Acta Cryst. (1998). D**54**, 1109–1118

The Role of Validation in Macromolecular Crystallography

Eleanor Dodson

Department of Chemistry, University of York, Heslington, York YO1 5DD, England. E-mail: ccp4@yorvic.york.ac.uk

(*Received 20 March 1998; accepted 18 May 1998*)

Abstract

The importance of validation techniques in X-ray structure determination and their relation to refinement procedures are discussed, with particular reference to atomic resolution structures. The requirements of deposition and publication, and the role of validation tools in this are analysed. The need for a rigorously defined file format is emphasized.

1. Introduction

Achievements in molecular-biology techniques and progress in macromolecular X-ray crystallography have led to an explosion in the number of experimentally determined three-dimensional (3-D) structures. There are already 7000 deposited coordinate sets in the Brookhaven Protein Data Bank (PDB) (Bernstein *et al.*, 1977; Abola *et al.*, 1987) and the Brookhaven team foresee the number of deposited structures (mostly proteins) reaching 15 000–25 000 in the year 2000, implying a growth rate of about 100–150 structures a month. Efficient recording, standardization and validation of these data have become major problems of modern structural molecular biology. Failing to address them would have serious consequences for progress in biological science, particularly in relation to the interpretation of data generated from the genome-sequencing projects.

Although all coordinate sets deposited in the PDB are presented with an apparent precision of 1/1000th of an angstrom, the true accuracy of the models and the level of detail they contain is a function of the quantity and quality of the experimental data on which they are based, and indeed the care and protocols used during their refinement. Historically, although some data on the fit of the model to the experimental measurements were often listed in the deposition, they could not in general be independently cross-checked since the experimental data were, in most cases, not deposited. In recent years, the *R* and free *R* factor (Brünger, 1992*a*) were usually given, but there was no guarantee that these were the values for the whole data set, or those from a set restricted by resolution or intensity range. Different refinement programs and protocols determined these rather differently, so that the absolute significance of the values reported were difficult to rationalize. In addition there was no requirement to give any assessment of the quality or completeness of the experimental information.

A number of excellent validation tools have been developed during the last decade for checking the geometry and stereochemistry of a set of coordinates against a target library and the expected parameters derived therefrom. These tools have been widely accepted by the community and are generally applied before submission of data to the PDB. However, these tools do not attempt to check the coordinates against the experimental data.

2. The validation network

To help address these problems an EU network titled *Integrated Procedures for Recording and Validating Results of Three-Dimensional Structural Studies of Biological Macromolecules* was established in 1993. Its membership included providers of high-resolution structures, writers of refinement programs, and groups working on assessment of geometric and stereochemical properties. A second contract was awarded in 1996, with a wider membership, listed in the acknowledgments.

The major target was to develop accepted procedures for validating 3-D structures. This required defining and testing criteria against which the quality and precision of 3-D structural models could be assessed. These criteria were to be based on statistical analyses of structural information already available as well as on evaluating agreement with experimental measures (*R* factors, how well the models explained electron density and observed structure factors, NMR-derived distance constraints and coupling constants). The group was also required to present the procedures and criteria for 3-D structure validation to the relevant scientific community for discussion.

I report here some of the achievements and findings of the members of the network, which are largely encapsulated in a paper published in the Journal of Molecular Biology, entitled *Who checks the Checkers? Four validation tools applied to eight atomic resolution structures* (EU 3-D Validation Network, Wilson *et al.*, 1998) and on the discussions on the whole question of validation which took place at the Porto Satellite

ISSN 0907-4449

Table 1. *Summary of the eight atomic resolution structures*

The abbreviations used here and in the text are Cytc6, cytochrome *c*6, PDB code 1CTJ (Frazão *et al.*, 1995); Cutinase, cutinase, PDB code 1CEX (Longhi *et al.*, 1997); Lysozyme, triclinic lysozyme, PDB code 3LZT (Walsh *et al.*, 1997); ProtG, fragment of protein G (Butterworth *et al.*, 1999); RNaseSa, ribonuclease Sa, PDB code 1RGG (Sevcik *et al.*, 1996); Ropm, a mutant of the repressor of primer protein (Vlassi *et al.*, 1998); RubrDv, rubredoxin from *Desulfovibrio vulgaris* (Butterworth, 1996) and RubrCp, rubredoxin from *Clostridium pasteurianum*, PDB code 1IRO (Dauter *et al.*, 1996). The data were recorded using synchroton radiation at EMBL Hamburg. The structures were refined using *SHELXL*93 or *SHELXL*96.

	Cytc6	Cutinase	Lysozyme	ProtG	RNaseSa	Ropm	RubrDv	RubrCp
PDB cooordinates	1CTJ	1CEX	3LZT	2IGD	1RGG	1NKD	XXX	1IRO
Temperature	RT	RT	110 K	RT	RT	RT	RT	RT
Space group	$R3$	$P2_1$	$P1$	$P2_12_12_1$	$P2_12_12_1$	$C2$	$P2_1$	$R3$
Packing density, V_M (Å^3 Da^{-1})	2.3	2.0	1.7	2.1	2.4	2.0	1.6	2.1
Resolution (Å)	25–1.2	15–1.0	20–0.925	10–1.1	10–1.2	23.1–1.1	20–0.92	10–1.1
Completeness (%)	99.9	93.3	90.1	98.6	95.3	98.2	98.5	94.0
$I/\sigma(I)$	25.6	16.5	29.1	39.7	8.6	18.5	9.6	23
$I/\sigma(I)$ outer shell	1.5	2.2	4.9	12.3	4.1	6.2	4.8	3.2
Solvent content (%)	47	43	36	46	48	35†	29	43
α-helix (%)	58	39	33	26	11	92	0	0
β-sheet (%)	0	19	15	43	29	0	18	23

† The percentage of solvent residues estimated for Ropm allows for the six C-terminal residues which are disordered, *i.e.* these are *not* included as disordered solvent.

Table 2. *A summary of properties checked using validation tools PROCHECK, PROVE, SQUID and WHATCHECK*

Target library for all programs	Engh & Huber
Reporting of main- and side-chain features	All programs
Nomenclature, structure summary and format checks	All programs except PV
Geometry checks, bond lengths, angles, planarity and chirality	All programs except PV
Conformation checks, torsion angles, non-bonded contacts	All programs except PV
Separate analysis of Gly and Pro	All programs except PV
Hydrogen-bond statistical analysis	All programs except PV
Suggested HNQ (His, Asn, Gln) flips	*SQUID* and *WHATCHECK*
Solvent distribution and listing of 'floating waters'	*SQUID* and *WHATCHECK*
ADP analysis including anisotropy	*SQUID*
Volumes and packing density	*PROVE* and *SQUID*
Global parameters	*PROCHECK* and *WHATCHECK*

meeting of the European Crystallographic Meeting held in August 1997.

Two general questions were addressed. (*a*) Do the atomic resolution structures imply changes in 'expected' stereochemical properties and are the target values used for restraints in the validation programs and the refinement protocol appropriate? (*b*) Can errors in models be detected and how reliable are the coordinates after refinement?

For the study, eight protein crystal structures, which have been refined against X-ray diffraction data extending to atomic resolution, 1.2 Å or better, were investigated using four different validation tools, *PROCHECK* (Laskowski, MacArthur *et al.*, 1993), *PROVE* (Pontius *et al.*, 1996), *SQUID* (Oldfield, 1992) and *WHATCHECK* (Vriend, 1990). All had been refined using *SHELX*93 or *SHELX*97 (Sheldrick & Schneider, 1997). Some restraints were imposed during the course of these refinements, but because there was a high data-to-parameter ratio, the experimental data could 'override' the imposed restraints, at least in well ordered parts of the molecules. The details of the structures and references are given in Table 1.

The validation tools all addressed properties residing in the coordinates alone. *PROCHECK*, *WHATCHECK* and *SQUID* examined such properties as bond lengths and angles, the Ramachandran plot (Ramachandran *et al.*, 1963) of peptide torsion angles, peptide planarity, chirality and χ angles. *PROVE* calculated and analysed the atomic volumes of atoms in the core of the structures. The coordinate indicators were used both to try to pinpoint local errors, and to attempt to 'score' each structure by its compliance with pre-established norms. A summary of properties checked is given in Table 2.

Electron-density maps for two of the structures (RNase and Rubr) were re-examined in great detail in the light of the validation reports. This gave us a 'feel' for what sort of deviations were real, where there was not enough evidence to decide whether there was a deviation from the expected value, and where the checks had pinpointed errors in interpretation.

Target values for stereochemistry and geometry were taken either from the Cambridge Structural Database (CSD) (Allen *et al.*, 1979), or from analysis of deposited, apparently reliable, structures in the PDB. The 3DB structures were selected on the basis of *R* factor and resolution. (For examples of selection criteria used see Morris *et al.*, 1992; Oldfield, http://www.yorvic.york.ac.uk/~oldfield/pdbsel). These are the same properties with the same target values which the refinement packages utilize to restrain the model.

Preliminary analysis by members of the network led to modifications both to the validation programs and to the refinement protocols. Initially there was a significant amount of ignorance of the different problems facing the three groups and one of the principal achievements of the network was to reduce this. The providers of structures were not sufficiently aware of the need for consistent formatting, and attention to detail, for example in naming unusual atoms, or space groups; the validators were not sufficiently aware of the extent of the use of geometric restraints during the refinement procedures and the effect of different protocols on the results. In addition even in a high-resolution structure some sections are less well determined than others, and many residues take up several conformations. Finally the software did not completely handle the consequences of crystal symmetry.

3. How had the target values been obtained

It became clear early on in the analysis that it was necessary to sub-divide the criteria used into two main classes which we referred to as 'geometric' and 'stereochemical' requirements.

The 'geometric' (or hard) properties are those such as bond lengths and angles, chirality and planarity of aromatic rings. These are 'unimodal', *i.e.* they can only take one value, assuming the chemical properties of the macromolecule are known, and the target values and their estimated standard uncertainty (s.u.) can reasonably be expected to be the same as for any organic molecule. The CSD contains over 100 000 such crystal structures, refined to a high degree of accuracy, so these values are extremely reliable. In fact all the programs use values based on the work of Engh and Huber (Engh & Huber, 1991) who analysed the small-molecule structural database in 1991. They assigned 17 atom types; such as C = carbonyl C, N = peptide N, NR = unprotonated N in histidine, OC = carbonyl O, OH1 = hydroxyl O, *etc.* and derived values for each of these classes. Their results were first included in the *X-PLOR* dictionary, but John Priestle (Priestle, 1994) converted this to formats suitable for use by most other refinement programs. Using these more realistic values improved the behaviour of refinement in general, and the only problems arise when the chemical properties of the macromolecule under study are assigned inappropriately. Examples are in the protonation state of histidine or carboxylate side chains, or in phosphate groups in DNA where the P—O bond length can depend on Ph.

The 'stereochemical' (or soft) checks include torsion angles, both φ and ψ used for the Ramachandran plots, the ω distribution which governs the peptide planarity, side-chain χ distributions, volumes, and other properties where there is no single solution. For these, distortions can be generated by the molecular environment, and the targets and standard uncertainties are not so easily obtained. The stereochemical targets are derived from a selection of already deposited macromolecular structures, but it is not clear that these can yield sufficiently reliable information. Almost all target values from macromolecules must be biased by the restraints applied

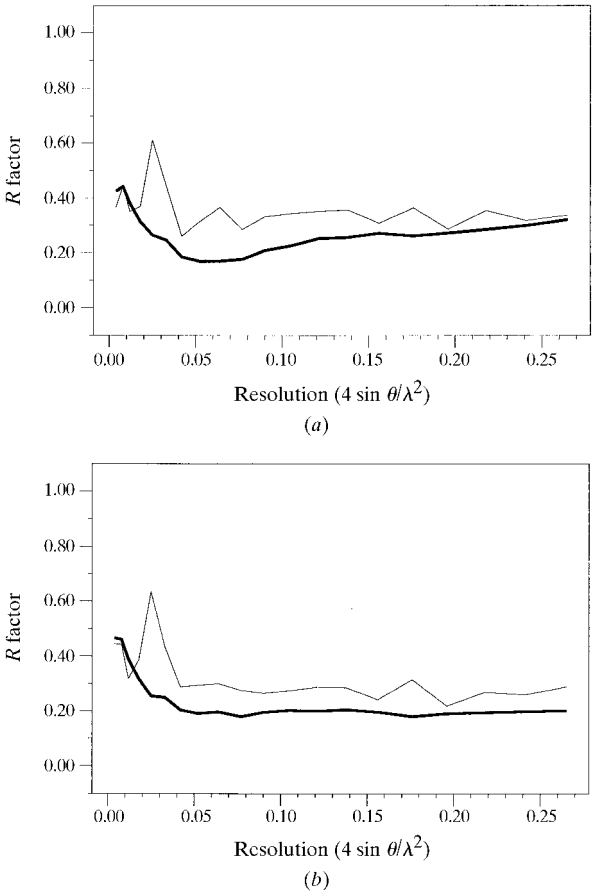

Fig. 1. The behaviour of the *R* and free *R* factors *versus* resolution (*a*) before and (*b*) after maximum-likelihood refinement with *REFMAC*. Bold lines show the *R* factor and the thin lines the free *R* factor. This structure had been refined to an *R* factor of 24% and a free *R* of 34% using unweighted least squares. *REFMAC* reduced the *R* factor by 4% and the free *R* by 6%. The plot shows that the greatest improvement was for the high-resolution range. The reason for this is that the fit for the high-resolution structure factors depend on accurate parameters for the atoms, but any improvement at low resolution will require large scale movements, or a more complete description of the model, *e.g.* more solvent atoms.

Table 3. *Refinement for the atomic resolution structures*

	Cytc6	Cutinase	Lysozyme	ProtG	RNaseSa	Ropm	RubrDv	RubrCp
Program	*SHELXL*96	*SHELXL*93	*SHELXL*96	*SHELXL*93	*SHELXL*93	*SHELXL*93	*SHELXL*93	*SHELXL*93
Residues	89	197 (213)	129	61	2*96	59	52	53 (54)
Ligands	Haem	—	3AcOH, $5NO_3$	—	—	—	FeS_4, SO_4	FeS_4
Observations/ parameters	4.0	5.6	4.5	4.4	3.5	3.8	5.6	7.0†
R factor (%)	14.0	9.4	9.5	9.4	10.6	10.1	7.9	9.0
R free (%)‡	18.8	11.9	11.3	12.5	n/a	12.3	11.0	n/a
R.m.s. deviations								
Protein bond lengths (Å)	0.013	0.023	0.017	0.021	0.024	0.058	0.019	0.029
Angles (°)	2.55	2.32	2.70	1.96	2.17	3.94	1.90	2.60
ω angles and s.u.s	180.0 (5.5)	179.5 (5.6)	179.5 (5.0)	178.3 (7.3)	178.8 (6.7)	178.2 (3.6)	179.6 (5.8)	178.1 (8.5)
Core Ramachandran (%)	85	94	91	94	92	97	90	92
χ_1 s.u.	9.2	10.7	8.6	7.2	10.2	8.5	7.3	11.6
χ_2 *trans* s.u.	11.3	9.5	8.6	17.4	13.3	10.8	8.3	8.8

† RubrCp was refined with the Friedel pairs treated independently, giving almost two times the number of observations. ‡ These free *R* factors do not correspond to the final models discussed in the text as those were refined using all measured data. For Ropm, only 59 of the 65 residues are included in the final model.

Table 4. *Comparison of the expected values for stereochemical parameters as detemined by Morris et al. (1992) with the actual values observed in the eight atomic resolution structures*

Evaluated using the Kabsch & Sander (1983) method.

	Original parameters			Atomic resolution structures		
Stereochemical parameter	Mean	s.u.	N_{obs}	Mean	s.u.	N_{obs}
χ_1 dihedral angle (°)						
gauche (−)	64.1	15.7	3240	66.1	8.0	90
trans	183.6	16.8	6015	183.2	9.9	192
gauche (+)	−66.7	15.0	9635	−65.1	9.6	346
χ_2 dihedral angle	177.4	18.5	5476	175.5	11.1	176
Proline (φ) torsion angle	−65.4	11.2	1038	−61.3	7.5	37
Helix (φ) torsion angle	−65.3	11.9	6675	−66.2	13.0	245
Helix (ψ) torsion angle	−39.4	11.3	6675	−38.8	9.8	245
ω dihedral angle (°)	179.6	4.7	23895	179.0	5.6	812
CA chirality: ζ 'virtual' torsion angle (CA−N−C−CB)	33.9	3.5	21950	33.8	2.42	752
% (φ, ψ) in most favoured regions of Ramachandran plot			> 90			92.1

during refinement and there is a great danger of circularity – accepted knowledge is applied, and lo and behold, all new structures conform with and hence reinforce this.

4. Refinement procedures

All the commonly used programs, *X-PLOR*/*CNS* (Brünger, 1992*b*; Brünger *et al.*, 1998), *TNT* (Tronrud, 1992), *PROTIN*/*PROLSQ* (Hendrickson & Konnert, 1980), *NUCLSQ* (Westhof *et al.*, 1985), *REFMAC* (Murshudov *et al.*, 1997), and *SHELX* use expected macromolecular properties as well as the fit to the experimental data to steer the refinement. It is sometimes not completely transparent to the user how this has been implemented, but several authors have demonstrated that it is possible to identify the dictionary and program used from the structural properties (Laskowski, MacArthur *et al.*, 1993; Parthasarathy & Murthy, 1999). Not surprisingly there are differences in detail between the models obtained from the different programs (Cruickshank, 1996; Daopin *et al.*, 1992). In addition the relative weighting of the experimental and prior information can usually be controlled by the user, and indeed in some programs it is easy to redefine the weighting of the different restraints. On the whole this is very undesirable; the statistical analysis which provided the target values will have provided a standard uncertainty, and these should be used together. To date, the PDB deposition has not required that details of the prior information used be provided.

Although this paper is principally addressing the problems of validation, it is nonsense to discuss this without some reference to the actual reliability of the coordinate parameters. This was discussed at the Daresbury Workshop *Macromolecular Refinement* in

1996, and the paper by Durward Cruickshank (Cruickshank, 1996) highlighted many of the problems. Refinement programs often attempt to give an overall estimate of expected coordinate error. Many of these are based on the Luzzati plot (Luzzati, 1952), which Cruickshank reminded the audience is a measure of convergence, NOT of accuracy. It is interesting to note that when maximum-likelihood weighting is used the Luzzati plot often becomes flat, showing that convergence of the current model has been achieved, and that any further improvement requires rebuilding and extension of the model (Fig. 1). Cruickshank suggested a precision indicator based on the amount of experimental data available, its completeness, and the free *R* factor. This is a better indicator, but it is still a global one. Theoretical statistical estimation methods incorporate routines for providing estimates of both the parameters and their reliability. These involve the calculation and inversion of a second derivative matrix, costly both in time and computer memory, but as the technology develops, this is becoming feasible. However, there is still a problem in deciding how to deal with the prior information incorporated into the equations which should not be treated as completely independent. There are interesting developments under way (Murshudov & Dodson, 1997; Tickle *et al.*, 1998), which will make obsolete much of this discussion! Atomic resolution structures can 'defeat' the restraints to some extent, and are probably the only way we can update our knowledge. There are at present only 13 structures with resolution better than 1.2 Å available in the PDB, but as this number increases it will be a most valuable new resource. It is still essential though for users to be aware of the properties of even the best ordered crystal *viz.* that there is almost always a large percentage of the surface in contact with solvent, and that residues on this surface often adopt several conformations or may not even be visible in the electron density. Indeed, if the temperature factor, more correctly called the atomic dispersion factor (ADP), of an atom is much higher than the mean for the core of the structure, Cruickshank showed that it will not contribute significantly to the atomic resolution data range.

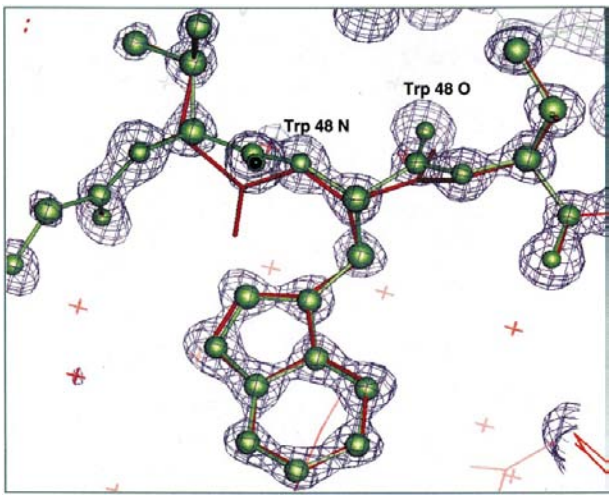

Fig. 2. The $3F_o - 2F_c$ electron density for the region around Trp48 in ProtG with the model superimposed. Additionally marked are the carbonyl O atoms in the idealized positions where strict planarity of the peptides, *i.e.* an ω angle of 180°, have been artificially imposed. Those positions clearly lie out of the centre of the density, by about 0.3 Å.

5. Is it possible to use these criteria for error detection?

The precision of the 'geometric' properties is clearly a direct result of the refinement protocol used, and as such is not much help in detecting errors. All the structures examined had a wider spread of bond lengths and angles than many of lower resolution examples deposited, but this does not indicate that they are less accurate (Table 3). *SHELX* imposes a weighting on distance targets which reflects the s.u.s obtained from the analysis of the CSD, *viz.* about 0.02 Å on bonds.

'Stereochemical' checks are better indicators, but can also be influenced in unpredictable ways by the restraints applied. None of the eight structures contained any gross errors [which in fact can often be detected from the Ramachandran plot alone (Kleywegt & Jones, 1995, 1996)]. Table 4 shows that the distribution of all torsion angles clustered more tightly for these structures, and indeed the χ distribution seemed the best indicator of quality. For the atomic resolution structures these were more clustered around the *gauche*−, *gauche*+ and *trans* positions than in previous analyses of deposited structures. However, Janet Thornton showed at a discussion meeting in Brussels in 1997 that the mean values for different residue types vary around these values. It is not clear whether this is a real phenomenon or a result of conflicts between different types of restraints. For instance serine is often involved in

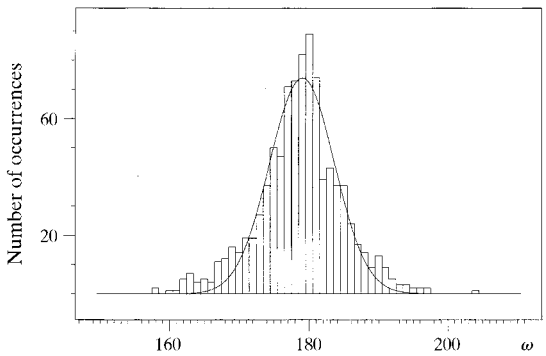

Fig. 3. The histogram of ω-angle distribution for the eight atomic resolution structures. It is clear that the Gaussian distribution does not fully describe the distribution, and the deviant values at the tails can usually be explained by external forces, as in the case illustrated in Fig. 2.

Table 5. *A selection of the residue-by-residue comments*

(*a*) Residue-by-residue comments on RubrCp

Anisotropy was noted if the principal axes of adjacent atoms were very different. The hydrogen-bond flag indicates a 'missing' hydrogen bond. Volumes were flagged if they deviated by more than 2.4σ from the mean.

Residue	Comment
Lys2	Main-chain density OK, CE small volume, side chain weak density. Also ω = 166.2°
Thr5	Two conformations. ADP for CG2A is high
Thr7	No hydrogen bond for O atom
Pro15	Two conformations. CD CG abnormal. A problem in naming each conformer
Pro20	Underpuckered. Good density however
Asp29	CA—CB—CG angle 120°, good density. Minor second conformation possible
Lys31	CE, NZ have high ADP's
Asp36	3.2 Å hydrogen bond, outside limits, no hydrogen bond for N
Pro40	Underpuckered but with good density
Leu41	N—H—SG hydrogen bond longer than target. C—O bond has smeared density
Glu50	Density OK. Very tight turn. Also ω = 167°
Glu51	All programs find the planarity wrong for CG—OE1—OE2. High ADPs, maybe anisotropic model inappropriate? Restraints should have prevented this. OE1 3.4 Å from H_2O, probably needs moving. C small volume. In fact all of residues 51–53 have relatively poor density

(*b*) Comments on RNaseSa (two chains, *A* and *B*)

Residue	Comment
1-3*A*, 1-3*B*	Surface residues. Several unusual volumes, *e.g.* 2*B* CA, but residues very poorly defined
Pro13*A*, 13*B*	Two conformations for CG only. Density OK
Pro27*A*, 27*B*	Underpuckered. Possibly two conformations? CD and CG have unusually high ADPs and poor density
Arg40*A*, 40*B*	For 40B, poor planarity of NE, CZ, NH1 NH2. High ADPs 59, 61, 67, 65. Surface residues with very poor density at the end of the side chain
Glu54*A*, 54*B*	Both have two conformations. 54*A* C small volume, but density seems OK
Arg63*A*, 63*B*	63*B* CD—NE short, twisted and non-planar. All ADPs are low and have very good density. 63A makes a strong salt bridge to the sulfate
Arg65*A*, 65*B*	65A N large volume. Near to double conformation of Glu54*A*. Smeared density suggesting possible multiple conformation
Ile71*A*, 71*B*	Odd angles CB—CG1—CD1: 128°, CB—CG1—CD1 129°. ADPs low
Gln77*A*, 77*B*	OE1 ADP 41, NE2 24. 2 conformations? Tight twist, but good density
Tyr81*A*, 81*B*	81*B* CE1 large volume, near to 57*B*. Density very good
His85*A*, 85*B*	85*A* and 85*B* CA large volume. Multiple conformations
Tyr86*B*, 86*B*	86*A* OH small volume. Some χ abnormality, tightly packed, good density.

hydrogen-bond networks, and if the hydrogen-bond distance were too tightly restrained this would influence the serine χ values.

Getting the refinement protocols right is a separate issue, but they will undoubtedly influence the properties of any given structure. It is an obvious problem that as an indicator is chosen to assess quality, it will become a target for minimization during refinement, and thus lose its usefulness as a check. George Sheldrick recommends excluding any multi-modal phenomena from the restraint list used for refinement and reserving them for validation analyses. This is partly because of the complexities of minimizing a property against an uncertain target. (If the χ value is 120° should it be pushed towards the *trans* or the *gauche*+ orientation?) However, when the experimental data is limited, it may be essential to invoke all available prior knowledge. The torsional angles refinement developed for *X-PLOR* (Rice & Brünger, 1994) has been used successfully with limited data, where it would not have been possible to parameterize the structure any other way. Gerard Kleywegt reports in his contribution that when the

model is more or less correct the method works well, but that when it is very poor, the shifts are cosmetic rather than genuine.

However, there are also deviations imposed by the protein fold and the crystal environment; properties that Gerard Kleywegt refers to as 'interesting features or errors'; and it is important for validation tools to indicate these but not to pre-judge that they are due to error. An interesting example was seen in the distribution of ω angles. (Figs. 2 and 3). These cluster near to 178 or 182° depending on the 'twist' of the peptide unit, but can be severely distorted by model contacts. (MacArthur & Thornton, 1996; Butterworth, 1996; Deacon *et al.*, 1997) Further examples have come to light in other high-resolution refinements (*e.g.* Guogong Lu, CCP4 Bulletin Board discussion). The histogram of ω angles illustrated in Fig. 3 plots the distribution for the available atomic resolution structures.

Any unusual feature in an X-ray structure should be checked against the maps derived from the experimental observations, and in fact the validation criteria are best used during the building of a structure. Tools such as *Oops* (Kleywegt & Jones, 1996), or the X-build validation in *QUANTA* (MSI, 1997) which run standard tests and list residues with unexpected features, make it easy to check maps at suspect points. This means that even large structures can be checked quickly for obvious mistakes. RNase and Rubr were checked residue by residue against electron density and some of the observations are given in Table 5. Some errors and oversights were highlighted. For example, some of the Asn, Gln and His residues were inverted to generate better hydrogen bonding (half of those where the validation queried the orientation). The terminal O atom of the Rubr chain had been labelled incorrectly, and showed up with an incorrect volume. The *WHATCHECK* target values for proline pucker at that time were shown to be rather unlikely.

When torsion angles were queried, in some cases the electron density clearly showed the orientation was correct. In other more mobile regions where the temperature factors were high, the electron density was less clear. It is debatable what to do here, but the crystallographers felt that it is unjustifiable to move parameters away from their refined minima without some indication from the experimental observations.

6. How can this expertise be used by the wider community?

During the EU and ECM sponsored workshop in Porto there was extensive discussion on the role of validation in structure determination and publication. Quite properly most journals now require that *both* experimental data and coordinates are deposited in the PDB before publication, but it is permissible to delay release to the community for a time. The PDB has mounted validation tests, including *WHATCHECK*, and runs them routinely to generate lists of 'errors' for depositors to verify. It became clear in Porto that the crystallographic community still feels some hostility towards this, and that there needs to be a clearer understanding of the role of validation in PDB deposition. It is important to understand why people feel threatened by the 'validation' checks.

Up until now the journals and their nominated referees have assumed full responsibility for judging the correctness and quality of any structure determination. However, there is now a perceived need for the PDB to take a more proactive role. Historically it has seen its responsibility to maintain an archive of structures, with the associated experimental data if available, and has eschewed any requirement that the PDB itself should act as a *de facto* referee.

Some reasons for the change of attitude are listed below.

It is accepted that referees cannot completely vet the structures without details of the model, but that authors are reluctant to release their coordinates before publication. This is perhaps regrettable, but I believe inevitable. The field has become much more competitive, and the referee may well be that competitor. Another factor plays a role in many first publications; solving a macromolecular structure is usually part of a considerable investment of effort by a team of scientists, ranging from molecular biologists to physicists, and these scientists may wish to digest the detailed insights following from the structure themselves before the coordinates are widely distributed. Similar factors come into play when the investigation has been performed in collaboration with an industrial group who may be willing to publish some parts of the work, but not to release the coordinates before they have patented resulting ideas.

However, it was agreed that nobody benefits if there are structures with substantial (and correctable!) errors in the PDB. Such incorrect structures can and should be trapped at this stage. They are often indicated by gross deviants in the Ramachandran plot, by serious clashes generated by crystal symmetry, or by wild deviations in expected temperature factors where there should be some correlation between low values and the atom position relative to the core of the structure. It was felt that at least this level of checking should be carried out by the PDB, and the depositors requested to comment on anomalies.

This was felt to be appropriate for the following reasons.

(*a*) There is a better understanding of validation. No one questioned the basic tenet that macromolecules have fairly predictable conformations and that violation of the established probable conformations can indicate potential local and global errors. (Everyone agreed that

they use such stereochemical checking routinely during structure building and refinement.)

(*b*) It has been realised that the structural information underpins other branches of science, and that these users of the data bank need some guarantees as to the quality of the entries.

However, there was no consensus on how much of this information should be available to a referee. Peter Lindley pointed out that since referees cannot carry out a full validation themselves, perhaps they should be sent *a précis* from the PDB *via* the journal? This would require that the preliminary deposition occurred when the paper was submitted, not as now after a paper is accepted. A better and more practical alternative might be for authors to supply validation output for the referees at the time of submission, even if it is not to be included in the paper. This is already performed for small molecules published in *Acta Crystallographica Section C*.

The crystallographers emphasized two key points in evaluating structures that seem to be overlooked in the reports currently returned from the PDB.

(1) Structures are solved, published and deposited for different reasons, and it is not always possible or desirable to refine all of them with exquisite care. Structures and data will often be deposited when not fully refined. The important question for referees and users of such structures to be able to answer is: 'Do the data presented justify the conclusions drawn?'

(2) The quality of any crystal structure is finally limited by the quality and quantity of the data. It is governed by the number of observations per parameter, which is related to the resolution, solvent content, and the existence and nature of non-crystallographic symmetry. The *R* and free *R* factor, plus various precision indicators can help estimate quality, but they are almost all global indicators, variable from structure to structure and unable to distinguish between well ordered and poorly ordered atoms. If all the experimental data and the current coordinates are deposited, any determined user would be able to check the quality of the structure themselves against the generated map. This is the best possible validation tool. Depositors need to supply *all* observed data and associated s.u.s plus statistics on quality of this data. It should be easy for the depositor to update the first model, but even if this is not performed, at least preliminary results are available.

7. Suggestions for the future

Should the PDB support a separate 'deposition quality' index? *i.e.* an associated file with the validation comments?

(1) Several people said 'yes', provided depositors have a chance to address the 'problems'. The first step is, as now, to return a list of 'aberrant' features to the depositor. The primary responsibility for deposition is with the authors. Correction should be made as simple as possible to encourage the depositor!

(2) What should such an index hold? Some suggestions were as follows.

(*a*) Information on experimental data quality, *R* factors, free *R* factors, number of reflections, and completeness *in resolution shells.*

(*b*) Residue-by-residue quality indicator. Flags re Ramachandran plot quality. χ-angle analysis. R.m.s. bonds/angles.

(*c*) These 'deposition quality' indexes would be updated as other indicators are agreed. Much of the information could be extracted automatically by data harvesting tools such as Kim Henrick is developing.

(*d*) Some crystallographers were not very happy about this, but would probably be persuaded that it is useful, especially if there was less use of the word error, and more of some such synonym as 'standard uncertainty'.

The PDB is used by many people who want to extract all sorts of information and such a 'deposition quality' index would provide a better criteria for choosing structures than the current reliance on *R* factor and resolution.

8. More robust formats

It has become clear that a tighter syntax must be used for depositing the structures. Database creators require this, and clear definitions of the entries are needed. Much time has been wasted in the past deconvoluting cryptic 'remarks' in PDB entries. These were quite adequate when each structure was analysed as a unit, but not for procedures which scan the whole range of depositions. mmCIF is designed to address this.

For example

```
save__refine.ls_percent_reflns_R_free
_item_description.description
;       The number of reflections that satisfy the resolution
        limits established by _refine.ls_d_res_high and
        _refine.ls_d_res_low and the observation limit
        established by _reflns.observed_criterion, and that
        were used as the test (i.e., excluded from refinement)
        set when refinement included calculation of a 'free'
        R factor, expressed as a percentage of the number of
        geometrically observable reflections that satisfy the
        resolution limits.
;
```

The syntax sometimes looks clumsy, but it is computer readable, and allows a large number of structures to be parsed automatically.

9. A user-friendly interface

To make the information useful to the non-expert user, a well designed interface to steer the user through likely

queries is needed. Many of these are discussed in this issue.

10. Conclusions

This paper can only touch on the problems of validation, deposition and publication, which are now being widely discussed. It is clear that validation cannot be a static procedure, but must always be finding new criteria against which to evaluate structures. In the first set of validation tools properties such as the r.m.s. deviation in bond lengths were flagged, but it is now realised that this simply reflects the refinement protocol used. At present it is more informative to check properties such as torsion angles, or the Ramachandran plot which are not now used during refinement; but before checking expected values must be assigned, which may well then be incorporated as targets into the next generation of refinement programs.

Publication will probably soon require the deposition of both coordinates and experimental data. The validation procedures are designed to test whether the refinement is complete, but it is difficult to define such 'completeness', or even whether it is always an appropriate target. The PDB will be enriched by containing a greater percentage of the solved structures, but it is inevitable that many of these will never be completely refined since the results are not needed by the scientist. So it is necessary for referees and the PDB to decide what is acceptable for a deposition, and the consensus at the Porto meeting was that this should be 'sufficient to justify the conclusions drawn in the publication'. Providing the data is available expert users could re-refine such a structure to answer their own questions at some later date. However, for non-crystallographic usage, it will be important to carefully flag the PDB entries with the currently agreed set of criteria which can then be updated in the future. There is a vigorous ongoing debate on these questions, and this paper cannot hope to predict future policy. The ultimate goal must be to associate an s.u. with each parameter based on proper statistical analysis. In the short term, Albert Podjarny summed up the fundamental need: 'we need to break the cycle of distrust between depositors and validators'. I feel that meetings like this help to do so.

Tom Oldfield, Garib Murshudov and Keith Wilson have made many contributions to the discussion and the text. Much of the work discussed results from an EU network of laboratories supported by the EC Framework III BIOTECHNOLOGY program, Contract BIO2-CT92-0524 entitled *Integrated Procedures for Recording and Validating Results of Three-Dimensional Structural Studies of Biological Macromolecules*. The original 3-D validation group has been continued under EC CT96-0189, 'CRITQUAL' with the following Partners: K. S. Wilson (University of York), G. Vriend (EMBL, Heidelberg), J. Thornton (University College, London), V. S. Lamzin (EMBL Hamburg), R. Kaptein (University of Utrecht), S. Wodak (Free University of Brussels), T. A. Jones (University Uppsala). Christian Cambilleau (CNRS, Marseilles), Metaxia Vlassi (IMBB, Heraklion) and Jozef Sevcik (IMB, Bratislava) are thanked for providing models and data prior to publication or release from the PDB.

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *Crystallographic Databases – Information Content, Software Systems, Scientific Applications*, edited by F. Allen, G. Bergerhoff & R. Sievers, pp. 107–132. Bonn/Cambridge/Chester: IUCr.

Allen, F. H. S, Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *Acta Cryst.* B**35**, 2331–2339.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Brünger, A. T. (1992*a*). *Nature (London)*, **355**, 472–475.

Brünger, A.T. (1992*b*). *X-PLOR Manual*, Version 3.1, Yale University, New Haven, Connecticut, USA.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Butterworth, S. (1996). DPhil thesis, University of York, England.

Butterworth, S., Lamzin, V. S., Wigley, D., Derrick, J. & Wilson, K. S. (1999). *Acta Cryst.* D**55**. In the press.

Cruickshank, D. (1996). *Protein precision re-examined: Luzzati plots do not estimate final errors. Proceedings of the CCP4 Study Weekend on Macromolecular Refinement*, *4–5 January 1996*, DL/Sci/R35, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp 11–22. Warrington: Daresbury Laboratory.

Daopin, S., Davies, D. R., Schlunegger, M. P. & Grutter, M. G. (1992). *Acta Cryst.* D**50**, 85–92.

Dauter, Z., Wilson, K. S., Sieker, L. C., Moulis, J.-M. & Meyer, J. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 8836–8840.

Deacon, A., Gleichmann, T., Gilboa, A. J., Price, H., Raftery, J., Bradbrook, G., Yariv, J. & Helliwell, J. R. (1997). *J. Chem. Soc. Faraday Trans.* **93**, 4305.

Engh, R. A. & Huber, R (1991). *Acta Cryst.* A**47**, 392–400.

Frazão, C., Soares, C. M., Carrondo, M. A., Pohl, E., Dauter, Z., Wilson, K. S., Hervas, M., Navarro, J. A., De la Rosa, M. A. & Sheldrick, G. M. (1995). *Structure*, **3**, 1159–1169.

Hendrickson, W. A. & Konnert, J. H. (1980). *Computing in Crystallography*, edited by R. Diamond, S. Ramasechan & K. Venkatesan, pp. 13.01–13.25. Bangalore: Indian Institute of Science.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Kleywegt, G. J. & Jones, T. A. (1995). *Proceedings of the CCP4 Study Weekend on Making the Most of Your Model, 6–7 January 1995*, DL/Sci/R35, edited by W. Hunter, J. Thornton & S. Bailey, pp. 11–24. Warrington: Daresbury Laboratory.

Kleywegt, G. J. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 829–832.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Laskowski, R. A., Moss, D. S. & Thornton, J. M. (1993). *J. Mol. Biol.* **231**, 1049–1067.
Longhi, S., Czjzek, M., Lamzin, V., Nicolas, A. & Cambillau, C. (1997). *J. Mol. Biol.* **268**, 779–799.
Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
MacArthur, M. W. & Thornton, J. M. (1996). *Protein Eng.* **8**, 217–224.
Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Proteins*, **12**, 345–364.
MSI (1997). *QUANTA*, MSI, 9685 Scranton Road, San Diego, CA 92121–3752, USA.
Murshudov, G. N. & Dodson, E. J. (1997). *CCP4 Newsletter*, Vol. 33, 2nd ed. Warrington: Daresbury Laboratory.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Oldfield, T. J. (1992). *J. Mol. Graphics*, **10**, 247–252.
Parthasarathy, S. & Murthy, M. R. N. (1999). *Acta Cryst.* D**55**. In the press.
Pontius, J., Richelle, J. & Wodak, S. (1996). *J. Mol. Biol.* **264**, 121–136.
Priestle, J. P. (1994). *Structure*, **2**, 911–913.
Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). *J. Mol. Biol.* **7**, 95–99.
Rice, L. M. & Brünger, A. T. (1994). *Proteins: Structure, Function, Genetics* **19**, 277–290.
Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). *Acta Cryst.* D**52**, 327–344.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.
Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *Acta Cryst.* D**54**, 243–252.
Tronrud, D. E. (1992). *Acta Cryst.* A**48**, 912–916.
Vlassi, M., Dauter, Z., Wilson, K. S. & Kokkinidis, M. (1998). *Acta Cryst.* D**54**, 1245–1260.
Vriend, G. (1990). *J. Mol. Graphics*, pp. 52–56.
Walsh, M. A., Schneider, T. R., Sieker, L. C., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1997). *Acta Cryst.* D**54**, 522–546.
Westhof, E., Dumas, P. & Moras, D. (1985). *J. Mol. Biol.* **184**, 119–145.
Wilson, K. S., Butterworth, S., Dauter, Z., Lamzin, V. S., Walsh, M., Wodak, S., Pontius, J., Richelle, J., Vaguine, A., Sander, C., Hooft, R. W. W., Vriend, G., Thornton, J. M., Laskowski, R. A., MacArthur, M. W., Dodson, E. J., Murshudov, G., Oldfield, T. J., Kaptein, R., Rullmann, J. A. C. (1998). *J. Mol. Biol.* **276**, 417–436.