Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

ValLigURL: a server for ligand-structure comparison and validation

Gerard J. Kleywegt* and Mark R. Harris

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 14 May 2007
Accepted 6 July 2007

A new web-based tool called ValLigURL is described. It can be used by practising crystallographers to validate the geometry of a ligand and to compare the conformation of a ligand with all instances of that ligand in the structural database (wwPDB). In addition, it can be used by structural bioinformaticians to survey the quality or conformational diversity of any ligand across the entire structural database. The server is freely accessible at the URL http://eds.bmc.uu.se/eds/valligurl.php.

1. Introduction

With the rapid advances in the area of biomacromolecular structure determination, both by X-ray crystallography and NMR spectroscopy, has come an explosive growth of the worldwide Protein Data Bank (wwPDB; Berman *et al.*, 2007). The total number of structures in the wwPDB has grown from ~5000 in 1996 to over 40 000 at the start of 2007. This expansion of the structural database has in turn stimulated and necessitated the development of tools (programs, databases, servers) for the analysis of biomacromolecular structures. Indeed, the field of structural bioinformatics has blossomed, as witnessed, for instance, by the many tools for analysis, classification, comparison, prediction and validation of biomacromolecular structural data described in the annual database and web-server issues of the journal *Nucleic Acids Research* (Roberts, 2006; Bateman, 2007). Traditionally, most of these tools have been developed for the study of protein structures and relatively few deal with other kinds of biologically important molecules (*e.g.* nucleic acids, carbohydrates, small-molecule ligands). For example, given a newly determined protein structure, many programs exist that can answer the question whether the fold of the protein is related to that of any other proteins and, if so, which residues are in corresponding positions in two related structures (Novotny *et al.*, 2004; Sierk & Kleywegt, 2004). However, if a new structure contains a bound ligand there are precious few tools to find out whether this ligand has previously been observed in this binding mode or conformation. There are some tools that can be used to study the binding environments of ligands in the wwPDB, such as MSDsite (Golovin *et al.*, 2005), SitesBase (Gold & Jackson, 2006), PDB-Ligand (Shin & Cho, 2005) and others (Guruprasad *et al.*, 2005). Other tools can be used to retrieve general information about these ligands, such as MSDchem (Golovin *et al.*, 2004), HIC-Up (Kleywegt & Jones, 1998), Ligand Depot (Feng *et al.*, 2004) and SuperLigands (Michalsky *et al.*, 2005). However, many of these tools are closed in the sense that they cover the structures that have been deposited in the wwPDB but do not allow the uploading of new structures for comparison to existing structures. Similarly, when it comes to structure validation there are vastly more tools for proteins than for other types of molecules (Kleywegt, 2000). The geometry of some small-molecule structures can be validated by running them through the *PRODRG* server (Schüttelkopf & van Aalten, 2004), certain structural aspects (such as close contacts) are covered by *WhatCheck* (Hooft *et al.*, 1996) and real-space fit values (Jones *et al.*, 1991) can be calculated by programs such as *O* (Jones *et al.*, 1991) and *MAPMAN* (Kleywegt & Jones, 1996).

doi:10.1107/S090744490703315X

In this communication, we report on a web server that can be used both by practising crystallographers who, for example, wish to compare the three-dimensional structure and geometry of a ligand in a newly determined structure with all its instances in the wwPDB and by structural bioinformaticians who wish to survey the conformational diversity or quality of one or more ligands across the entire wwPDB. The web server is called ‘ValLigURL’, reflecting its original purpose as a server for validating ligand structures, but the name is, like, totally pronounced ‘Valley Girl’.

2. Server description

The server consists of three major components. At its core lies a program called *LIGCOM* (Kleywegt, 2007; http://xray.bmc.uu.se/usf), which compares and superimposes two different structures of a ligand and reports statistics such as root-mean-square distance (r.m.s.d.) of common atoms, r.m.s. differences of bond lengths and angles *etc.* The second component is a shell script that retrieves a list of all wwPDB entries that contain the ligand of interest using *OCA* (http://oca.ebi.ac.uk) and fetches the coordinates of these wwPDB entries. The script then invokes *LIGCOM* to compare each of the instances with the user’s ligand, report the statistics of the comparison and superimpose the instances on the user’s ligand. Finally, a web server, implemented in *php*, handles the user input, runs the script and collects and presents the results.

The server presents a form through which the user can upload a structure (as a PDB file) and provide the three-letter code of the ligand of interest. If no code is provided, the first residue in the file is used. If no file is provided, the ‘ideal’ coordinates of the ligand, as generated by *CORINA* (Gasteiger *et al.*, 1990) at MSDchem, are retrieved from HIC-Up. This mechanism enables one to retrieve or compare the geometry of all instances of a ligand in the wwPDB.

Once the ligand has been identified and extracted from the uploaded file, the corresponding MSDchem two-dimensional structure diagram is retrieved from HIC-Up, the structure is converted into a SMILES string (Weininger, 1988; Weininger *et al.*, 1989) by Babel (http://www.eyesopen.com/; through the BabelWeb server at ChemDB; Chen *et al.*, 2005) and a list of all wwPDB entries that contain the ligand (based on its three-letter code) is fetched from *OCA*. Optionally, NMR structures may be excluded, as may any instances of the ligand that do not contain the same number of atoms as the uploaded structure, and the number of wwPDB entries to be retrieved can be limited to 200. Using a local mirror of the wwPDB, the relevant wwPDB entries are collected and the various instances of the ligand are extracted. Each of the instances of the ligand in turn is compared with the uploaded (or ideal, whichever the case may be) structure with *LIGCOM* and superimposed onto it using all atoms (Kearsley, 1989). The *LIGCOM* log files and the superimposed coordinates are stored and relevant statistics of the structural comparisons are extracted from the log files. The header of the ValLigURL results page contains the following.

(i) The number and ID codes of the wwPDB entries that contain the ligand.

(ii) The two-dimensional structure diagram of the ligand (from MSDchem).

(iii) The SMILES string of the ligand and a hyperlink to search MSDchem with that string, which can be useful to identify related molecules if the ligand itself does not yet occur in the wwPDB.

(iv) Hyperlinks to the HIC-Up and MSDchem entries of the ligand.

(v) A hyperlink to download a compressed archive file with all the ValLigURL result files (log files and superimposed coordinates).

The remainder of the results page consists of a table that lists the following information for the ideal coordinates and every instance of the ligand in the wwPDB (see Fig. 1 for an example).

(i) The ID code of the parent wwPDB entry. If the entry also occurs in the Uppsala Electron Density Server (EDS; Kleywegt *et al.*, 2004), the ID code is a hyperlink to that EDS entry; otherwise it links to the entry in the MSD database (Golovin *et al.*, 2004).

(ii) The resolution of the parent wwPDB entry (if applicable).

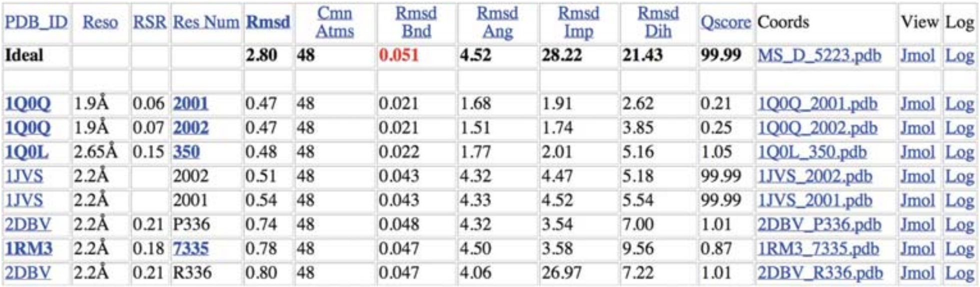
(iii) If the entry occurs in EDS, the real-space *R* value (Jones *et al.*, 1991; Kleywegt *et al.*, 2004) of the ligand in that entry is reported and the residue number of the ligand is provided as a hyperlink that will launch the EDS viewer (the EBI version of the *AstexViewer*; Hartshorn, 2002) showing the structure and its σ_A-weighted electron density (Read, 1990) and centred on the ligand. If there is no EDS entry, only the residue number is shown.

(iv) The number of common atoms between the uploaded or ideal structure and the wwPDB instance is shown (in red if it differs from the total number of atoms in the uploaded or ideal structure), along with the r.m.s.d. of those atoms after least-squares superimposition.

(v) The r.m.s. differences of the bond lengths and angles are reported (and shown in red if they are considered to be high, namely >0.05 Å for bond lengths and >5° for bond angles).

(vi) The r.m.s. differences of all improper torsion angles and of all dihedral angles are listed (but note the caveat discussed below).

(vii) A rough measure, Qscore, of the quality of the ligand instance is calculated by the script. If no ligand was uploaded and all instances are thus compared with the ideal coordinates, this score is calculated as follows: Qscore = $d^2\rho(10\beta + 0.1\alpha)$, where d is the resolution, ρ is the real-space R value from EDS, β is the r.m.s. deviation of bond lengths from ideal values and α is the r.m.s. deviation from ideal bond angles. If any of these values is unavailable, or if there are missing or extra atoms, Qscore is set to 99.99 instead. This purely empirical score tends to order the ligands by their crystallographic and geometric quality, with very good ligands having Qscore values of <0.1 and poor or low-resolution ones having values of >1.0. If the instances are compared with an uploaded ligand instead, the geometric deviations are with respect to that ligand, which is unlikely to have ideal bond lengths and angles. Therefore, in such cases the geometric part is omitted from the formula and Qscore is calculated as Qscore = $d^2\rho$.



PDB_ID	Reso	RSR	Res Num	Rmsd	Cmn Atms	Rmsd Bnd	Rmsd Ang	Rmsd Imp	Rmsd Dih	Qscore	Coords	View	Log
Ideal				**2.80**	**48**	**0.051**	**4.52**	**28.22**	**21.43**	**99.99**	MS_D_5223.pdb	Jmol	Log
1Q0Q	1.9Å	0.06	**2001**	0.47	48	0.021	1.68	1.91	2.62	0.21	1Q0Q_2001.pdb	Jmol	Log
1Q0Q	1.9Å	0.07	**2002**	0.47	48	0.021	1.51	1.74	3.85	0.25	1Q0Q_2002.pdb	Jmol	Log
1Q0L	2.65Å	0.15	**350**	0.48	48	0.022	1.77	2.01	5.16	1.05	1Q0L_350.pdb	Jmol	Log
1JVS	2.2Å		2002	0.51	48	0.043	4.32	4.47	5.18	99.99	1JVS_2002.pdb	Jmol	Log
1JVS	2.2Å		2001	0.54	48	0.043	4.33	4.52	5.54	99.99	1JVS_2001.pdb	Jmol	Log
2DBV	2.2Å	0.21	P336	0.74	48	0.048	4.32	3.54	7.00	1.01	2DBV_P336.pdb	Jmol	Log
1RM3	2.2Å	0.18	**7335**	0.78	48	0.047	4.50	3.58	9.56	0.87	1RM3_7335.pdb	Jmol	Log
2DBV	2.2Å	0.21	R336	0.80	48	0.047	4.06	26.97	7.22	1.01	2DBV_R336.pdb	Jmol	Log

Figure 1
List of some of the NADP(H) instances in the wwPDB that are most similar to an early model of an NADPH in *M. tuberculosis* DXR (Henriksson *et al.*, 2007) as generated by ValLigURL. The first line (‘Ideal’) compares the uploaded structure to the ideal NADPH structure from MSDchem. Refer to the text for more details.

(viii) Finally, three hyperlinks are provided to (1) the superimposed coordinates of the ligand, (2) a page on which *Jmol* (http://www.jmol.org/) is used to display the original and superimposed structures (Fig. 2) and (3) the *LIGCOM* log file with additional details and information about the structure comparison.

If the user uploaded a ligand structure, the table with results is sorted by increasing value of the r.m.s.d. However, by clicking on the header of any numerical column the table can be sorted on the corresponding criterion. Thus, the server can be used to rapidly find the instance of a ligand that is most similar (*i.e.* has the smallest r.m.s.d.) to the uploaded one or to find the highest resolution instance of a ligand or the instance that displays the best fit to its own electron density. If no ligand was uploaded, the ideal coordinates are used instead and thus the server can reveal quickly which instance of a ligand in the wwPDB has the best or the poorest bond lengths or bond angles. In that case, the results are initially sorted by increasing value of Qscore.

The weakest point at present is the perception of ligand chemistry in *LIGCOM* since this is based on the coordinates and not in a very sophisticated fashion at that. Bonds are deduced using a simple distance cutoff (2.0 Å by default), angles are defined by any two pairs of bonded atoms that share one atom and improper torsions are calculated for all atoms with three or more neighbours without any attempt to check whether they are chiral (and thus whether or not their sign is important). *LIGCOM* also calculates differences separately for torsion angles that are close to 0° or 180°, but since this is too unreliable a method to find nonconformational torsion angles only the r.m.s.d. for all torsion angles is reported. We plan to remedy this by replacing *LIGCOM* by a new program that uses XML dictionary descriptions of ligands from MSDchem, which include definitions of bonds, stereocentres and planar groups (M. Hong & G. J. Kleywegt, work in progress).

LIGCOM does not carry out graph matching to detect corresponding atoms in two ligand structures, but simply uses the atom names for this purpose. This implies that inconsistencies or errors in atom naming will lead to unrecognized atoms or high values of the r.m.s. deviations of bond lengths and angles.

If a ligand is uploaded that does not yet occur in the wwPDB, the server still produces the SMILES string and a hyperlink that launches a database query at MSDchem to find chemically similar ligands that do occur in the database.

The server can be accessed at the URL http://eds.bmc.uu.se/eds/valligurl.php and there are no restrictions on its use other than a limit of 20 runs per IP address per 48 h period (to reduce problems with robots). Results are stored on the server for a period of 24 h and then deleted. Typical queries require around a minute of processing time, although this increases linearly with the number of instances of a ligand in the wwPDB.

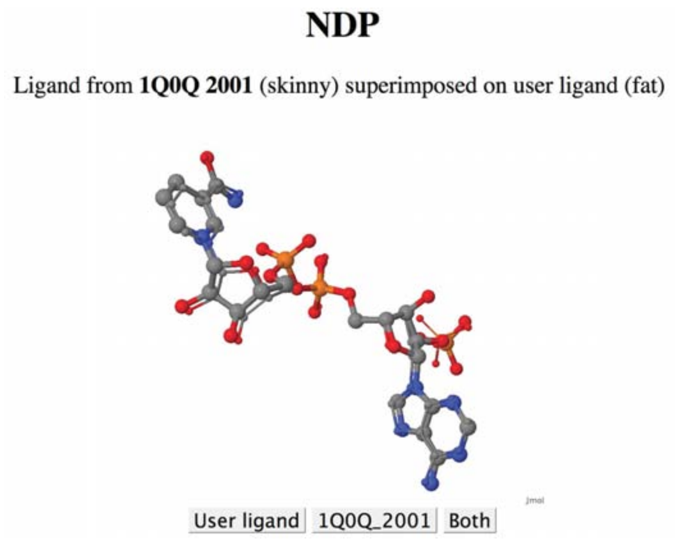
3. Applications

ValLigURL has been used to investigate whether the conformation of an NADPH molecule in the structure of *Mycobacterium tuberculosis* 1-deoxy-D-xylulose 5-phosphate reductoisomerase (DXR) in complex with Mn^{2+}, NADPH and the inhibitor fosfidomycin (Henriksson *et al.*, 2007) was unusual or not (the refined structure is now available as wwPDB entry 2jcz; the NADPH molecule is A1391). It is known that NAD(P) molecules display a wide range of conformations in complex with proteins (Carugo & Argos, 1997; Stockwell & Thornton, 2006). ValLigURL finds that there are indeed instances of NADP(H) in the wwPDB that are similar to the conformation observed in DXR (Figs. 1 and 2). At the time of writing, there are 408 copies of NADP(H) (three-letter code 'NDP') in 188 distinct wwPDB entries. Of these 408 instances, five have an r.m.s.d. to the NADPH in DXR of less than 0.6 Å calculated over all 48 atoms. As Fig. 2 shows, such values are indicative of very similar conformations. All five instances occur in structures of *Escherichia coli* DXR (wwPDB entries 1q0q, 1q0l and 1jvs). Comparison with the ideal NADP(H) coordinates reveals an r.m.s. deviation of 0.051 Å on bond lengths and 4.5° on bond angles. There are five bond lengths that differ by more than 0.05 Å from the ideal values and six angles that differ by more than 10° from their ideal counterparts. Most of these outliers involve a P atom. However, this comparison was carried out using an early model of the cofactor. If the refined and deposited coordinates are used instead, the r.m.s. deviations are only 0.016 Å for bonds and 2.5° for angles, with no outliers.

Instead of running ValLigURL with a single structure of a ligand, one could also run it with a number of alternative models, for instance if density for a ligand is poor or ambiguous or featureless or if disorder is suspected. In such cases, a number of candidate conformations (either hand-built or generated with automatic ligand-building tools; Zwart *et al.*, 2004; Aishima *et al.*, 2005; Terwilliger *et al.*, 2006; Wlodek *et al.*, 2006) could be processed with the server to assess how common they are.

If the identity of a ligand is not certain, a number of refined candidate ligands could be processed with the server to find out how common their conformations are, what their density typically looks like and perhaps also which interactions they tend to have with their host molecules (using tools cited in §1).

Finally, ValLigURL can be used as a simple validation tool by comparing the geometry of a ligand to that of its ideal counterpart, as exemplified in the discussion above about the NADPH in DXR. As has been pointed out a number of times previously (van Aalten *et al.*, 1996; Kleywegt & Jones, 1998; Kleywegt, 2000, 2007; Boström, 2001; Nissink *et al.*, 2002; Davis *et al.*, 2003; Kleywegt *et al.*, 2003; Schüttelkopf & van Aalten, 2004; Lütteke & von der Lieth, 2004), examples of 'unusual' ligand stereochemistry abound in the wwPDB. Probable causes of this phenomenon include the omission of necessary restraints, the imposition of incorrect restraints and the use of inappropriate restraint targets or weights (Kleywegt, 2007). Inspec-



NDP

Ligand from **1Q0Q 2001** (skinny) superimposed on user ligand (fat)

Figure 2
Comparison of an early model of an NADPH molecule in *M. tuberculosis* DXR (Henriksson *et al.*, 2007; displayed with fat atoms and bonds) and the most similar instance identified in the wwPDB by ValLigURL (residue 2001 in wwPDB entry 1q0q; displayed with thin atoms and bonds). Refer to the text for more details. The molecular display was generated using *Jmol* (http://www.jmol.org/).

tion of the *LIGCOM* log file for the comparison of a ligand and its ideal partner from MSDchem quickly reveals whether there are any unusual bond lengths or bond angles and may also help in detecting any instances of incorrect chirality or problems with atom naming.

The applications listed above are chiefly of interest to practising crystallographers. However, ValLigURL can also be used for wwPDB-wide data mining and analysis. For instance, the server provides a rapid way to identify and superimpose all instances of a ligand in the wwPDB. The coordinates can then be downloaded and the structures subjected to further analysis, clustering *etc*., for instance with *GAMUT* (Stockwell & Thornton, 2006). If one is interested in finding reliable instances of a ligand (*e.g.* as a starting point for modelling or to include in a set of validated docking targets; Hartshorn *et al.*, 2007), the consideration of more specific statistics than resolution alone is imperative. The structures can be selected from the ValLigURL results table taking into account such statistics as the r.m.s. deviations of bond lengths and angles from ideal values, the real-space fit and the resolution of the structure. The Qscore statistic is provided in an attempt to combine these statistics into a single number.

The authors are grateful to Lena Henriksson and Alwyn Jones (Uppsala University) for providing the problem that sparked this work, and to Alwyn Jones and Kim Henrick (EBI) for useful comments, discussions and suggestions. They are also grateful to the authors and curators of wwPDB, MSDchem, *OCA*, *Jmol* and BabelWeb for providing their services free of charge to the community. This work was supported by a European Union grant for the Sixth Framework Programme Vizier (LSHG-CT-2004-511960) to T. A. Jones (MRH) and by Uppsala University (GJK).

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.

Aishima, J., Russel, D. S., Guibas, L. J., Adams, P. D. & Brünger, A. T. (2005). *Acta Cryst.* D**61**, 1354–1363.

Bateman, A. (2007). *Nucleic Acids Res.* **35**, D1–D2.

Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. (2007). *Nucleic Acids Res.* **35**, D301–D303.

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

Carugo, O. & Argos, P. (1997). *Proteins*, **28**, 10–28.

Chen, J., Swamidass, S. J., Dou, Y., Bruand, J. & Baldi, P. (2005). *Bioinformatics*, **21**, 4133–4139.

Davis, A. M., Teague, S. J. & Kleywegt, G. J. (2003). *Angew. Chem. Int. Ed.* **42**, 2718–2736.

Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H. M. & Westbrook, J. (2004). *Bioinformatics*, **20**, 2153–2155.

Gasteiger, J., Rudolph, C. & Sadowski, J. (1990). *Tetrahedron Comput. Methods*, **3**, 537–547.

Gold, N. D. & Jackson, R. M. (2006). *Nucleic Acids Res.* **34**, D231–D234.

Golovin, A. *et al.* (2004). *Nucleic Acids Res.* **32**, D211–D216.

Golovin, A., Dimitropoulos, D., Oldfield, T., Rachedi, A. & Henrick, K. (2005). *Proteins*, **58**, 190–199.

Guruprasad, K., Savitha, S. & Babu, A. V. (2005). *Int. J. Biol. Macromol.* **37**, 35–41.

Hartshorn, M. J. (2002). *J. Comput. Aided Mol. Des.* **16**, 871–881.

Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T., Mortenson, P. N. & Murray, C. W. (2007). *J. Med. Chem.* **50**, 726–741.

Henriksson, L. M., Unge, T., Carlsson, J., Åqvist, J., Mowbray, S. L. & Jones, T. A. (2007). *J. Biol. Chem.* **282**, 19905–19916.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.

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

Kearsley, S. K. (1989). *Acta Cryst.* A**45**, 208–210.

Kleywegt, G. J. (2000). *Acta Cryst.* D**56**, 249–265.

Kleywegt, G. J. (2007). *Acta Cryst.* D**63**, 94–100.

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. (1996). *Acta Cryst.* D**52**, 826–828.

Kleywegt, G. J. & Jones, T. A. (1998). *Acta Cryst.* D**54**, 1119–1131.

Lütteke, T. & von der Lieth, C. W. (2004). *BMC Bioinformatics*, **5**, 69.

Michalsky, E., Dunkel, M., Goede, A. & Preissner, R. (2005). *BMC Bioinformatics*, **6**, 122.

Nissink, J. W., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C. & Taylor, R. (2002). *Proteins*, **49**, 457–471.

Novotny, M., Madsen, D. & Kleywegt, G. J. (2004). *Proteins*, **54**, 260–270.

Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.

Roberts, R. J. (2006). *Nucleic Acids Res.* **34**, W1.

Schüttelkopf, A. W. & van Aalten, D. M. (2004). *Acta Cryst.* D**60**, 1355–1363.

Shin, J. M. & Cho, D. H. (2005). *Nucleic Acids Res.* **33**, D238–D241.

Sierk, M. L. & Kleywegt, G. J. (2004). *Structure*, **12**, 2103–2111.

Stockwell, G. R. & Thornton, J. M. (2006). *J. Mol. Biol.* **356**, 928–944.

Terwilliger, T. C., Klei, H., Adams, P. D., Moriarty, N. W. & Cohn, J. D. (2006). *Acta Cryst.* D**62**, 915–922.

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.

Wlodek, S., Skillman, A. G. & Nicholls, A. (2006). *Acta Cryst.* D**62**, 741–749.

Zwart, P. H., Langer, G. G. & Lamzin, V. S. (2004). *Acta Cryst.* D**60**, 2230–2239.