Structure, Vol. 11, 1051–1059, September, 2003, ©2003 Elsevier Science Ltd. All rights reserved. DOI 10.1016/S0969-2126(03)00186-2

Pound-Wise but Penny-Foolish: How Well Do Micromolecules Fare in Macromolecular Refinement?

Gerard J. Kleywegt,[1,*] Kim Henrick,[2] Eleanor J. Dodson,[3] and Daan M.F. van Aalten[4]
[1]Department of Cell and Molecular Biology
Uppsala University
Biomedical Centre
Box 596
SE-751 24 Uppsala
Sweden
[2]EMBL Outstation
The European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD
United Kingdom
[3]Structural Biology Laboratory
Department of Chemistry
University of York
Heslington, York Y010 5YD
United Kingdom
[4]Division of Biological Chemistry and Molecular Microbiology
School of Life Sciences
University of Dundee
Dundee DD1 5EH
Scotland

Summary

For the refinement of protein and nucleic acid structures, high-quality geometric restraint libraries are available. Unfortunately, for other compounds, such as physiological ligands, lead compounds, substrate analogs, etc., the situation is not as favorable. As a result, the structures of small molecules found in complexes with biomacromolecules are often less reliable than those of the surrounding amino or nucleic acids. Here, we briefly review the use of geometric restraints in structure refinement (be it against X-ray crystallographic or NMR-derived data) and simulation. In addition, we discuss methods to generate both restraint libraries and (idealized) coordinates for small molecules and provide some practical advice.

Introduction

In the past decade, structural studies of biomacromolecules using crystallographic and NMR-spectroscopic techniques have paved the way for an increased understanding of biological processes at the atomic level. As the technology improved and interest in structural studies increased, the number of publicly available biomacromolecular structures has grown rapidly. Data provided by the Protein Data Bank (PDB) (Bernstein et al., 1977; Berman et al., 2000) indicate that the growth is roughly exponential (http://rcsb.rutgers.edu/pdb/holdings.html). In 1992, only ~1000 PDB entries were available, but this number has grown to over 20,000 today. However, fewer than half of these entries are unique at the level of the sequence. The remaining entries are mostly complexes of proteins (or nucleic acids) with other biomacromolecules or with small molecules. The latter category of compounds (also known as "hetero compounds") includes cofactors, physiological and nonphysiological ligands, inhibitors, potential drugs, substrate analogs, metal clusters, ions, molecules that are part of the crystallization or cryo-cooling solution, etc. Structural genomics initiatives the world over are expected to contribute thousands of new protein structures over the next five to ten years. It is, however, quite likely that the number of structures that will be determined of complexes between small molecules and proteins whose native structure is already known will be many times higher. In the pharmaceutical industry, where structural biology has become a key component of the drug-design process (Davis et al., 2003), often large numbers of such complexes between target proteins and lead compounds are studied. Such studies are necessary in order to increase our understanding of the atomic basis of the interactions between biomacromolecules and small molecules, to assess how a ligand (or a protein) should be modified in order to strengthen or weaken binding or to modulate specificity, etc.

Restrained Structure Refinement

Experimental structural biologists need to refine their models in order to render these sufficiently accurate for them to confidently draw detailed conclusions regarding chemistry, structure, and function. Unfortunately, there is rarely sufficient experimental data available (both quantitatively and qualitatively) to warrant refinement of structures without any restraints. The restraints embody empirical knowledge, for instance, concerning geometry and stereochemistry, and they are either derived by direct application of chemical knowledge or based on statistical analysis of structural databases. The building blocks of proteins and nucleic acids are small in number and well studied, and, therefore, a complete and accurate description of them can be obtained from database analysis. However, when it comes to small molecules, the situation is rather drastically different. Here, we will briefly review the refinement of complexes between biomacromolecules and small molecules using X-ray crystallographic data, although similar considerations apply to structure refinement using NMR spectroscopic data and to simulation studies (e.g., using molecular-dynamics approaches).

In 1973, the structure of the protein rubredoxin was the first to be refined against crystallographic data (Watenpaugh et al., 1973). This refinement used techniques developed in small-molecule crystallography and did not include knowledge of the geometry of amino acids. As a result, the refined model fit the X-ray data well, but (measured by today's standards) it had rather poor geometry. Hendrickson and Konnert were the first to use a priori geometric information as restraints in the

*Correspondence: gerard@xray.bmc.uu.se

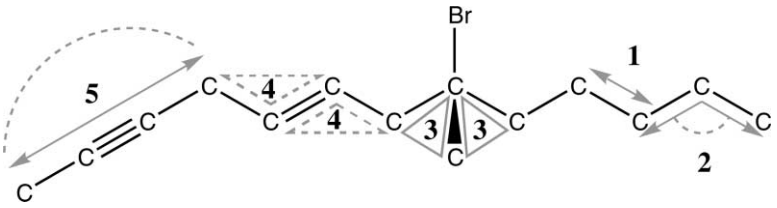

Figure 1. Overview of Different Types of Geometric Restraints

A hypothetical molecule, hypotheticin, is shown, labeled 1–5 to indicate the various restraint types. (1) bond length, distance between two atoms, (2) bond angle, angle between to bond vectors, (3) and (4) improper torsion (or dihedral) angle, angle between the normal vectors of two planes defined by two groups of three atoms each, (5) linearity restraint: two atoms connected by a triple bond together with their neighbors must form a linear group.

refinement of protein structures (Hendrickson and Konnert, 1980), although the ideas go back to Waser (1963). Restrained and constrained refinement were subsequently incorporated in all major macromolecular crystallographic refinement programs. During refinement, the molecular model is adjusted (by altering the coordinates, etc., of the atoms) in order to improve its ability to predict (or "fit") the experimental data. In the simplest (least-squares) approach, this is done by minimizing a function ϕ (the so-called target function) that consists of the (weighted) sum over all observed reflections of the squared discrepancies between observed and calculated (from the model) structure-factor amplitudes ($|F_{obs}|$ and $|F_{calc}|$, respectively, with w_r being the weight of a reflection):

$$\phi = \phi_{X\text{-}ray} = \sum_{reflections} w_r \, (|F_{obs}| - |F_{calc}|)^2$$

Geometric and other restraints can be imposed on the model by adding appropriate terms to the target function (Figure 1). For instance, in order to obtain an atomic model that displays chemically reasonable bond lengths, a term consisting of the (weighted) sum of the squared discrepancies between the bond lengths in a model (d_{model}) and their corresponding "ideal" values and weights (d_{ideal} and w_d, respectively, defined in some external library) could be added to the target function:

$$\phi = W_{X\text{-}ray} \, \phi_{X\text{-}ray} + W_{bonds} \, \phi_{bonds} = W_{X\text{-}ray} \, \phi_{X\text{-}ray} + W_{bonds} \sum_{bonds} w_d \, (d_{model} - d_{ideal})^2$$

Since structure factor amplitudes and bond lengths have different units, a conversion and weighting factor W_{bonds} needs to be applied. The weighting factor also determines how strong the restraints will be. If the restraint function outweighs the structure-factor-based function, the result will be a model that displays reasonable bond lengths but may not predict the experimental data very well. If, on the other hand, the experimental data is given more weight, one may well end up with a model that appears to predict the experimental data well but has chemically implausible geometry.

Apart from bond lengths, several other geometric features are usually restrained (Hendrickson, 1985). Bond angles are most commonly restrained by minimizing the sum of squared differences between model and ideal angles (in degrees or radians). However, some programs treat angles indirectly by restraining 1-3-distances (one should keep in mind that the latter measure also depends on the two bond lengths 1-2 and 2-3). In both cases, a properly weighted term consisting of the sum of the squared deviations from ideal values can be added to the target function for refinement.

A tetrahedral atom with four different substituents is a chiral center (i.e., not superimposable with its own mirror image), the stereochemistry of which needs to be defined and restrained during refinement. Again, there are two ways in which this can be done. Some programs restrain so-called improper torsion angles (they are called improper, or pseudo, or virtual, because they do not restrain torsions around a chemical bond). For instance, to impose proper chirality on the Cα carbon of an L-alanine residue, one could restrain the improper torsion angle Cα-N-C-Cβ to lie near +35°. An alternative method relies on the calculation of the chiral volume. For the same Cα carbon discussed above, the chiral volume is defined using the position vectors ($\underline{r}_X$) of the atom and three of its neighbors as:

$$V = (\underline{r}_N - \underline{r}_{C\alpha}) \cdot [(\underline{r}_C = \underline{r}_{C\alpha}) \times (\underline{r}_{C\beta} - \underline{r}_{C\alpha})]$$

For an L-alanine residue, this value should be ~2.5 Å^3. Just like 1-3-distances, chiral volumes also implicitly depend on the values of the 1-2-distances that are involved in their definition.

Some moieties, such as carboxylate groups and aromatic rings, are planar, and this property can also be restrained during refinement. Again, there are two ways of doing this. The first entails calculation of a best-fitting plane through the group of atoms and minimizing the sum of squared distances of the atoms to that plane. The other method relies on the definition of a set of torsion angles that have values of either 0° or 180°. In the case of a benzene ring, for instance, all six ring torsions should be 0° to ensure flatness. In the case of a carboxylate group, an improper torsion can be defined involving the central, *sp*^{2}-hybridized carbon and its three neighboring atoms. Again, a value of 0° indicates that the group is perfectly flat. Note that *cis* and *trans* configurations around a double bond can only be enforced with torsion angle restraints.

Conformational torsion angles (e.g., around single carbon-carbon bonds) are not usually restrained, although it is known from studies on both small molecules and biomacromolecules that certain values are more favorable than others (e.g., staggered versus eclipsed). Indeed, empirically observed distributions of (combinations of) torsion angles can be used as restraints during refinement (Kuszewski et al., 1996, 1997; Bertini et al., 2003). On the other hand, if they are left unrestrained,

such torsion angles provide an excellent means for a posteriori structure validation instead (Kleywegt and Jones, 1998; Kleywegt, 2000).

In addition to the restraint types discussed above, refinement programs may apply additional restraints, e.g. van der Waals or antibumping restraints, electrostatic repulsion and attraction, hydrogen-bonding restraints, harmonic restraints (to keep a moiety close to its starting position or conformation), as well as restraints particular to the structure-determination technique (e.g., on temperature factors of bonded atoms, or on the geometry of noncrystallographically related molecules). However, these additional restraints are usually not specific for small molecules and therefore tend not to require explicit definition.

In order to describe interactions between atoms in terms of lists of force constants for bond, angle, and dihedral restraints (a force field), it has been useful to divide atoms in certain classes (atom types) depending on atomic properties. For example, for carbon atoms, a very simple division would be to assume that there are three types: sp^3 (tetrahedral), sp^2 (flat with bond angles of flat with bond angles of 120°), and sp^1 (bond angle of $\sim$180°). This would be an oversimplification, however, as the precise bond angles depend on the chemical environment of the atom, i.e., whether it is bonded to other carbon, nitrogen, or oxygen atoms. In practice, most force fields are based on approximately 40 atom types (van Gunsteren and Berendsen, 1987; Murshudov et al., 1997; Brünger et al., 1998). Many of the force field parameters in use today have been based on a study by Engh and Huber (Engh and Huber, 1991) who used a limited set of atom types (31). More modern approaches, however, define different atom types depending on the chemical environment up to two or three bonds removed (Boutselakis et al., 2003). This will undoubtedly result in a more accurate description of small molecule geometry, once the information has been translated into force fields for macromolecular refinement programs.

Parameters and Programs

By the 1980s, all macromolecular crystallographic refinement programs used a priori geometric information as restraints. However, every program used its own set of "ideal" values (for bond lengths, etc.), and these different libraries left their mark on the resulting models. This was demonstrated convincingly by Laskowski et al. (1993), who showed that, by inspecting the covalent geometry of a protein model, it was possible to infer (with 95% accuracy) with which program the model had been refined. In 1991, Engh and Huber (Engh and Huber, 1991) performed an analysis on fragments of compounds found in the Cambridge Structural Database (CSD [Allen et al., 1979], vide infra) that resembled parts of amino acid residues. This enabled them to formulate a set of improved target values for the bond lengths and angles encountered in proteins, based on experimental observations. They also calculated weights for these restraints based on the statistical distribution of the values encountered in the CSD, and in most cases these weights made the restraints considerably tighter than they had been previously. Nowadays, their set of target values and weights is used by most crystallographic refinement and modeling programs. In 1996, a similar analysis was carried out for nucleic acids, based on structures from both the CSD and the Nucleic Acid Database (NDB [Berman et al., 2002]) solved at a resolution of 1.0 Å or better (Parkinson et al., 1996). More recently, Engh and Huber have repeated their analysis and published an updated version of their set of atom types and target parameters (Engh and Huber, 2001). The Engh and Huber analysis only considered bond lengths and angles, but Priestle has shown that the target values and weights of several dihedral angle restraints remain incorrect or inappropriate (Priestle, 2003).

Refinement programs come in many flavors. They generally differ in two major aspects, namely the formulation of the target function and the method(s) by which the target function is minimized (the latter aspect is less relevant to our discussion). In energy-based methods, such as X-PLOR (Brünger et al., 1987) and CNS (Brünger et al., 1998), the target function is expressed as an energy, and the weights of the restraints (usually called force constants) are estimates of the energy penalty associated with deviations from the target values. Such force constants are often derived from experimental observations (e.g., infrared spectral data). The complete set of parameters and force constants is called a force field. In nonenergy-based methods, such as TNT (Tronrud, 1997), PROLSQ (Hendrickson and Konnert, 1980; Hendrickson, 1985), REFMAC (Murshudov et al., 1997), and SHELX (Sheldrick and Schneider, 1997), the weight of a restraint is related to the standard deviation of the distribution of the restrained parameter. In the parameter sets of Engh and Huber (1991) and Parkinson et al. (1996), the force constants are derived from the distributions observed in high-resolution crystal structures. Because of this, energy-based and nonenergy-based methods have in practice been unified.

As described earlier, a simple X-ray target function is the weighted sum over all observed reflections of the squared discrepancies between observed and calculated structure factor amplitudes. The weight of each reflection is estimated rather crudely or in many cases omitted altogether, giving every reflection the same unitary weighting factor, regardless of its reliability. However, use of a maximum likelihood-based target function tends to result in smoother refinement. The available programs minimize suitably weighted differences between the observed and calculated structure factors taking into account both the differences in amplitude and the likely differences in phase. Unfortunately, estimating better weights for each observation is a major problem, only partially addressed by these methods. An expected error for the current model is estimated from the overall fit of the calculated and observed amplitudes for a set of observations that is not used for actual refinement (the free-R or test set). A figure-of-merit for each reflection is then derived taking into account this expected error, the magnitude of the difference, and the estimated measurement error. The details are not our prime concern in this paper, but are given in the original literature (Hendrickson, 1985; Murshudov et al., 1997; Brünger et al., 1998). If the elements of both the crystallo-

graphic target and the restraints are weighted properly, the model adapts smoothly to fit the available information. Initially, when the agreement between the X-ray observations and the calculated structure factors is poor, the X-ray target is weighted down, especially for the highest resolution observations, while the restraints are satisfied. However, as the model improves the X-ray target becomes more important and steers the model to fit the available data better.

The Problem with Small Molecules

Unfortunately, the various efforts to provide better restraint libraries for proteins and nucleic acids have not had a counterpart for small molecules. The reason is most likely the boundless diversity of small molecules compared to proteins and nucleic acids that are conveniently made up of a small set of building blocks (amino acids and nucleotides). This means that structural biologists who need to deal with small molecules in their structure determinations or simulations are out on their own. First, in order to include a molecule into their model, they need access to a reliable set of coordinates. Then, in order to include it in the refinement, they need a description of its ideal geometry (such descriptions are variably known as libraries, dictionaries, force fields, restraint sets, or topology and parameter files). In practice, many structural biologists have difficulties with both processes, but in particular with the latter. If a library contains errors, these will propagate into the model as well, since refinement programs are both ignorant of chemistry and very good at minimizing their target function.

The process of defining a library description of a molecule consists of two steps: definition of the restraints, and definition of target values (and weights) for the restraints. In principle, the rules for defining these are fairly simple (Figure 1), for instance:

(1) Every pair of chemically bonded atoms generates a bond-length restraint. Target values can be obtained by measuring them in a reliable set (or multiple sets) of coordinates of the compound or from tabulated listings (e.g., single carbon-carbon bonds should be $\sim$1.53 Å). Weights would be chosen to be of the same order of magnitude as those of the bond-length restraints for proteins and nucleic acids.

(2) Two pairs of bonded atoms that have one atom in common generate a bond-angle restraint. Target values and weights can be obtained in the same way as those for bond lengths.

(3) A tetrahedral carbon with four different substituents generates a chirality restraint. Depending on the definition and the hand, the target value will be close to $+35^\circ$ or -35° if the restraint is defined through an improper torsion.

(4) A (partial) double bond generates a planarity restraint involving the two atoms that are involved in the double bond, plus all their other nearest neighbors (note that this rule also covers carboxylate groups and aromatic rings). When expressed in terms of (possibly improper) torsion angles, one restraint is required for every atom that participates in a (partial) double bond, and the target values of these restraints will be 0° or 180°.

(5) A triple bond generates a linearity restraint for the two atoms involved in the bond and their other nearest neighbors. Such bonds are relatively rare, but they do occur and should then be properly restrained. Similar considerations apply to atoms with two double bonds (such as in aza groups).

Unfortunately, in practice such libraries are difficult to construct manually as they require a thorough understanding of the chemistry of the molecules involved. As a consequence, the quality of the geometry of more than a few small molecules in the PDB is rather poor (van Aalten et al., 1996; Kleywegt and Jones, 1998). With respect to the definition of restraints, the most common error is the omission of restraints that are chemically necessary (in particular those due to rules [3] and [4] above). For instance, it is often forgotten that not only the ring atoms of aromatic rings should lie in one plane, but that the nearest bonded nonring neighbor of all ring atoms should lie in that same plane as well. Boström (2001) presented an example of this, in which an aromatic carbon atom is somewhere in between planar and pyramidal (with an improper torsion angle of 17°). This suggests that there was no restraint (or a very weak one) to force the nonring neighbor atom of that carbon atom to lie in the plane of the phenyl ring. Another cause of problems is the introduction of inappropriate restraints. An example of this cited by Boström (2001) involves the sulfur atom of a methanesulfinyl moiety that has been restrained to be planar, whereas in fact it ought to be pyramidal. Nissink et al. (2002) have also identified a number of cases in which ligand coordinates were deemed to be unreliable due to dubious ligand geometry (or other problems such as significant clashes between ligand and protein atoms, a poor fit to the electron density, etc.). Several other, related pitfalls have recently been discussed by Davis et al. (2003).

However, the most common cause of problems appears to be the use of inappropriate target values for restraints (in particular for bond lengths and angles), although it should be noted that "unusual" bond lengths could also be due to the omission of a restraint or underweighting of a restraint. To illustrate these problems, we have determined ranges of observed values for bond lengths, etc., of the small molecule ATP (adenosine triphosphate; Figure 2). ATP is similar to adenosine, a compound that is present in most standard libraries, and, therefore, a smart structural biologist could have "stolen" (adapted) the restraints of adenosine and used them in the refinement of ATP. We looked at 39 ATP molecules found in crystal structures determined at better than 2.0 Å resolution, and 100 ATP molecules found in structures with resolution worse than 2.0 Å. The atom-naming convention of ATP is shown in Figure 2, and a selection of the results is shown in Table 1. This reveals that the ranges of values observed for bond lengths and angles are considerably larger than would be expected on the basis of the standard deviations of their distributions observed in atomic resolution structures (typically, 0.01–0.02 Å for bond lengths and 0.5–2.0° for angles [Engh and Huber, 1991; Parkinson et al., 1996]). The examples of the (improper) torsion angles reveal that chemically necessary restraints are often omitted, far too weak, or have incorrect target values. Somewhat

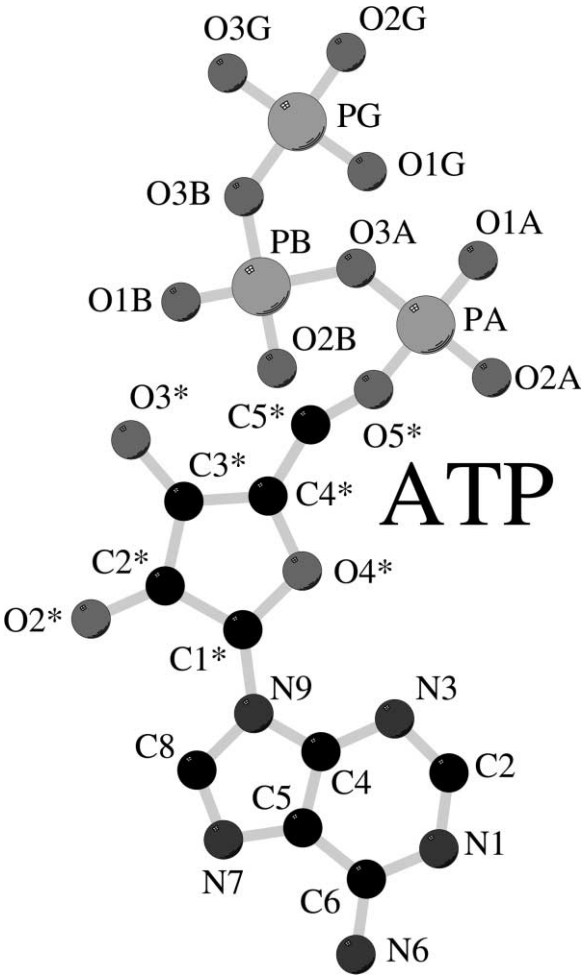

Figure 2. ATP

Atom-naming convention for ATP (adenosine triphosphate), referred to in Table 1. Figure created with LIGPLOT (Wallace et al., 1995).

surprisingly, the quality of the high-resolution structures does not appear to be much better than that of the low-resolution ones. Although in most cases the observed ranges of values are somewhat smaller for high-resolution structures, the improper torsions for C2* and C3* reveal that there are high-resolution structures (but not low-resolution ones) where these atoms even have incorrect stereochemistry (something which has also been observed for amino acids [Hooft et al., 1996; Mitchell and Smith, 2003] and nucleotides [Schultze and Feigon, 1997]). Table 1 reveals that some bonded atoms are very far apart, but the opposite phenomenon also occurs, i.e., nonbonded atoms are in too close proximity. Using a cutoff of 2.0 Å in the high-resolution structures, this only occurs for the O1A-O5* distance that has a lowest observed value of 1.91 Å. For the low-resolution structures, on the other hand, no fewer than seven pairs of nonbonded atoms sometimes occur at such short distances (O1A-O5* 1.90 Å; C2*-O3* 1.77 Å; C3*-O2* 1.73 Å; O2*-O3* 1.54 Å; O3B-PG 1.53 Å; O1B-PG 1.47 Å; O2G-O3G 1.44 Å). In these cases, restraints that prevent nonbonded atoms from bumping into one another (e.g., van der Waals repulsion terms) have apparently not been used or they were not weighted sufficiently heavily, or incorrect torsional restraints have been imposed.

Obtaining Libraries

The examples in the previous section demonstrate that there are several problems associated with the handling of small molecules in the refinement of their complexes with biomacromolecules. The only way to improve this situation is to use high-quality libraries for refinement (even if the starting geometry is poor, the restraints can rectify that during refinement). Libraries can be obtained in a number of ways, and for each method resources are available (many through the Internet and many free of charge). In this section, we discuss a number of such resources and describe how they can be of help in the process of obtaining high-quality restraint libraries for refinement programs. URLs for these resources are listed in Table 2.

Manual Labor

Restraint libraries can be generated manually. This entails analyzing the chemistry of the compound at hand carefully to define the restraints (e.g., following the rules given in the previous section). The second task is to define proper target values for the restraints. Tabulations of "ideal" bond lengths and angles can be found in books (e.g., the Handbook of Chemistry and Physics [Lide, 1995]), in papers (e.g., Allen et al., 1987), and sometimes on the Internet (e.g., the tabulation of metal-coordination distances provided by Marjorie Harding at http://tanna.bch.ed.ac.uk/targets.html [Harding, 2001]).

Colleagues

Many small molecules occur in many different structures. Hence, a lot of duplication of work can be avoided if libraries are shared. Indeed, some workers have gone to great lengths to produce and distribute libraries. A prime example is the work of Roy Lancaster who did a similar analysis as Engh and Huber for the cofactors encountered in his photosynthetic reaction center (Lancaster and Michel, 1997) and made his libraries available to colleagues (http://www.mpibp-frankfurt.mpg.de/lancaster/publications/structure1997/supmat.html). However, in general, one should be very skeptical when using other people's libraries and verify that no restraints have been omitted, that no superfluous, erroneous, or contradictory restraints have been included, and that the target values and weights are reasonable.

A La Mode

A La Mode is an environment for building models of ligand and monomer molecular components (Clowney et al., 1999). For every compound, it contains information about chiral centers, bond lengths, bond angles, and torsion angles, based on an analysis of structures from the CSD. Unfortunately, at present it only contains information for nucleic acid-like compounds.

CHEMPDB

The CHEMPDB service (Boutselakis et al., 2003) provides web access to a consistent and enriched library of all the small molecules and monomers that are referred to in any macromolecular structure of a PDB entry, such as bound molecules and standard and modified amino acids. The ligand dictionary explicitly maintains

Table 1. Ranges of Geometric Parameters Encountered for ATP Molecules in Protein Crystal Structures

Bond	Ideal Value (Å)	Range Low-Resolution (Å)	Range High-Resolution (Å)
O5*-C5*	1.440	1.23 ... 1.57	1.25 ... 1.49
O4*-C4*	1.453	1.41 ... 1.61	1.42 ... 1.48
C3*-C4*	1.524	1.38 ... 1.63	1.45 ... 1.58
C2*-O2*	1.413	1.39 ... 1.53	1.40 ... 1.48
C5*-C4*	1.510	1.35 ... 1.67	1.42 ... 1.62
C2*-C1*	1.528	1.42 ... 1.60	1.50 ... 1.64
N3-C4	1.344	1.24 ... 1.49	1.29 ... 1.43
Angle	**Ideal Value (°)**	**Range Low-Resolution (°)**	**Range High-Resolution (°)**
PG-O3B-PB	—	64 ... 146	122 ... 141
PB-O3A-PA	—	108 ... 147	114 ... 152
C3*-C2*-O2*	113.3	68 ... 126	101 ... 121
C2-N3-C4	110.6	102 ... 124	108 ... 120
O3*-C3*-C2*	110.0	82 ... 129	95 ... 117
Torsion	**Ideal Value (°)**	**Range Low-Resolution (°)**	**Range High-Resolution (°)**
C1*-N9-C4-N3	0	−32 ... +29	−11 ... +21
N1-C2-N3-C4	0	−28 ... +19	−11 ... +11
C2-N3-C4-C5	0	−12 ... +53	−10 ... +9
Improper Torsion	**Ideal Value (°)**	**Range Low-Resolution (°)**	**Range High-Resolution (°)**
C2*-C3*-O2*-C1*	−35	−43 ... −36	−55 ... +29
C3*-C4*-O3*-C2*	−35	−51 ... −30	−46 ... +66
N9-C1*-C8-C4	0	−12 ... +8	−13 ... +33

Refer to Figure 2 for atom-naming convention. Ideal values of bond lengths and angles are taken from the library for adenosine as defined by Parkinson et al. (1996). Ideal values of (improper) torsions are dictated by the chemistry of the molecule.

the properties that define the chemical identity of the molecules, such as the connectivity and bond orders, as well as stereocenter descriptors for atoms and bonds, that allow the exact identification of the stereoisomer that the molecule corresponds to. The definition of a small molecule is a distinct stereoisomer but not a conformation isomer. All stereoisomers are treated as different ligands with different three-letter codes. The dictionary also contains classifications of the atoms of the small molecules into energy types, and associates them with reference dictionaries. The set of coordinates that are used by default are idealized coordinates generated by CORINA (Gasteiger et al., 1990).

SMILES2DICT and SKETCHER

Many organic chemists describe molecules using SMILES strings (Weininger, 1988; Weininger et al., 1989). The SMILES syntax is extremely simple, defining connectivity and atom types by a string of letters, and a few symbols that define the bonding pattern of the molecule completely. Upper case letters denote aliphatic atoms and lower case aromatics, and ring closures are denoted by a numerical suffix, e.g., c1ccccc1 is benzene. It is also possible to indicate branching, double bonds, and chirality. This information is sufficient to describe large molecules. Software such as SMILE2DICT (Greaves et al., 1999) can interpret this formalism to create coordinate sets and lists of restraints. After energy minimization and chemical checking, library descriptions in suitable formats for refinement programs are generated.

A graphical program called SKETCHER is available within the CCP4 suite (Collaborative Computational Project, 1994). It allows the user to draw a molecule from scratch or to display and modify an existing molecule, described either as a list of restraints or as a set of

Table 2. Internet Resources for Small Molecule Coordinates and Libraries and Related Software

Resource	URL
A La Mode	http://ndbserver.rutgers.edu/alamode
CHEMPDB	http://www.ebi.ac.uk/msd-srv/chempdb
ConQuest	http://www.ccdc.cam.ac.uk/prods/quest.html
CORINA	http://www2.chemie.uni-erlangen.de/software/corina/free_struct.html
CSD	http://www.ccdc.cam.ac.uk/
HIC-Up	http://xray.bmc.uu.se/hicup
HIC-Up web server	http://xray.bmc.uu.se/cgi-bin/gerard/hicup_server.pl
NCIscreen	http://131.188.127.153/services/nciscreen
NIST Webbook	http://webbook.nist.gov/chemistry
PDB	http://www.pdb.org/
PRODRG	http://davapc1.bioch.dundee.ac.uk/prodrg
SPARTAN	http://wavefun.com/
SWEET	http://www.dkfz.de/spec
XPLO2D	http://xray.bmc.uu.se/usf/xplo2d_man.html

coordinates. The restraints are then generated and refined using the programs MAKECIF and LIBCHECK (Vagin et al., 1998) and displayed for the user to check.

PRODRG

PRODRG (van Aalten et al., 1996; A. Schuettelkopf and D.M.F. van Aalten, submitted) was conceived as a tool to generate topologies for the popular molecular dynamics program GROMOS87 (van Gunsteren and Berendsen, 1987) and uses the GROMOS87 force field libraries to model the bonded forces in small molecules. This force field has 37 atom types—bond, angle, and (improper) dihedral force constants are based on different types of data ranging from empirical to experimental. PRODRG determines a connectivity table, including *sp*-hybridization and chirality, using either PDB coordinates, or a schematic drawing in ASCII format as input. Using the connectivity table, GROMOS87 atom types are determined, which are then used to determine bond, angle, and dihedral types and associated constants. Although this is a rather crude approach, a test set of 40,000 molecules from the CSD minimized with the PRODRG topologies in GROMOS87, yielded RMSDs of approximately 0.04 Å on bond lengths, 2.5° on angles and 2° on improper torsions compared to the starting crystal structures (A. Schuettelkopf and D.M.F. van Aalten, submitted). Over the past years, PRODRG has been expanded to translate the GROMOS87 topology into topologies for CNS (Brünger et al., 1998), SHELX (Sheldrick and Schneider, 1997), O (Jones et al., 1991), REFMAC (Murshudov et al., 1997), and GROMACS (Lindahl et al., 2001).

HIC-Up

The Hetero-compound Information Centre in Uppsala (HIC-Up) (Kleywegt and Jones, 1998) is a database that is derived from the PDB. It has evolved from a repository of topology and parameter files (G.J.K., unpublished data) for the refinement program X-PLOR (Brünger et al., 1987). Nowadays, its purpose is to provide various kinds of information, images, and links to other resources for all small molecules that are found in the PDB. As part of that effort, restraint libraries are provided for the refinement programs CNS (Brünger et al., 1998) and TNT (Tronrud, 1997), as well as for the crystallographic modeling program O (Jones et al., 1991). These libraries are derived from the copy of each small molecule that occurs in the highest-resolution crystal structure (or in any NMR structure, if no crystal structure is available). Since the libraries are derived from other macromolecular crystal structures, and the result of refinement using some other library, their quality can be poor. They are intended as a last resort, and therefore a link is provided to both the PRODRG server and CHEMPDB for every compound in the database.

XPLO2D

XPLO2D (Kleywegt and Jones, 1998) is the program that generates most of the libraries in the HIC-Up database. However, for compounds that are not in the database, the program can be run through the HIC-Up web server to produce appropriate libraries. These are based on users providing one or more sets of coordinates of a compound. If more than one coordinate set is provided, all bond lengths, etc., will be averaged. On the server, the program is run with a set of default parameters and options. Users who wish to run the program with different parameters, or who do not wish to submit coordinates over the Internet, can run the program in-house instead.

Obtaining Coordinates

When first including a small molecule into a model, one needs a reliable set of coordinates. As discussed above, such coordinates can also be used to generate a library for the compound. Most of the library resources of the previous section can also supply or generate coordinates. However, there are other resources available, which will be reviewed briefly here. URLs for these resources are also listed in Table 2.

Experimental Coordinates

Experimental coordinates can be retrieved from a number of sources. Usually, structures determined by small-molecule crystallography will be the most accurate and reliable. These structures can be retrieved from the CSD (Allen et al., 1979), for instance, with the search and retrieval program ConQuest (Bruno et al., 2002). This sophisticated program provides a full range of textual and numeric database search options, in addition to more complex search functionality, including chemical substructure searching.

Coordinates of molecules that have been determined in complex with macromolecules previously can of course also be retrieved from the PDB (Bernstein et al., 1977; Berman et al., 2000), HIC-Up (Kleywegt and Jones, 1998), or CHEMPDB (Boutselakis et al., 2003). However, one should keep in mind that these coordinates are the result of refinement against comparatively low-resolution data where the small molecule constituted only a minute fraction of the total scattering matter. This makes these coordinates inherently much less accurate than those obtained from the CSD. In addition, the coordinates may contain errors due to the use of incorrect restraints. Hence, such coordinate sets should only be used as a last resort, and only after verification that they are reliable. The latter can be facilitated by inspection of the electron density for the compound in question, for instance at the Uppsala Electron-Density Server (http://fsrv1.bmc.uu.se/eds) (G.J.K. et al., submitted).

Ab Initio and Semiempirical Calculations

Several powerful and low-cost packages exist for carrying out ab initio and semiempirical calculations, which provide an alternative means of generating coordinate sets. It is beyond the scope of this work to review these packages, but a number of them have been evaluated and compared recently by Boström (2001). A related program, developed specifically to generate force fields for use in refinement, is Hess2FF (Nilsson et al., 2003). This method is based on a Hessian (force constant) matrix estimation. The method can handle metal complexes and can automatically assign a separate energy type to each atom.

Conversion Software

CORINA (Gasteiger et al., 1990) is a rule and data-based system that automatically generates 3D atomic coordinates from the constitution of a molecule as expressed in a connection table or linear code (such as a SMILES

string [Weininger, 1988; Weininger et al., 1989]). Stereochemistry for double-bond configurations (*cis/trans*) and tetrahedral chirality are supported, as are allene-like bonds. The program has been compared to a variety of other conversion programs (Ricketts et al., 1993; Sadowski et al., 1994) and was found to give the highest conversion rate and reproduce the largest fraction of X-ray structures.

There are also a number of chemical-name-to-structure conversion programs, with the most widely known being AutoNom (Beilstein Informationssysteme GmbH), Chem4D Draw's Nomenclator (ChemInnovation Software), Bio-Rad's IUPAC NameIt and ACD/Name to Structure (Advanced Chemistry Development, Toronto, Canada). These products generate (energy-minimized) 3D structures from systematic and trivial chemical names that include the stereoconfiguration of chiral centers and double bonds. One should keep in mind that many structure generation programs have difficulties in producing accurate coordinates for metal-coordination complexes.

SWEET (Bohne et al., 1999) is a system for conversion of sequence information for complex carbohydrates directly into a preliminary but reliable 3D model. Conformational space for each glycosidic linkage is explored to give a favorable conformation. Coordinates can be generated in PDB format.

Finally, there are several publicly accessible websites that contain large collections of structures generated by various kinds of software (see Table 2), including NCIscreen (230,000 compounds generated with CORINA) and the NIST Chemistry Webbook (30,000 compounds generated by the molecular mechanics program Alchemy 2000 [Tripos, St. Louis, MO]). The CHEMPDB site provides CORINA-generated coordinate sets for all small molecules that occur in the PDB.

Other Issues

Besides the definition of proper target values for geometric restraints, their weights (sigmas or force constants) also need some consideration. It is of paramount importance that the weights of all refined entities (proteins, nucleic acids, and small molecules) are on the same scale. If this is not the case, e.g., if the weights for a small molecule are an order of magnitude smaller than those used for a protein, errors will build up in the geometry of the former.

Before a set of restraints is used in refinement of the complex with the biomacromolecule, it is worthwhile to refine the geometry of the compound in isolation and without any experimental data. The result of this refinement will show the conformation that the refinement program will attempt to attain. If there are gross errors in the library, these will become apparent at this stage rather than after a possible lengthy refinement of the entire complex. It is also worthwhile to inspect any restraints that are seriously violated in the refined structure, since these may indicate erroneous target values for one or more restraints, or the presence of erroneous or incongruous restraints. The latter can occur, for instance, if the target values for the angles inside a flat ring are off by even a few degrees. It may then well be impossible to satisfy both the angle restraints and the planarity restraint(s) on the ring simultaneously.

Libraries are used not only in refinement programs but also in model-building programs such as O (Jones et al., 1991), Quanta (Oldfield, 2001), and XtalView (McRee, 1999). It is a good idea to use similar libraries for refinement and model building, as an incorrect restraint in a model-building program's library may bring about erroneous changes to a model that are beyond the radius of convergence of the refinement program. Priestle (1994) has made Engh and Huber-style libraries for several other programs, including O. In addition, the PRODRG and HIC-Up web servers can produce libraries for small molecules for use with O.

Acknowledgments

E.J.D. thanks Drs. E. Potterton, A. Vagin, R. Greaves, and G. Murshudov for useful discussions. G.J.K. 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, Uppsala University, and the Swedish Structural Biology Network. K.H. gratefully acknowledges support from the Wellcome Trust (GR062025MA), EU (TEMBLOR, NMRQUAL, SPINE, AUTOSTRUCT, and IIMS), CCP4, BBSRC, MRC, and EMBL. E.J.D. is a research Professor funded by the Wellcome Trust. Financial support by a Wellcome Trust Career Development Research Fellowship to D.M.F.v.A. is gratefully acknowledged.

Received: June 23, 2003
Revised: July 25, 2003
Accepted: August 1, 2003
Published: September 2, 2003

References

van Aalten, D.M.F., Bywater, R., Findlay, J.B.C., Hendlich, M., Hooft, R.W.W., and Vriend, G. (1996). PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. J. Comput. Aided Mol. Des. *10*, 255–262.

Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., et al. (1979). The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. Acta Crystallogr. *B35*, 2331–2339.

Allen, F.H., Kennard, O., Watson, D.G., Brammer, L., Orpen, A.G., and Taylor, R. (1987). Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. J. Chem. Soc. Perkin Trans. II *1987*, S1–S19.

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

Berman, H.M., Westbrook, J., Feng, Z., Iype, L., Schneider, B., and Zardecki, C. (2002). The Nucleic Acid Database. Acta Crystallogr. *D58*, 889–898.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. *112*, 535–542.

Bertini, I., Cavallaro, G., Luchinat, C., and Poli, I. (2003). A use of Ramachandran potentials in protein solution structure determinations. J. Biomol. NMR *26*, 355–366.

Bohne, A., Lang, E., and von der Lieth, C.W. (1999). SWEET—WWW-based rapid 3D construction of oligo- and polysaccharides. Bioinformatics *15*, 767–768.

Boström, J. (2001). Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. J. Comput. Aided Mol. Des. *15*, 1137–1152.

Boutselakis, H., Dimitropoulos, D., Fillon, J., Golovin, A., Henrick,

K., Hussain, A., Ionides, J., John, M., Keller, P.A., Krissinel, E., et al. (2003). E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. Nucleic Acids Res. *31*, 458–462.

Brünger, A.T., Kuriyan, J., and Karplus, M. (1987). Crystallographic *R* factor refinement by molecular dynamics. Science *235*, 458–460.

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., et al. (1998). Crystallography and NMR System: a new software suite for macromolecular structure determination. Acta Crystallogr. *D54*, 905–921.

Bruno, I.J., Cole, J.C., Edgington, P.R., Kessler, M., Macrae, C.F., McCabe, P., Pearson, J., and Taylor, R. (2002). New software for searching the Cambridge Structural Database and visualizing crystal structures. Acta Crystallogr. *B58*, 389–397.

Clowney, L., Westbrook, J.D., and Berman, H.M. (1999). CIF applications. XI. A La Mode: a ligand and monomer object data environment. I. Automated construction of mmCIF monomer and ligand models. J. Appl. Crystallogr. *32*, 125–133.

Collaborative Computational Project (1994). The *CCP4* suite: programs for protein crystallography. Acta Crystallogr. *D50*, 760–763.

Davis, A.M., Teague, S.J., and Kleywegt, G.J. (2003). Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. Angew. Chem. Int. Ed. Engl. *42*, 2718–2736.

Engh, R.A., and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr. *A47*, 392–400.

Engh, R.A., and Huber, R. (2001). Structure quality and target parameters. In International Tables for Crystallography. Volume F: Crystallography of Biological Macromolecules, M.G. Rossman and E. Arnold, eds. (Dordrecht, The Netherlands: Kluwer), 382–392.

Gasteiger, J., Rudolph, C., and Sadowski, J. (1990). Automatic generation of 3D-atomic coordinates for organic molecules. Tetrahedron Comp. Method. *3*, 537–547.

Greaves, R.B., Vagin, A.A., and Dodson, E.J. (1999). Automated production of small-molecule dictionaries for use in crystallographic refinements. Acta Crystallogr. *D55*, 1335–1339.

van Gunsteren, W.F., and Berendsen, H.J.C. (1987). GROMOS Manual. BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry (Groningen, The Netherlands: University of Groningen Press).

Harding, M.M. (2001). Geometry of metal-ligand interactions in proteins. Acta Crystallogr. *D57*, 401–411.

Hendrickson, W.A., and Konnert, J.H. (1980). Incorporation of stereochemical information into crystallographic refinement. In Computing in Crystallography, R. Diamond, S. Ramaseshan, and K. Venkatesan, eds. (Bangalore, India: Indian Academy of Science), 13.01–13.25.

Hendrickson, W.A. (1985). Stereochemically restrained refinement of macromolecular structures. Methods Enzymol. *115*, 252–270.

Hooft, R.W.W., Vriend, G., Sander, C., and Abola, E.E. (1996). Errors in protein structures. Nature *381*, 272.

Jones, T.A., Zou, J.Y., Cowan, S.W., and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr. *A47*, 110–119.

Kleywegt, G.J. (2000). Validation of protein crystal structures. Acta Crystallogr. *D56*, 249–265.

Kleywegt, G.J., and Jones, T.A. (1998). Databases in protein crystallography. Acta Crystallogr. *D54*, 1119–1131.

Kuszewski, J., Gronenborn, A.M., and Clore, G.M. (1996). Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Sci. *5*, 1067–1080.

Kuszewski, J., Gronenborn, A.M., and Clore, G.M. (1997). Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. J. Magn. Reson. *125*, 171–177.

Lancaster, C.R.D., and Michel, H. (1997). The coupling of light-induced electron transfer and proton uptake as derived from crystal structures of reaction centres from *Rhodopseudomonas viridis* modified at the binding site of the secondary quinone, QB. Structure *5*, 1339–1359.

Laskowski, R.A., Moss, D.S., and Thornton, J.M. (1993). Main-chain bond lengths and bond angles in protein structures. J. Mol. Biol. *231*, 1049–1067.

Lide, D.R. (1995). CRC Handbook of Chemistry and Physics, 75th Edition (Boca Raton, FL: CRC Press).

Lindahl, E., Hess, B., and van der Spoel, D. (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. J. Mol. Model. *7*, 306–317.

McRee, D.E. (1999). XtalView/Xfit—a versatile program for manipulating atomic coordinates and electron density. J. Struct. Biol. *125*, 156–165.

Mitchell, J.B., and Smith, J. (2003). D-amino acid residues in peptides and proteins. Proteins Struct. Funct. Genet. *50*, 563–571.

Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. *D53*, 240–255.

Nilsson, K., Lecerof, D., Sigfridsson, E., and Ryde, U. (2003). An automatic method to generate force-field parameters for hetero-compounds. Acta Crystallogr. *D59*, 274–289.

Nissink, J.W., Murray, C., Hartshorn, M., Verdonk, M.L., Cole, J.C., and Taylor, R. (2002). A new test set for validating predictions of protein-ligand interaction. Proteins Struct. Funct. Genet. *49*, 457–471.

Oldfield, T.J. (2001). A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent. Acta Crystallogr. *D57*, 82–94.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A.T., and Berman, H.M. (1996). New parameters for the refinement of nucleic acid-containing structures. Acta Crystallogr. *D52*, 57–64.

Priestle, J.P. (1994). Stereochemical dictionaries for protein structure refinement and model building. Structure *2*, 911–913.

Priestle, J.P. (2003). Improved dihedral-angle restraints for protein structure refinement. J. Appl. Crystallogr. *36*, 34–42.

Ricketts, E.M., Bradshaw, J., Hann, M., Hayes, F., Tanna, N., and Ricketts, D.M. (1993). Comparison of conformations of small molecule structures from the Protein Data Bank with those generated by Concord, Cobra, ChemDBS-3D, and Converter and those extracted from the Cambridge Structural Database. J. Chem. Inf. Comput. Sci. *33*, 905–925.

Sadowski, J., Gasteiger, J., and Klebe, G. (1994). Comparison of automatic three-dimensional model builders using 639 X-ray structures. J. Chem. Inf. Comput. Sci. *34*, 1000–1008.

Schultze, P., and Feigon, J. (1997). Chirality errors in nucleic acid structures. Nature *387*, 668.

Sheldrick, G.M., and Schneider, T.R. (1997). SHELXL: high-resolution refinement. Methods Enzymol. *277*, 319–344.

Tronrud, D.E. (1997). The TNT refinement package. Methods Enzymol. *277*, 306–319.

Vagin, A.A., Murshudov, G.N., and Strokopytov, B.V. (1998). BLANC: the program suite for protein crystallography. J. Appl. Crystallogr. *31*, 98–102.

Wallace, A.C., Laskowski, R.A., and Thornton, J.M. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng. *8*, 127–134.

Waser, J. (1963). Least-squares refinement with subsidiary conditions. Acta Crystallogr. *16*, 1091–1094.

Watenpaugh, K.D., Sieker, L.C., Herriott, J.R., and Jensen, L.H. (1973). Refinement of the model of a protein: rubredoxin at 1.5 Å resolution. Acta Crystallogr. *B29*, 943–956.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. *28*, 31–36.

Weininger, D., Weininger, A., and Weininger, J.L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. J. Chem. Inf. Comput. Sci. *29*, 97–101.