Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Crystallographic refinement of ligand complexes

Gerard J. Kleywegt

Department of Cell and Molecular Biology, Uppsala University, Biomedical Centre, Box 596, SE-751 24 Uppsala, Sweden

Correspondence e-mail: gerard@xray.bmc.uu.se

Received 3 March 2006
Accepted 13 June 2006

Model building and refinement of complexes between biomacromolecules and small molecules requires sensible starting coordinates as well as the specification of restraint sets for all but the most common non-macromolecular entities. Here, it is described why this is necessary, how it can be accomplished and what pitfalls need to be avoided in order to produce chemically plausible models of the low-molecular-weight entities. A number of programs, servers, databases and other resources that can be of assistance in the process are also discussed.

1. The null hypothesis

When the crystal structure of a complex between a macromolecule and a small molecule is determined, the null hypothesis is usually: 'My crystal contains the compound I soaked in or cocrystallized with and it has ideal geometry'. The assumption of ideal geometry is usually warranted, although one has to keep in mind that deviations may occur owing to steric strain, unexpected effects of pH or ionic strength *etc*. However, the first assumption should indeed be that the geometry is 'ideal' and only very convincing density in high-resolution maps should be allowed to tempt one to depart from that assumption. In most cases, the major problem will be to define the restraints that are necessary to impose the ideal geometry, as well as to find the appropriate ('ideal') target values for those restraints. This issue is discussed in detail below. However, before discussing restraints we should briefly examine the other assumption that is made in the null hypothesis, namely that the crystal contains the expected compound. There are a number of circumstances that can invalidate this assumption. A trivial one is the fact that the crystal is bound to contain much more than just the macromolecule and the small molecule of interest: any molecules retained during purification, components of the crystallization soup, cryoprotectant *etc*. In many cases, therefore, interpreting density features can be a major obstacle in and of itself. Some of the automated methods described elsewhere in this issue may be of assistance in such cases (Evrard *et al.*, 2007; Terwilliger *et al.*, 2007). In addition, as the examples below show, compounds (known and unknown) may undergo chemical reactions, 'known' compounds may turn out to be something completely different and sometimes a putative ligand simply does not bind or binds with too low an occupancy to give a clear feature in the electron density.

Depending on the nature of the small molecule and the environment inside the protein, a ligand may be reduced or oxidized (*e.g.* sulfur- or metal-containing compounds), it may form dimers (*e.g.* β-mercaptoethanol may dimerize to form

2-hydroxyethyldisulfide), it may turn out to be an unexpected substrate or it may react with the protein or other components present in the crystal. An interesting example of an unexpected reaction taking place in a crystal (albeit with an unusual amino acid rather than a ligand) was encountered in the structure determination of a methanogen methyltransferase, the first known protein to contain a copy of the 22nd naturally occurring amino acid, L-pyrrolysine (Hao *et al.*, 2002). Crystals were obtained with both sodium chloride and ammonium sulfate. However, the unusual amino acid had undergone a spontaneous addition reaction (of an amine group; 60% occupancy) in the crystals grown with ammonium sulfate (Hao *et al.*, 2002).

A communication problem caused confusion during the refinement of a complex of cellular retinoic acid-binding protein II with a synthetic retinoid that was supposed to be TTNPB (Fig. 1; Kleywegt *et al.*, 1994). However, persistent features in subsequent difference maps suggested that the ligand was something else. After consultation with the synthetic chemists half a world away, it turned out that the compound they had supplied was in fact a different synthetic retinoid, 'compound 19' (Fig. 1; Kleywegt *et al.*, 1994; Davis *et al.*, 2003).

Sometimes a ligand simply does not bind (or binds with too low occupancy or with too much disorder) and this may explain what happened in the structure determination of a complex between botulinum neurotoxin type B protease and an inhibitor (Hanson *et al.*, 2000). Close inspection of the maps after publication convinced the authors that these did 'not support the placement of the inhibitor as stated in the paper' and the structure was retracted (Hanson *et al.*, 2002; Fig. 2).

2. The need for restraints

Macromolecular X-ray crystallography is a notoriously poor method for determining the structure of small molecules that are bound to macromolecules and it has been pointed out by a number of people that the stereochemical quality of more than a few small-molecule structures encountered in the worldwide Protein Data Bank (wwPDB; Berman *et al.*, 2003) is less than overwhelming (van Aalten *et al.*, 1996; Kleywegt & Jones, 1998; Kleywegt, 2000; Boström, 2001; Nissink *et al.*, 2002; Davis *et al.*, 2003; Kleywegt *et al.*, 2003; Schüttelkopf & van Aalten, 2004; Lütteke & von der Lieth, 2004). Part of the explanation of this phenomenon lies in the general limitations of macromolecular crystallography, namely limited resolution (and information content) and weak data (leading to a low signal-to-noise ratio). This means that in typical cases the data-to-parameter ratio is of the order of 0.5–5, where one would prefer to have values in excess of 10. The lack of data can to some extent be compensated for by the use of prior knowledge in the model refinement process. The data-to-parameter ratio can be improved by reducing the number of model parameters (by applying constraints) or by increasing the number of observations (in the form of restraints). A constraint imposes an exact condition and thereby removes one or more parameters from the model. Examples of constraints include the use of strict noncrystallographic symmetry (NCS), rigid-body refinement, refinement of overall or grouped temperature factors and model parameterization in torsion-angle space (in which bond lengths and angles can be kept fixed during refinement). A restraint expresses empirical knowledge (or expectations) regarding the chemistry or physics of a system in the form of a condition on one or more parameters (often in the form of a target value for a single parameter, with some indication of the allowed deviations from that value). Examples include restraints on bond

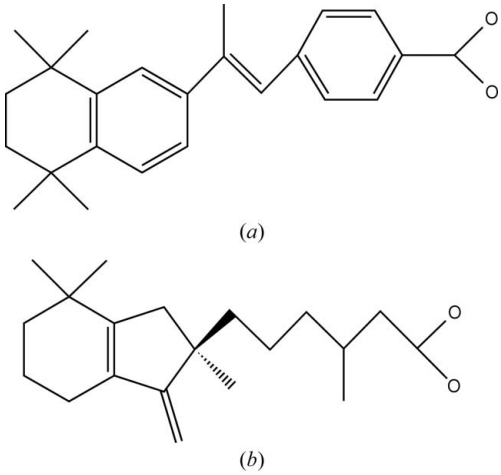

Figure 1
Chemical structure diagrams of (*a*) TTNPB and (*b*) 'compound 19'.

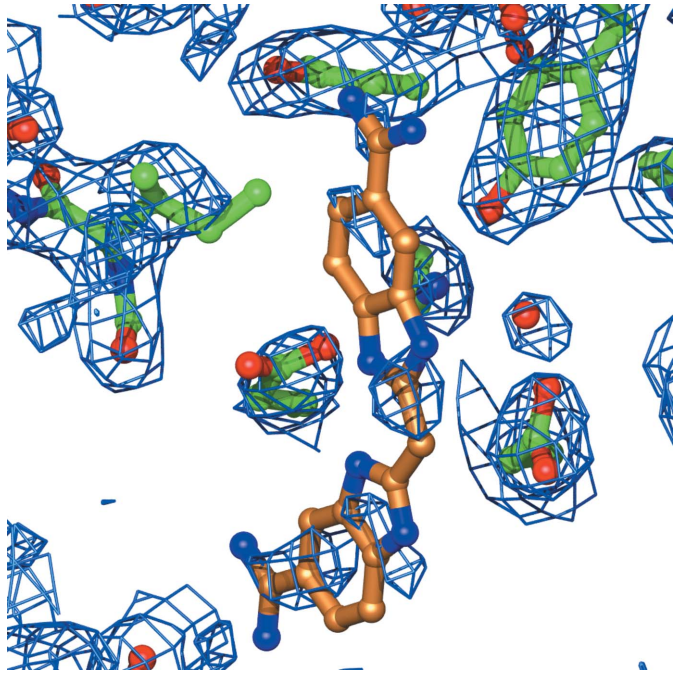

Figure 2
Electron density for the inhibitor BABIM (shown with gold C atoms) in its complex with botulinum neurotoxin type B protease (Hanson *et al.*, 2000, 2002). The map is a $2mF_o - DF_c$ synthesis, calculated with all deposited data (2.5 Å), and taken from EDS (Kleywegt *et al.*, 2004). Figs. 2 and 3 were created with *O* (Jones *et al.*, 1991) and *MolRay* (Harris & Jones, 2001).

lengths (either by specifying a target length or by specifying that all bonds of a certain type should have roughly the same length), bond angles, certain 'fixed' torsion angles, planar groups, repulsion between non-bonded atoms and temperature-factor differences between related atoms.

Refinement programs incorporate restraints into the target function (*i.e.* the function that is minimized, which can be a least-squares, maximum-likelihood or energy-based function) by adding empirical restraint functions that take different functional forms depending on the nature of the restraints. For instance, bond-length restraints are conveniently implemented by adding a quadratic penalty or cost function of the type

$$\varphi_{\rm bonds} = \sum_{\rm bonds} \omega (d_{\rm model} - d_{\rm ideal})^2,$$

where ω is a weight, $d_{\rm model}$ is a bond length in the model, $d_{\rm ideal}$ is the target value for that bond and the sum extends over all covalent bonds in the model. For restrained refinement of a model three things are needed: a set of definitions (atom types, bonds, angles, planar groups *etc.*), a set of target ('ideal') values for the restraints and appropriate weights for the individual restraints and for the restraint functions (to determine the relative importance of the experimental data and the restraints). In the following, this collection of items will be called a restraint set, but it goes by many other names: (stereochemical) dictionary, library, force field or topology and parameter definitions. The various types of (stereochemical) restraints and their use in refinement will not be discussed here. Instead, the reader is referred to the paper by Evans in this issue (Evans, 2007), to other review papers (Hendrickson, 1985; Kleywegt *et al.*, 2003; Tronrud, 2004) and to standard textbooks.

3. Intelligent design

Biomacromolecules are (mostly linear) polymers composed of a limited repertoire of units (amino acids, nucleotides *etc.*). For the purposes of restraint-set definition, this means that only a limited set of restraint specifications are required to cover most cases. Indeed, after the seminal work of Engh & Huber (1991, 2001) such specifications are now available for all popular refinement and model-building programs, both for proteins (for bond lengths and angles, although Priestle has also derived restraints for some torsion angles; Priestle, 2003) and nucleic acids (Parkinson *et al.*, 1996). For other entities ('heterocompounds') the situation is less favourable, although restraints for common compounds are often provided with the programs. In principle, there is an infinite variety of possible compounds that can be complexed with biomacromolecules and for every one of these the crystallographer will have to obtain a sensible set of restraints and a sensible starting model. The problem is alleviated somewhat through the use of atom-typing techniques where atoms with similar physical and chemical properties are treated the same (*i.e.* they have the same restraint target values and weights) in all compounds they occur in. Atom types depend on the chemical element type, the hybridization state, the charge, the number of attached H atoms (implicit or explicit) and the chemical environment. For instance, in many restraint sets for the programs *X-PLOR* (Brünger, 1992) and *CNS* (Brünger *et al.*, 1998), an sp^3-hybridized C atom with two (implicit) H atoms attached to it is assigned the atom type CH2E, although Engh & Huber define two extra types, namely CH2P (in prolines) and CH2G (in glycines). Many bond-length and bond-angle restraints have already been defined for such atom types, which reduces the onus on the crystallographer when creating restraint sets for new compounds. For instance, the compound benzene can simply be specified to consist of six C atoms of type CR1E that form a six-membered ring. Since this atom type also occurs in phenylalanine residues, the target values and weights for the bond lengths and angles will automatically be the same as those defined by Engh & Huber.

The use of high-quality restraint sets is especially important for small-molecule ligands since the determination of their conformation, binding mode and interactions with the macromolecule is typically the main reason for determining the crystal structure of the complex in the first place. Although some crystallographers recycle the restraint sets of colleagues, in general the evolution of such sets does not lead to a high level of quality. This is one area where 'intelligent design' is to be preferred. The key both to generating and validating restraint sets (*a priori*) and to validating the resulting geometry (*a posteriori*) is a thorough understanding of the chemistry of the compound. This enables one to define the types of all atoms and to specify all the necessary restraints. The general rules for specifying stereochemical restraints are fairly straightforward (Kleywegt *et al.*, 2003; Evans, 2007).

(i) Each pair of bonded atoms yields one bond-length restraint.

(ii) Two pairs of bonded atoms that have one atom in common yield one bond-angle restraint.

(iii) A tetrahedral C atom with four different neighbours (possibly including an implicit H atom) yields one chirality restraint.

(iv) A (partial) double bond (as in carboxylate groups, aromatic rings, conjugated systems, peptide bonds *etc.*) implies that the atoms involved, as well as all their direct neighbours, lie in one plane. They thus require planarity restraints and, in some cases, a specification of whether an arrangement is *cis* or *trans*.

(v) A triple bond (or two consecutive double bonds, as in some aza compounds) requires a linearity restraint.

Particular attention is required when covalent links between distinct entities are to be defined. This occurs, for instance, when a suicide inhibitor has reacted with a catalytic residue, when a post-translational modification has occurred on an amino-acid residue or when a ligand consists of multiple hetero-entities (such as oligosaccharides). In such cases, bond lengths, angles and torsion angles need to be defined that involve atoms from two separate entities (*e.g.* an amino acid and a carbohydrate). In addition, a C atom that is achiral in the isolated compound may become chiral when it is linked to another entity. A related phenomenon may explain why there are a few dozen instances of 2-(acetylamino)-2-deoxy-α-D-glucopyranose in the wwPDB (where it is labelled NDG); the

compound is identical to *N*-acetyl-D-glucosamine (NAG), with the exception of the chirality of the C1 atom that links it to an asparagine residue. It seems likely that some or all of these are really NAGs that have been refined without a chirality restraint (or with the wrong target value).

It is important to realise that some restraints are interdependent and others are even redundant (Kleywegt, 2000; Tronrud, 2004). For instance, if bond angles are restrained by the corresponding 1–3 distances, the restraints implicitly also restrain the two 1–2 bond lengths that are involved, and a similar situation arises when 1–4 distances are used to restrain torsion angles.

Prior knowledge regarding restraint target values (in particular, for bond lengths and angles) can be obtained from different sources, for instance by recycling previously defined atom types or by looking them up in compilations in papers, books or websites, or by calculating them from a high-quality (crystal) structure of the compound of interest. Conformational torsion angles are not usually restrained and target values for other restraints (chirality, planarity) tend to follow immediately from the chemistry of the system (*e.g.* a torsion-angle restraint to enforce a *trans* arrangement around a double bond implies a restraint target of 180°).

The proper way to define restraint sets is to perform a detailed analysis *à la* Engh & Huber. Besides appropriate target values, such an analysis also yields reasonable estimates of the standard deviations of these values. One of the few examples of such an analysis is the work of Lancaster & Michel (1997) on the cofactors encountered in the photosynthetic reaction centre. For energy-based methods, weights (or, rather, 'force constants') have sometimes been derived from experimental data (*e.g.* from infrared spectra). However, the most common method for defining weights is to simply use values that are in the same ballpark as those used for proteins. For bond lengths, the standard deviation is typically set to 0.02 Å (or the corresponding force constant to 1000 kCal mol^{-1} Å^{-2}); for bond angles a value of 2° is often used (or a force constant of 500 kCal mol^{-1} deg^{-2}).

4. The twilight zone

Since the construction of high-quality restraint sets is not trivial, it should come as no surprise that examples of 'unusual' ligand stereochemistry abound in the wwPDB (van Aalten *et al.*, 1996; Kleywegt & Jones, 1998; Kleywegt, 2000; Boström, 2001; Nissink *et al.*, 2002; Davis *et al.*, 2003; Kleywegt *et al.*, 2003; Schüttelkopf & van Aalten, 2004; Lütteke & von der Lieth, 2004). A small number of examples are shown in Fig. 3. Manual inspection of a large number of such anomalies suggests that there are a number of different problems that may occur.

(i) Restraints that should have been applied have been omitted (or had too low a weight to have had any impact). This may explain many large distortions of bond lengths and angles, as well as unexpected deviations from planarity and incorrect chirality.

Table 1
URLs for some of the resources mentioned in the text.

Resource	URL
A La Mode	http://ndbserver.rutgers.edu/alamode/
MSDChem	http://www.ebi.ac.uk/msd-srv/msdchem/cgi-bin/cgi.pl
HIC-Up	http://xray.bmc.uu.se/hicup/
PRODRG	http://davapc1.bioch.dundee.ac.uk/programs/prodrg/prodrg.html
Ligand Depot	http://ligand-depot.rutgers.edu/
CSD	http://www.ccdc.cam.ac.uk/products/csd/
ICSD	http://icsd.ill.fr/dif/icsd/
COD	http://www.crystallography.net/
Reciprocal Net	http://www.reciprocalnet.org/
NCI Open Database	http://cactus.nci.nih.gov/ncidb2/
SWEET	http://www.dkfz-heidelberg.de/spec/sweet2/doc/index.php
RESID	http://www.ebi.ac.uk/RESID/

(ii) Restraints that should not have been included have been applied. This can result in such anomalies as C atoms in aromatic rings having a tetrahedral arrangement of their neighbour atoms, of phosphates being trigonal or tetragonal pyramids *etc.*

(iii) Restraints have been applied with incorrect target values. This may, for example, lead to carbon–carbon 'double' bonds with lengths of 1.5 Å.

(iv) Finally, there are many errors that cannot easily be explained in terms of incorrect restraints and that are unlikely to have been the result of a refinement run. Examples of non-bonded contacts shorter than 1 Å and of covalent bond lengths in excess of 5 Å can be found. These are possibly the result of *a posteriori* modifications to the model (either with a text editor or by dragging atoms around in a modelling program) which have not subsequently been regularized by a refinement program.

It is worth noting that errors in ligand stereochemistry occur in structures in essentially the entire resolution spectrum (Kleywegt *et al.*, 2003). This merely demonstrates that the X-ray data alone are insufficient to define the stereochemistry of small molecules (although incorrect restraints hardly help either, of course).

It is important to realise that a restraint set is in essence a specification of the ideal stereochemistry of a compound. In the best of worlds all the restraints will be satisfied, but the old adage 'garbage in, garbage out' applies. If there are incorrect restraints or restraints with incorrect target values, one should not be surprised to find that the refinement program produces a chemically implausible model. Similarly, where freedom is given (*i.e.* where necessary restraints are omitted or given too low a weight), liberties will be taken: a refinement program cannot be more intelligent than its user (yet). Consequently, the best way to prevent errors in the first place is to make sure that both the restraint set for a compound and its starting model are of high quality. A useful way to validate a restraint set is to randomize the coordinates of a ligand and subsequently refine it in isolation (*i.e.* without protein *etc.* and without use of the X-ray data). If the resulting geometry is chemically implausible this means that the restraints are incomplete, erroneous or conflicting.

5. Tools of the trade

Fortunately, there are a number of resources available to crystallographers who need plausible starting models and reasonable restraint sets for small molecules. A few resources will be discussed here (links are listed in Table 1); several others can be found in a previous review (Kleywegt *et al.*, 2003).

Restraint sets can be specified by anyone with a good knowledge of chemistry, but the process is tedious, time-consuming and error-prone (Pähler & Hendrickson, 1990). Restraint sets from colleagues should be shunned as a rule, unless there are strong indications that the colleague is considerably more skilled, patient and conscientious than oneself. A collection of validated dictionaries for various nucleotide units is available from the A La Mode website (Clowney *et al.*, 1999). Restraint sets for *REFMAC* (Murshudov *et al.*, 1997) can be generated from SMILES strings (Weininger, 1988; Weininger *et al.*, 1989) with *AFITT* (Peat *et al.*, 2005) and with *CCP*4 software (Greaves *et al.*, 1999; Vagin *et al.*, 2004). These programs can also be used to draw two-dimensional diagrams of ligands that can be converted into structures and restraint sets.

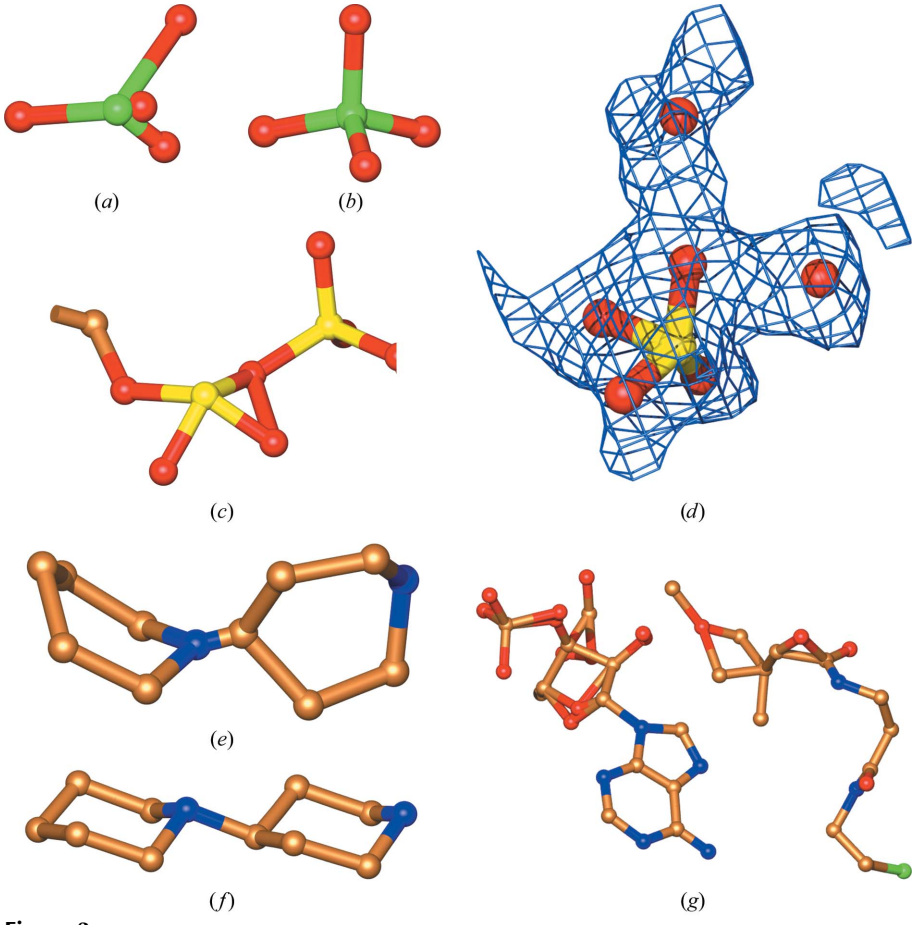

Figure 3
Examples of errors in heterocompounds encountered in contemporary wwPDB entries. (*a*) A sulfate ion as found in a 1.65 Å structure from 1999. One of the O atoms lies in an obviously impossible location. (*b*) Geometry of an 'ideal' sulfate from MSDChem. (*c*) Detail of an FAD molecule found in a 2.3 Å structure from 2005. One of the two phosphates has been subjected to incorrect restraints (in both copies in the asymmetric unit), forcing it into a tetragonal pyramidal structure. The other phosphate has its neighbouring O atoms in the proper tetrahedral arrangement. (*d*) A different, but equally wrong, phosphate. This 2.0 Å structural genomics structure from 2002 contains a phosphate forced into a trigonal pyramidal arrangement, with all four P—O bonds shorter than 1.5 Å (suggesting, incorrectly, that all four are double bonds). In the vicinity of this phosphate there is a large unoccupied density feature that looks as if it could also accommodate a phosphate ion (not shown). A nearby residue has density features that show that its peptide bond needs to be flipped (not shown). These uninterpreted yet obvious density features suggest that the maps have not been inspected with a great degree of enthusiasm. (*e*) The N atom in this ligand (found in a 2.5 Å structure from 2001) appears to have been forced to be planar. In addition, the bond from the N to the C atom in the other ring is implausibly short (0.8 Å). (*f*) The 'ideal' structure of the ligand in (*e*), taken from MSDChem. The r.m.s. deviation from ideal values of the bond lengths in the experimental structure is 0.2 Å and the r.m.s. deviation of the angles is 8°. (*g*) This poor impersonation of a coenzyme A molecule is found in a 2.25 Å structure from 2003. It contains non-bonded distances as short as 0.54 Å, bonded distances as long as 6.7 Å and bond angles as small as 18°.

The MSD database contains a component called MSDChem that holds a wealth of information about all hetero-entities that occur in any wwPDB entry (Golovin *et al.*, 2004). Atom types are available for *CCP*4 and *CNS*, the order, length and stereochemistry of bonds is described, both experimental and 'ideal' coordinate sets are available, *REFMAC* dictionaries can be exported *etc*. The ideal structures have been generated from SMILES strings with the program *CORINA* (Gasteiger *et al.*, 1990).

HIC-Up (Kleywegt & Jones, 1998) is a repository of information about hetero-entities that occur in the wwPDB. It began in the mid-1990s as a collection of restraint files for use with *X-PLOR* that had been derived from coordinate sets taken from wwPDB entries. Nowadays, restraint sets are available for *X-PLOR*/*CNS*, *O* and *TNT* and most of them have been derived from the ideal coordinate sets from MSDChem (as these are often of higher quality than those taken directly from the wwPDB entries). In addition to these coordinate and restraint sets, HIC-Up also provides a number of links to external sites for every entry as well as statistics derived from data stored at the Electron Density Server (EDS; Kleywegt *et al.*, 2004); for an example of the latter, see Table 2. The links from HIC-Up to EDS enable crystallographers to assess quickly how the fit of their ligand to the density compares with what has been observed in other structures at similar resolution. They may also be of use in cases where interpretation of the density is ambig-

Table 2
Example of statistics derived from the Uppsala Electron Density Server (EDS; Kleywegt *et al.*, 2004) that are available from HIC-Up (Kleywegt & Jones, 1998).

The table shows real-space *R* value (RSR) statistics (Jones *et al.*, 1991) for the heterocompound NAG (*N*-acetyl-D-glucosamine). The sample statistics enable one to assess whether a particular instance of this compound fits well or poorly at a given resolution in comparison to other structures at similar resolution (note that lower RSR values indicate a better fit of the model to the density). On the HIC-Up pages, the PDB codes in the last two columns are in fact links to the corresponding entries in EDS. This enables one to quickly access electron-density maps for particularly well or particularly poorly fitting instances of this compound in any resolution range for which a sufficient number of instances have been observed. The *Z* scores indicate how many sample standard deviations the observed values lie removed from the sample average [$Z_i = (\mathrm{RSR}_i - \langle\mathrm{RSR}\rangle)/\sigma_{\mathrm{RSR}}$]. In addition to RSR statistics, HIC-Up also lists EDS-derived statistics pertaining to the real-space correlation coefficient and isotropic temperature factors.

Resolution range (Å)	Instances	Average RSR	St. dev. RSR	Minimum (PDB code; *Z* score)	Maximum (PDB code; *Z* score)
5.0–3.0	435	0.30	0.13	0.11 (1ism; −1.4)	0.81 (1rer; 3.9)
3.0–2.8	452	0.29	0.12	0.10 (1h15; −1.6)	0.84 (1i1a; 4.5)
2.8–2.6	456	0.27	0.10	0.10 (1fq4; −1.7)	0.89 (1ht5; 6.0)
2.6–2.4	367	0.27	0.13	0.05 (1ncb; −1.7)	0.94 (1rd3; 5.2)
2.4–2.2	344	0.26	0.13	0.09 (1dp5; −1.4)	0.73 (1jmj; 3.7)
2.2–2.0	371	0.22	0.11	0.07 (1rem; −1.4)	0.69 (1re2; 4.3)
2.0–1.8	485	0.19	0.11	0.06 (1ur9; −1.2)	0.61 (1u2y; 3.8)
1.8–1.6	266	0.20	0.12	0.05 (1lmq; −1.2)	0.75 (1h4p; 4.6)
1.6–1.4	168	0.17	0.12	0.04 (1oc6; −1.1)	0.60 (1qwo; 3.6)
1.4–1.2	31	0.19	0.11	0.05 (1jnd; −1.3)	0.45 (1lk2; 2.3)
1.2–0.0	14	0.15	0.08	0.04 (1uwc; −1.4)	0.31 (1qwn; 2.1)

uous. A separate server is available to generate restraint sets directly from coordinate files; this can be used for compounds that are not yet covered by HIC-Up.

PRODRG (van Aalten *et al.*, 1996; Schüttelkopf & van Aalten, 2004) is a versatile server for generating coordinates and restraint sets for a wide variety of refinement, docking, modelling and molecular-dynamics programs. *PRODRG* can handle C, N, O, S, P, Cl, I, Br and F atoms, which covers a large fraction of all ligands in the wwPDB as well as most pharmaceutically relevant compounds. Input to the server can be provided as a two-dimensional chemical diagram or an ASCII text drawing. A set of three-dimensional coordinates can also be supplied but this is actually discouraged.

There are many resources that can be used to obtain a chemically reasonable starting model of small-molecule ligands. Experimental coordinates can be extracted from the wwPDB (with all the associated caveats), either from wwPDB entries directly or from derived databases such as MSDChem, HIC-Up and Ligand Depot (Feng *et al.*, 2004). Potentially more reliable experimental coordinates can be found in databases that contain small-molecule crystal structures. Traditionally, chemical databases have not been in the public domain and this is how two crystallographic databases, the Cambridge Structural Database (CSD) and the Inorganic Crystal Structure Database (ICSD), still operate today. However, in recent years at least two databases have been set up that make such structures available free of charge: the Crystallography Open Database (COD) and Reciprocal Net. Although their coverage is considerably smaller than that of the CSD, they are a good starting point, in particular for macromolecular crystallographers without access to the other databases.

Structures of small molecules can also be calculated without resorting to crystallographic data. Many packages are available that calculate structures using *ab initio*, semi-empirical or molecular-mechanics methods. A program that is specifically tailored to producing restraint sets and that can handle metals is *Hess2FF* (Nilsson *et al.*, 2003). There are also many programs that can convert one-dimensional representations (such as SMILES strings) or two-dimensional diagrams into a set of plausible three-dimensional coordinates, including *CORINA*, *PRODRG* and *AFITT*. *CORINA* has also been used in the construction of the NCI Open Database, a freely accessible database containing three-dimensional coordinates for more than a quarter of a million compounds. Two useful specialized resources are RESID (Garavelli, 2004), a database with (model) structures and information for around 400 types of modified and cross-linked amino-acid residues, and *SWEET* (Bohne *et al.*, 1999), a server to generate model structures of simple and complex carbohydrates. A companion resource to *SWEET* is *PDB-CARE*, which is designed to validate carbohydrate structures (Lütteke & von der Lieth, 2004).

Note added in proof. After this paper had been accepted, the author found out about two more useful resources, similar to the NCI Open Database. ChemDB (http://cdb.ics.uci.edu/CHEM/Web/; Chen *et al.*, 2005) and ZINC (http://blaster.docking.org/zinc/; Irwin & Shoichet, 2005) both provide calculated coordinates for more than 4 million compounds.

The author wishes to acknowledge the many people with whom he has discussed issues of refinement and validation of ligand complexes over the past 15 years and on whose shoulders he was trampling while writing this contribution. These giants include, among many others, Alwyn Jones (Uppsala), Eleanor Dodson (York), Phil Evans (Cambridge), Kim Henrick (Hinxton), Axel Brunger (Stanford), Paul Adams (Berkeley), Dale Tronrud (Eugene), Alexander Schüttelkopf and Daan van Aalten (Dundee), Andy Davis and Simon Teague (Charnwood), Gert Vriend (Nijmegen) and Roman Laskowski (Hinxton). The author is a Royal Swedish Academy of Sciences (KVA) Research Fellow, supported through a grant from the Knut and Alice Wallenberg Foundation. He is supported by KVA, the Swedish Structural Biology Network (SBNet), Uppsala University and its Linnaeus Centre for Bioinformatics.

References

Aalten, D. M. F. van, Bywater, R., Findlay, J. B. C., Hendlich, M., Hooft, R. W. W. & Vriend, G. (1996). *J. Comput. Aided Mol. Des.* **10**, 255–262.

Berman, H., Henrick, K. & Nakamura, H. (2003). *Nature Struct. Biol.* **10**, 980.

Bohne, A., Lang, E. & von der Lieth, C. W. (1999). *Bioinformatics*, **15**, 767–768.

Boström, J. (2001). *J. Comput. Aided Mol. Des.* **15**, 1137–1152.

Brünger, A. T. (1992). *X-PLOR. Version* 3.1. *A System for X-ray Crystallography and NMR.* Yale University, Connecticut, USA.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Chen, J., Swamidass, S. J., Dou, Y., Bruand, J. & Baldi, P. (2005), *Bioinformatics*, **21**, 4133–4139.
Clowney, L., Westbrook, J. D. & Berman, H. M. (1999). *J. Appl. Cryst.* **32**, 125–133.
Davis, A. M., Teague, S. J. & Kleywegt, G. J. (2003). *Angew. Chem. Int. Ed.* **42**, 2718–2736.
Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Engh, R. A. & Huber, R. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 382–392.
Evans, P. R. (2006). *Acta Cryst.* D**63**, 58–61.
Evrard, G. X., Langer, G. G., Perrakis, A. & Lamzin, V. (2006). *Acta Cryst.* D**63**, 108–117.
Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H. M. & Westbrook, J. (2004). *Bioinformatics*, **20**, 2153–2155.
Garavelli, J. S. (2004). *Proteomics*, **4**, 1527–1533.
Gasteiger, J., Rudolph, C. & Sadowski, J. (1990). *Tetrahedron Comput. Methods*, **3**, 537–547.
Golovin, A. *et al.* (2004). *Nucleic Acids Res.* **32**, D211–D216.
Greaves, R. B., Vagin, A. A. & Dodson, E. J. (1999). *Acta Cryst.* D**55**, 1335–1339.
Hanson, M. A., Oost, T. K., Sukonpan, C., Rich, D. H. & Stevens, R. C. (2000). *J. Am. Chem. Soc.* **122**, 11268–11269.
Hanson, M. A., Oost, T. K., Sukonpan, C., Rich, D. H. & Stevens, R. C. (2002). *J. Am. Chem. Soc.* **124**, 10248.
Hao, B., Gong, W., Ferguson, T. K., James, C. M., Krzycki, J. A. & Chan, M. K. (2002). *Science*, **296**, 1462–1466.
Harris, M. & Jones, T. A. (2001). *Acta Cryst.* D**57**, 1201–1202.
Hendrickson, W. A. (1985). *Methods Enzymol.* **115**, 252–270.
Irwin, J. J. & Shoichet, B. K. (2005). *J. Chem. Inf. Model.* **45**, 177–182.
Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kleywegt, G. J. (2000). *Acta Cryst.* D**56**, 249–265.
Kleywegt, G. J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K. & Jones, T. A. (1994). *Structure*, **2**, 1241–1258.
Kleywegt, G. J., Harris, M. R., Zou, J.-Y., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
Kleywegt, G. J., Henrick, K., Dodson, E. J. & van Aalten, D. M. (2003). *Structure*, **11**, 1051–1059.
Kleywegt, G. J. & Jones, T. A. (1998). *Acta Cryst.* D**54**, 1119–1131.
Lancaster, C. R. & Michel, H. (1997). *Structure*, **5**, 1339–1359.
Lütteke, T. & von der Lieth, C. W. (2004). *BMC Bioinformatics*, **5**, 69.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Nilsson, K., Lecerof, D., Sigfridsson, E. & Ryde, U. (2003). *Acta Cryst.* D**59**, 274–289.
Nissink, J. W., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C. & Taylor, R. (2002). *Proteins*, **49**, 457–471.
Pähler, A. & Hendrickson, W. A. (1990). *J. Appl. Cryst.* **23**, 218–221.
Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *Acta Cryst.* D**52**, 57–64.
Peat, T. S., Christopher, J. & Schmidt, K. (2005). *Acta Cryst.* A**61**, C165.
Priestle, J. P. (2003). *J. Appl. Cryst.* **36**, 34–42.
Schüttelkopf, A. W. & van Aalten, D. M. F. (2004). *Acta Cryst.* D**60**, 1355–1363.
Terwilliger, T. C., Adams, P. D., Moriarty, N. W. & Cohn, J. D. (2006). *Acta Cryst.* D**63**, 101–107.
Tronrud, D. E. (2004). *Acta Cryst.* D**60**, 2156–2168.
Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F. & Murshudov, G. N. (2004). *Acta Cryst.* D**60**, 2184–2195.
Weininger, D. (1988). *J. Chem. Inf. Comput. Sci.* **28**, 31–36.
Weininger, D., Weininger, A. & Weininger, J. L. (1989). *J. Chem. Inf. Comput. Sci.* **29**, 97–101.