum of independent log-probabilities is followed but the individual probability density functions do not have a Gaussian form. As is the
4.2 Methods

4.2.1 A prototype procedure

The flow chart of the basic procedure developed for ligand building is depicted in Figure 4.1. The protein part of a macromolecular model is refined as a rigid body and the resulting phases and figures of merit are used to generate a difference electron density map. An orthogonal grid is constructed from which points are selected that are likely to belong to the ligand. The geometrical features from the ligand prototype and the geometry of the grid are used to construct an error model for the positional parameters of the ligand atoms. A search algorithm designed to optimise the constructed scoring function should not be seen as a log-probability, but rather an approximation to that. For this reason we will use the generic term *scoring function* rather than log-likelihood.

At present no suitable distributions or classifier systems for an incorrect interpretation (noise) has been constructed. The availability of these would be useful in disregarding (partial) interpretations during the construction of a large set of hypotheses, thus effectively limiting the size of the search space.

Although from various topological and geometrical descriptors, distance distributions for each atomic pair could be generated, we will only use information from bonded atoms (1-2 distances), bonding angles (1-3 distances), the chirality of the atoms and information describing van der Waals repulsions. Although a large amount of other information, such as planarity restraints, cis-trans specifications, possible intramolecular hydrogen bonding patterns and unfavourable combinations of specific torsion angles are ignored, the limited amount of geometrical information specified in combination with the electron density heights contains enough information to obtain a suitable estimate of the position, orientation and conformation of the ligand.

The methodology for automated ligand building presented in this chapter has been successfully applied to a number of examples as will be shown in section 4.3. An implementation of the presented methods will be made available in the ARP/wARP suite for automated model building in due course.
4.2. Methods

Figure 4.1: A prototype procedure. Solid lines represent implemented parts. Dotted lines denote parts which can be implemented in the future.

function finally results in ligand atom names assigned to grid points. A chemically sensible ligand model can be then obtained by restrained refinement or idealisation of the ligand-protein complex. Phase improvements obtained from a (partially) build ligand can be used to iterate the procedure in order to look for other ligands. This iterative scheme is reminiscent to the warpNtrace protocol in ARP/wARP.

4.2.2 Trial atom generation

Given a difference electron density map that might contain a ligand or other molecular fragment, we parameterise it by an orthogonal grid with a minimum spacing of $d_{\text{grid}}$ between two grid points of 0.50 Å. The choice of a grid spacing of 0.5 Å is linked to the error model used, as will be explained in section 4.2.3. The orthogonal grid is constructed in such a way that it covers the complete macromolecule, with an added border of appropriate size. Symmetry relations are ignored at this stage. The positional parameters of each grid point are associated with three properties: 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 if the grid point will be used at specific points in the trial atom generation algorithm. The cluster numbers divide the set of grid points in clusters in which the elements are path-connected.

The size of the set of grid points is defined by a threshold value of the electron density. Grid points with a density height lower than the threshold value $\rho_{\text{thres}}$ are discarded. The threshold $\rho_{\text{thres}}$ is determined by a method suggested by Lamzin & Wilson (1997). A
Gaussian distribution is fit to the electron density histogram of the difference map and a threshold is chosen to be equal to that density value where the relative difference between the observed and modelled cumulative distribution is 10%. The selected grid points are clustered using an approach that is related to well-known skeletonisation procedures (Greer, 1974; Swanson, 1994):

0. Set occupancy of all grid points to zero
1. Set occupancy of all grid points with an associated density height larger than \( \rho_{\text{thres}} \) equal to 1; Set the cluster number of all grid points to 0.
2a. Move to the next grid point with an occupancy of 1 that has a neighbouring grid point with an occupancy equal to zero
2b. 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 this grid point disconnects the neighbouring grid points
3. Go to 2a, until all grid points have been visited
4. Set occupancies of the 'to be removed' grid points to 0 go to 2a until nothing changes
5. Assign different cluster numbers to each grid point with a non-zero occupancy

This algorithm, known as constrained erosion, delivers a number of isolated grid points. It can be shown that these remaining grid points have not been connected geodesically, given the definition of the neighbourhood in step 2. In the present implementation, two grid points are defined as neighbours when their distance is smaller than or equal to \( \sqrt{3}d_{\text{grid}} \). The inverse of this algorithm, geodesic reconstruction (Heijmans, 1992), is be applied and yields the original set of grid points grouped into connected clusters:

1. Initialise C to 0.
2. \( C = C + 1 \)
3. Move to the next grid point with occupancy equal to 1 and cluster number equal to C.
4. Select all neighbours of this grid point, set the cluster numbers and occupancies to C and 1 respectively.
5. Go to 3 until no changes occur.
6. Go to 2 until all clusters are constructed.

A similar implementation is discussed by Hunt et al. (1997). The clustering algorithm is illustrated in Figure 4.2.

The number of grid points grouped in a connected cluster, can be used to estimate the volume of the cluster of the difference density. The distribution of these cluster volumes in a typical 2.0 Å difference map that only contains density belonging to solvent atoms
and noise, given a threshold value of twice the standard deviation of the difference map, is shown in Figure 4.3.

The observation that noise and solvent atoms in general have a cluster volume that is smaller than 15 Å³ is used as an additional filter for the grid point selection. In practice, the largest cluster is assigned to the ligand to be build, or in a case of multiple ligands, the volume ordered list of clusters is matched to the list of ligands ordered on their size. Building multiple ligands with similar sizes requires multiple iterations or more elaborate classification schemes. These will not be considered here.

Although the choice of the density threshold used in the clustering algorithm can be obtained via the analysis of the density histogram, another option is to determine it on the basis of the sizes of the obtained clusters. Initial implementations give promising results but require further testing.

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

1. Move to grid point with highest density and occupancy 1.
2. select all grid points within a distance of $d_\rho$ and set their occupancy to 0.
3. go to 1 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 should reflect the bonding distances present in the ligand which is sought and the choice of the grid spacing $d_{grid}$. Setting $d_\rho$ to 1.1 Å for the grid spacing of 0.5 Å gives satisfactory results. Using a peak search algorithm to select possible locations of atoms is another possibility but has the major drawback that at about 2.3 Å resolution atoms are not resolved anymore. Peak picking a difference map with a nominal resolution of 1.5 Å showed that even at that resolution not all ligand atoms correspond to peaks in the difference map, thus resulting in the construction of an incomplete ligand. The grid based selection procedure is insensitive to the shape or topological properties of the electron density around an atom but has a disadvantage of generating a large surplus of initial trial atoms.

4.2.3 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 in section 4.1. The positional parameters of the trial atoms are not continuously distributed as was the case the in previous chapters but follow a discrete, the so-called lattice distribution (Abramovicz & Stegun, 1974; Bricogne, 1974). We assume that the best possible interpretation is the one that maps the ligand atoms to their closest neighbours on the grid. The proposed error model of the positional parameters consists
Figure 4.2: An example of constrained erosion (top) and geodesic reconstruction (bottom). Filled nodes represent grid points with occupancy 1, empty nodes represent grid points with occupancy 0. Numbers inside the nodes represent assigned cluster numbers. Nodes are connected by vertices. The successive states of both algorithms are linked by arrows.

Figure 4.3: The distribution of cluster volumes for a typical 2.0 Å difference map containing only solvent and noise contribution.
thus of a rounding-off operation of the positional parameters of the ‘true’ ligands atoms to positional parameters of the grid. The distribution of interatomic distances after the rounding-off operation can be obtained via simulation. Inclusion of any uncertainty or ‘natural spread’ of a given interatomic distance can be taken into account as well using the following Monte Carlo approach:

1. Draw a random distance $d$ from a distribution $f(d_{jk})$ modelling the natural spread of the distance between the ligand atoms $j$ and $k$.
2. Pick a random point $d$ on a sphere with radius $d$.
3. Apply a random shift on the ‘stick’ $d$.
4. Move the the positional parameters of the ends of the stick to the closest grid neighbour and calculate the distance.

This sampling procedure is applied a large number of times (about a hundred thousand) and the resulting rounded-off distances are stored. The sampling of a point distributed on a sphere is carried out using rejection sampling, the algorithm is outlined in Appendix III-1. The resulting distribution is normalised and stored for further use.

The present implementation of the algorithm outlined in section 4.2.3 uses the Mersenne Twister pseudo random number generator (Matsumoto & Nishimura, 1998) and allows the generation of approximately 150 thousand random distances on a grid per second. Empirical, clearly non-Gaussian, distributions for a distance of 1.5 Å and an orthogonal grid of 0.5 and 0.8 Å are shown in Figure 4.4. The choice of a grid spacing of 0.5 Å is made on the observation that none of the distances between bonded atoms will a probability of becoming equal to zero after the rounding-off operation. For a grid spacing larger than 0.5 this probability increases resulting in a likely event that two atoms are mapped to the same grid point. Furthermore, the 0.5 Å grid spacing ensures that the distance of a ‘true’ ligand atom to the nearest grid point, is smaller then or equal to 0.47Å. This is an error that should lie well within the radius of convergence of restrained refinement procedures for the ligand in question.

4.2.4 The distribution of Chirality

The distribution of the chirality of an atom is constructed in a way similar way to the construction of the distance distributions. Chiral atoms and their bonded neighbours are randomly oriented and placed on a grid. After rounding off the positional parameters to the nearest grid points the chirality is re-computed.

The chirality of an atom is defined by the sign of scalar triple product of the interatomic vectors between the chiral atom $j$ and 3 bonded neighbours $k$, $l$ and $m$:

$$C_j = \text{sign}[d_{jk} \cdot (d_{jl} \times d_{jm})]$$  \hspace{1cm} (4.2)

$d_{jx}$ 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 required as an internal standard. Generation of random orientations is done by sampling from a uniform distribution of points on a 4 dimensional unit sphere. These 4 numbers can be considered to form a quaternion and are used to reorient the fragment under consideration, Appendix III-1. Application of this procedure for a large number of times produces an approximate distribution of the chirality for a given error model. The distribution is obtained after 10,000 trials and takes approximately 1 CPU second. An example distribution of the sign of the chiral volume is shown in Figure 4.5.
4.2.5 Repulsion

Another essential source of information on the internal geometry of a molecular fragment are 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 recognised as possible molecular fragments. The repulsion term used has the following form:

\[ W(d|a,b) = \frac{1}{2}(1 + \tanh((d - a)b)) \]  

(4.3)

By varying \(a\) and \(b\), the location of the inflection point and shape of the repulsion function can be modified, as shown in Figure 4.6. From a probabilistic viewpoint this function could be seen as an improper prior (Bernado & Smith, 2000) on the 1-n (\(n \geq 4\)) 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.

4.2.6 Searching and Scoring

A graph of the known ligand is constructed by assuming that 1-2 distances are uniformly distributed 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 prior to a grid by using the Monte Carlo procedure described in section 4.2.3.

The search procedure starts with the generation of a set of partial interpretations by assigning the label of a given ligand atom to each possible grid points 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. The top \(N_{store}\) partial interpretations are stored. When all possible one atom expansions have been tried, the stored partial interpretations are used for further expansions until completion of the ligand. We call the order in which specific atoms of the ligand are assigned to the grid points the expansion order. By default the first atom to be assigned is the one 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 is gained by adding this atom. The larger the amount of geometrical information available on a partial structure, the easier it is to recognise it as a correct fragment. For this reason, atoms are added to a partial structure in such a way that the increase of information available to score an interpretation is maximum. Conceptually this procedure should minimise the chances that a correct interpretation falls
outside the Nstore best partial interpretation, thus risking convergence to an incorrect interpretation.

The partial interpretations are scored in the following way:

\[ Q(\text{Grid}|\text{Ligand}) = w_{\text{prior}} \sum_m \ln[P_{\text{prior}}(d_m^{\text{obs}}|d_m^{\text{tar}})] + w_c \sum_n \ln[P_C(C_n^{\text{obs}}|C_n^{\text{tar}})] + w_{\text{rep}} \sum_o \ln[W(d_o^{\text{obs}}|a,b)] + w_{\text{dens}} \sum_j \exp[v_L\rho_j] \] (4.4)

\( P_{\text{prior}}(d_m^{\text{obs}}|d_m^{\text{tar}}) \) denotes the probability of the observed distance given the assigned target distance. \( P_C(C_n^{\text{obs}}|C_n^{\text{tar}}) \) gives the probability of the observed chirality after assignment of the labels given the target chirality. These distributions are obtained as described in sections 4.2.3 and 4.2.4. \( W(d_o^{\text{obs}}|a,b) \) denotes the repulsion terms discussed in section 4.2.5. The \( \exp[v_L\rho_j] \) term accounts for the density values on the absolute scale of the atoms \( j \). \( v_L \) has been set to 7.0 and gives satisfactory results. The multipliers \( w_{\text{prior}}, w_c, w_{\text{rep}} \) and \( w_{\text{dens}} \) are relative weights for the contributions of the four features.

Global optimisation 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 optimisation 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. Preliminary implementations of these algorithms did however show successes, but required considerably longer time and fine tuning of parameters in order to converge to the correct solution.

### 4.3 Results

A number of tests has 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 \( b \), equation (4.3), were set to 2.5 and 2.0. The weights \( w_{\text{prior}}, w_c, w_{\text{rep}} \) and \( w_{\text{dens}} \) were set to 2, 10, 5 and 1. 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 with the computed distance limits (section 4.2.6). The maximum number of partial structures stored during each expansion cycle was 5 times the number of putative 1-2 neighbours observed in the set of trial atoms. The characteristics of the used structures and X-ray data sets are summarised in Table 4.1. The procedure has been run with the specified parameters unless stated otherwise. Electron density thresholds in the building were determined by the procedure outlined in section 4.2.1. By default, the interpretation with the highest score has been used to
4.3. Results

Figure 4.6: Repulsion function $W(d|a,b)$ with various choices of location parameter $a$ and shape parameter $b$

Table 4.1: Test data characteristics

<table>
<thead>
<tr>
<th>Structure</th>
<th>$d_{min} (\text{A})$</th>
<th>$B_{Wil}/B_{lig}^+$ ($\text{Å}^2$)</th>
<th>Ligand (non hydrogen atoms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1EE2</td>
<td>1.5</td>
<td>15/13</td>
<td>NADH (44), Cholic Acid (29)</td>
</tr>
<tr>
<td>1OBD</td>
<td>1.4</td>
<td>14/27</td>
<td>ATP (31), AMP (23)</td>
</tr>
<tr>
<td>102D</td>
<td>2.2</td>
<td>38/34</td>
<td>Propamidine (23)</td>
</tr>
<tr>
<td>1A28</td>
<td>1.8</td>
<td>24/25</td>
<td>Progesteron (23)</td>
</tr>
<tr>
<td>1CBS</td>
<td>1.8</td>
<td>13/12</td>
<td>Retinoic Acid (22)</td>
</tr>
<tr>
<td>1LD8</td>
<td>1.8</td>
<td>18/24</td>
<td>FDP$^\dagger$ (24), IC49$^\ddagger$ (33), Sucrose (23)</td>
</tr>
<tr>
<td>Aldolase*</td>
<td>2.1</td>
<td>19/18</td>
<td>DHAP$^#$ (9)</td>
</tr>
</tbody>
</table>

$^+$ Average B value of the ligand atoms

$^\dagger$ Farnesylidiphosphate

$^\ddagger$ Inhibitor compound 49

* Courtesy of E. Lorenztzen and E. Pohl.

# Dihydroxyacetone phosphate
validates the procedure. Table 4.2 summarises the results of the building procedure for all the test structures. Detailed descriptions of the building for each case is given in the next subsections.

4.3.1 1EE2: Cholic acid and NADH

The X-ray data and atomic model of SS-LADH (Adolph et al., 2000), PDB id:1EE2, have been downloaded from the PDB. The asymmetric unit contains two times 374 residues, two Cholic acid molecules, two NADH molecules 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. Initial cluster construction using a grid spacing of 1.5 Å reveals 4 clusters of connected density with a volume larger than 80 Å³. The clusters with the approximate volumes of 150 Å³ are interpreted as NADH, the clusters with volumes around 85 are assigned as possible Cholic acids. Constructing a 0.5 Å grid around each cluster resulted in a set of trial atoms that were interpreted given the assigned ligand type. Figure 4.7 shows the initial difference density with the placed grid atoms and the model after refinement with REFMAC5 for one of the the Cholic acid clusters. The rmsd of the build model to the deposited model is 0.09 Å.

The building of NADH resulted in a structure with an rmsd of 0.11 Å to the deposited
Figure 4.7: Trial atoms (left) and refined interpretations (right) of Cholic Acid (top) and NADH (bottom) in the original difference density.
coordinates, Figure 4.7. In order to prevent the algorithm to discard 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 4 from the default value.

4.3.2 1OBD: ATP and AMP

The atomic model of saicar synthetase, PDB id: 1OBD, contains AMP and ATP. Because of the ill-defined difference density the described clustering procedure was unable to determine the locations of the ligands within a reasonable amount of time. For this reason, knowledge of the positions of the ATP and AMP have been used in the interpretations. The building and subsequent refinement of ATP resulted in a structure matching the deposited coordinates (rmsd = 0.09 Å), Figure 4.8. Building of AMP was unsuccessful due to the ill-defined/absent difference density for the phosphate and sugar moiety. The deposited AMP structure has an occupancy of 0.5. A 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.
4.3.3 102D: Propamidine

Locating Propamidine in 102D, a double stranded DNA structure, has been carried out using 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 idealisation of the geometry resulted in the placement of the ligand with a different conformation as compared to the deposited structure, Figure 4.9. In the same Figure the best 6 geometrised interpretations are shown. Restrained refinement did not improve the fit of the interpretation to the deposited structure. The relatively weak density of a part of the Propamidine molecule possibly explains the difference between the deposited and automatically build structure.

4.3.4 1A28: Progesteron

The position of the steroid in 1A28, a human progesteron receptor, was located using default parameters. The built and deposited model differ the orientation of the keto group, Figure 4.10. 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 considerations are however not taken into account in our scoring function. The rmsd of the build and refinement structure from the deposited coordinates (apart from the flipped keto group) is equal to 0.17 Å.
Figure 4.10: Difference density with trial positions (left), non-geometrised interpretation (middle) and deposited structure of Progesteron (right). The * denotes the position of the keto oxygen that is different in the interpretation and deposited structures.

4.3.5 1CBS: Retinoic acid

The retinoic acid in the difference electron density of 1CBS, a retinoic acid transport protein, was built using default parameters, Figure 4.11. The rmsd of the build model to the deposited model was 0.22 Å.

4.3.6 1LD8: Farnesylphosphate (FDP), Inhibitor compound 49 (IC49) and Sucrose

Location of each ligand in the difference density of 1LD8, human Farnesyltransferase, was carried as follows. The largest 3 difference density clusters could be assigned to the individual ligands on the basis of the volume rankings. Once one ligand has been build, the protein-ligand complex was re-refined and the new density map has been subsequently used to build the remaining ligands. Due to the size of the ligands, the number of intermediate partial interpretations was increased by a factor of 2. Whereas IC49 was build and refined to an rmsd of 0.28 Å, FDP was build in a cis rather than trans conformation compared to the deposited structure, Figure 4.12. Attempts to build sucrose failed under various settings. This is ascribed 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.
Figure 4.11: Difference density with trial positions (left) and refined interpretation (right) of Retenoic acid (right).

Figure 4.12: Difference densities and refined interpretations for IC49 (left) and an overlay of the interpretation and deposited FDP model (right). The * marks the incorrectly build part of FDP.
4.3.7 Aldolase: Dihydroxyacetone phosphate (DHAP)

The location of the ligand was found using the clustering algorithm around the residues where the ligand was known to bound a priori. Inclusion of density terms was crucial in this example. Without this information, the ligand could not have been built automatically. The refined interpretation is shown in Figure 4.13. Inclusion of protein-ligand interactions would have made the interpretation easier, as DHAP is covalently bound to the protein.

4.4 Discussion and Conclusions

The modelling of the distribution of distances via a marginalisation of a lattice distribution proved an adequate tool in modelling prior geometrical knowledge in grid-based model building routines. Least-squares or non-central Maxwell target functions perform worse, as can be expected from Figure 4.4. Although analytical distributions can possibly be derived, the approximate distributions are fairly quickly obtained via the Monte Carlo simulations. Approximate distributions of relative complex quantities, such as the distribution of the sign of the chiral volume of an atom, can be obtained using simulations in a relative straightforward way. It must be noted that the constructed error model on the positional parameters is an approximation. First of all, distances are not independent and correlations should in principle be taken into account. If efficient ways of storing and handling multidimensional distributions of geometric features can be implemented, one could attempt to obtain the joint probability distribution of all geometric features within the search molecule. The second approximation made is that 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 has most likely a distribution that is different from the one constructed in our simulations and probably depends on the spatial distribution of density heights within the cluster. However, the designed classifier proved good enough to recognise the correct solution. A similar situation is present for handling the available prior knowledge. Currently, a minimalistic approach is pursued: our prior knowledge is limited to 1-2 and 1-3 distances, chiral signs and repulsions. Even with this limited amount of information one is able to recognise 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 position so REFMAC5 is able to optimise the geometry of the ligand. Ideally, a real space refinement optimising both the positional parameters and the fit to the electron density, should be carried out on all proposed interpretations. Tuning the parameters of the scoring function for the density to enhance the performance of the algorithm is currently being investigated. As seen from the progesteron example and, to a certain extent, the Dihydroxyacetone Phosphate 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 placements of ligands and could possibly limit the search space. A similar approach would also be useful for building structure such as glycosilation sites or other molecules with internal repeats. 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 research program may be its speed: most ligands were build in approximately 10 minutes, whereas ATP took 15 minutes and NADH about half an hour. In future implementations the building algorithm will be optimised for CPU efficiency, increasing the speed substantially. 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 section 4.3.2, to be carried out automatically. The present version of the ligand construction algorithm will be distributed as a β-test module within ARP/wARP version 6.1.