Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

PRODRG: a tool for high-throughput crystallography of protein–ligand complexes

Alexander W. Schüttelkopf and Daan M. F. van Aalten*

Division of Biological Chemistry and Molecular Microbiology, Wellcome Trust Biocentre, School of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, Scotland

Correspondence e-mail: dava@davapc1.bioch.dundee.ac.uk

Received 25 February 2004
Accepted 13 May 2004

The small-molecule topology generator *PRODRG* is described, which takes input from existing coordinates or various two-dimensional formats and automatically generates coordinates and molecular topologies suitable for X-ray refinement of protein–ligand complexes. Test results are described for automatic generation of topologies followed by energy minimization for a subset of compounds from the Cambridge Structural Database, which shows that, within the limits of the empirical *GROMOS*87 force field used, structures with good geometries are generated. X-ray refinement in *X-PLOR*/*CNS*, *REFMAC* and *SHELX* using *PRODRG*-generated topologies produces results comparable to refinement with topologies from the standard libraries. However, tests with distorted starting coordinates show that *PRODRG* topologies perform better, both in terms of ligand geometry and of crystallographic *R* factors.

1. Introduction

With the rise of structure-based drug-design techniques (reviewed in Davis *et al.*, 2003), it is important to have software available which supports the ligand/inhibitor throughout the entire design process. Firstly, coordinates for the drug need to be built or an existing molecule modified, followed by docking of the drug into the active site and/or refinement of a protein–drug complex against X-ray diffraction data. The protein–drug interaction then needs to be examined in terms of detailed hydrogen-bonding geometry or other scoring functions (reviewed in Brooijmans & Kuntz, 2003). During this process, the drug interacts with different types of software and for each of these types a wide variety of packages are available (Davis *et al.*, 2003). Making these computer programs understand the topology of the drug involved is often a laborious process and, when no structural information is available, prone to errors as bond lengths and angles often have to be guessed (Kleywegt *et al.*, 2003). In the current drive towards high-throughput crystallography, a large number of protein–inhibitor complexes need to be refined and evaluated, which increases the need for a high level of automation (Blundell *et al.*, 2002). Similarly, significant effort is currently being invested into virtual screening of small-molecule libraries using docking methods (Richards, 2002). To be able to create, dock and refine large libraries of small molecules, a fast, accurate and publicly available program is needed to create topological information from a variety of input formats (two-dimensional and three-dimensional representations) for a wide range of computer packages used in this process. Here, a new version of the program *PRODRG* is described which performs all these tasks. The program is tested against the Cambridge Structural

Database (CSD; Allen, 2002) and a number of protein–ligand complexes.

2. Details of the *PRODRG* algorithm

2.1. *PRODRG* basics

The basics of the *PRODRG* algorithm have been described previously (van Aalten *et al.*, 1996); hence only a short overview will be given here. The main aim of *PRODRG* is to provide topological information for small molecules that can be used in X-ray refinement, molecular-dynamics simulations, molecular modelling and docking studies. *PRODRG* is currently limited to molecules containing H, C, N, O, P, S, F, Cl, Br or I atoms; also, atoms with more than four bonds and certain types of bonds between halogens and non-C atoms are not supported.

Previously, *PRODRG* only accepted coordinates in PDB format (PDB mode) as input (van Aalten *et al.*, 1996). This has now been expanded, with two additional input modes. The first allows description of molecules as a simple ASCII drawing (TXT mode), illustrated in Fig. 1. The TXT mode represents a portable description of the molecule (it can be created and edited in any text editor in any operating system) that is easily interpreted by humans as well as machines. Single, double and triple bonds can be drawn between atoms and chirality (discussed below) is indicated by the case of the letter describing the atom. The second new mode is the popular MDL Molfile/SDfile format (MOL mode), which is used in programs such as *ChemDraw* (CambridgeSoft, Massachusetts, USA) and *ISIS*/*Draw* (MDL Information Systems, California, USA) and is also written out by the Java-based JME editor (Ertl & Jacob, 1997).

The net result after initial processing by *PRODRG* is a connection table, containing the bonds between non-H atoms, the hybridization states and information on chirality (see van Aalten *et al.*, 1996 for a full description). All further information, such as all coordinates and the H atoms in the input, are ignored. This has the advantage of *PRODRG* entering the subsequent steps with the same information regardless of what this was determined from: a small molecule input *via* TXT mode will thus lead to the same topology and derived information as the same molecule supplied *via* a high-resolution crystal structure.

2.2. Determination of protonation state

After the initial connection table has been generated, probable amide N atoms are identified and the presence/extent of aromatic systems is determined. The aromaticity detection is based on Hückel's $4n + 2$ rule, but is not limited to single-ring systems. With this information it is then possible to add H atoms so that the expected valencies are satisfied, even though in some cases the program will add fewer or more H atoms, so that *e.g.* carboxylates remain deprotonated while guanidinium groups are fully protonated. *PRODRG* offers three statements for modifying the input or generated structure. Two of them, `INSHYD <atom>` and `DELHYD <atom>` allow modification of the protonation state of any atom by either adding or removing a hydrogen to/from it (Fig. 2). The third command, `PATCH <atom> <value>`, is used to force the hybridization of an atom (`value` = 1, 2, 3 for *sp*, sp^2 or sp^3 hybridization) or to invert a chiral centre (`value` = −1). It thus provides an easy tool to modify an existing structure on the fly, but the ability to modify hybridization assignments is also useful in case *PRODRG* misinterprets poor input coordinates.

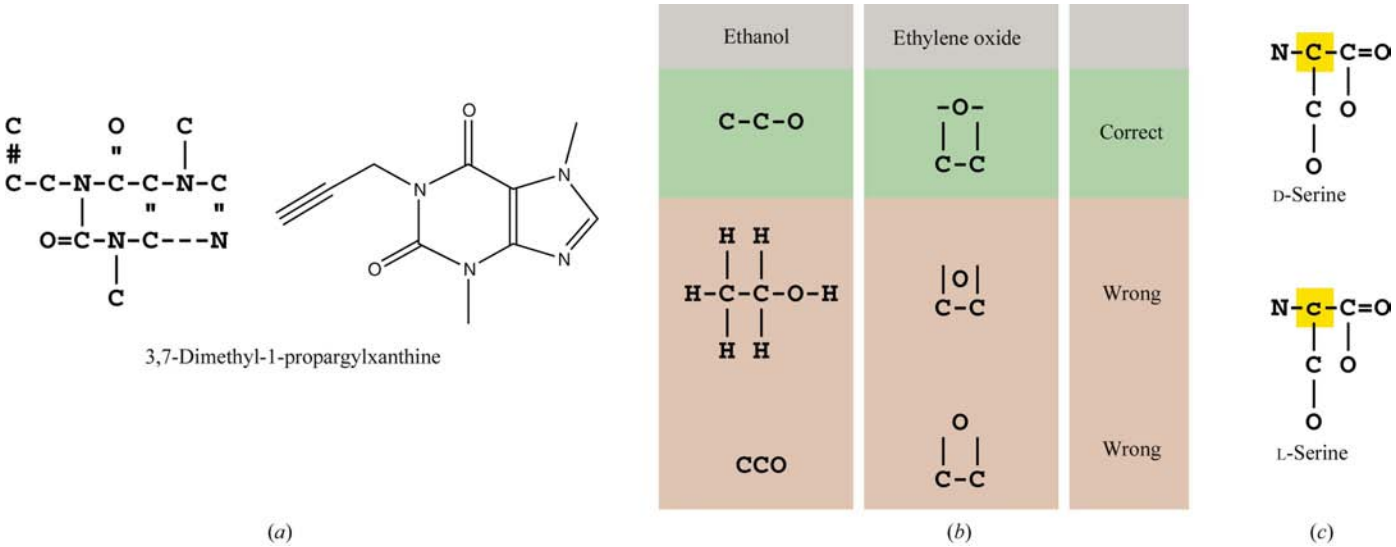

Figure 1
The TXT input mode. (*a*) 3,7-Dimethyl-1-propargylxanthine as an example for TXT input. Single bonds can be input as – or |, double bonds as = or " and triple bonds as #. Atoms must be separated by bonds, while bonds/atoms that are not connected must be separated by white space; 'diagonal' connections are not accepted. H atoms may be included but will be ignored. (*b*) Common mistakes when entering TXT drawings. Left (ethanol) from top to bottom: correct drawing; useless inclusion of H atoms; missing bonds. Right (ethylene oxide) from top to bottom: correct drawing; no space between O and the C—C bond; diagonal connection to O. (*c*) The chirality of atoms can be changed by using lower-case element symbols.

2.3. Coordinate generation and energy minimization

The H-atom assignment is followed by the generation of a topology for use with *GROMACS* (Berendsen *et al.*, 1995; Lindahl *et al.*, 2001). If desired, *PRODRG* can then use *GROMACS* to either generate coordinates *ab initio* for the molecule or energy-minimize user-provided coordinates. Energy minimization is performed by steepest descent for at most 50 000 steps, with the ffgmx *GROMACS* force field, extended by 11 additional atom types to accommodate halogens, *sp*-hybridized atoms and other chemical features. Parameters for the new atom types have been determined from about 47 000 experimentally determined small-molecule structures from the CSD (see below).

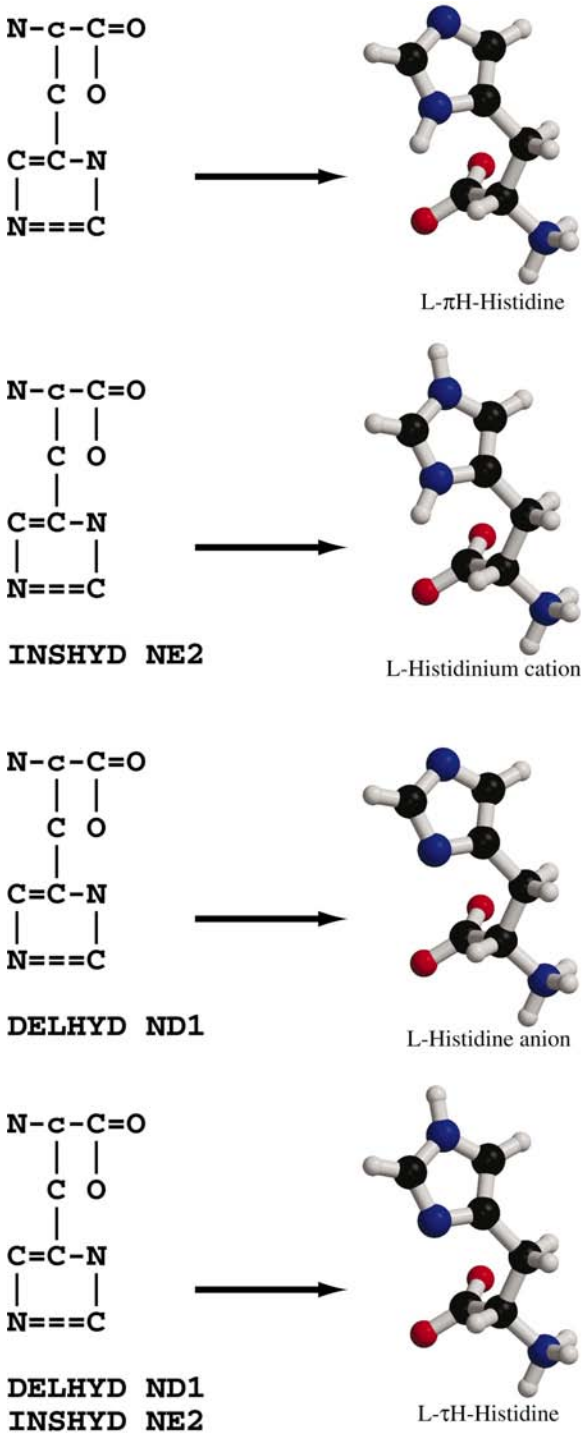

Figure 2
Use of INSHYD and DELHYD to generate different protonation states of histidine. For some simple molecules *PRODRG* will automatically generate meaningful/standard atom names, which in this case allows the two N atoms of the imidazole ring to be addressed as ND1 and NE2.

Table 1
Statistics for the *PRODRG* run on ~47 000 small-molecule X-ray structures selected from the CSD.

Failure owing to *PRODRG* limitations includes structures containing atoms with more than four connections, unsupported non-carbon–halogen bonds and molecules consisting of fewer than three atoms. Structures are considered too complex if repeated attempts at energy minimization fail to yield results of acceptable geometry in terms of the ffgmx *GROMACS* force field. 'Bad input geometry' summarizes structures of unusual geometry, the interpretation of which led to unresolvable inconsistencies, forcing *PRODRG* to fail.

	No. compounds	Time (s)	Time/ compound (s)
Success	46144 (98.3%)	37220.8 (96.6%)	0.81
Overall failure	820 (1.7%)	1309.6 (3.4%)	1.60
PRODRG limitations	539 (1.1%)	16.6 (0.0%)	0.03
Input too complex	175 (0.4%)	1289.2 (3.3%)	7.37
Bad input geometry	106 (0.2%)	3.8 (0.0%)	0.04

2.4. Program output

Apart from the *GROMACS* topology and molecular coordinates, which are written out in PDB format, *GROMOS*/*GROMACS* format and as an MDL Molfile, *PRODRG* now generates topologies for use with numerous other programs. This includes crystallographic refinement/model-building programs [*X-PLOR* (Brünger, 1988), *CNS* (Brünger *et al.*, 1998), *REFMAC*5 (Murshudov *et al.*, 1997), *SHELX* (Sheldrick & Schneider, 1997) and *O* (Jones *et al.*, 1991)] as well as docking programs [*AutoDock* 2.4/3.0 (Morris *et al.*, 1996, 1998), *Hex* (Ritchie & Kemp, 2000)]. Furthermore, *PRODRG* writes out SYBYL2 files, which can be read by numerous computational chemistry and ligand-design programs. Particularly useful is the topology for the molecular-modelling program *WHAT IF* (Vriend, 1990), which allows the precise and automatic determination of protein–ligand hydrogen bonding geometry with *WHAT IF*'s HB2 algorithm (Hooft *et al.*, 1996; Rao *et al.*, 2003).

3. Results and discussion

3.1. Testing on compounds in the CSD

A set of compounds was selected from the CSD to perform a large-scale test of *PRODRG* topology quality. Compounds were selected if they did not contain atoms other than C, H, N, O, P, S, F, Cl, Br and I. In the case of entries containing multiple molecules, the largest molecule was chosen. This resulted in 46 964 compounds which were processed by *PRODRG* in less than 11 h on an 2.0 GHz AMD Athlon-based Linux system. For each compound, the full topological

information was generated, followed by energy minimization with the generated topology in the *GROMACS* package. Of the 46 964 *PRODRG* runs, 820 failed for the reasons described in Table 1. The 46 144 successfully processed structures were then compared with the starting structures in terms of bond lengths, bond angles, improper dihedral angles and coordinate r.m.s.d. (Fig. 3). The average r.m.s.d.s between crystallographic and *PRODRG*-generated structures are 0.040 Å on bonds, 2.99° on angles, 1.97° on improper dihedrals and 0.26 Å on aligned coordinates. These reasonable results reflect both *PRODRG*'s ability to extract topological information from coordinates only and the quality of the *GROMOS*87 force-field-based limited parametrization used.

There are numerous other programs that generate three-dimensional coordinates from connection-table data (reviewed in Sadowski *et al.*, 1994, and updated in Gasteiger *et al.*, 1996). The aim of these programs is to predict accurately the 'real' conformation of a compound for use in *e.g. 3D-QSAR* (quantitative structure–activity relationship) studies. *PRODRG*-generated structures, on the other hand, while generally of low energy and chemically meaningful, are neither guaranteed nor intended to represent the absolute energy minimum of an input compound. This is not necessary, as *PRODRG*-produced structures will normally be used as the starting point for other procedures such as model building, crystallographic refinement, molecular dynamics or docking, which will determine the final conformation.

3.2. Testing in X-ray refinement

PRODRG writes out topology information which can be used in *X-PLOR*/*CNS*, *REFMAC*5 or *SHELX* to properly model small-molecule compounds during refinement against X-ray crystallographic data. The quality of the automatically generated topologies was evaluated using a number of refined structures, in which the previously used small-molecule topology was substited with a *PRODRG* topology generated from a TXT drawing (Figs. 4*a* and 4*b*). Refinement was then continued and initial and final *R* factors compared, together with an indication of conformational change in the small molecule introduced by switching the topology, expressed as the r.m.s.d. on the atomic positions. In the *PRODRG*-

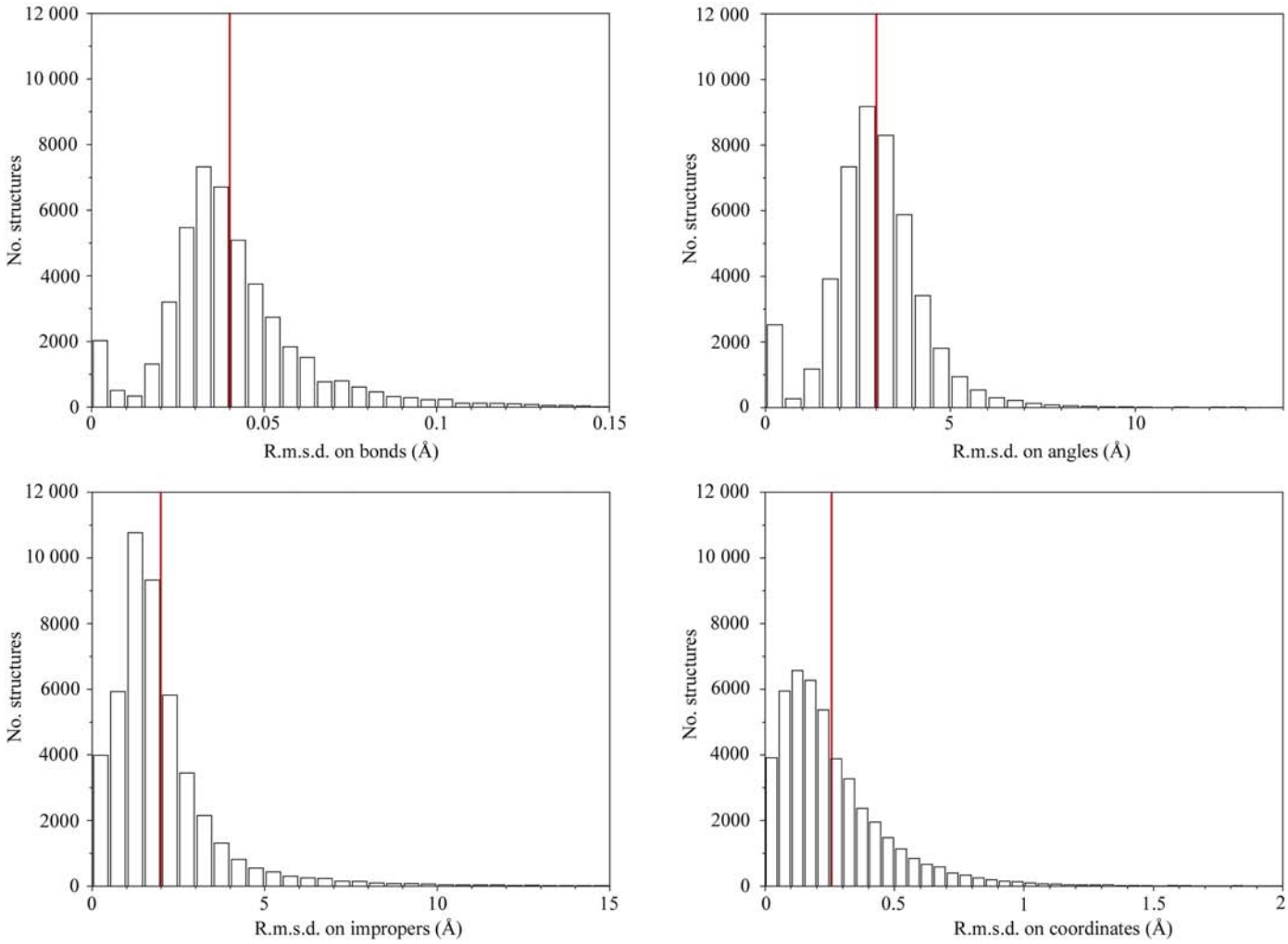

Figure 3
Comparison of crystallographic and *PRODRG*-treated small-molecule structures. Histograms are shown for r.m.s.d. values on bond lengths, angles, improper dihedral angles and coordinates; the average r.m.s.d. is indicated by a red line.

Table 2
Details of X-ray refinement tests of protein–ligand complexes using *PRODRG* topologies.

All measured data were included in the refinement. The source of the original topology is indicated (S, standard library of the refinement program; M, manually made topology; L, topology made with *LIBCHECK* and validated manually). The additional refinement consisted of two cycles of 100 steps of positional refinement followed by 20 steps of temperature-factor refinement (*CNS*) or ten steps (*REFMAC*5). The real-space *R* factor was calculated using *O* with standard settings. ChiB, *Serratia marcescens* chitinase B (van Aalten *et al.*, 2001); SCP-2L, sterol carrier protein type 2-like domain of human multifunctional enzyme type 2 (Haapalainen *et al.*, 2001); ACBP, acyl-CoA binding protein (not published); PTR1, *Leishmania major* pteridine reductase 1 (Gourley *et al.*, 2001; Schüttelkopf, 2003); PYP, *Ectothiorhodospira halophila* photoactive yellow protein (van Aalten *et al.*, 2002); n/a, not applicable; n/d, not deposited.

Protein	ChiB	SCP-2L	ACBP	PTR1	PTR1	PYP
Refinement program	*CNS*	*CNS*	*CNS*	*CNS*	*REFMAC*5	*SHELX*
Original topology	S	M	M	M	L	M
PDB code	1e6n	1itk	n/d	1e92	n/d	1kou
Resolution (Å)	2.25	1.75	1.48	2.20	2.70	1.16
Ligand(s)	NAG_5	Triton X-100	Badan	$NADP^+$ and DHB	NADPH and pterin derivative	Caffeic acid
Ligand atoms	142	25	16	260	332	20
Initial R_{work}	0.189	0.192	0.200	0.198	0.205	0.162
Initial R_{free}	0.239	0.216	0.221	0.227	0.240	0.206
Initial real-space *R*	0.123	0.060	0.064	0.132	0.190	0.099
Final R_{work}	0.189	0.192	0.200	0.199	0.205	0.162
Final R_{free}	0.240	0.217	0.222	0.228	0.239	0.206
Final real-space *R*	0.121	0.060	0.063	0.132	0.190	0.104
Ligand W_{CNS}	4.00	0.25	4.00	0.25	n/a	n/a
R.m.s.d. ligand (Å)	0.14	0.16	0.15	0.06	0.08	0.06

generated *X-PLOR*/*CNS* topologies, the bonded forces are scalable with a separate weight factor and values of 0.25, 0.5, 1.0, 2.0 and 4.0 were tested for all systems to obtain an optimum weight of the geometrical restraints *versus* X-ray data for the small molecule in terms of the smallest separation between *R* and R_{free}. The results are presented in Table 2 and Figs. 4(*a*) and 4(*b*), showing that *PRODRG* topologies perform well in crystallographic refinement.

In addition to our own tests described here, a number of recent studies describing refinement of protein–ligand complexes have successfully employed *PRODRG* for description of the ligand geometry (*e.g.* Ekstrom *et al.*, 2002; Evans *et al.*, 2002; Gadola *et al.*, 2002; Hall *et al.*, 2002; Matern *et al.*, 2003; Nicolet *et al.*, 2003; Zavala-Ruiz *et al.*, 2003; Dong *et al.*, 2004).

3.3. Comparison with similar programs

3.3.1. *XPLO2D*/*HIC-Up*. The Uppsala Software Factory program *XPLO2D* (Kleywegt, 1995) can be used to generate topologies for use with, amongst others, *X-PLOR*/*CNS* and *O* from small-molecule coordinates. For small molecules present in PDB entries, the *HIC-Up* service (Kleywegt & Jones, 1998) provides the required coordinates (gathered from the PDB) as well as pregenerated *XPLO2D* topologies. Unlike *PRODRG*, which always uses its own *GROMOS*87-derived parameters, *XPLO2D* derives topology parameters from the input coordinates, thus implicitly assuming these are correct (Kleywegt *et al.*, 2003).

To compare the performance of *XPLO2D*- and *PRODRG*-generated topologies for refinement with *CNS*, several high-resolution structures (≤1.2 Å) were obtained from the PDB and re-refined after truncating the data to 2.8 Å resolution, optionally after slight perturbation (by an average random coordinate shift of 0.1 Å), with topologies produced from the original ligand coordinates either by *PRODRG* or *XPLO2D*. In all cases, the crystallographic weight was optimized to give the lowest R_{free}. Table 3 shows that the coordinate r.m.s.d.s between the original high-resolution ligand(s) and the re-refined ligand(s) do not differ significantly between the two topology sources. This is remarkable considering that *XPLO2D*, unlike *PRODRG*, acquires its parameters from the 'perfect' input structure and thus its topologies might be expected to present a better model of this perfect structure. The values of R_{work} as well as the real-space *R* factor computed with *O* are generally similar for *PRODRG*- and *XPLO2D*-based refinement runs; on the other hand, R_{free} is consistently lower when using *PRODRG*-generated topologies. The r.m.s.d.s for the runs with perturbed or unperturbed coordinates are essentially identical in all cases, showing that the quality of the results is not significantly influenced by either topology being 'too loose'.

Next, the impact of the quality of the input coordinates was investigated. The refinement of HGPRT (PDB code 1fsg) was repeated several times with *XPLO2D*- and *PRODRG*-generated topologies produced from ligand coordinates to which an increasing random coordinate shift (from 0.05 to 0.25 Å) had been applied (Fig. 4*e*). As expected, the *XPLO2D*-dependent refinement deteriorates steadily with increasing ligand coordinate error. Because *PRODRG* uses tabled parameters, its topologies are less sensitive to the quality of the input coordinates, even though above an average shift of 0.15 Å atom-type misassignments begin to occur (intriguingly though in this case these lead to a minimal improvement in the refined ligand geometry). For comparison the results obtained with topologies generated independently of input coordinates are also shown (empty diamonds in Fig. 4*e*). In *PRODRG*, topologies produced from two-dimensional descriptions can be expected to perform equally well or better than those derived from PDB input, as the drawings allow greater precision in the specification of a compound. Indeed, in the test case the TXT-produced

topology performs slightly better than the ligand PDB-generated topology. Alternatively, topologies were obtained from *HIC-Up*: this relies on the required ligands being available in a PDB-deposited structure of reasonably high quality.

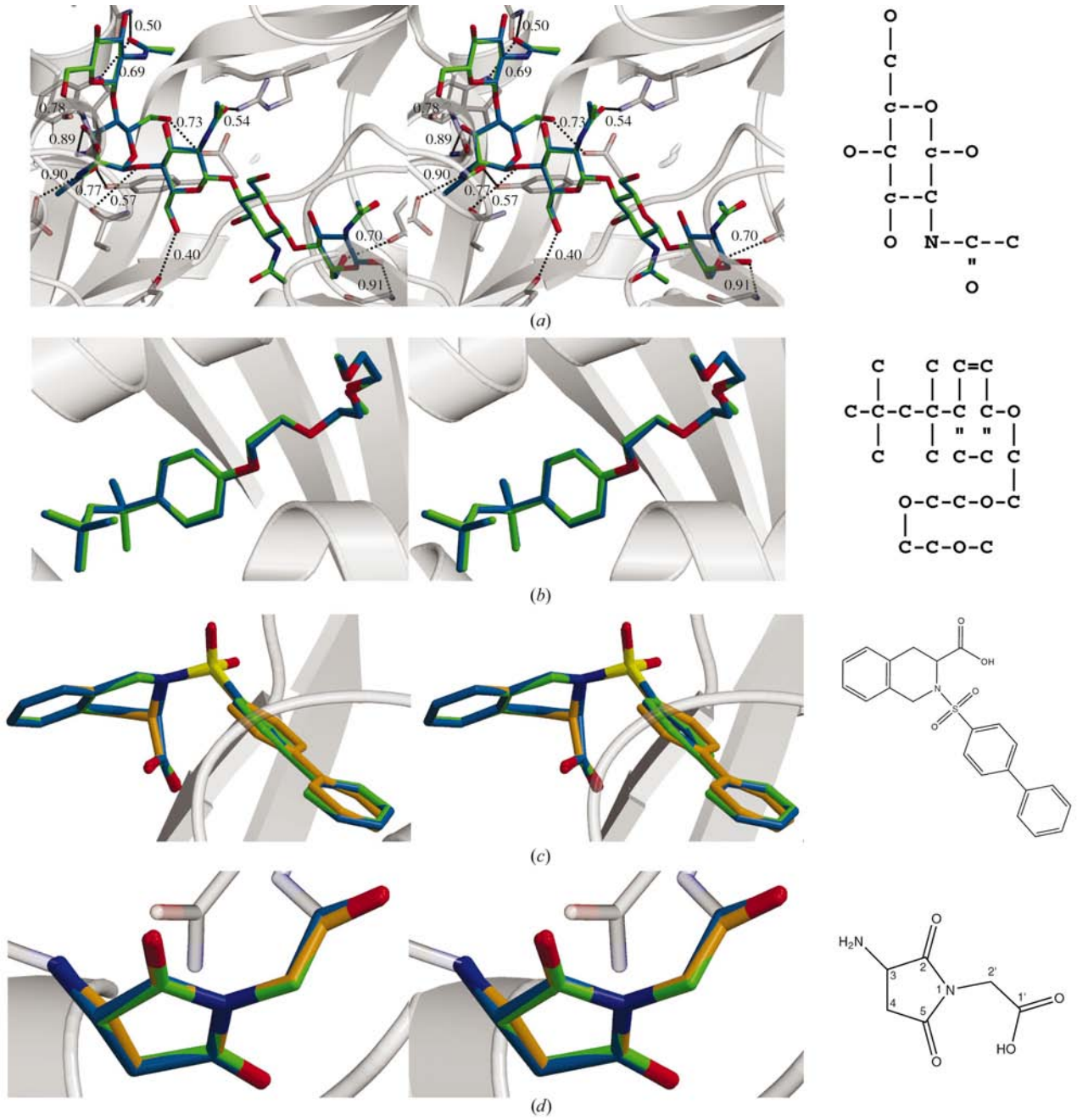

Figure 4
Use of *PRODRG*-generated topologies. (*a*) $GlcNAc_5$ in ChiB (van Aalten *et al.*, 2001). Left, stereo diagram of the ligand molecule before (cyan) and after (green) refinement with a *PRODRG*-generated topology. The surrounding protein is shown as a semitransparent cartoon. Right, text drawing used to generate the topology. (*b*) As (*a*) for Triton X-100 in SCP-2L (Haapalainen *et al.*, 2001). (*c*) Ligand from a high-resolution structure (cyan molecule) of human neutrophil collagenase (Gavuzzo *et al.*, 2000) re-refined at lower resolution with topologies generated either with *PRODRG* (green molecule) or with *LIBCHECK* (orange molecule). Again, the protein is shown as a semitransparent cartoon. To the right, the chemical structure of the ligand [2-(biphenyl-4-sulfonyl)-1,2,3,4-tetrahydroisoquinoline-3-carboxylic acid] is given. (*d*) As (*c*) for (3-amino-2,5-dioxo-1-pyrrolidinyl)-acetic acid in *Cryphonectria parasitica* endothiapepsin (Erskine *et al.*, 2003). (*e, f*) Effect of poor input geometries on the quality of generated topologies as indicated by the r.m.s.d. between small-molecule coordinates from the 'ideal' starting structure and the same structure after refinement at lower resolution. The refinement of HGPRT as described in Table 3 is repeated with topologies generated from coordinates perturbed by a given random shift (filled squares). In addition, the corresponding refinement results using topologies produced in a coordinate-independent manner are given (empty diamonds). For *PRODRG* this means topologies were generated from TXT-mode drawings; for *XPLO2D* the topologies available from *HIC-Up* were used and for *LIBCHECK* the ligands were drawn in *SKETCHER*. Weights are kept at the values given in Table 3. (*e*) shows the results for refinement with *CNS* and (*f*) for *REFMAC*5.

Table 3
Low-resolution (2.8 Å) re-refinement of high-resolution structures.

The *CNS* refinement protocol comprised two cycles of 30 steps of positional refinement followed by 30 steps of temperature-factor refinement; refinement with *REFMAC*5 proceeded for ten steps. All refinements were carried out both on the original structure and on coordinates perturbed by an average shift of 0.1 Å. R_{work}, R_{free}, real-space *R* (calculated with *O* using default settings) and final ligand r.m.s.d. values are given for the unperturbed and perturbed case separated by a slash. HGPRT, *Toxoplasma gondii* hypoxanthine-guanine phosphoribosyltransferase (Heroux *et al.*, 2000); CBM29-2, *Piromyces equi* family 29 carbohydrate-binding module (Charnock *et al.*, 2002); HNC, human neutrophil collagenase (Gavuzzo *et al.*, 2000); DERA, *E. coli* D-2-deoxyribose-5-phosphate aldolase (Heine *et al.*, 2001); DHFR, human dihydrofolate reductase (Klon *et al.*, 2002); EAPA, *Cryphonectria parasitica* endothiapepsin (Erskine *et al.*, 2003); PRPP, phosphoribosylpyrophosphate; BSI, 2-(biphenyl-4-sulfonyl)-1,2,3,4-tetrahydroisoquinoline-3-carboxylic acid; LIH, 6-[(5-quinolylamino)methyl]-2,4-diamino-5-methylpyrido(2,3-D)pyrimidine; LOV, 5-amino-4-hydroxy-2-isoproyl-7-methyl-octanoic acid; SUI, (3-amino-2,5-dioxo-1-pyrrolidinyl)-acetic acid.

Protein	HGPRT	CBM29-2	HNC	DERA	DHFR	EAPA
PDB code	1fsg	1gwm	1i76	1jcj	1kms	1oex
Resolution (Å)	1.05	1.15	1.20	1.10	1.09	1.10
Ligand(s)†	PRPP and 9-deazaguanine	β-D-glucose	BSI	D-2-deoxyribose-5-phosphate	LIH	LOV and SUI
Ligand atoms	66	66	28	24	25	26
Refinement with *CNS* (*XPLO2D*)						
Overall W_{CNS}	0.4	1.0	2.0	2.0	0.4	0.4
R_{work}	0.182/0.181	0.185/0.186	0.171/0.170	0.163/0.162	0.186/0.190	0.166/0.165
R_{free}	0.205/0.207	0.235/0.238	0.226/0.221	0.220/0.224	0.224/0.225	0.196/0.194
Real-space *R* factor	0.261/0.262	0.194/0.196	0.081/0.080	0.093/0.092	0.088/0.087	0.097/0.097
R.m.s.d.$_{i-f}$ (Å)	0.16/0.16	0.32/0.32	0.13/0.14	0.11/0.12	0.22/0.23	0.10/0.10
Refinement with *CNS* (*PRODRG*)						
Overall W_{CNS}	0.4	1.0	2.0	2.0	0.4	0.4
R_{work}	0.172/0.174	0.176/0.177	0.168/0.172	0.163/0.164	0.180/0.180	0.157/0.159
R_{free}	0.187/0.192	0.214/0.210	0.205/0.209	0.215/0.220	0.213/0.214	0.182/0.184
Real-space *R* factor	0.261/0.264	0.192/0.193	0.081/0.079	0.093/0.092	0.088/0.087	0.096/0.097
R.m.s.d.$_{i-f}$ (Å)	0.14/0.15	0.35/0.35	0.10/0.10	0.14/0.15	0.17/0.17	0.09/0.10
Refinement with *REFMAC*5 (*LIBCHECK*)						
Overall W_{MAT}	0.010	0.040	0.100	0.300	0.007	0.007
R_{work}	0.144/0.148	0.152/0.154	0.139/0.141	0.127/0.128	0.167/0.171	0.136/0.140
R_{free}	0.151/0.157	0.161/0.168	0.174/0.178	0.190/0.194	0.178/0.180	0.146/0.145
Real-space *R* factor	0.042/0.042	0.031/0.031	0.011/0.010	0.014/0.014	0.012/0.012	0.014/0.013
R.m.s.d.$_{i-f}$ (Å)	0.07/0.08	0.38/0.38	0.18/0.18	0.10/0.12	0.15/0.15	0.08/0.10
Refinement with *REFMAC*5 (*PRODRG*)						
Overall W_{MAT}	0.010	0.040	0.100	0.300	0.007	0.007
R_{work}	0.145/0.152	0.152/0.154	0.138/0.140	0.127/0.128	0.167/0.171	0.136/0.141
R_{free}	0.148/0.158	0.162/0.168	0.175/0.180	0.190/0.194	0.178/0.179	0.146/0.146
Real-space *R* factor	0.042/0.042	0.031/0.031	0.011/0.010	0.014/0.014	0.012/0.012	0.013/0.013
R.m.s.d.$_{i-f}$ (Å)	0.06/0.08	0.39/0.38	0.06/0.07	0.11/0.12	0.15/0.15	0.05/0.07

† Ligands that were not refined with a *PRODRG*-generated topology (*e.g.* metal ions or molecules with a full description in the *REFMAC*5 libraries) are not listed.

In the test case we obtain the most favourable results possible in terms of coordinate r.m.s.d., as the *HIC-Up* versions of both ligands used come from structure 1fsg and thus are identical to the 'ideal' structures.

3.3.2. *REFMAC*5/*LIBCHECK*. *REFMAC*5 comes with a library containing topologies and parameters for several common small molecules and topologies only ('minimal descriptions') for a large number of additional molecules (Murshudov *et al.*, 1997). Upon encountering a small molecule for which no or only a minimal description is available, *REFMAC*5 (using the associated program *LIBCHECK*; Murshudov *et al.*, 1997) will generate a complete description

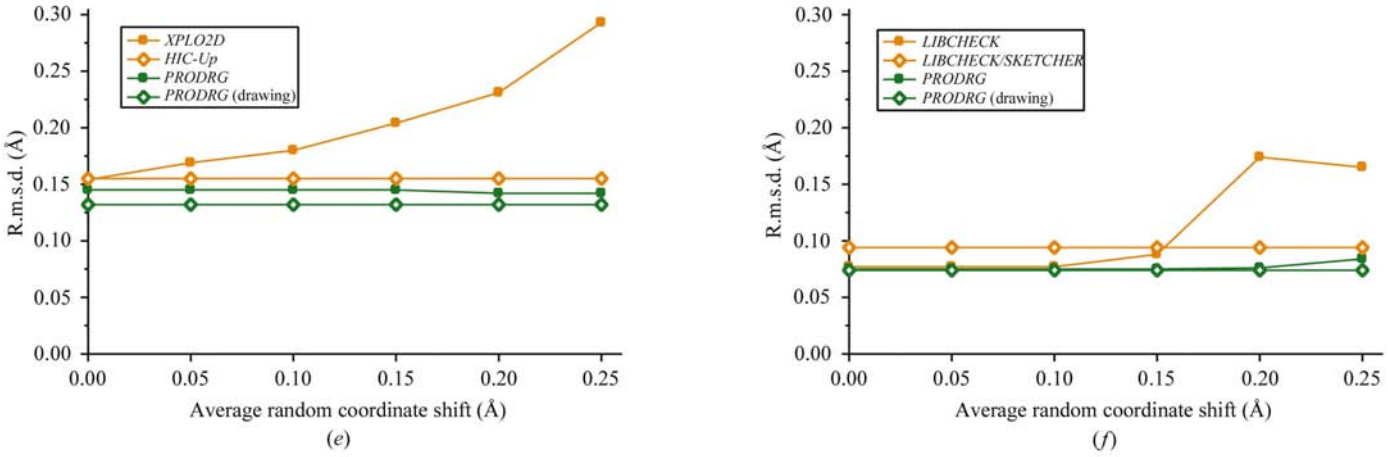

Figure 4 (continued)

which, after inspection by the user, can be used in refinement. In addition, it is possible to enter a compound description interactively by drawing it in *SKETCHER*, which, like *REFMAC*5, is part of the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994), and from it generate both a topology and three-dimensional coordinates.

Refinement with *REFMAC*5 using *PRODRG*- or *LIBCHECK*-generated topologies was compared in a similar fashion to that described for *CNS*/*XPLO2D*: the weight of the X-ray data was varied between 0.001 and 0.5. Results are shown in Table 3. As with *CNS*, the real-space R, $R_{\rm work}$ and in this case also $R_{\rm free}$ are similar for both topologies. The differences in r.m.s.d. resulting from using unperturbed or perturbed starting structures are slightly larger for both topologies than in the tests with *CNS*, but still small compared with the average coordinate perturbation applied.

The differences in performance between *PRODRG*- and *LIBCHECK*-produced topologies are small in four of the six test cases. In the remaining two cases [HNC (Gavuzzo *et al.*, 2000) and EAPA (Erskine *et al.*, 2003)] the refinement using *PRODRG* topologies gives significantly better ligand conformations, with r.m.s.d.$_{LIBCHECK}$/r.m.s.d.$_{PRODRG} \geq 1.5$. A closer look shows that in the case of HNC the large conformational difference introduced by refinement with the *LIBCHECK* topology is a consequence of an inappropriate planarity restraint covering the entire biphenyl moiety of the ligand, even though in the high-resolution structure the two phenyl rings are, as would be expected, at an angle of $\sim 22^{\circ}$ (Fig. 4*c*); in EAPA the geometry of a residue representing a cyclized Asp-Gly dipeptide is somewhat distorted by the *LIBCHECK*-generated topology owing to two atom-type misassignments: C2′ and C3 are incorrectly typed as sp^2-hybridized, which results in bond lengths that are too short (Fig. 4*d*). It should be pointed out that neither of the two poorly performing compounds exist in the *REFMAC*5-distributed library and thus *LIBCHECK* had to generate the topologies without the help of a minimal description.

The relative performance of *LIBCHECK* and *PRODRG* with lower-quality ligand coordinates was again assessed for the case of HGPRT; the results are shown in Fig. 4(*f*). As pointed out above, the *PRODRG*-generated topologies show some deterioration above a random coordinate shift of 0.15 Å, which can be avoided by instead defining the ligands through TXT drawings or other two-dimensional descriptions. *LIBCHECK* performs similarly to *PRODRG* in this case, even though its atom-type detection seems to be more sensitive to coordinate error. Like *PRODRG*, *LIBCHECK* allows the production of topologies in a truly coordinate-independent fashion by drawing them interactively in the S*KETCHER* program. While this obviates the need for high-quality ligand coordinates, the GUI-based procedure is relatively tedious and incompatible with high-throughput approaches.

3.4. Current limitations

The dependence of *PRODRG* on *GROMACS* (the *GROMOS*87 force field) leads to a number of limitations in the scope of compounds that *PRODRG* can handle. The most notable restriction is the comparatively small number of elements that the program supports. While the current selection (H, C, N, O, P, S, F, Cl, Br, I) allows processing of a wide range of biomolecules and potential drugs, further elements covering at least B, As, Se and common metal cations such as $Fe^{2+/3+}$ or Mg^{2+} would greatly extend this range. The other force-field-related problem is the limited number of atom types available for the supported elements, occasionally leading to a poor representation of phosphorus/sulfur chemistry and of *sp*-hybridized atoms. While many of these issues have been addressed in the current version of *PRODRG*, further improvements could be achieved with the addition and parametrization of more atom types.

Further limitations include the inability to detect certain aromatic systems such as pyrene, which possess $4n$ π-electrons. Also, *PRODRG* currently does not store information on bond types provided in Molfiles or text drawings: all computation is based solely on the hybridization state of individual atoms. Keeping bond-type data would be helpful in resolving certain ambiguities, *e.g.* in hydrogen placement.

3.5. Conclusions

PRODRG provides fast, automated and, within the given limitations, reliable access to small-molecule topologies and coordinates for use with high-throughput protein–ligand crystallography. Tests in crystallographic refinement show that *PRODRG*-generated topologies are generally of equal quality or better than topologies obtained by other means. *PRODRG* obviates the requirement for high-quality input coordinates or other additional data in generating topologies, as it can operate even on two-dimensional representations of a molecule, such as the industrial standard MDL Molfile/SDfile. It should also be noted that the variety of topologies generated by *PRODRG* allows the use of consistent descriptions of a given molecule in all steps of the inhibitor-design process, from crystallographic refinement and visualization through structure analysis to molecular-dynamics or docking studies.

Additional extensions of *PRODRG* with applications in automated ligand design and optimization are currently being developed, as well as *PRODRG*-based algorithms for automated identification and fitting of small molecules in electron-density maps. Development on the core *PRODRG* application aims to overcome the limitations in terms of atom types and force field. A particular focus is the implementation of a new coordinate-generating mechanism which will remove the dependency on *GROMACS* from *PRODRG*, speed up coordinate production and, most importantly, open a path towards the use of different/novel force fields. This in turn will then allow support for additional atom types, thus extending the applicability of *PRODRG*.

Financial support by a Wellcome Trust Senior Fellowship and an EMBO Young Investigator Fellowship (to DvA) is gratefully acknowledged. We would like to thank Charlie Bond for valuable discussions and critical reading

of the manuscript. For academic research purposes *PRODRG* is freely available as a WWW service at http://davapc1.bioch.dundee.ac.uk/prodrg/. Binaries for Linux, IRIX, FreeBSD or Windows are available upon request.

References

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

Aalten, D. M. F. van, Crielaard, W., Hellingwerf, K. & Joshua-Tor, L. (2002). *Acta Cryst.* D**58**, 585–590.

Aalten, D. M. F. van, Komander, D., Synstad, B., Gåseidnes, S., Peter, M. G. & Eijsink, V. G. H. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 8979–8984.

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. (1995). *Comput. Phys. Commun.* **91**, 43–56.

Blundell, T. L., Jhoti, H. & Abell, C. (2002). *Nature Rev. Drug. Discov.* **1**, 45–54.

Brooijmans, N. & Kuntz, I. D. (2003). *Annu. Rev. Biophys. Biomol. Struct.* **32**, 335–373.

Brünger, A. T. (1988). *J. Mol. Biol.* **203**, 803–816.

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.

Charnock, S. J., Bolam, D. N., Nurizzo, D., Szabo, L., McKie, V. A., Gilbert, H. J. & Davies, G. J. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 14077–14082.

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

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

Dong, C., Huang, F., Deng, H., Schaffrath, C., Spencer, J. B., O'Hagan, D. & Naismith, J. H. (2004). *Nature (London)*, **427**, 561–566.

Ekstrom, J. L., Pauly, T. A., Carty, M. D., Soeller, W. C., Culp, J., Danley, D. E., Hoover, D. J., Treadway, J. L., Gibbs, E. M., Flettrick, R. J., Day, Y. S. N., Myszka, D. G. & Rath, V. L. (2002). *Chem. Biol.* **9**, 915–924.

Erskine, P. T., Coates, L., Mall, S., Gill, R. S., Wood, S. P., Myles, D. A. & Cooper, J. B. (2003). *Protein Sci.* **12**, 1741–1749.

Ertl, P. & Jacob, O. (1997). *Theochem*, **419**, 113–120.

Evans, J. C., Huddler, D. P., Jiracek, J., Castro, C., Millian, N. S., Garrow, T. A. & Ludwig, M. L. (2002). *Structure*, **10**, 1159–1171.

Gadola, S. D., Zaccai, N. R., Harlos, K., Shepherd, D., Castro-Palomino, J. C., Ritter, G., Schmidt, R. R., Jones, E. Y. & Cerundolo, V. (2002). *Nature Immunol.* **3**, 721–726.

Gasteiger, J., Sadowski, J., Schuur, J., Selzer, P., Steinhauer, L. & Steinhauer, V. (1996). *J. Chem. Inf. Comput. Sci.* **36**, 1030–1037.

Gavuzzo, E., Pochetti, G., Mazza, F., Gallina, C., Gorini, B., D'Alessio, S., Pieper, M., Tschesche, H. & Tucker, P. A. (2000). *J. Med. Chem.* **43,** 3377–3385.

Gourley, D. G., Schüttelkopf, A. W., Leonard, G. A., Luba, J., Hardy, L. W., Beverley, S. M. & Hunter, W. N. (2001). *Nature Struct. Biol.* **8**, 521–525.

Haapalainen, A. M., van Aalten, D. M. F., Merilinen, G., Jalonen, J. E., Wierenga, R. K., Hiltunen, J. K. & Glumoff, T. (2001). *J. Mol. Biol.* **313**, 1127–1138.

Hall, D. R., Bond, C. S., Leonard, G. A., Watt, I., Berry, A. & Hunter, W. N. (2002). *J. Biol. Chem.* **277**, 22018–22024.

Heine, A., DeSantis, G., Luz, J. G., Mitchell, M., Wong, C.-H. & Wilson, I. A. (2001). *Science*, **294**, 369–374.

Heroux, A., White, E. L., Ross, L. J., Juzin, A. P. & Borhani, D. W. (2000). *Structure*, **8**, 1309–1318.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996). *Proteins*, **26**, 363–376.

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

Kleywegt, G. J. (1995). *Jnt CCP4/ESF–EACBM Newsl. Protein. Crystallogr.* **31**, 45–50.

Kleywegt, G. J., Henrick, K., Dodson, E. J. & van Aalten, D. M. F. (2003). *Structure*, **11**, 1051–1059.

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

Klon, A. E., Heroux, A., Ross, L. J., Pathak, V., Johnson, C. A., Piper, J. R. & Borhani, D. W. (2002). *J. Mol. Biol.* **320**, 677–693.

Lindahl, E., Hess, B. & van der Spoel, D. (2001). *J. Mol. Med.* **7**, 306–317.

Matern, U., Schleberger, C., Jelakovic, S., Weckesser, J. & Schultz, G. E. (2003). *Chem. Biol.* **10**, 997–1001.

Morris, G. M., Goodsell, D. S., Halliday, R., Huey, R., Hart, W. E., Belew, R. K. & Olson, A. J. (1998). *J. Comput. Chem.* **19**, 1639–1662.

Morris, G. M., Goodsell, D. S., Huey, R. & Olson, A. J. (1996). *J. Comput. Aided. Mol. Des.* **10**, 293–304.

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

Nicolet, Y., Lockridge, O., Masson, P., Fontecilla-Camps, J. C. & Nachon, F. (2003). *J. Biol. Chem.* **278**, 41141–41147.

Rao, F. V., Houston, D. R., Boot, R. G., Aerts, J. M. F. G., Sakuda, S. & van Aalten, D. M. F. (2003). *J. Biol. Chem.* **278**, 20110–20116.

Richards, W. G. (2002). *Nature Rev. Drug Discov.* **1**, 551–555.

Ritchie, D. W. & Kemp, G. J. L. (2000). *Proteins*, **39**, 178–194.

Sadowski, J., Gasteiger, J. & Klebe, G. (1994). *J. Chem. Inf. Comput. Sci.* **34**, 1000–1008.

Schüttelkopf, A. (2003). PhD thesis, University of Dundee, Scotland.

Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enyzmol.* **277**, 319–343.

Vriend, G. (1990). *J. Mol. Graph.* **8**, 52–56.

Zavala-Ruiz, Z., Sundberg, E. J., Stone, J. D., DeOliveira, D. B., Chan, I. C., Svendsen, J., Mariuzza, R. A. & Stern, L. J. (2003). *J. Biol. Chem.* **278**, 44904–44912.