*J. Mol. Biol.* (1998) **276**, 417–436

JMB

# Who Checks the Checkers? Four Validation Tools Applied to Eight Atomic Resolution Structures

## EU 3-D Validation Network

Eight protein crystal structures, which have been refined against X-ray diffraction data extending to atomic resolution, 1.2 Å or better, were inspected using four different validation tools, PROCHECK, PROVE, SQUID and WHATCHECK. Two general questions were addressed. (1) Do the structures imply changes in ''expected'' stereochemical properties and are the target values used for restraints in the validation programs and the refinement protocol appropriate? (2) Can errors in models be detected and how reliable are the coordinates after refinement? Preliminary analysis by members of the network led to modifications both to the validation programs and to the refinement protocols. The results of the final analyses are reported here. Apparent discrepancies in cell dimensions were identified. Most stereochemical properties are shown to be more tightly clustered than for lower resolution analyses. In contrast the ω angle has a wider distribution. The validation software is generally available and can be accessed at servers listed at the end of the paper.



*Keywords:* X-ray crystal structure; structure validation; atomic resolution; protein stereochemistry; stereochemical restraints

## Introduction

X-ray data on proteins rarely extend to atomic resolution and refinement of the structures therefore requires the X-ray observations to be supplemented by stereochemical or energetic restraints. These can be divided into two groups. The first includes the ''hard'' unimodal restraints of bond lengths, angles, planarity of conjugated groups and chiral volumes. The second includes the conformational torsion angles of the backbone and side-chains, ring pucker and van der Waals repulsions, some of which are multimodal. Here we refer to the first group as ''geometric'' restraints and the second as ''conformational'' restraints and discuss the stereochemistry of a molecule in terms of its geometric and conformational attributes.

The target values for the geometric restraints are most commonly based on an analysis of the X-ray structures of amino acids and peptides (Engh & Huber, 1991) in the Cambridge Structural Data Bank (CSD, Allen *et al*., 1979, 1983). This assumes that the stereochemistry in proteins is the same as that in small peptides. Although this is likely to be true for the mean values of bond lengths and angles, it is not obvious that their natural variability, i.e. the distribution about their means, should be the same given the different specific environment in proteins. In principle ''true'' targets could

The network was part of the EC Framework III BIOTECHNOLOGY program, Contract BIO2-CT92-0524 titled ''Integrated Procedures for Recording and Validating results of 3D Structural Studies of Biological Macromolecules''. The Partners in the network are: (1) K. S. Wilson, S. Butterworth, Z. Dauter, V. S Lamzin, M. Walsh, EMBL Hamburg Outstation, c/o DESY, Notkestrasse 85, 22603 Hamburg, FRG; (2) S. Wodak, J. Pontius, J. Richelle, A. Vaguine, Free University, Brussels; (3) C. Sander, R. W. W. Hooft, G. Vriend, EMBL, Meyerhofstrasse, 1, 69117 Heidelberg, FRG; (4) J. M. Thornton[1,2], R. A. Laskowski[1,2], M. W. MacArthur[1], [1]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London, WC1E 6BT, UK; and [2]Department of Crystallography, Birkbeck College, Malet Street, London, WC1E 7HX, UK; (5) E. J. Dodson, G. Murshudov, T. J. Oldfield, University of York, York, YO1 5DD, UK; and (6) R. Kaptein, J. A. C. Rullmann, Department of NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands.

Abbreviations used: PDB, Protein Data Bank; CSD, Cambridge Structural Data Bank; 3-D, three-dimensional; ADP, atomic displacement parameter; s.u., standard uncertainty.

0022–2836/98/070417–20 $25.00/0/mb971526 

be obtained from known protein structures in the Protein Data Bank (PDB; Bernstein *et al*., 1977). However their values would be heavily biased both by the current targets used for the geometric restraints in structure refinement and by the limited amount of experimental X-ray data from which they were generally derived (Laskowski *et al*., 1993b): indeed the worse the resolution of the data, the greater the bias.

The conformational attributes, on the other hand, such as torsion angles and packing volumes, are not generally restrained during refinement, so their statistical distribution can justifiably be derived from data bases such as the PDB. These attributes have been used in the development of a number of validation packages, including PROCHECK (Laskowski *et al*., 1993a), SQUID (Oldfield, 1992), WHATCHECK (Hooft *et al*., 1996d) and PROVE (Pontius *et al*., 1996), which are discussed below. The purpose of the first three packages is to (a) verify the syntax of the file, (b) check the consistency of an atomic model with the current library and identify outliers for further investigation, (c) detect gross errors in the structures, such as mistracing of the chain, (d) check for local abnormalities of stereochemistry and (e) produce global stereochemical quality criteria. The fourth program PROVE evaluates deviations from standard atomic volumes.

Many groups are currently working in the validation field and other validation packages that consider unrestrained attributes are described by MacArthur *et al*. (1994). Examples are the ERRAT program of Colovos & Yeates (1993), which considers the relative frequencies of non-bonded interactions between C, N and O atoms; the empirically derived threading potentials of PROSA-II (Sippl, 1993); distributions of polar residues (Novotny *et al*., 1988; Baumann *et al*., 1989; Luthardt & Frömmel, 1994); three-dimensional profiles (Luthy *et al*., 1992); and atomic solvation parameters (Holm & Sander, 1992), the real space *R* factor (Bränden & Jones, 1990), the free *R* factor (Brünger, 1992a) and the recent work by the Uppsala group, for example, Kleywegt & Jones (1996).

Together, the validation programs cover both the geometric and the conformational properties of the refined models. For the former, they tend to use the same dictionaries as those used to set up restraints during refinement. Not surprisingly, these properties tend to agree with the standard dictionaries and the degree of scatter is merely a consequence of the relative weighting imposed on compliance to the X-ray data and to the restraints (although this is not strictly true for refinement programs using energy rather than geometric restraints, i.e. molecular dynamics). However it is the checks made on the conformational properties, which are independent as far as possible of the restraints applied, that are of the greatest use in validation. For example, torsion angles, if not restrained during refinement, provide the basis for an excellent validation check. Parts which have unusual conformations warrant further investigation; they are possibly wrongly interpreted, or may be at the core of the structure's active site, where strained conformations could be extremely interesting. If the crystallographer can make the structure more ''normal'' without degrading the fit to the X-ray data, this suggests that it might be an error. However, there is a real danger of negative feedback; structures which have been erroneously forced into conformations to pass the validation checks then enter the data base and thereby artificially reinforce the expectations and keep the door closed to novel conformational features. This leads to the question in the title: who checks the checkers, based on an idea published earlier; ''Sed quis custodiet ipsos custodes – But who is to guard the guards themselves?'' (Juvenal, 117).

One answer to the question is to let the protein structures themselves check whether our validation criteria are correct, or tell us what they should be. With improved techniques of crystallisation and data collection using synchrotron radiation and cryogenic cooling, X-ray data for macromolecules can now sometimes be measured to atomic (1.2 Å or higher) resolution (Dauter *et al*., 1995). This allows structures to be refined, imposing less strict compliance with prior knowledge of expected geometry for the well ordered parts to attain an accuracy better than 0.02 Å. Although some restraints still need to be imposed, especially to deal with more mobile regions, and hence some bias remains, we might expect the structures to provide more precise information about conformational properties such as torsion angles. Atomic resolution structures move us one step closer to an understanding of the ''true'' geometrical and conformational properties of proteins in general. Ideally we would like to derive these parameters anew using only such structures, but currently there are too few to provide sufficient statistical data.

The aim of the present work was to use atomic resolution structures to check the conformational parameters that have been derived from large data sets of protein structures at low as well as high resolutions. Some of these parameters vary with resolution, so how well do their extrapolated values agree with the observed properties of the atomic resolution structures? In addition, we wished to use the structures to assess the data collection techniques, refinement protocols and stereochemical target libraries plus the level of restraints applied, in the elucidation of 3-D X-ray structures: this particularly relates to which parameters should be restrained. We wished to assess the validation tools themselves, in terms of the target values and especially the tightness of the distribution imposed on them. In particular can the validation programs handle the information provided and can we identify which of the tools are most informative? Can we suggest means by which either the data collection and refinement or the validation protocols may be improved in the future?

There were two inputs to this project. One was a set of eight structures refined against atomic resolution data. The second was the set of four validation programs previously calibrated against the 3-D models in the PDB, peptides in the CSD or quantum chemically calculated parameters.

## The experimental data

Details of the eight atomic resolution structures are given in Table 1, with abbreviations used throughout the text. Data were recorded at EMBL, DESY, Hamburg, using synchrotron radiation and a MAR research imaging plate scanner. For seven structures data were recorded at room temperature; only for lysozyme were data recorded at cryogenic temperature. The crystals diffracted extremely well; nevertheless the atoms in the structures show a wide range of atomic displacement parameters (ADPs), ranging from those typical for small molecules at the core of the protein, to substantially greater values at the surface. In most of the models there were regions with high ADPs, with multiple conformations of side-chains especially at the protein-solvent interface and substantial regions of disordered solvent.

The structures were all first refined with an isotropic atomic model and restraints based on the Engh & Huber (1991) set, either using the CCP4 (1994) suite of programs or X-PLOR (Brünger, 1992b). Refinement was continued using SHELXL-93 or 96 (Sheldrick & Schneider, 1997) with anisotropic ADPs and including H-atoms riding at their calculated positions.

It was still essential to impose stereochemical restraints to maintain satisfactory geometry for poorly ordered regions. Similarly to other refinement programs, SHELXL allows the use of target restraints on geometric parameters. Torsion angles were not restrained. The weightings of different properties have built-in default values; however, these can be altered by the user. In addition SHELXL allows the geometry of structural moieties, e.g. residues of the same type, to be restrained to be similar. Such restraints were in general not imposed in the refinement of the current structures: they were only applied to discretely disordered side-chains with both alternative conformations restrained to have similar geometrical characteristics. The chemically equivalent units of the cytochrome haem group were also restrained in this way (Frazão *et al.*, 1995).

The anisotropic ADP restraints took three forms. (1) A strict restraint was applied to the anisotropic ADPs of atoms bonded to one another to ensure their vibration along the bond was the same. Two weaker restraints were: (2) atoms should not be too anisotropic, i.e. to restrict the degree of anisotropy of the atomic displacement parameter tensor, the atoms were split over two sites if this became too large; and (3) adjacent atoms in the structure should have a similar degree of anisotropy.

All refinements were carried on until no further improvement in the model could be made. It can be assumed that the resulting deviations of stereochemical parameters from their respective target values will largely reflect the relative weighting of each attribute. The final values are influenced by the weights assigned automatically within the program both to the contributions of the X-ray terms and to the different stereochemical and thermal parameters. The resulting deviations are listed in Table 2. The atomic parameters are more accu-

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

| | Cytc6 | Cutinase | Lysozyme | ProtG | RNaseSa | Ropm | RubrDv | RubrCp |
|---|---|---|---|---|---|---|---|---|
| PDB cooordinates | 1CTJ | 1CEX | 3LZT | 2IGD | 1RGG | 1NKD | 1RB9 | 1IRO |
| Temperature | RT | RT | 110 K | RT | RT | RT | RT | RT |
| Space group | *R*3 | $P2_1$ | *P*1 | $P2_12_12_1$ | $P2_12_12_1$ | *C*2 | $P2_1$ | *R*3 |
| Cell (Å) | 52.11 | 35.20 | 26.65 | 34.78 | 64.73 | 47.06 | 19.99 | 64.04 |
| | 52.11 | 67.30 | 30.80 | 40.28 | 78.56 | 37.88 | 41.51 | 64.04 |
| | 81.02 | 37.10 | 33.63 | 42.19 | 38.99 | 31.65 | 24.40 | 32.51 |
| Cell (°) | 90 | 90 | 89.3 | 90 | 90 | 90 | 90 | 90 |
| | 90 | 94.1 | 107.4 | 90 | 90 | 100.8 | 107.6 | 90 |
| | 120 | 90 | 112.2 | 90 | 90 | 90 | 90 | 120 |
| 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 |
| $R_{merge}$ (%) | 5.8 | 3.9 | 2.8 | 3.7 | 3.9 | 4.5 | 3.1 | 4.6 |
| $I/\sigma(I)$ | 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[a] | 29 | 43 |
| α-Helix (%) | 58 | 39 | 33 | 26 | 11 | 92 | 0 | 0 |
| β-Sheet (%) | 0 | 19 | 15 | 43 | 29 | 0 | 18 | 23 |

The abbreviations used here and in the text are Cytc6: cytochrome *c*6, (Frazão *et l.*, 1995); Cutinase: cutinase (Longhi *et al.*, 1997); Lysozyme: triclixic lysozyme (Walsh *et al.*, 1998); ProtG: fragment of protein G (Butterworth *et al.*, 1997); RNaseSa: ribonuclease Sa (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* (Dauter *et al.*, 1996). The data were recorded using synchroton radiation at EMBL Hamburg. The structures were refined using SHELXL93 or 96.

[a] 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.

rately defined in well ordered parts than in the less ordered regions, which do not contribute to the high angle X-ray observations (Cruickshank, 1996). In addition refinement of two structures at the same nominal resolution will generate parameters whose reliability is influenced by factors such as the solvent content and average ADP.

The structures are all fairly small and several have a low solvent content. In addition Ropm has an unusually high proportion of α-helices. However, some useful generalisations about details of protein structure can be deduced from the sample.

## Validation programs for stereochemistry

The four sets of software, PROCHECK, PROVE, SQUID and WHATCHECK address various aspects of structure validation and details of the algorithms and their implementation can be found in the publications referenced in the Introduction. An overview of the type of checks carried out is presented in Table 3. The programs exploit different but to some extent complementary aspects of the structures, although there is a set of properties common to all.

(1) PROCHECK makes use of properties originally derived from a set of 119 non-homologous protein crystal structures at a resolution of 2.0 Å or higher and having an *R*-factor no greater than 20% (Morris *et al*., 1992). Table 4 includes the current benchmark values and their standard uncertainties. The standard uncertainties of several unrestrained parameters were shown to have a clear correlation with resolution. For example, the standard deviation in a protein's main-chain hydrogen bond energies decreases with improving resolution, as does the variation of χ angles which is discussed below.

(2) PROVE computes standard volumes of atoms from a set of 64 high quality X-ray structures of proteins with low sequence homology having a resolution of 2.0 Å or better and an *R* factor of at most 20%. These standard volumes correspond to the mean values of the volume distributions computed for 178 atom types, each being defined by the residue type and the IUPAC standard atom nomenclature (Table 1 of Pontius *et al*., 1996). The atomic volume is computed using the classical Voronoi procedure (Voronoi, 1908), where the dividing plane is placed midway between the atoms. Here only buried atoms are considered and hydrogen atoms, water molecules and non-protein groups are completely excluded.

(3) The parameters used in SQUID can be derived for any chosen subset from the PDB using the program PDBSEL. For the present study statistics were derived from 186 structures selected using the criteria: (a) X-ray structures determined after 1982, (b) at resolution better than 2 Å, (c) excluding those with many outliers in the Ramachandran plot. Duplicate structures with more than 90% sequence identity were excluded.

(4) WHATCHECK calculates most of its expected properties from a data base of about 300 sequence-unique structures with the prime selection criterion being the quality of the structures (Hooft *et al*., 1996a). The data base is regenerated two to three times per year and so the associated expected properties change with this frequency as well. Exceptions are: the bond-length and angle geometries have been taken from Engh & Huber (1991) for protein residues and from Parkinson *et al*. (1996) for DNA/RNA residues. Planarities have been deduced (Hooft *et al*., 1996b) from the CSD (Allen *et al*., 1983). Hydrogen bond energies have been deduced from CSD statistics (Hooft *et al*., 1996c; Hooft, Kanters & Kroon, unpublished).

## Results and Discussion

### Discrepancies in restrained stereochemistry

As expected, given the atomic resolution of these structures, there were no gross errors in the structures, such as mistracing of the chain. All the validation packages would easily identify such gross errors, usually at the simple level of an appalling Ramachandran plot. Similarly, the ''threading potentials'' (Luthy *et al*., 1992; Jones *et al*., 1995; Lemer *et al*., 1995; Jones & Thornton, 1996; Vajda *et al*., 1997; and many references therein) and Directional Atomic Contact Analysis (DACA; Vriend & Sander, 1993) calculated in WHATCHECK are very sensitive to any gross misinterpretation.

Nevertheless the programs flagged some deviations in stereochemistry in all the structures. For example, SQUID identified deviations >4 s.u. for several types of restraint: chiral (six proteins), planarity (three proteins), bond/angle (all proteins), unexpected anisotropy (in the five coordinate sets where the anisotropic ADPs were retained) and close van der Waals contacts (bumping) in most of them. The other programs identified essentially the same deviations, with the exception of the anisotropy, which is only treated by SQUID. A deviation of 4 s.u. in any normally distributed value has an expected probability of less than 1 in a 1000, so that it is advisable to attempt to pinpoint whether they are statistical fluctuations or real errors. The way to do this of course is to refer to experimental data.

### Detailed inspection of the electron density for RNaseSa and RubrCp

With this in mind two structures, RubrCp and RNaseSa, already available from the PDB with codes 1IRO and 1RGG, respectively, were selected for a detailed inspection of the electron density in the light of the outliers flagged by the programs. They were inspected at every residue where

**Table 2.** Refinement for the atomic resolution structures

| Program | Cyt c6 Shelxl_96 | Cutinase Shelxl_93 | Lysozyme Shelxl_96 | ProtG Shelxl_93 | RNaseSa Shelxl_93 | Ropm Shelxl_93 | RubrDv Shelxl_93 | RubrCp Shelxl_93 |
|---|---|---|---|---|---|---|---|---|
| Residues | 89 | 197(213) | 129 | 61 | 2 × 96 | 59 | 52 | 53(54) |
| Molecular mass (kDa) | 9.3 | 22.0 | 14.5 | 6.9 | 2 × 10.5 | 7.0 | 6.0 | 6.1 |
| Ligands | Haem | – | 3 AcOH,5 $NO_3$ | – | – | – | $FeS_4$, $SO_4$ | $FeS_4$ |
| Solvent sites | 151 | 332 | 286 | 130 | 338 | 114 | 111 | 110 |
| Partial solvent sites of these | 59 | 107 | 44 | 0 | 0 | 49 | 47 | 0 |
| Observations/parameters | 4.0 | 5.6 | 4.5 | 4.4 | 3.5 | 3.8 | 5.6 | 7.0[a] |
| *R* factor (%) | 14.0 | 9.4 | 9.5 | 9.4 | 10.6 | 10.1 | 7.9 | 9.0 |
| $R_{free}$ (%)[b] | 18.8 | 11.9 | 11.3 | 12.5 | n/a | 12.3 | 11.0 | n/a |
| RMS deviations in: | | | | | | | | |
| 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. (°) | 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 |
| Backbone score (ZR) | −0.8 | 0.4 | 0.5 | 1.7 | 1.6 | 3.4 | 2.3 | 1.1 |
| (H-bond energies from PROCHECK: there were a total of 520 backbone H-bonds with a mean energy of −2.03 kcal $mol^{-1}$ and s.u. 0.64 kcal $mol^{-1}$) | | | | | | | | |
| Mean H-bond (kcal $mol^{-1}$) | −2.1 | −2.0 | −2.0 | −2.2 | −2.0 | −2.3 | −2.0 | −1.9 |
| s.u. H-bond (kcal $mol^{-1}$) | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.4 | 0.7 | 0.9 |
| $\chi_1$ s.u. | 9.2 | 10.7 | 8.6 | 7.2 | 10.2 | 8.5 | 7.3 | 11.6 |
| $\chi_2$ *trans*. s.u. | 11.3 | 9.5 | 8.6 | 17.4 | 13.3 | 10.8 | 8.3 | 8.8 |
| Dihedral angle *G*-factors | −0.06 | −0.03 | 0.01 | 0.04 | −0.09 | 0.35 | −0.07 | −0.15 |

[a] RubrCp was refined with the Friedel pairs treated independently, giving almost two times the number of observations.

[b] These $R_{free}$ values do not correspond to the final models discussed in the text as those were refined using all data including those used for $R_{free}$. For Ropm, only 59 of the 65 residues are included in the final model.

**Table 3.** An overview of the type of checks carried out

| Program | PROCHECK | PROVE | SQUID | WHATCHECK |
|---|---|---|---|---|
| Target library | Engh & Huber | | Engh & Huber | Engh & Huber |
| Main/side-chain reporting | Y | Y | Y | Y |
| Nomenclature, structure summary and format | | | | |
| Molecular mass and volume | | | Y | Y |
| Space group and symmetry | | | | Y |
| Cell and orthogonalisation matrix consistency | | | | Y |
| No. of atoms, residues, solvent, chains, ligands | | | Y | |
| Atom name - IUPAC standards | Y | | Y | Y |
| IUCr standard side-chain torsions | Y | | Y | Y |
| Missing or suspect atoms | | | | Y |
| L/D amino acids | Y | | Y | Y |
| Cell dimension check | | Y | Y | Y |
| Geometry | | | | |
| Bond lengths and angles | Y | | Y | Y |
| Planarity | Y | | Y | Y |
| Chirality | Y | | Y | Y |
| Conformation | | | | |
| Torsion angles ($\phi$,$\psi$): Ramachandran | Y | | Y | Y |
| Torsion angles ($\omega$) | Y | | Y | Y |
| Torsion angles (CA) | | | Y | |
| Torsion angles ($\chi_1$,$\chi_2$) | Y | | Y | Y |
| Peptide flip | | | Y | |
| Non-bonded contacts | Y | | Y | Y |
| Special residues | | | | |
| Gly and Pro Ramachandran | Y | | Y | |
| Proline puckering | | | Y | |
| Hydrogen bonds | | | | |
| Statistics with and without waters | | | Y | |
| Donor-donor and acceptor-acceptor contacts | | | Y | Y |
| Unsatisfied donors and acceptors | | | Y | Y |
| Suggested HNQ (His, Asn, Gln) flips | | | Y | Y |
| H-type assignments:HisD, HisE, HisH | | | | Y |
| Solvent | | | | |
| Water protein distribution | | | Y | |
| Floating water molecules | | | Y | Y |
| Atomic displacement parameters | | | | |
| Anisotropic *B* values | | | Y | |
| Volumes/packing density | | Y | Y | |
| Global parameters | | | | |
| Position specific rotamer score | | | | Y |
| Backbone normality | | | | Y |

WHATCHECK, SQUID, PROVE or PROCHECK had flagged something as unusual, i.e. where the 4 s.u. limit for the parameter distributions was exceeded in the model. The results of this inspection are summarised in Table 5.

The first structure, RubrCp, is a small protein of 54 residues. Despite the excellent diffraction data several residues were disordered and not visible in the map, in particular the C-terminal Glu54, which had resulted in an OXT atom being incorrectly generated for residue Glu53. The protein contains an $FeS_4$ cluster involving four cysteine residues. There are four main-chain N–H$\cdots$S hydrogen bonds to the cluster. Both SQUID and WHATCHECK originally indicated missing H-bonds for the main-chain N atoms involved in the long (3.5 to 3.8 Å) N–H$\cdots$S hydrogen bonds. The current releases of both programs handle these particular bonds correctly. This sort of feature is only possible for validation programs to treat properly if a set of ''structural entity'' records are deposited in the file. In principle such problems are not difficult to treat, provided ''ligands'' are described in an accepted standard format.

The second structure, RNaseSa, has two molecules in the asymmetric unit each with 96 amino acid residues. These have been refined independently and outliers frequently occurred in both molecules at the same point in the sequence, indicating that the features are inherent in the protein fold. There were some bump problems with multiple conformers in the structure. Residues 86 and 54 collide; 54 has two alternative conformations in both A and B molecules and probably 86 should also. SQUID found a surprising 2 Å water shell as well as the more populated 2.8 Å one. All water molecules were assigned unit occupancy in this model, a most unlikely situation in reality. The

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

| Stereochem. parameter | Original parameters Mean | s.u. | $N_{obs}$ | Atomic resolution structures Mean | s.u. | $N_{obs}$ |
|---|---|---|---|---|---|---|
| $\chi_1$ 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 |
| $\chi_2$ 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 |
| $\chi_3$ (S-S bridge): | | | | | | |
| Right-handed | 96.8 | 14.8 | 124 | 87.0 | 13.1 | 2 |
| Left-handed | −85.8 | 10.7 | 103 | −86.4 | 10.6 | 6 |
| Disulphide bond (Å) | 2.0 | 0.1 | 227 | 2.0 | 0.0 | 8 |
| ω Dihedral angle (°) | 179.6 | 4.7 | 23895 | 179.0 | 5.6 | 812 |
| Main-chain hydrogen bond energy (kcal/mol)[a] | −2.03 | 0.75 | 15597 | −2.03 | 0.64 | 520 |
| CA chirality: ζ ''virtual'' torsion angle (CA–N–C–CB) | 33.9 | 3.5 | 21950 | 33.8 | 2.42 | 752 |
| %age (ϕ,φ) in most favoured regions of Ramachandran plot | | | >90 | | | 92.1 |

[a] Evaluated using the Kabsch & Sander (1983) method.

problem of partial solvent occupancy will be addressed in a separate study.

Several potential abnormalities in planar groups of Gln, Glu, Asn, Asp and Arg were flagged by the programs. It is not clear whether this is a consequence of the refinement protocol. In several cases they have high ADPs and maybe required stricter restraints. They may, for example, have been distorted by any restraints to form hydrogen bonds imposed by the program.

SQUID found a single cavity in RNaseSa with *x,y,z* coordinates 56,9,10 Å. However the map had no feature at this position to indicate the presence of a solvent molecule and the significance of the cavity appears to be marginal. It is perhaps of interest that it lies close to, but not at, the active site and may reflect a degree of loose packing and flexibility in this region.

Both SQUID and WHATCHECK check the H-bond network around His, Asn and Gln (HNQ residues) to see if reorientation could improve the theoretical fit. SQUID's quick distance check found possible H bonds to CE and CD atoms for His53 A and B, and suggested checking the electron density for these residues. There are also good H-bonds to His53 ND and NE, so no correction was required. WHATCHECK suggested inverting the orientations of Gln94 A and Gln47 B. At high resolution it is often possible to detect the difference between N with seven electrons, and O with eight, as incorrect assignment of the atom types leads to very different ADPs for the pair. This was supported by inspection of maps or by plots of density *versus B* factor (Sevcik *et al*., 1996) and was flagged by SQUID.

## Application of the validation programs to the atomic resolution structures and what they reported

### *Cell dimension errors*

One clear problem that the validation programs identified is the accuracy of the cell dimension estimated from the synchrotron experiment. WHATCHECK, SQUID and PROVE all suggested that the cell dimensions used for determining the atomic resolution structures were in error by up to 0.5%. The three programs detect likely errors of this form using different methods. (1) WHATCHECK uses direction-dependent systematic deviations observed in the bond lengths to calculate a ''cell transformation matrix'' (Vriend *et al*., 1986). Applied to the given unit cell, this transformation minimises the systematic deviations in bond lengths from target values. (2) SQUID reports the packing density and correlated errors in the CA–CA distances. (3) PROVE identifies systematic shifts in average atomic volumes. Neither SHELXL nor any other refinement programs adjust the cell dimensions during minimisation, so errors in the dimensions will lead directly to systematic deviations of the refined atomic parameters, which is precisely what the validation programs reported.

To check whether the validation programs were right and whether there were indeed errors in the cell parameters, an independent calculation of corrections to cell dimensions was developed at EMBL Hamburg as part of the beam line wavelength calibration system (V. S. Lamzin, personal communication). The cell parameters for the eight atomic resolution models were post-refined. At atomic resolution the contribution of the X-ray

**Table 5.** Residue by residue comments

| A. *RubrCp* Residue | Comment |
|---|---|
| Lys2 | Main-chain density OK, CE small volume, side-chain weak density. Also $\omega = 166.2^\circ$ |
| Thr5 | Two conformations. ADP for CG2A is high |
| Thr7 | No H-bond for O atom |
| Val8 | N–H–SG H-bond longer than target |
| Tyr11 | N–H–SG H-bond longer than target |
| Ile12 | Two conformations, both have high ADP for CG1 |
| Pro15 | Two conformations. CD CG abnormal. A problem in naming each conformer: CB and CG mixed up |
| Pro20 | Underpuckered. Good density however |
| Asp29 | CA–CB–CG angle 120°, good density. Minor second conformation possible |
| Lys31 | CE, NZ have high ADPs |
| Asp36 | 3.2 Å H-bond, outside limits, no H-bond for N |
| Pro40 | Underpuckered but with good density |
| Leu41 | N–H–SG H-bond longer than target. C-O bond has smeared density |
| Val44 | N–H–SG H-bond |
| Asp47 | CB, CG. high ADP, hard to fit |
| Glu50 | Density OK. Very tight turn. Also $\omega = 167^\circ$ |
| 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 to 53 have relatively poor density |
| Val52 | O large volume |
| Glu53 | Side-chain problems. $\chi_1$ bad, CB–CG–CE angle 97° |

| B. *RNaseSa* Residue | Comment |
|---|---|
| 1-3A, 1-3B | Surface residues. Several unusual volumes, e.g. 2B CA, but residues very poorly defined |
| Pro13A, 13B | Two conformations for CG only. Density OK |
| Asp25A, 25B | CA–CB–CG angle <109°. ADPs 20, 32 and 40. Poor density for the carboxylate group in both molecules |
| Pro27A, 27B | Underpuckered. Possibly two conformations? CD and CG have unusually high ADPs and poor density |
| Gln38A, 38B | CA of 38A large volume, maybe related to poor planarity at the end of the side-chain. Bad bond length. Smeared density. Almost certainly more than a single conformation, but difficult to model |
| Arg40A, 40B | 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 |
| Glu41A, 41B | 41A. Strange CB–CD–CG angle. ADPs 16, 22, 28 with poor density |
| Glu54A, 54B | Both have two conformations. 54A C small volume, but density seems OK |
| Arg63A, 63B | 63B. CD-NE short, twisted and non-planar. All ADPs are low and have very good density. 63A makes a strong salt bridge to the sulphate |
| Arg65A, 65B | 65A N large volume. Near to double conformation of Glu54A. Smeared density suggesting possible multiple conformation |
| Ile71A, 71B | Odd angles CB–CG1–CD1, 128°; CB–CG1–CD1, 129°. ADPs are 16, 26, 11, 17, 20. Closely packed region |
| Gln77A, 77B | OE1 very weak ADP 41, NE2 24. Maybe two conformations. Tight twist, but good density |
| Tyr81A, 81B | 81B CE1 large volume, near to 57B. Density very good |
| His85A, 85B | 85A and 85B CA large volume. Multiple conformations |
| Tyr86B, 86B | 86A OH small volume. Some $\chi$ abnormality, tightly packed, good density |

Anisotropy was noted if the principal axes of adjacent atoms were very different. Only residues which were flagged are listed. The H-bond flag indicates a ''missing'' H-bond. Volumes were flagged if they deviated by more than 2.4 $\sigma$ from the mean.

term is dominant and thus if there is an inaccuracy in the cell parameters, interatomic distances will deviate from their expected values. In the presence of restraints these deviations will not provide precise information about the true cell parameters but rather show a tendency towards them. The residual between the interatomic distances in an orthogonal coordinate frame and the targets, taken from the Engh & Huber (1991) set, was minimised by least-squares. The parameters of the unit cell, which define the transformation tensor from fractional to orthogonal frames, were refined taking into account symmetry constraints. Estimated uncertainties for the parameters were obtained from the inversion of the normal matrix.

For Cytc6 the experimental cell dimensions had been estimated as an average from a set of about ten crystals measured in Hamburg or on a conventional source with a FAST TV detector in Lisbon. The Cutinase cell had been determined from a different crystal using a diffractometer. For the other six structures, the dimensions had been determined during data reduction using DENZO, using the estimated values of wavelength and sample to detector distance. For all of the structures except Cytc6 and Cutinase, there are significant deviations between the dimensions suggested by the validation programs and the experimental values. For these six, the experimental lengths seem to be too short, by about 0.5% on average, leading to an underestimate of cell volume of 1 to 2% (Table 6). The geometric bond lengths in the target library are accurate to about 0.1% (Lamzin *et al.*, 1995). We therefore believe the corrections suggested by the programs to be valid, but we are uncertain about their absolute accuracy. There are three important points. (1) How can such problems be avoided in the future? The errors almost cer-

**Table 6.** Re-estimation of the cell dimensions for the eight atomic resolution structures

| | | Cytc6 | Cutinase | Lysozyme | ProtG | RNaseSa | Ropm | RubrDv | RubrCp |
|---|---|---|---|---|---|---|---|---|---|
| Experimental | *a* (Å) | 52.11 | 35.20 | 26.65 | 34.78 | 64.73 | 47.06 | 19.99 | 64.04 |
| | *b* (Å) | 52.11 | 67.30 | 30.80 | 40.28 | 78.56 | 37.88 | 41.51 | 64.04 |
| | *c* (Å) | 81.02 | 37.10 | 33.63 | 42.19 | 38.99 | 31.65 | 24.40 | 32.51 |
| | α (°) | 90 | 90 | 89.3 | 90 | 90 | 90 | 90 | 90 |
| | β (°) | 90 | 94.1 | 107.4 | 90 | 90 | 100.8 | 107.6 | 90 |
| | γ(°) | 120 | 90 | 112.2 | 90 | 90 | 90 | 90 | 120 |
| Cell changes (%) | | | | | | | | | |
| WHAT-CHECK | *a* | None | None | 0.3 | 0.7 | −0.1 | 0.6 | 0.7 | 0.7 |
| | *b* | | | 0.3 | 0.4 | 0.3 | 0.5 | −0.1 | 0.7 |
| | *c* | | | 0.5 | 0.3 | 0.6 | 0.2 | 0.5 | 0.7 |
| | Volume | None | None | 1.1 | 1.4 | 0.9 | 1.3 | 1.1 | 1.9 |
| Hamburg | *a* | 0.02 | 0.22 | 0.39 | 0.78 | 0.07 | 0.44 | 1.04 | 0.69 |
| | *b* | 0.02 | 0.10 | 0.51 | 0.55 | 0.35 | 0.62 | −0.06 | 0.69 |
| | *c* | −0.02 | 0.13 | 0.42 | 0.42 | 0.69 | 0.35 | 0.61 | 0.77 |
| | α | – | – | 0.13 | – | – | – | – | – |
| | β | – | −0.14 | −0.05 | – | – | 0.24 | −0.14 | – |
| | γ | – | – | −0.01 | – | – | – | – | – |
| | Volume | 0.01 | 0.47 | 1.34 | 1.76 | 1.10 | 1.35 | 1.68 | 2.17 |
| PROVE | Shrinkage | −0.03 | −1.26 | 3.10 | 2.88 | 0.45 | 2.19 | 2.09 | 2.60 |

Cell dimensions for six of the structures were obtained directly from the synchrotron data. For Cytc6, the cell was obtained from an average of about ten different crystals; see the text. The cutinase cell was determined using a four-circle diffractometer from a different crystal. Abbreviations as in Table 1

tainly arose from rather poor definition of both the absolute wavelength used at the synchrotron (unlike conventional home sources) and the crystal to detector distance. We estimate that for several of the data sets the error was up to 0.5%. In EMBL Hamburg, this problem has recently been addressed by improved wavelength calibration and distance estimates. It should therefore disappear provided care is taken during data collection.

(2) How to treat the errors for the present structures? The structures can in principle be re-refined with the cell dimension changes indicated by the programs. However, it would be ideal to recollect the data using the improved calibration, so that the results are based directly on experiment rather than on derived data. Hopefully new structures will not suffer from this aberration.

The ProtG, RubrCp and Lysozyme (and probably Ropm) structures will be further refined with the adjusted parameters before deposition in the PDB. Cytc6 and Cutinase require no corrections. Comments will be introduced to the deposited files for RNaseSa and RubrCp, after assessing the effects of refining with the adjusted parameters.

(3) It is clearly essential to determine accurate experimental cell parameters at the time of data collection if we are to produce a set of models for creation of a future library.

### *Proline pucker*

WHATCHECK reported that several of the 37 proline rings in the structures display an unusually low pucker. The program verifies the ring conformation of proline residues by calculating the Cremer & Pople (1975) ring puckering parameters. The parameters for five-membered rings are the puckering amplitude $Q$, which gives the r.m.s. out-of-plane deviation of the five ring atoms, and the puckering phase $\phi$, which is the phase of the sine wave describing the deviations of the atoms from the plane of the ring. The values obtained are compared to those regularly observed in high-quality structures, namely with $0.20 < Q < 0.45$ Å and in one of the two conformations: $\phi = 76°$ (s.u. 23°), corresponding to an approximate envelope conformation with CG above the plane of the ring, or $\phi = -89$ (s.u. 23°), corresponding to an approximate twist conformation with CG below and CB above the plane of the ring.

The electron density of the proline residues in RubrCp and RNaseSa was examined carefully to see if their apparently unusual ring pucker could be a consequence of environment, or of some other effect (Table 7). In those cases where the proline was surprisingly flat, the maps showed no evidence of two conformations (Figure 1). Some other proline rings displayed unusual puckering phases, a twist having the CG and CD atoms out of the plane of the ring instead of the much more usual CG and CB. One was due to a nomenclature problem with two conformations. Only one proline in RubrCp and two in RNaseSa showed evidence for two conformations, and indeed those in RNaseSa differed between the two independent molecules. There was also substantial variation in the degree of puckering and this also varied between RNaseSa molecules A and B.

SQUID's analyses of the puckering distributions for structures in the PDB and for the present structures are shown in Figure 2. The torsion angle chosen to represent the puckering angle is CA–CB–CD–CG, i.e. it is the dihedral angle between the

**Table 7.** The pucker of the prolines in RubrCp and RNaseSa

| *RubrCp* Residue | No. of conf. | Comment |
|---|---|---|
| Pro15 | 2 | One conformation almost flat, the second puckered |
| Pro20 | 1 | Not very puckered, good density |
| Pro26 | 1 | Puckered, good density |
| Pro34 | 1 | Beautifully puckered, very good density |
| Pro40 | 1 | Very flat, with good density |

| *RNaseSa: two molecules, A and B in the asymmetric unit* Residue | No. conf. (A,B) | Comment |
|---|---|---|
| Pro12 | 1,1 | A: reasonably puckered. B: good pucker |
| Pro13 | 2,1 | A: two puckered conformations, both with CG out of plane B: reasonable pucker |
| Pro27 | 1,1 | A: very little pucker. B: highly puckered, poor density |
| Pro29 | 1,1 | A and B: single well-puckered conformation with good density |
| Pro45 | 1,1 | A and B: single well-puckered conformation with good density |
| Pro60 | 1,2 | A: puckered, good density. B: 2 clear puckered conformations |

CA,CB,CD and CB,CG,CD planes. The distribution is trimodal for the PDB set, with peaks at 144, 181 and 218°, corresponding to positive envelope, flat ring and negative envelope conformations. In contrast, for the admittedly small sample of atomic resolution structures the sharp negative envelope peak remains, the positive one is broader and lower, and overlaps with that of the flat conformation. Taken together, these results all suggest that the lower limit used in WHATCHECK might need to be adjusted. Indeed a more thorough examination of proline residues is required when more atomic resolution structures become available.

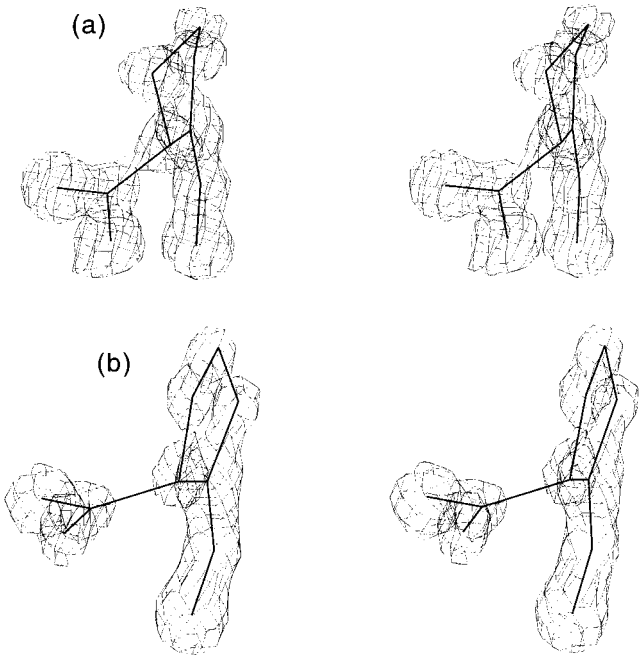


**Figure 1.** The $3F_o - 2F_c$ electron density contoured at the 1 σ level around (a) Pro34 and (b) Pro40 in RubrCp. Both have single conformations modelled in good density. The ring in the former is puckered, in the latter essentially flat.

### *Hydrogen bonds*

Hydrogen bonds are validated in different ways by the programs. PROCHECK calculates

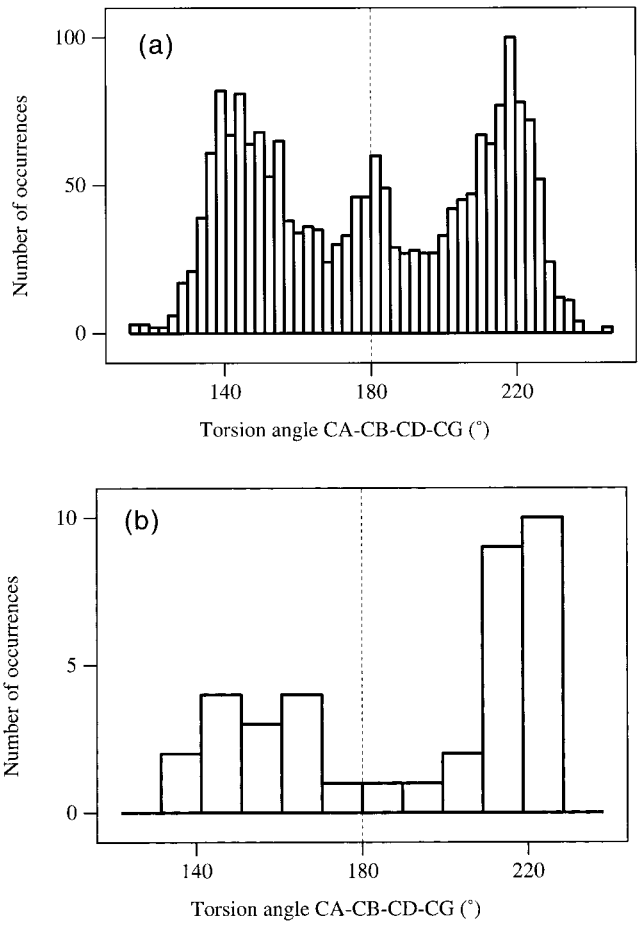


**Figure 2.** Distributions of the CA–CB–CD–CG torsion angle in the proline residues of (a) the structures in the PDB and (b) the eight current structures. The angle represents the positive or negative pucker of the proline ring around the flat conformation corresponding to 180°.

the main-chain hydrogen bond energies using the method of Kabsch & Sander (1983) and finds the s.u. about the mean value of −2.03 kcal/mol (Table 2). For the eight structures, s.u. values are approximately those expected at resolution around 1.0 Å. The Ropm structure appears to have more tightly clustered hydrogen-bond energies and this might be a consequence of it being an all α-helical protein.

SQUID and WHATCHECK perform detailed analyses of H-bonding (Baker & Hubbard, 1984; Hooft *et al.*, 1996c) both within the molecule and to the solvent, but with somewhat different definitions of what an H-bond is. The mean H-bond length, excluding water molecules, was 3.0 Å (s.u. of 0.32 to 0.35 Å), but the mean is of course biased by the upper cut-off selected. When water molecules are included in the calculation, the mean H-bond length is generally 0.08 Å longer with a smaller s.u. of around 0.27 Å. This trend is not seen in low resolution structures where the H-bond lengths tend to be very variable.

All eight structures show a number of unsatisfied buried H-bond donors. Some of these might be due to the validation programs not taking into account alternative residue orientations. Also all structures appear to have some donor/donor and acceptor/acceptor non-bonded contacts, but some apparent violations might be expected for folded proteins. These structures have around 10 such contacts per 1000 polar protein atoms. The data base of lower resolution structures has many more.

An H-bond analysis, disregarding the SHELXL hydrogen positions and using positions calculated either by the HBPLUS program (McDonald & Thornton, 1994) or from an optimised hydrogen network (Hooft *et al.*, 1996c), highlighted a few problems in the histidine, glutamine and asparagine orientations, as discussed above.

### Solvent shells

SQUID analysed the waters in the atomic resolution structures and found them distributed in two shells around the proteins. The first shell is well defined at 2.8 Å (s.u. 0.24) and the second less clear at 3.7 Å (s.u. 0.4; Figure 3). This sort of analysis is not possible with lower resolution structures in the PDB where the solvent is often very poorly described.

There were apparent discrepancies in the solvent model in most of the structures, ranging from water molecules closer than 2 Å to fully occupied protein atoms, to an apparently spurious solvent shell at 2 Å observed in RNaseSa. The latter was because all the water molecules were assigned unit occupancy during refinement. Hence partially occupied solvent atoms were treated as having full occupancy but with anti-bumping restraints forcing them to be at least 2 Å from their nearest neighbour.

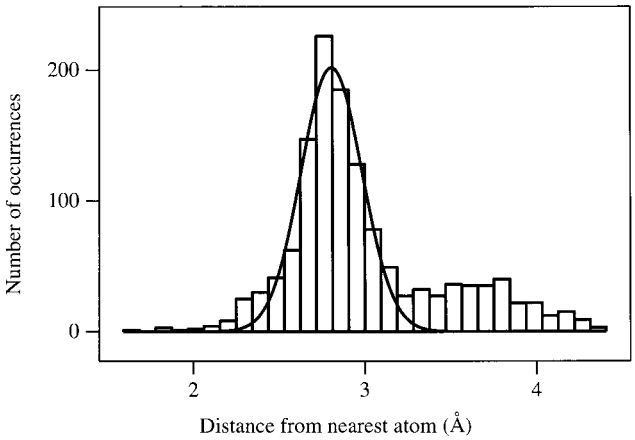


**Figure 3.** The solvent shell for seven of the atomic resolution structures. RNaseSa was excluded due to its odd solvent positions. The distance from each water molecule to the nearest protein atom was computed. The distribution was generated by binning the data between 1.9 and 4.3 Å and fitting a gaussian to determine the first water shell. It is clear there is a second solvent shell at approximately 3.8 Å from the nearest protein atom, but the data were not sufficient to fit a second gaussian.

### Volumes

PROVE computes the volumes of buried atoms using Voronoi polyhedra and analyses their deviations from standard values derived from the PDB. For each atom, the magnitude of the deviation from the expected volume is computed as a volume *Z*-score, defined as the difference between the actual and expected volume divided by the s.u. of the volume distribution. An absolute and relative volume difference are computed as well as mean *Z*-scores and *Z*-score r.m.s. values for groups of atoms and for the entire protein (Pontius *et al.*, 1996). The mean *Z*-score indicates whether the atomic volumes tend on average to be smaller or larger than expected, while the *Z*-score r.m.s. measures the deviation from the expected volumes. The *Z*-score r.m.s. tends to decrease with increasing resolution and thus provides a global indicator of the structure's quality.

The atomic resolution structures were found to have only slight volume irregularities (Table 8). The *Z*-score r.m.s. values ranged from 0.88 to 1.03, as expected for well resolved structures (Pontius *et al.*, 1996). Five of the eight structures had average atomic volumes more than 2% smaller than the standards (Table 8). In Lysozyme (the only cryogenic structure), scored atoms were on average more than 3% (0.5 $Å^3$) smaller than the standards, with an average *Z*-score of −0.33. The volumes of the CA atoms for Ropm and Cytc6 were larger than the standards, while for the other structures the CA atoms were smaller. More than half of the scored atoms in Ropm and Cytc6 have helical secondary structure as assigned by DSSP (Kabsch & Sander, 1983). Recalculation of the standard

**Table 8.** Analysis of polyhedron volumes of buried atoms using PROVE

| Protein | (a) | (b) | (c) | (d) | (e) | (f) | (g) |
|---|---|---|---|---|---|---|---|
| Cytc6 | 1.00 | −0.02 | 0.02 | 0.03 | 212 | 2 | 0.94 |
| Cutinase | 1.03 | 0.12 | 0.25 | 1.26 | 678 | 8 | 1.18 |
| Lysozyme | 1.03 | −0.33 | −0.56 | −3.10 | 437 | 4 | 0.92 |
| ProtG | 0.93 | −0.31 | −0.47 | −2.88 | 129 | 0 | 0.00 |
| RNaseSa | 0.95 | −0.06 | −0.03 | −0.45 | 532 | 5 | 0.94 |
| Ropm | 0.88 | −0.26 | −0.27 | −2.19 | 134 | 0 | 0.00 |
| Rubrdv | 0.94 | −0.23 | −0.33 | −2.09 | 141 | 1 | 0.71 |
| RubrCp | 1.01 | −0.31 | −0.40 | −2.60 | 154 | 2 | 1.30 |

(a) PROVE *Z*-score r.m.s.; (b) average deviation from standard volumes using atomic *Z*-score; (c) average deviation of observed from standard volumes (Å$^3$); (d) average deviation from standard volumes (%); (e) number of scored atoms; (f) number of atoms found to be outliers with scores greater than 2.5, or less than −2.5; (g) scored atoms found to be outliers (%).

volumes per secondary structure class indicates that volumes of backbone atoms, and to a lesser extent side-chain atoms, vary with the type of secondary structure to which the residue belongs. Indeed the CA atoms have a larger apparent volume in α-helical residues.

For the 46 atom types for which more than 20 (taken to be statistically meaningful) buried atoms were found in the atomic resolution structures, the distributions were significantly narrower than those computed for the 64 protein reference set from the PDB. For many of the hydrophobic atoms and for some of the polar atoms, there was also a slight shift of the computed distribution towards smaller volumes (Figure 4), in agreement with the observations made above on the trend of the atomic volumes to be somewhat smaller in these structures. The structures with volumes smaller than the standards also had average bond lengths shorter than the Engh & Huber targets and were flagged as having too-short cell dimensions (see above). Table 6 includes the shrinkage in volume estimated by PROVE, which is in general accordance with the shifts indicated for the cell dimensions.

Atoms which had an absolute *Z*-score in PROVE greater than 2.5 were consistently associated with poorly defined regions in the electron density maps as well as with irregularities in other parameters, such as departure from planarity and unusual bond angles, as determined by PROCHECK, WHATCHECK or SQUID. Not unexpectedly, atoms in contact with neighbouring side-chains with alternate conformations have larger absolute volume *Z*-scores, suggesting that a proper treatment of multiple conformations is still necessary.

## Analyses of the conformational properties of the atomic resolution structures and what they tell us about proteins in general

### *The ω torsion angle*

Having seen what the validation programs had to say about the structures, what could the structures tell us about the conformational properties of proteins in general? A particularly interesting property is the ω torsion angle. Although ω is not usually directly restrained during refinement, the planarity restraint on the peptide group in practice restrains ω to a target of 180° and its variability largely reflects how tightly this restraint was imposed.

ω in the atomic resolution structures has a mean value of 179.0° (s.u. 5.6°). The mean is significantly lower than the 179.6° previously observed in protein structures, while the s.u. is larger than the 4.7° observed before (Morris *et al.*, 1992; MacArthur & Thornton, 1996). A remarkable feature of the ω angles, borne out by the atomic resolution structures, is the bimodality of their distribution

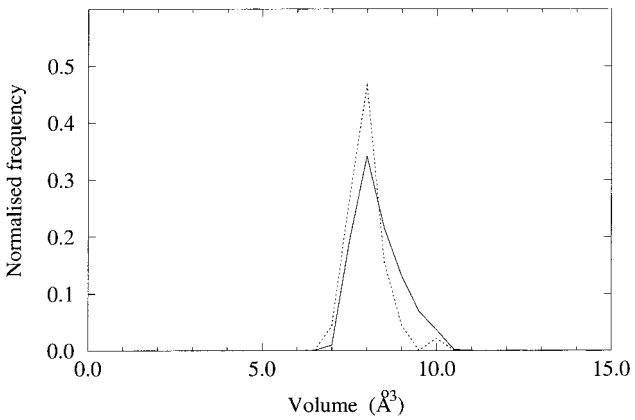


**Figure 4.** The volume distributions for atoms of type Ala C. The Figure shows the standard range used in PROVE (continuous line) and the distribution calculated using the atomic resolution structures (broken line). The distribution using the atomic resolution structures is narrower ($\sigma = 0.54$ *versus* 0.66), with a smaller average volume (8.23 *versus* 8.57 Å$^3$).

between residues of left-handed and right-handed chain chirality (i.e. above and below the top left to bottom right diagonal in the conventional $\phi,\psi$ plot). For all 489 *trans* peptide bonds in residues of right-handed twist, the mean $\omega$ is 179.9° (s.u. 5.0°) while for the 323 residues of left-handed twist, the mean value is lower at 177.7° (s.u. 6.2°). For individual structures the mean values of $\Delta\omega$, where $\Delta\omega$ is the difference in the mean values of $\omega$ for right and left-handed chain twist, range from 0.82° in Cytc6 to 3.49° in ProtG. This is observed for all residues irrespective of secondary structure and is consistent with the left/right dichotomy observed in proteins previously by MacArthur & Thornton (1996). Indeed, for the eight atomic resolution structures the effect is even more pronounced ($\Delta\omega$ in the 85 protein set is 1.3°) and is observed for all eight structures without exception. In RNaseSa, the two molecules in the asymmetric unit were refined without non-crystallographic symmetry restraints and had deviations of $\omega$ from planarity which correlated with a coefficient of 0.88 (Sevcik *et al.*, 1996).

The distribution of $\omega$ angles is similar to that observed in a sample of 287 small linear peptides taken from the CSD (MacArthur & Thornton, 1996), which had a mean value of 178.8° (s.u. 5.6°). The distribution of values in the atomic resolution structures is intermediate in character between that for the Morris *et al*. (1992) protein set and for the small peptides (Figure 5). The pronounced V-shape of the energy well, calculated from protein structures, is clearly a consequence of the different type and degree of restraints applied in the course of refinement. The energy well for the atomic structures shows a more rounded shape, although it is still somewhat sharper than that of the peptides. The latter approaches a classical parabolic curvature. These results combine to suggest that the target values for $\omega$ and its s.u. should be modified to about 6°. This value is independent of the resolution of the analysis as it reflects an intrinsic property of protein folds.

The effect of over-restraining peptide planarity is shown in Figure 6. Trp48 in ProtG lies between the two peptides with the greatest deviation from planarity, with $\omega$ values of 195° (before) and 162° (after) this residue (Butterworth *et al.*, 1998). The bulky aromatic side-chain is held in a hydrophobic cleft inducing considerable strain on the main chain. The density corresponds to that of the refined model. The Figure shows the effect of artificially imposing absolute planarity, which causes the O atoms to lie well outside the density, i.e. well away from their true position. This is an extreme view of the effect of too-tight restraints on parameters such as $\omega$, and confirms that in reality the peptide bond can deviate substantially from planarity. We do not suggest removing the restraints on planarity ($\omega$) at this or lower resolutions. The restraints should however reflect the expected s.u., which in the case of $\omega$ appears to be about 6° rather than the often used 3°.

### *Core regions of the Ramachandran $\phi,\psi$ plot*

The Ramachandran plot (Ramachandran *et al.*, 1963) is a representation of the conformation of the main chain of the protein. Expected distributions of $\phi,\psi$ angles were originally generated theoretically by checking where van der Waals steric clashes would restrict the conformational space available for an Ala dipeptide unit. It cannot be sufficiently emphasised that the Ramachandran plot is the best indicator of the global correctness of a structure, because the main-chain torsion angles are not usually restrained. A quick glance at the plot would instantly reject all the structures with gross errors found in the PDB; such models are not part of the present analysis.

The version of the plot most commonly reported in the literature is that output by PROCHECK. Also commonly reported is the percentage of residues lying in the ''core'' regions of this plot. The

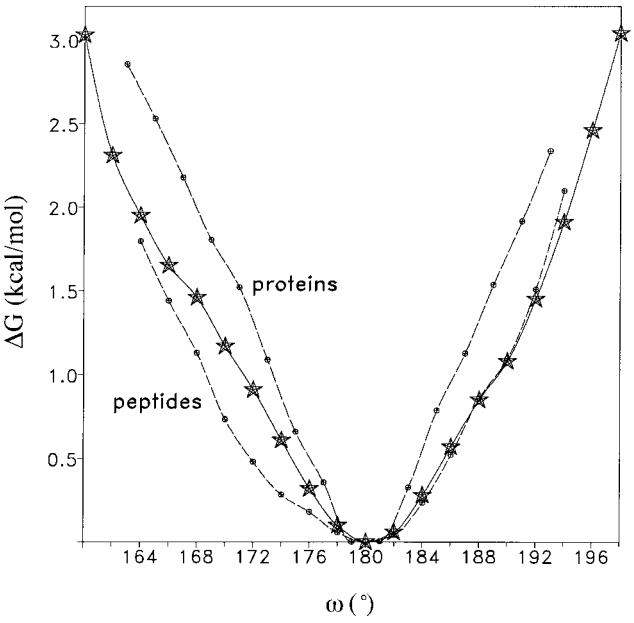


**Figure 5.** Peptide bond energy wells in the region around $\omega = 180°$ derived from the $\omega$ angle distributions and converted to ''energies'' using the Maxwell-Boltzmann equation:

$$n_i = n_0 e^{-\Sigma\Delta E/kT}$$

where $n_i$ is the number of observations in state $i$, $n_0$ is the number of observations in some reference state, $k$ is the Boltzmann constant, $T$ is the temperature of the system and $\Delta E$ is the energy difference between the two energy states. The protein data were obtained from 7953 residues from a set of 85 non-homologous chains which were solved to a resolution of 2.0 Å or better and refined to an $R$-factor no worse than 0.20. The peptide data were taken from 552 *trans* peptide bonds from a combined set of 166 linear and cyclic peptides from the CSD. The atomic resolution set of eight structures is marked by asterisks. The minima have been centred on 180° in order to illustrate the different shapes more clearly.

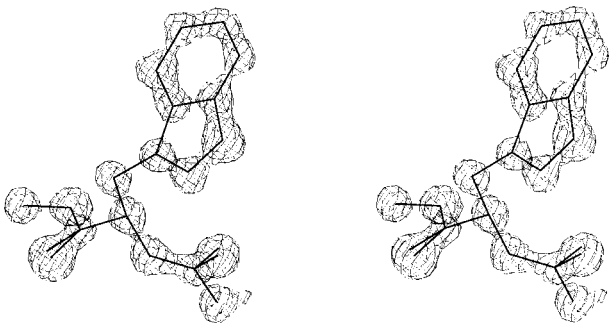

**Figure 6.** The $3F_o - 2F_c$ electron density for the region around Trp48 in ProtG with the model superimposed. Additionally marked are the carbonyl oxygen atoms in the idealised positions where strict planarity of the peptides, i.e. an ω angle of 180°, have been artificially imposed. Those positions lie clearly out of the centre of the density, by about 0.3 Å.

expected distribution of residues on the plot was derived empirically (Morris *et al*., 1992) by dividing ϕ,ψ space into 10° by 10° pixels and counting the number of representatives within each pixel from all protein structures in the October 1990 release of the PDB. The set included low (including those poorer than 2.0 Å) as well as high resolution structures. The core regions were defined as those pixels containing more than 100 residues each, which corresponds to only 14% of ϕ,ψ space. The percentage of a protein's residues within these Ramachandran core regions was found to increase with improving resolution, suggesting it as a useful measure of protein quality. By extrapolation, very high resolution structures are expected to have over 90% of their non-glycine residues in these core regions.

Figure 7 shows (a) the PROCHECK and (b) the SQUID Ramachandran plots for the residues of all eight atomic resolution structures, with the core regions shadowed. The percentage of residues within the core regions is high, 92.1%, reflecting the accuracy of the coordinates (Table 2). The points in the α-helix region in Figure 7 (a) are very tightly clustered extending into the $3_{10}$ helix region diagonally to the top left. The β-strand region also clusters tightly within the core region. The clustering in these two regions is so tight that many of the data points are obscured by others in the plot. Some parts of the core regions are barely occupied, particularly the bottom right-hand section of the α-helix and the upper left of the β-strand region, suggesting that the original Morris *et al*. (1992) divisions may need to be redefined. Another region that appears to need adjustment is the α-left region (labelled L) on the right-hand side of the plot. Here the data points do not cluster within the core region, but rather are more spread out. However, there are too few of these points, coming from just three of the eight structures, to allow definite conclusions to be drawn about this region. Figure 8 shows a plot of individual core percentages as a function of resolution for each of the eight structures, together with those obtained for lower resolution structures in the PDB. This clearly shows that the atomic resolution structures confirm the trend to higher core percentage with increasing resolution.

Although ϕ,ψ angles are not restrained during refinement, their values depend to a certain degree on the value of ω (Ramakrishnan & Balasubramanian, 1972; Balasubramanian & Ramakrishnan, 1972) and thus are to some extent affected by the restraints imposed on the planarity of the ω peptide angle (see above). Deviation from planarity, i.e. values of ω other than 180°, clearly influence details of the core region of the Ramachandran plot. Two other recent analyses (Karplus, 1996; Kleywegt & Jones, 1996) are consistent with the experimental distributions derived from the eight atomic resolution structures and the theoretical distributions.

WHATCHECK includes a procedure for calculating a Ramachandran *Z*-score on the basis of observed ϕ,ψ values classed according to residue type and DSSP secondary structure assignment (Hooft *et al*., 1997). Table 2 gives WHATCHECK ϕ,ψ *Z*-scores, which range from −0.8 to 2.3, except for Ropm, which has an anomalously high score of 3.4. The overall average positive values indicate that these structures confirm the trend for residues in higher resolution structures to cluster more tightly and that this is a useful measure of protein quality. In other words the atomic resolution structures confirm the core regions of the Ramachandran plot previously identified and indeed suggest that these core regions are even tighter than those obtained from the lower resolution structures in the PDB.

### χ angles

Other conformational features not generally restrained are the side-chain χ torsional angles. PROCHECK, SQUID and WHATCHECK all analyse these, the first two reporting the mean and s.u. from the closest rotamer, while the third calculates a *Z*-score for the $\chi_1,\chi_2$ distributions in a series of bins, in a comparable way to its treatment of the ϕ,ψ distributions. Morris *et al*. (1992) found that the $\chi_1$ s.u. showed a good correlation with resolution (Figure 9). The χ angles in these atomic resolution structures are tightly clustered about the three preferred rotameric states (Table 4 summarises the PROCHECK output) as predicted for structures determined at this resolution. This answers, to some extent, the reservations raised by Schrauber *et al*. (1993) about the paucity of data at higher resolutions in the Morris *et al*. (1992) analyses, and shows that the trends reported by the latter do indeed hold at higher resolutions. WHATCHECK's *Z*-score has values between 1.0 and 2.3, significantly better than the mean value of 0.0 with expected deviation of 1.0 for the 300 proteins in the WHATIF (Vriend, 1990) calibration data base.

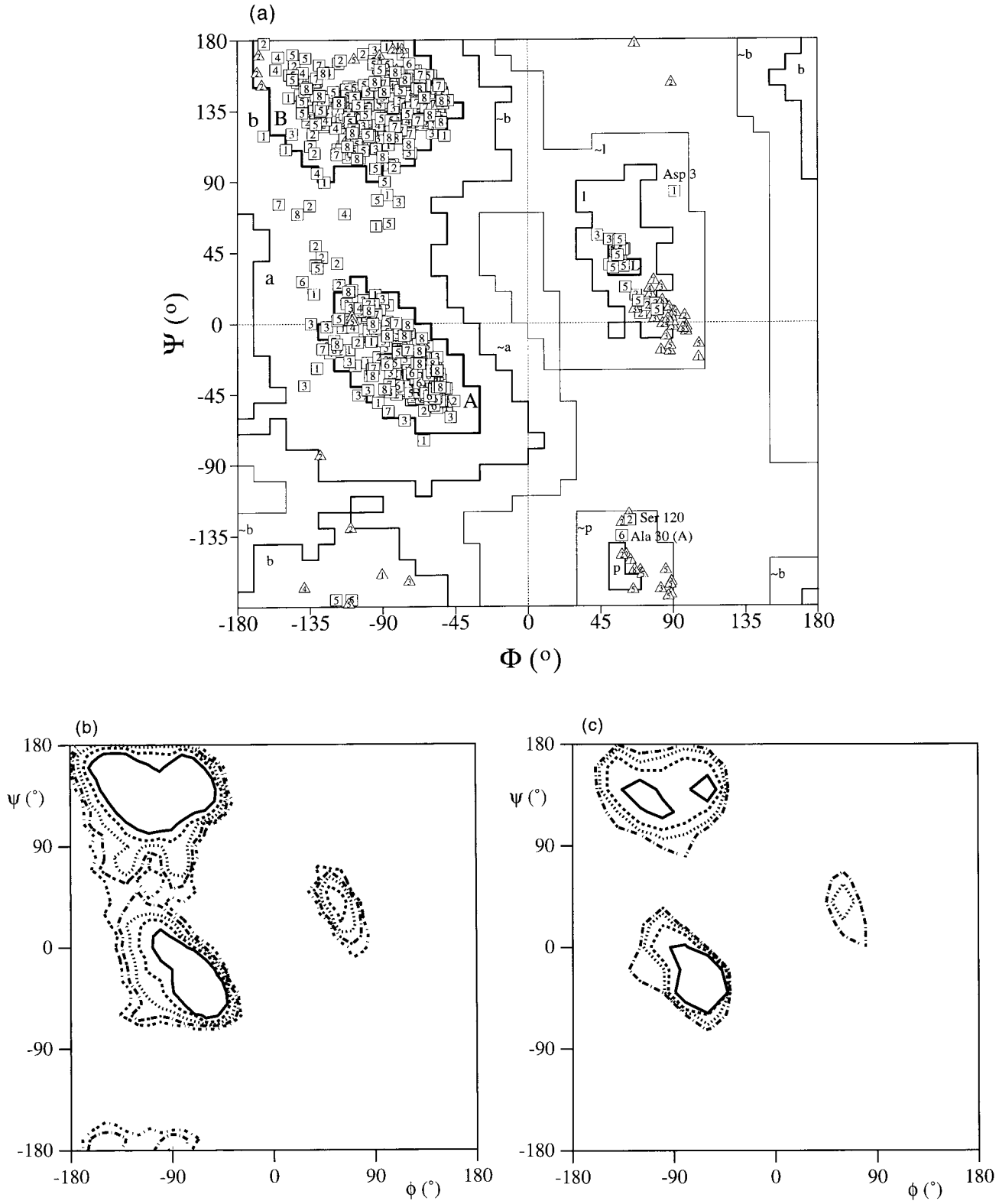


**Figure 7.** (a) Ramachandran plot for the eight atomic resolution structures as derived by PROCHECK. The numbers inside each data point indicate from which structure it comes: 1, Cytc6; 2, Cutinase; 3, Lysozyme; 4, ProtG; 5, RNaseSa; 6, Ropm; 7, RubrDv; 8, RubrCp. Glycine residues are represented by triangles while all other residues are shown by squares. The most favoured core regions are outlined in bold, labelled A for α-helix, B for β-strand and L for α-left. Around them, progressively thinner lines surround the allowed and generous regions, as defined by Morris *et al*. (1992). (b) Ramachandran plots generated using SQUID from 186 proteins in the PDB selected on the basis of: sequence homology <90%, data after 1982, resolution better than 2.0 Å. The data for all the (ϕ,ψ) torsion angles were tabulated and 2-D probability surfaces generated. A bin size of 10 × 10° was used. The plots are contoured at levels corresponding to the number of occurrences: 100, 50, 25, 12, 6. (c) As (b) for the eight atomic resolution structures with a bin size of 20 × 20°. There is a sharpening of features associated with the α-helix and the two peaks of the β-sheet region for the atomic resolution structures. The plots are contoured at levels corresponding to the number of occurrences: 16, 8, 4, 2.

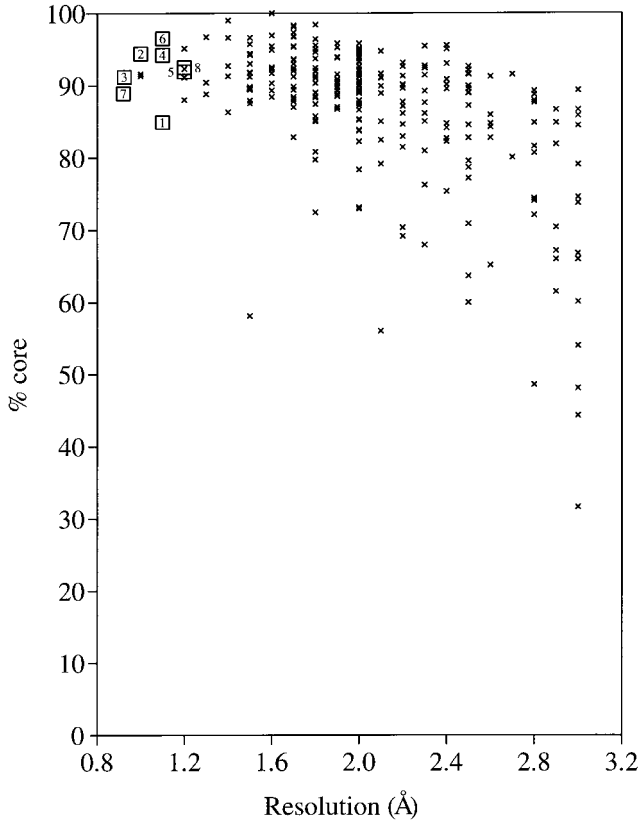


**Figure 8.** Comparison of the percentages of residues in the core of the Ramachandran plot for the eight atomic resolution structures (boxes) with the percentages for 278 lower resolution structures in the PDB. The atomic resolution structures confirm the trends obtained from the lower resolution data with the percentage of core residues increasing with higher resolution.

## Conclusions

Eight protein structures with X-ray data extending beyond 1.2 Å resolution have been assessed using four different validation programs.

### The atomic resolution structures

What have we learned about the practice of the 3-D structure determination?

(1) The atomic resolution structures provide models of unprecedented accuracy and individual coordinate errors can be estimated from the inversion of the least-squares matrix. In the ordered parts of the structures the positional errors are less than 0.03 Å. Not surprisingly for such high resolution structures, no gross errors were found.

(2) A simple problem was identified in the earlier structures in the set, namely errors in the estimated cell dimensions ranging up to more than 0.5%, which led to comparable errors in the model. As a result of this study, in EMBL Hamburg the protocol in the experimental measurement of the cell dimensions has improved the accuracy to better than 0.1%.

(3) The modelling of the solvent structure is still less than ideal. Unit occupancy of the sites was used for several structures, in others the partial occupancy was estimated manually. This is clearly limiting and ways of describing overlapping water networks should be established in the future.

(4) All the analysed models were refined with the same software and with restraint protocols which were not completely transparent. This limited our ability to assess and compare refinement protocols.

### The validation software

What are the conclusions with regard to the validation programs? In summary:

(1) A number of syntax errors were identified which, although seemingly minor, are of great importance when comparing structures. For example, there are two equivalent descriptions of phenylalanine or tyrosine ring orientations and a consistent selection must be made.

(2) Partly as a result of this initiative, the validation programs make more complete use of the information in the coordinate files. For example, the atomic occupancy, space group symmetry and multiple conformations are handled more correctly.

(3) All programs report outliers from various target values and the distribution about them. With a large numbers of atoms, it is not surprising that a significant number of parameters deviate by 3 to 4$\sigma$. Better filtering of the output would be helpful.

(4) In several cases where unusual geometry was observed, there was evidence of two conformations.

(5) Target values are important. However, if the ''geometric'' targets in the Engh & Huber (1991) dictionaries are used in refinement restraints, then the results cannot provide updated validation targets. When a sufficient number of effectively unrestrained atomic resolution structures are available, then the targets may be properly checked.

(6) The results confirm that, for the most part, as the resolution of protein structures is improved the distributions of the conformational parameters become more tightly clustered, with smaller standard uncertainties about their mean values.

(7) On the Ramachandran plot, the atomic resolution structures strongly suggest that the PROCHECK core regions of the plot need to be re-estimated, in general reducing the area of $\phi,\psi$ space that defines them.

(8) One exception to the reduced standard uncertainties is the distribution of $\omega$ torsion angles, which has an increased standard uncertainty (Table 4). Its value of 5.6° appears to be more in line with the variability observed in small peptides than in previous sets of high resolution proteins. This suggests that this might be the value to be

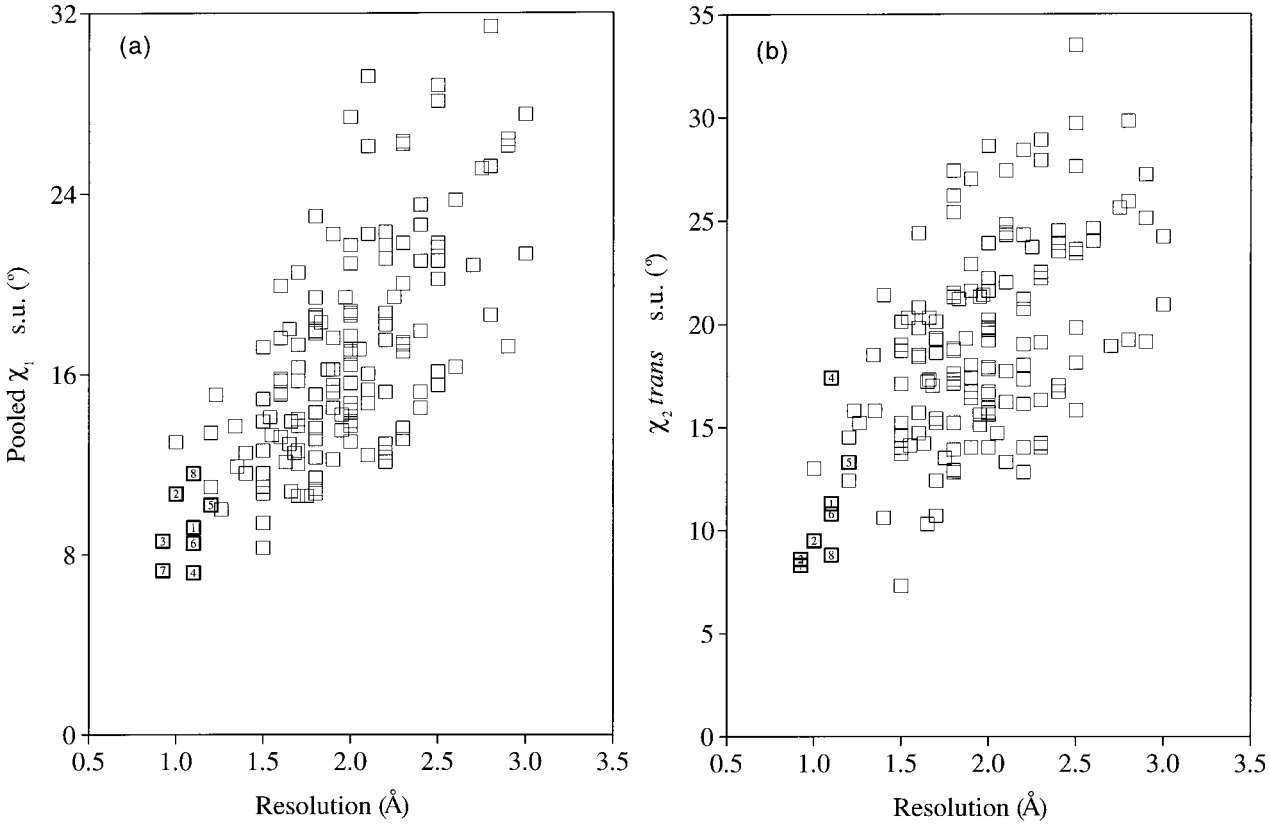


**Figure 9.** Comparisons of the (a) $\chi_1$ and (b) $\chi_2$ s.u.s *versus* resolution for the eight atomic resolution structures (numbered boxes in bold as in Figure 7a) and 278 lower resolution structures in the PDB (empty boxes).

used in refinement and that the previous results were just reflecting the restraint values used. Ideally we should not restrict ω in the refinements used to calibrate this value.

## Future suggestions

As seen from the above, some of the results have already led to improvements in current practice in both areas. In addition there remain further problems to be addressed. What are the suggestions for the future?

(1) The solution to the technical problems in estimating cell dimensions properly must be implemented generally.

(2) All the present data were recorded at EMBL Hamburg and refined with SHELXL. There is a need for a wider range of refinement protocols to be tested and a consideration of data collected at other sources may be relevant.

(3) The eight structures alone give insufficient data for accurately computing the stereochemical parameters and their variability, but, as part of a new collection of structures, will provide improved information at the high resolution end of the protein structures that are currently available. In addition the differing percentage of secondary structure elements and solvent is important: a more representative set of larger proteins is needed and is already becoming available. The expected distributions of a number of parameters should be updated as soon as sufficient atomic resolution structures become available. The most obvious candidates are the allowed regions of the Ramachandran plot and ω angles.

(4) It is important to incorporate some assessment of the relative accuracy of different parts of the structure. This is needed in the reporting provided by both the refinement and validation programs. In addition when setting up the target libraries for refinement and validation, only the ''good'', i.e. well-ordered parts of the structures, with low ADPs should be included.

(5) The formulation of the equations used to impose planar restraints needs further thought. The approach appropriate for ring structures may not be appropriate for peptide planes, which govern the distribution of ω angles.

(6) A number of parameters, e.g. torsion angles, should continue to be unrestrained as they provide the ideal validation tools. This is somewhat similar to the use of $R_{free}$ for cross-validation in reciprocal space.

(7) The validation tools used here only address the problems of the consistency of the coordinates. It is very limiting to assign a global quality indicator to a structure from stereochemical validation alone. Meaningful criteria will have to include information derived from the experimental X-ray

data, e.g. the $R$ and $R_{free}$ factors, precision indicators and the agreement of the model with the electron density. Some of these problems are already being addressed (Brändén & Jones, 1990: Kleywegt & Jones, 1996; Vaguine, A. A., Richelle, R. & Wodak, S. J. unpublished data).

(8) A full and proper treatment of alternative conformations needs to be introduced to the validation suites. This must check contacts for the alternatives, relative occupancies and alternating solvent sites with their occupancies. This is increasingly important as more and more high resolution structures appear.

The close interaction between the experimental and theoretical groups has already led to many improvements in the details of our work practice and this initiative has created a fertile ground for testing new developments both theoretical and experimental. Further developments of both of these will be critically tested.

## Software availability

The software is in general freely available. The PROCHECK source code is copyrighted and is freely available to academic and commercial users: there are restrictions on modifications and sale of the code. PROVE source code is also copyrighted and free to academia with a handling charge of $1000 for commercial users. WHATCHECK is freely available with source code to academia and industry alike. SQUID source code is freely available to academic but not commercial users.

A WWW interface to the PROCHECK, PROVE and WHATCHECK software was created by a collaborative effort of the authors of these programs. The resulting server can be used without restrictions by crystallographers and/or biologists needing a comprehensive structural analysis. The server can be reached at three locations:

http://biotech.embl-heidelberg.de:8400/;
http://biotech.ebi.ac.uk:8400/;
http://biotech.pdb.bnl.gov:8400/.

Details of the individual programs can be obtained *via* the servers.

SQUID and its associated PDBSEL can be obtained from:

http://www.yorvic.york.ac.uk/~oldfield/squid;
http://www.yorvic.york.ac.uk/~oldfield/pdbsel.

## Acknowledgements

This work results from an EU network of laboratories supported by the EC Framework III BIOTECHNOLOGY program, Contract BIO2-CT92-0524 titled ''Integrated Procedures for Recording and Validating results of 3D Structural Studies of Biological Macromolecules''. 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. G.V. thanks the BMFT for support through the RELIWE project.

## References

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). The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallog. sect. B,* **35**, 2331–2339.

Allen, F. H., Kennard, O. & Taylor, R. (1983). Systematic analysis of structural data as a research technique in organic chemistry. *Acc. Chem. Res.* **16**, 146–153.

Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.

Balasubramanian, R. & Ramakrishnan, C. (1972). Stereochemical criteria for polypeptide and protein chain conformations. VIII. Energy maps for a pair of non-planar peptide units having distortion of bond angle at the α-carbon atom. *J. Pept. Protein Res.* **4**, 91–99.

Baumann, A., Frömmel, C. & Sander, C. (1989). Polarity as a criterion in protein design. *Protein Eng.* **2**, 329.

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). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Brändén, C.-I. & Jones, T. A. (1990). Between objectivity and subjectivity. *Nature,* **343**, 687–689.

Brünger, A. T. (1992a). Free *R*-value – a novel statistical quantity for assessing the accuracy of crystal structures. *Nature,* **355**, 472–475.

Brünger, A. T. (1992b). *XPLOR Manual: Version 3. 1,* Yale University, New Haven.

Butterworth, S. (1996). Single crystal X-ray diffraction studies for small, medium and large molecules. D Phil thesis, University of York.

Butterworth, S., Lamzin, V. S., Wigley, D., Derrick, J. & Wilson, K. S. (1998). Anisotropic refinement of a protein G domain at 1. 1 Å resolution. *Acta Crystallog. sect. D,* in the press.

Collaborative Computational Project Number 4 (1994). The CCP4 Suite: Programs for protein crystallography. *Acta Crystallog. sect. D,* **50**, 760–763.

Colovos, C. & Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* **2**, 1511–1519.

Cremer, D. & Pople, J. A. (1975). A general definition of ring puckering coordinates. *J. Am. Chem. Soc.* **97**, 1354–1358.

Cruikshank, D. (1996). Protein precision re-examined: Luzzati plots do not estimate final errors. Proceedings of the CCP4 study weekend, 4-5 January 1996, DL/SCI/R35 (Dodson, E., Moore, M., Ralph, A. & Bailey, S., eds), pp. 11–22, Daresbury Laboratory, Warrington, UK.

Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1995). Proteins at atomic resolution. *Curr. Opin. Struct. Biol.* **5**, 784–790.

Dauter, Z., Wilson, K. S., Sieker, L. C., Moulis, J.-M. & Meyer, J. (1996). Zinc- and iron-rubredoxins from

*Clostridium pasteurianum* at atomic resolution: the first high precision model of a $ZnS_4$ coordination unit in a protein. *Proc. Natl Acad. Sci. USA,* **93**, 8836–8840.

Engh, R. A. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallog. sect. 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). *Ab initio* determination of the crystal structure of cytochrome c6 ; comparison with plastocyanin. *Structure,* **3**, 1159–1169.

Holm, L. & Sander, C. (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**, 93–105.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996a). The PDBFINDER data base: a summary of PDB, DSSP and HSSP information with added value. *CABIOS,* **12**, 525–529.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996b). Verification of protein structures: side-chain planarity. *J. Appl. Crystallog.* **29**, 714–716.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996c). Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins: Struct. Funct. Genet.* **23**, 363–376.

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

Hooft, R. W. W., Sander, C. & Vriend, G. (1997). Objectively judging the quality of a protein structure from a Ramachandran plot. *CABIOS,* **13**, 425–430.

Hooft, R. W. W. & Vriend, G. (1996). Improved coordinate reconstruction from stereo diagrams. *J. Mol. Graphics,* **14**, 168–172.

IUPAC-IUB Commission on Biochemical Nomenclature (1970). Abbreviations and symbols for the description of the conformation of polypeptide chains. *J. Mol. Biol.* **52**, 1–17.

Jones, D. T. & Thornton, J. M. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.* **6**, 210–216.

Jones, D. T., Miller, R. T. & Thornton, J. M. (1995). Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins: Struct. Funct. Genet.* **23**, 387–397.

Juvenal (117). *Satires,* 2(VI), line 347.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bond and geometrical features. *Biopolymers,* **22**, 2577–2637.

Karplus, P. A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci.* **7**, 1406–1420.

Kleywegt, G. J. & Jones, T. A. (1996). Phi/psi-chology: Ramachandran revisited. *Structure,* **4**, 1395–1400.

Lamzin, V. S., Dauter, Z. & Wilson, K. S. (1995). Dictionary of protein stereochemistry. *J. Appl. Crystallog.* **28**, 338–340.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993a). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallog.* **26**, 283–291.

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

Lemer, C. M. R., Rooman, M. D. & Wodak, S. J. (1995). Protein structure prediction by threading methods: evaluation of current techniques. *Proteins: Struct. Funct. Genet.* **23**, 337–355.

Longhi, S., Czjzek, M., Lamzin, V., Nicolas, A. & Cambillau, C. (1997). Atomic resolution (1.0 Å) crystal structure of *Fusarium solani* cutinase: stereochemical analysis. *J. Mol. Biol.* **268**, 779–799.

Luthardt, G. & Frömmel, C. (1994). Local polarity analysis: a sensitive method that discriminates between native proteins and incorrectly folded models. *Protein Eng.* **7**, 627–631.

Luthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature,* **356**, 83–85.

MacArthur, M. W. & Thornton, J. M. (1996). Deviations from planarity of the peptide bond in peptides and proteins. *Protein Eng.* **8**, 217–224.

MacArthur, M. W., Laskowski, R. A. & Thornton, J. M. (1994). Knowledge-based validation of protein-structure coordinates derived by X-ray crystallography and NMR-spectroscopy. *Curr. Opin. Struct. Biol.* **4**, 731–737.

McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen-bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Struct. Funct. Genet.* **12**, 345–364.

Novotny, J., Rashin, A. A. & Bruccoleri, R. E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins: Struct. Funct. Genet.* **4**, 19–30.

Oldfield, T. J. (1992). SQUID: A program for the analysis and display of data from crystallography and molecular dynamics. *J. Mol. Graphics,* **10**, 247–252.

Parkinson, G., Voitechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. (1996). New parameters for the refinement of nucleic acid-containing structures. *Acta Crystallog. sect. D,* **52**, 57–64.

Pontius, J., Richelle, J. & Wodak, S. (1996). Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.* **264**, 121–136.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99.

Ramakrishnan, C. & Balasubramanian, R. (1972). Stereochemical criteria for polypeptide and protein chain conformations. VII. Effect of non-planarity and bond angle distortion at the α-carbon atom on the contact map for a pair of peptide units. *J. Pept. Protein Res.* **4**, 79–90.

Schrauber, H., Eisenhaber, F. & Argos, P. (1993). Rotamers: To be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J. Mol. Biol.* **230**, 592–612.

Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). Ribonuclease from *Streptomyces aureofaciens* at atomic resolution. *Acta Crystallog. sect. D,* **52**, 327–344.

Sheldrick, G. M. & Schneider, T. R. (1997). SHELXL: high resolution refinement. **277**, 319–343.

Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins: Struct. Funct. Genet.* **17**, 355–362.

Vajda, S., Sippl, M. & Novotny, J. (1997). Empirical potentials and functions for protein folding and binding. *Curr. Opin. Struct. Biol.* **7**, 222–228.

Vlassi, M., Dauter, Z., Wilson, K. S. & Kokkinidis, M. (1998). 1.07 Å structure of a designed variant of the ColE1 Rop protein. *Acta Crystallog. sect. D,* in the press.

Voronoi, G. F. (1908). Nouvelles applications des paramétres continus à la théorie des forms quadratiques. *J. Reine Angew. Math.* 198–287.

Vriend, G. (1990). WHATIF: a molecular modelling and drug design program. *J. Mol. Graphics,* **8**, 52–56.

Vriend, G. & Sander, C. (1993). Quality control of protein models: directional atomic contact analysis. *J. Appl. Crystallog.* **26**, 47–60.

Vriend, G., Rossmann, M. G., Arnold, E., Luo, M., Griffith, J. P. & Moffat, K. (1986). Post-refinement of oscillation diffraction data collected at a synchrotron radiation source. *J. Appl. Crystallog.* **19**, 134–139.

Walsh, M. A., Schneider, T. R., Sieker, L. C., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1998). Refinement of triclinic lysozyme at atomic resolution. *Acta Crystallog. sect. D,* in the press.

***Edited by I. A. Wilson***

*(Received 21 March 1997; received in revised form 1 October 1997; accepted 3 October 1997)*