Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them?

Mariusz Jaskolski,[a] Miroslaw Gilski,[a] Zbigniew Dauter[b] and Alexander Wlodawer[c]*

[a]Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University and Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland, [b]Synchrotron Radiation Research Section, Macromolecular Crystallography Laboratory, NCI, Argonne National Laboratory, Biosciences Division, Building 202, Argonne, IL 60439, USA, and [c]Protein Structure Section, Macromolecular Crystallography Laboratory, NCI at Frederick, Frederick, MD 21702, USA

Correspondence e-mail: wlodawer@ncifcrf.gov

Received 20 December 2006
Accepted 28 February 2007

The Protein Data Bank and Cambridge Structural Database were analyzed with the aim of verifying whether the restraints that are most commonly used for protein structure refinement are still appropriate 15 years after their introduction. From an analysis of selected main-chain parameters in well ordered fragments of ten highest resolution protein structures, it was concluded that some of the currently used geometrical target values should be adjusted somewhat (the C—N bond and the N—C^{α}—C angle) or applied with less emphasis (peptide planarity). It was also found that the weighting of stereochemical information in medium-resolution refinements is often overemphasized at the cost of the experimental information in the diffraction data. A correctly set balance will be reflected in root-mean-square deviations from ideal bond lengths in the range 0.015–0.020 Å for structures refined to *R* factors of 0.15–0.20. At ultrahigh resolution, however, the diffraction terms should be allowed to dominate, with even higher acceptable deviations from idealized standards in the well defined fragments of the protein. It is postulated that modern refinement programs should accommodate variable restraint weights that are dependent on the occupancies and *B* factors of the atoms involved.

1. Introduction

During the first two or three decades after the structures of hemoglobin (Perutz *et al.*, 1960) and myoglobin (Kendrew *et al.*, 1960) were solved, protein crystallography was mostly practiced by scientists highly trained in the application of this technique. However, the situation has changed markedly in the last 15–20 years. A proliferation of synchrotron facilities has made data collection much easier and accessible even to beginning students, while the introduction of methods such as MAD (Hendrickson *et al.*, 1990) and SAD (Wang, 1985; Wang *et al.*, 2000; Dauter *et al.*, 2002), coupled with widespread use of integrated software packages such as *CCP*4 (Collaborative Computational Project, Number 4, 1994), *SHELX* (Sheldrick, 1998), *SHARP* (de La Fortelle & Bricogne, 1997), *SOLVE*/*RESOLVE* (Terwilliger, 2003), *CNS* (Brünger *et al.*, 1998) and *HKL*-3000 (Minor *et al.*, 2006), just to name a few, has eased the process of structure solution. With protein crystallography becoming more routine and automated, there is a tendency to rely on a set of standardized procedures, often without the participation of experienced crystallographers. Although in general this might be a positive trend, structural investigations are sometimes still less than straightforward and require nonstandard approaches to assure success (Dauter *et al.*, 2005). However, in the cases when the structures are successfully solved, they still need to be refined.

Table 1
Bond-length statistics for peptide structures deposited in the CSD.

For each main-chain bond, the sample mean and standard deviation (in Å) are given in the upper row. The lower row gives the sample size/number of structures (in parentheses) in each $R1$ range, which, from $R1 \leq 0.050$ to $R1 \leq 0.100$, include increasing numbers of less accurate structures.

$R1$ limit	$R1 \leq 0.050$	$R1 \leq 0.075$	$R1 \leq 0.100$
N—C$^{\alpha}$†	1.455 (7)	1.455 (12)	1.456 (19)
	(231/124)	(519/226)	(722/278)
C$^{\alpha}$—C‡	1.523 (11)	1.524 (17)	1.523 (25)
	(146/81)	(513/202)	(749/255)
C—N§	1.332 (8)	1.333 (12)	1.333 (17)
	(348/141)	(739/256)	(992/310)
C═O	1.231 (9)	1.230 (12)	1.230 (15)
	(480/157)	(1039/285)	(1361/343)

† Excluding glycine and proline residues. ‡ Excluding glycine residues. § Excluding Aaa-Pro peptides.

With a few rare exceptions, all macromolecular refinement procedures utilize standard stereochemical information (Evans, 2007), since the observation-to-parameter ratios are usually considered to be insufficient for unrestrained refinement. The restraint targets are derived primarily from very high resolution structures of small molecules. Initially, X-ray and neutron diffraction structures of individual amino acids were utilized for this purpose in programs such as *PROLSQ* (Wlodawer & Hendrickson, 1982; Hendrickson, 1985), but the restraints were later improved on the basis of large databases. Almost universally, the currently used refinement programs, such as *CNS* (Brünger *et al.*, 1998), *SHELXL* (Sheldrick & Schneider, 1997) and *REFMAC*5 (Murshudov *et al.*, 1997), use the parameters compiled over 15 y ago by Engh & Huber (1991) and subsequently updated by the same authors (Engh & Huber, 2001). These parameters were obtained by careful analysis of the Cambridge Structural Database (CSD; Allen, 2002). Although there is no compelling reason to suspect that extensive modifications to the refinement targets are required, a fresh look at them is warranted, especially taking into account that nearly 35 000 protein crystal structures have been deposited in the Protein Data Bank (PDB; Berman *et al.*, 2000) since these parameters were first introduced. Indeed, the number of atomic resolution protein structures (834 in December 2006), as defined by the 1.2 Å criterion (Sheldrick, 1990; Morris & Bricogne, 2003), exceeds the total number of PDB deposits (709) in January 1991.

In addition, a practical question to ask is 'How much deviation from idealized geometrical target values should be allowed in properly refined structures?' Surprisingly, this question is still asked quite often and in our experience the answer is not always quite correct. Although a number of previous studies have addressed the problem of the assessment of the quality of protein crystal structures (Kleywegt & Jones, 1995; Dodson *et al.*, 1996; EU 3-D Validation Network, 1998), we are not aware of a single reference that would answer this question in an unambiguous way. At best, suggestions such as 'The molecular geometry will be restrained with r.m.s.d.s of 0.01–0.02 Å on bond lengths, 2–4° on bond angles and 2–4° on improper dihedrals' are given without full explanation of how these choices were made. The overall level of the restraint weights can be validated by the use of the free R factor (Brünger, 1992, 1997). However, being a global parameter based on reflection amplitudes, $R_{\rm free}$ is not well suited for checking whether some individual selected geometrical features within the refined model are correct. A properly refined protein model should optimally predict the experimental structure-factor amplitudes and its geometrical features should correspond to the expected stereochemically reasonable targets. It is not trivial to satisfy both requirements simultaneously and neither should be sacrificed at the expense of the other.

Thus, the aim of this paper is twofold. Firstly, we analyzed the current holdings in both the PDB and CSD in order to check whether the stereochemical targets should be adjusted based on the additional data accumulated over the last 15 years. We found that although most of them do not need to be changed, some do require at least minor adjustments, even on top of the corrections introduced by Engh & Huber (2001), which were based on the CSD only and did not utilize the contents of the PDB. Secondly, based on the results of this analysis and of a number of previous analyses of the accuracy of protein crystal structures, we attempted to define rational values for the r.m.s. deviations of the refined parameters from their idealized targets, concluding that in many cases the restraints are unnecessarily tight in the well behaving parts of the macromolecule, whereas the more flexible or disordered fragments require more stringent restraining to enforce acceptable stereochemistry. In this spirit, we postulate that modern refinement programs should accommodate variable restraint weights that are dependent on the occupancies and B factors of the atoms involved.

It is not our aim in this paper to provide a comprehensive analysis of this complicated subject, but rather to indicate some practical guidelines. In this respect, we present a cookbook, with the intended audience being the cooks rather than the chefs.

2. Methods

This work was based on the PDB database release of 22 August 2006 (38 320 total structures, of which ∼34 000 were proteins). For the statistical analyses, the PDB structures were divided into the following resolution classes: 0.54–0.8, 0.8–0.9, 0.9–1.0, 1.0–1.1, 1.1–1.2, 1.2–1.3, 1.3–1.4, 1.4–1.5, 1.5–1.6, 1.6–1.7 and 1.7–1.8 Å. Only those entries that contained protein and had $R \leq 0.16$ (highest quality) were selected (see supplementary Table 1[1]). With the exception of the 0.54–0.8 Å resolution range, structures were rejected if they were reported without $R_{\rm free}$ (presumably old and possibly not up to the current standard). Structures were selected at random but with a preference for the low-$R_{\rm free}$ group to accumulate about

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: WD5076). Services for accessing this material are described at the back of the journal.

3000–5000 instances of a given parameter (about 1000 in the 0.54–0.8 Å range). The PDB entries were selected using the 'Advanced Search' PDB tool and their geometrical parameters were calculated using the 'Geometry' option or in *SHELXL*. For each parameter under investigation, the average value and sample standard deviation were calculated using *OpenOffice* and *MS Excel* tools.

The ten ultrahigh-resolution (defined as higher than 0.8 Å) structures used in this study were crambin (PDB code 1ejg; Jelsch *et al.*, 2000), subtilisin (1gci; Kuhn *et al.*, 1998), α-conotoxin (1hje; not published beyond deposition of coordinates), the PDZ2 domain of synthenin (1r6j; Kang *et al.*, 2004), antifreeze protein RD1 (1ucs; Ko *et al.*, 2003), aldose reductase (1us0; Howard *et al.*, 2004), PAK pilin (1x6z; Dunlop *et al.*, 2005), rubredoxin (1yk4; Bönisch *et al.*, 2005), hydrophobin HFBII (2b97; Hakanpää *et al.*, 2006) and a D,L-α1 designed peptide (3al1; Patterson *et al.*, 1999). All these structures were characterized by *R* factors of 0.14 or lower, with $R_{\rm free}$ not exceeding 0.16. The geometrical parameters discussed here were derived from only the well ordered regions, which were defined as having single conformation and all atomic isotropic *B* values below 40 Å^2. The threshold of 40 Å^2 was selected arbitrarily, according to our experience showing that fragments with higher *B* factors tend not to have confidently refined positional and displacement parameters and often display unacceptable stereochemistry. For comparison, corresponding sets of geometrical parameters were separately estimated for all atoms without any screening for disorder.

Average deviations of bond lengths from their target values were evaluated for structures refined at 1.0 Å or higher resolution, for structures at 1.5 Å and at just beyond 2 Å. The deviations of bond lengths from their targets reported for structures in the relevant resolution ranges were extracted from the PDB using a variety of keywords (since they are not coded in a consistent way). The resulting data were curated by hand in order to remove the sets that did not report any r.m.s. deviations for bond lengths or those that were clearly in error. Since the number of structures at exactly 2 Å resolution exceeded 2500, we utilized the range 2.02–2.08 Å instead, which yielded ∼500 structures.

Our analysis of peptide parameters in small-molecule structures was based on the CSD release of May 2006 (380 864 structures). Structures were selected, retrieved and analyzed using the *CCDC* software distributed with the database. Firstly, structures of peptides composed of α-amino acids were selected, excluding cyclic peptides, metal complexes and structures with disorder or with evident errors. No special attempt was made to select only L-forms or to limit the search to protein amino acids. To check the robustness of the results, statistics of the main-chain bond distances were calculated for structures in different *R*-factor categories, namely with $R1 \leq 0.050$, $R1 \leq 0.075$ and $R1 \leq 0.100$ ($R1$ is the conventional linear residual defined as $R1 = \sum \big||F_{\rm o}| - |F_{\rm c}|\big| / \sum |F_{\rm o}|$). In very few isolated cases, individual structures were deleted from a subset when the data points contributed by them were conspicuous outliers and were internally inconsistent (*i.e.* they appeared as low-end as well as high-end outliers). The statistics for the C$^\alpha$—C bond excluded the C-terminal residues and similarly N-terminal residues were excluded from the N—C$^\alpha$ statistics.

3. Results

3.1. Engh and Huber parameters and their application

Almost all currently used refinement programs utilize the Engh and Huber (EH) parameters (Engh & Huber, 1991, 2001) to define the targets for geometrical restraints. These parameters were derived from analysis of the CSD, with bond lengths defined for 59 different types of interatomic distances and bond angles for 108 bond pairs. Each parameter was accompanied by a standard deviation, varying for bond lengths from 0.010 to 0.059 Å for different bond types and from 1.0 to 5.0° for bond angles. Although Engh and Huber proposed to use these data as the basis for parameterization of force constants, they did not directly address the question of how much overall deviation from the target values should be expected in the refined structures. The values of less than 0.02 Å for the standard deviations of bond lengths and 2° for bond angles have been attributed to Hendrickson (1985), although the latter value is most likely misquoted, since early *PROLSQ* did not utilize bond angles as refinement targets. Other programs use similar default targets, for example *SHELXL* (0.02 Å; Sheldrick & Schneider, 1997) and *REFMAC*5 (0.021 Å; Murshudov *et al.*, 1997).

The standard deviations that accompany the original EH parameters reflect the intrinsic variation of these parameters in the selected small-molecule structures in the CSD, as well as uncertainties resulting from the limited samples. Although the average values of the standard deviations ascribed to different

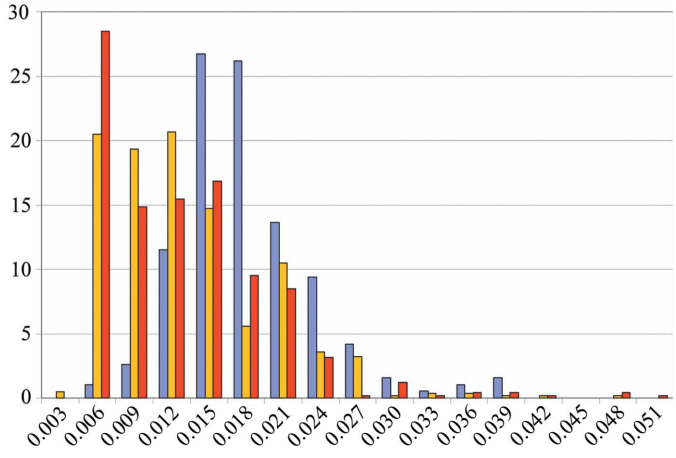

Figure 1
Distribution (%) of r.m.s. deviations from bond-distance targets (Å) reported in PDB-deposited structures determined at 2.0 Å (red), 1.5 Å (orange) and higher than 1 Å (blue) resolution. About 500 randomly selected PDB structures were used at the 2.0 and 1.5 Å resolution ranges. In the 0.54–1.0 Å resolution range, all 191 structures with reported r.m.s.d.s for bonds were included. The value ranges (and mean values) for the <1.0, 1.5 and 2.0 Å sets are 0.006–0.038 (0.017), 0.001–0.048 (0.012) and 0.004–0.053 (0.012) Å, respectively.

classes of bonds or angles may not be strictly valid in the statistical sense, they may nevertheless provide an indication of the expected accuracy of the whole set. Such averages, calculated by us from the data presented in Tables 2 and 3 of the original Engh and Huber paper, are 0.022 Å for bond lengths and 1.85° for bond angles, without application of any weights that would represent the relative frequency of different classes of bonds and angles.

In most refinement programs the relative weights for different parameter categories (bonds, angles *etc.*) are set within the program as defaults, although it is usually possible to change these default values. Typically, a user needs at most to adjust an overall weight of stereochemical information relative to the diffraction terms. This balance is often adjusted not quite correctly by giving too much weight to stereochemical data and in effect leading to regularization rather than optimization of the structure. This is illustrated by unusually small reported r.m.s.d. values, sometimes several times lower than the recommended library values. Some refinement programs tend to drive the r.m.s.d. values very low when used with the default weighting schemes and at medium resolution (see below).

3.2. R.m.s. deviations of bond lengths from target values in PDB structures

We have evaluated the PDB structures falling into three resolution ranges in order to find the average values of the r.m.s. deviations of bond lengths reported in the structures solved to date. For this purpose, we have analyzed all protein structures at a resolution of 1.0 Å or higher (191 structures), as well as structures near 1.5 Å resolution (590 structures) and near 2.0 Å resolution (505 structures). The r.m.s.d. values reported for different structures ranged from 0.0012 Å (a case at 1.5 Å) to 0.053 Å (a case at 2.0 Å), *i.e.* departing at both ends far from the values dictated by experience and even common sense. Another curiosity detected in the PDB is the reporting of r.m.s.d. values with exuberant precision. In two cases, r.m.s. deviations from bond targets were reported as 0.004361 Å, *i.e.* with precision better than one-tenth of the radius of an electron! The distributions of the r.m.s. deviations in the three resolution ranges are shown in Fig. 1. As expected, the average value of the r.m.s. deviations was the highest (0.017 Å) for the atomic resolution structures, but the average was the same (0.012 Å) for both ranges of lower resolution. However, the actual distribution of the deviations was quite different in the three resolution ranges, only approaching a Gaussian (with a long-end tail) for the highest resolution structures. At both 1.5 and 2 Å resolution the most highly populated intervals were found on the lower side of the spectrum, indicating that the geometry of the models was determined more by the restraints than by the diffraction data. The effect is especially pronounced at 2 Å, where nearly 30% of the structures are reported with r.m.s. deviations from idealized bond lengths within 0.006 Å.

An analysis of the r.m.s. deviations from bond-length targets shows a clear correlation of this parameter with the program used for structure refinement. We have calculated the average of the r.m.s. deviations of bond lengths for four sets, each consisting of 25 structures refined at resolutions between 1.95 and 2.05 Å with *CNS*, *REFMAC*, *SHELXL* and *TNT*. We also calculated the average *R* factors to check whether different programs seem to be yielding structures of different quality. The structures were selected at random from among the most recent structures available in the PDB, avoiding the inclusion of redundant entries. Although we did not intend to provide a full analysis of the spread of the r.m.s. deviations from bond-length targets, the results are instructive (see Supplementary Table 2). The average values of the *R* factors are quite comparable, ranging from 0.189 for *TNT* to 0.202 for *REFMAC*. However, the spread of the mean r.m.s. deviations of bond lengths is much larger. Structures refined with *CNS* and *SHELXL* seem to be much closer to the EH targets, with r.m.s.d. (bonds) of 0.0085 Å (range 0.005–0.023 Å) and 0.0090 Å (range 0.004–0.029 Å), respectively. *TNT* structures are intermediate, with deviations of 0.0130 Å (range 0.005–0.026 Å), whereas *REFMAC* deviations are 0.0165 Å (range 0.005–0.027 Å), about twice those of *CNS*. These results are not surprising, since we suspect that most of the refinements were performed with default weights, which tend to be much tighter in *CNS* than in *REFMAC*. This confirms our experience that structures refined with *CNS* at medium resolution and with default weights tend to end up with r.m.s. deviations for bond lengths of ∼0.006 Å. Although the default target r.m.s. bond deviation in *SHELXL* is 0.02 Å, the structures refined at 2 Å resolution seem to be more idealized, most likely through the influence of the overall weight of the diffraction terms, which is more attuned to refinement at higher resolution, where this program is used more often. To summarize, we have shown that the average r.m.s. deviations of bond lengths are correlated to the programs used to refine protein structures, although the spread of these parameters among individual structures can be quite large.

3.3. Sample bond-length data from current CSD

Although we have made no special attempt to faithfully reproduce the procedures of data selection and statistical analysis employed by Engh and Huber, the data in Table 1 can generally be considered to represent a choice of protein structure parameters (limited to main-chain bond distances) evaluated from a substantially expanded CSD database. Comparison of the results obtained with systematically enlarged (but concomitantly less accurate, as measured by *R*1) subsets of structures shows that the corresponding mean values remain practically unchanged. The sample standard deviations, however, increase with the inclusion of less accurate structures, indicating an inflated scatter of values as the quality of the data deteriorates. For practical purposes, we would recommend using as most representative the values obtained for $R1 \leq 0.075$ (Fig. 2), where the standard deviations are still not inflated by lack of accuracy but the samples are sufficiently large (at least 500 entries). Comparing the standard deviations obtained for the individual bond lengths,

Table 2
Bond-length statistics for the highest resolution structures in the PDB (resolution higher than 0.8 Å).

The calculations (in *SHELXL*; Sheldrick & Schneider, 1997) were carried out twice. Firstly, all atoms with $B_{iso} \geq 40$ Å^2 or in multiple-conformation fragments were excluded (first row). Next, all protein atoms were included without any screening (second row). Except for R/R_{free}, all numerical values are in Å.

	1ejg	1ucs	1us0	1yk4	1r6j	1hje	3al1	2b97	1gci	1x6z	⟨*d*⟩ (s.d.)†	EH‡
Resolution	0.54	0.62	0.66	0.69	0.73	0.75	0.75	0.75	0.78	0.78		
R	0.090	0.137	0.094	0.100	0.075	0.127	0.130	0.130	0.099	0.143		
R_{free}	0.094	0.155	0.103	0.108	0.087	§	0.145	0.148	0.103	0.157		
R.m.s.d. (*d*)¶	0.011	0.014	0.014	0.018	0.012	0.011	0.016	0.027	0.014	0.017		
	0.022	0.015	0.067	0.022	0.016	0.023	0.036	0.028	0.016	0.019		
	0.023	0.012	0.016	0.026	0.019	—	0.038	0.027	0.012	—		
N—C$^\alpha$††	1.456 (11)	1.455 (10)	1.453 (9)	1.462 (9)	1.454 (8)	1.453 (6)	1.454 (8)	1.455 (23)	1.457 (9)	1.449 (13)	1.454 (12)	1.458 (19)
	1.460 (28)	1.459 (10)	1.455 (8)	1.467 (25)	1.455 (9)	1.463 (26)	1.454 (8)	1.455 (23)	1.457 (10)	1.449 (13)	1.456 (15)	
C$^\alpha$—C‡‡	1.530 (9)	1.528 (10)	1.524 (10)	1.534 (11)	1.525 (10)	1.524 (13)	1.519 (8)	1.532 (21)	1.528 (10)	1.523 (11)	1.527 (13)	1.525 (21)
	1.529 (13)	1.528 (11)	1.524 (11)	1.531 (22)	1.525 (9)	1.517 (22)	1.519 (8)	1.532 (22)	1.527 (13)	1.523 (11)	1.526 (14)	
C—N§§	1.337 (9)	1.337 (11)	1.334 (10)	1.337 (12)	1.334 (9)	1.330 (9)	1.333 (8)	1.334 (35)	1.336 (10)	1.331 (12)	1.334 (13)	1.329 (14)
	1.336 (10)	1.336 (12)	1.333 (15)	1.336 (12)	1.333 (13)	1.334 (11)	1.333 (8)	1.333 (38)	1.336 (11)	1.331 (12)	1.334 (18)	
C═O	1.234 (7)	1.235 (12)	1.231 (9)	1.240 (10)	1.234 (10)	1.229 (9)	1.229 (6)	1.229 (18)	1.237 (9)	1.236 (13)	1.234 (12)	1.231 (20)
	1.237 (15)	1.235 (12)	1.231 (11)	1.240 (17)	1.234 (10)	1.233 (15)	1.229 (6)	1.229 (18)	1.237 (9)	1.236 (13)	1.234 (13)	
Long (+)/ short (−)¶¶				+		−	−			−		
Remarks	*MOLLY*†††					Unrestrained L/D aa‡‡‡				*REFMAC*§§§		

† Bond length averaged over all structures, with sample standard deviation in parentheses. ‡ Stereochemical targets (and standard deviations) of Engh & Huber (1991). § R_{free} test was not used. ¶ The third row shows the r.m.s. deviation from targets for bonds as reported in the PDB entry. †† Excluding glycine and proline residues. ‡‡ Excluding glycine residues. §§ Excluding Aaa-Pro peptides. ¶¶ Ordering of the structures from cases where the bond distances are systematically the longest (+) to cases where they are systematically the shortest (−). ††† Deformation density study: refinement was carried out with *MOPRO* (Jelsch *et al.*, 2005), a version of *MOLLY* (Hansen & Coppens, 1978). ‡‡‡ Centrosymmetric structure composed of L- and D-amino acids. §§§ The only case, except 1ejg, of structure refinement with *REFMAC* (Murshudov *et al.*, 1997); all other structures were refined with *SHELXL* (Sheldrick & Schneider, 1997).

one notices that the values for the C$^\alpha$—C bonds are systematically higher (by about 50%) than for the remaining three main-chain bonds. Within the set of EH parameters this trend was much less pronounced and instead the C—N bond was characterized by a reduced scatter (see last column of Table 2).

3.4. Main-chain bond distances from ultrahigh-resolution protein structures

The most accurately determined protein structures in the PDB, refined at ultrahigh resolution (here defined as higher than 0.8 Å), provide a wealth of information about the expected deviations of the geometrical parameters from their assumed target values. For the purpose of our analysis, we will concentrate primarily on the main-chain geometry of well ordered fragments, defined as single-conformation models with atomic (equivalent) $B_{iso} < 40$ Å^2. With the main chain being usually the best determined part of any structure, the bond lengths and angles, as well as their standard deviations, observed in the ordered parts should represent the 'best-case' scenario for the definition of the restraints.

Table 2 reports the main-chain bond lengths in ultrahigh-resolution PDB structures (also summarized in Fig. 3). Several general remarks are possible. Firstly, it is surprising that at such high resolution some of the structures appear to be less well refined than many structures at less 'ultra' atomic resolution. The presence of structures 1ejg and 1hje in this set is of special significance, because the former represents an extremely careful study at nearly the ultimate resolution (0.54 Å) aimed at mapping deformation density distribution, with stereochemical restraints applied in the disordered parts of the structure, while the latter illustrates unrestrained protein structure refinement and thus might provide some hints about the restraint target values themselves. The mean values for the individual main-chain bonds are systematically higher for 1yk4 and indeed the set-wide averages almost uniformly fall in between the two extremes determined for 1hje and 1yk4. This may indicate unit-cell scaling problems in some of these data sets (see below).

Structure 3al1 is characterized by unusually narrow distributions of the listed bond distances. This is not a result of over-restraining, as the r.m.s. deviation from idealized values is relatively high (Table 2). Such narrow distributions of observations could provide very precise numbers, but unfortunately this structure probably suffers from a systematic error in unit-cell determination (see below). Comparison of the last two columns of Table 2 indicates a non-uniform relationship between the currently used EH targets and the intrinsic properties of protein structures. The average C$^\alpha$—C bond is practically identical to the target, with lower variance. The largest departure from the EH standards is found for the amide C—N bond and this case will be discussed separately. Strangely, the carbonyl C═O bond has a narrow distribution somewhat above the target value, despite the relatively large standard deviation (0.02 Å) ascribed to this target by Engh and Huber.

3.5. Sample main-chain bond angle

It has been indicated several times (Esposito, Vitagliano, Sica *et al.*, 2000; Esposito, Vitagliano, Zagari *et al.*, 2000; Addlagatta *et al.*, 2001) that the N—C$^\alpha$—C valence angle has a wide spread and may have a bimodal distribution correlated with secondary structure. Table 3 shows an analysis of the

Table 3
Distribution of the N—C^{α}—C bond angles (°) (separately for glycine, proline and all other residues) in protein structures determined at resolution higher than 0.8 Å.

Type	Gly	Pro	Other
EH	112.50, σ = 2.90	111.80, σ = 2.50	111.20, σ = 2.80
PDB†	113.91, σ = 2.23	112.48, σ = 2.19	110.72, σ = 2.22
PDB‡	113.80, σ = 2.28	112.44, σ = 2.31	110.61, σ = 2.41

† Only well ordered fragments included. ‡ Using all atoms.

distribution of the N—C^{α}—C angle (separately for proline, glycine and all other residues) in all PDB structures with higher than 0.8 Å resolution. The distributions are wide, but do not have bimodal character, as illustrated for the case of non-Gly/non-Pro residues in Fig. 4. This cursory analysis suggests that the EH targets for N—C^{α}—C angles probably need adjustments of up to 1°.

3.6. The ω torsion angle of *trans*-peptides

It has been pointed out in numerous studies that Pauling's postulate of the planarity of the peptide group should not be enforced too strictly and that deviations of up to 20° from exact planarity should be treated as normal (MacArthur & Thornton, 1996; EU 3-D Validation Network, 1998; Edison, 2001). Consequently, the weight applied to the 180° target value for the *trans* C^{α}—N—C—C^{α} (ω) torsion angle should correspond to a rather large standard deviation. The *trans* ω torsion angles in the well ordered fragments of the ten highest resolution structures in the PDB are distributed with a mean of 179.36° and a standard deviation of 6.00° (179.43 and 6.30° without disorder elimination), while in the 1.7–1.8 Å interval the mean value is practically the same (179.44°) but the standard deviation (5.83°) is as in the EH definition (180°, standard deviation 5.80°) (Fig. 5). This result suggests a value of 6.0° for the standard deviation of this target.

3.7. Structural parameters from ultrahigh- and medium-resolution protein structures

To investigate whether the statistics of the structural parameters obtained in restrained crystallographic refinements depend on resolution, we have analyzed the behavior of the mean values of two main-chain bond lengths in different resolution intervals. The results are presented in Table 4. The PDB mean values for the C—N peptide bond (excluding Aaa-Pro peptides) depart from the EH target [1.329 (14) Å][2] in pace with increasing resolution. Although taken individually the differences are statistically not significant, as a trend they seem to suggest that a slightly modified target, perhaps 1.334 (18) Å, should be used to restrain nonproline C—N peptide bonds. Although nonproline *cis*-peptides may slightly bias the statistics, according to Jabs *et al.* (1999) protein models in the PDB contain only 0.026% of such bonds. Therefore, we did not search for and discriminate them from the overall statistics.

[2] Throughout this paper, wherever possible we use the crystallographic shorthand for noting the standard deviation in parentheses immediately after the number to which it refers and in the units of the last digit to which that number has been rounded. Thus, 1.329 (14) Å denotes a distance of 1.329 Å accompanied by σ = 0.014 Å.

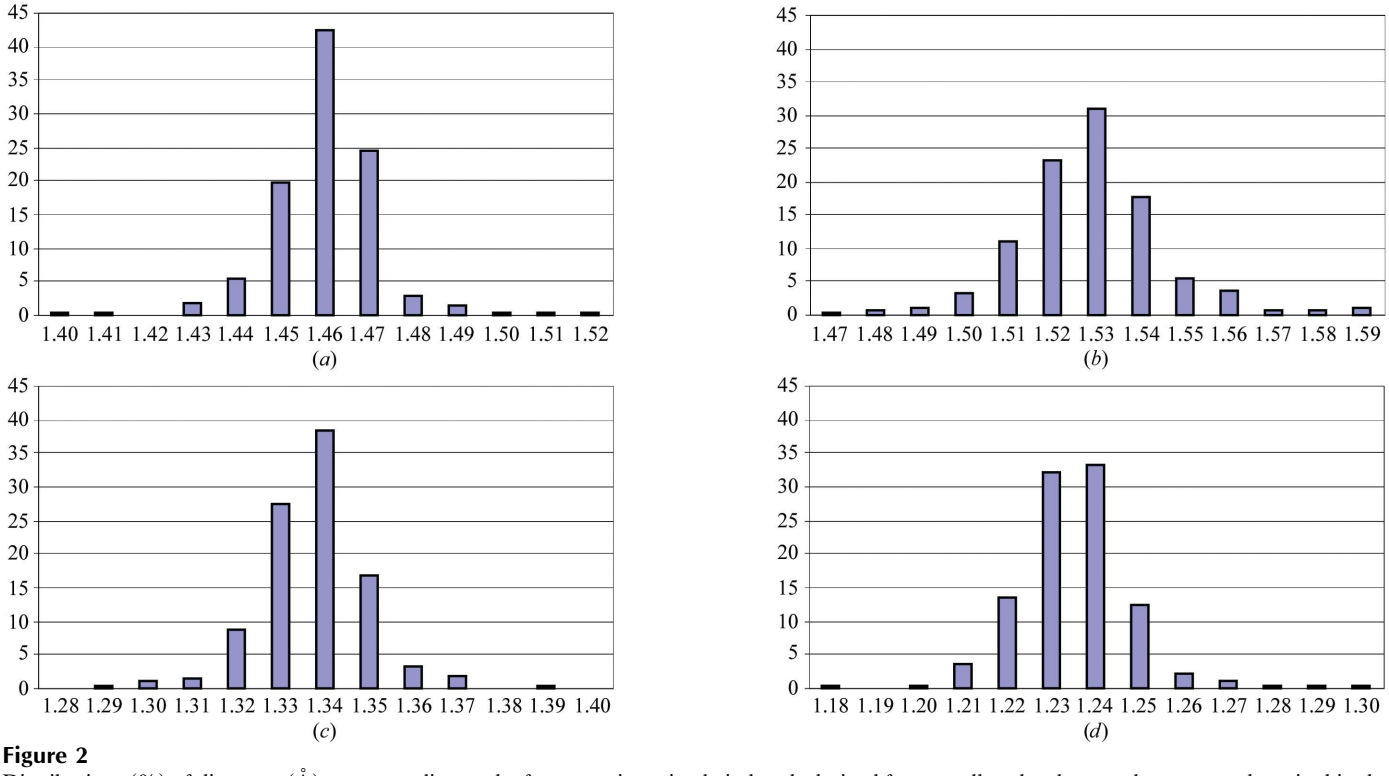

Figure 2
Distributions (%) of distances (Å) corresponding to the four protein main-chain bonds derived from small-molecule crystal structures deposited in the CSD with $R \leq 0.075$. (*a*) N—C^{α} excluding Gly and Pro residues, (*b*) C^{α}—C for non-Gly residues, (*c*) C—N excluding Aaa-Pro peptides, (*d*) C═O.

In addition, we examined the main-chain carbonyl C=O bond because chemical intuition suggested that the C=O bond length might exhibit elevated variability resulting from differing degrees of π-electron decoupling in correlation with deviations from peptide-group planarity. The effect of C=O elongation in strictly planar peptide groups could be further amplified (through conservation of Pauling's bond number) by the increased likelihood of polarized $C\cdots O^{\delta-}$ groups to participate in stronger hydrogen bonds. [In fact, an analogous argument is also true about the C—N bond and hydrogen-bond donor capability of the amide group. The negative correlation between the C=O and C—N bond lengths in proteins has been pointed out before (Esposito, Vitagliano, Zagari *et al.*, 2000).] We found out that in all resolution ranges shown in Table 4, the C=O bond distance has definite unimodal distribution and the average values slightly exceed the EH target. Surprisingly, the standard deviations of the experimental distributions are about half the EH value (indicating a lower spread of the C=O distances in proteins than observed in small-molecule structures) and in the high-resolution range are not very different from those determined for the C—N bond. It is noted, however, that the standard deviation does not fall off at lower resolution but remains at a constant level of 0.011 Å.

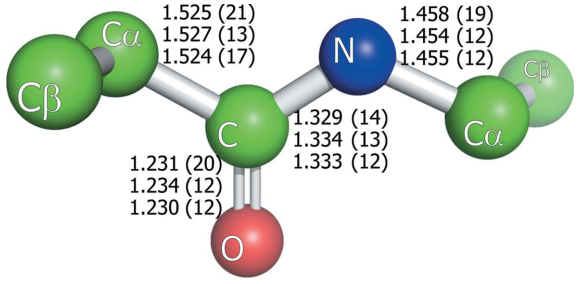

Figure 3
The peptide group with standard atom labeling, showing the main-chain bond distances (Å) as defined by Engh and Huber (top), as determined from ultrahigh-resolution protein structures in the PDB (middle, this study) and as determined from the current version of the CSD using organic structures with $R \leq 0.075$ (bottom, this study). The PDB-derived values are based on ten protein structures determined to resolutions higher than 0.8 Å, from which atoms in fragments with multiple conformation or with $B_{iso} \geq 40$ Å^2 have been excluded.

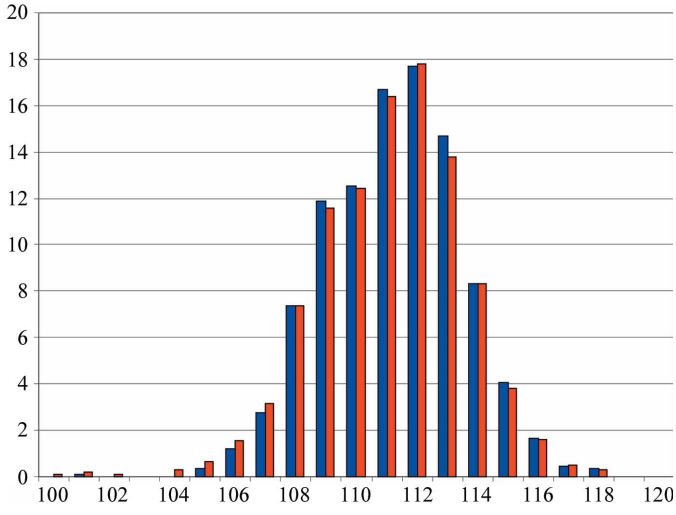

Figure 4
Distribution (%) of the N—C^{α}—C angles (°) (excluding Gly and Pro residues) in PDB structures determined to resolutions higher than 0.8 Å. Red, all data; blue, well ordered regions.

Table 4
Mean values and sample standard deviations calculated for the C—N bond (excluding Aaa-Pro peptides) and for the carbonyl C=O bond of the main-chain peptides in protein structures determined at different resolutions.

Resolution (Å)	C—N (Å)	C=O (Å)
0.54–0.8	1.334 (18)	1.234 (13)
0.8–0.9	1.333 (16)	1.236 (14)
0.9–1.0	1.332 (14)	1.236 (13)
1.0–1.1	1.329 (13)	1.233 (13)
1.1–1.2	1.330 (12)	1.236 (13)
1.2–1.3	1.329 (10)	1.233 (12)
1.3–1.4	1.329 (9)	1.232 (11)
1.4–1.5	1.329 (16)	1.232 (11)
1.5–1.6	1.329 (7)	1.234 (11)
1.6–1.7	1.329 (7)	1.233 (11)
1.7–1.8	1.329 (7)	1.233 (11)
EH value	1.329 (14)	1.231 (20)

3.8. Incorrect linear scaling of interatomic distances

Even at very high resolution, it is possible to obtain inaccurate interatomic distances if the axial lengths of the unit cell have a systematic error caused, for example, by poor wavelength calibration or an incorrect sample-to-detector distance during reduction of diffraction data. Should such an error occur, all interatomic distances would be systematically shorter or longer than expected. The problem of incorrect scaling of unit-cell parameters has been noted previously and its detection (and possible *post factum* correction) is implemented in analytical tools such as *WHATCHECK* (Hooft *et al.*, 1996) and *ELAST* (Yeates, 1990). With regard to the data presented in Table 2, by simple comparison of the sets of mean

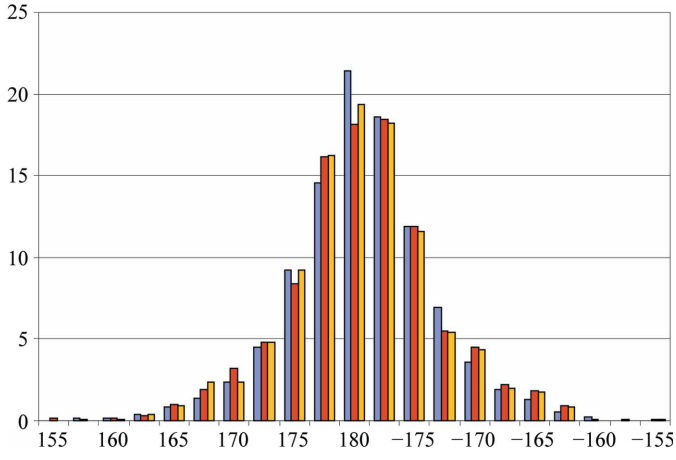

Figure 5
Distribution (%) of ω *trans* torsion angles (°) in PDB structures determined at 1.7–1.8 Å resolution (blue) and at higher than 0.8 Å resolution. In the latter case, data are presented for all atoms (red) and using only well ordered fragments (orange).

bond distances calculated for each of the structures, one notes cases (1yk4) where they uniformly exceed the set-wide mean values (as well as the EH targets) or are nearly uniformly below the mean values (1hje, 3al1, 1x6z). The most likely explanation is that the experimental axial lengths were overestimated in the former case or underestimated in the latter cases. An analysis of 1yk4 with the program *ELAST* (Yeates, 1990) suggests that the unit-cell parameters are overestimated by ~0.5%, although similar analysis of 1x6z does not indicate a unit-cell related problem and the 3al1 and 1hje structures are too small to be analyzed using this program. The remaining structures do not show any obvious bias in unit-cell parameter determination. While for individual structures such experimental errors may misguide structural interpretations, one hopes that in averages calculated over many structures these effects will cancel out, possibly only influencing the spread of the observed values around the average.

3.9. Robustness of protein geometry determination from PDB data

To test whether our results depended on the size and choice of the selected examples, the statistics obtained for the 0.8–0.9 Å subset of C—N bonds (3602 cases) were recalculated using an extended 0.8–0.9 Å subset with 5662 bonds, which also included some structures without $R_{\rm free}$. The result [1.332 (15) Å] is practically identical.

Since the sample standard deviation for the C—N bond in the 1.4–1.5 subset in Table 4 appeared to be elevated, we have looked closely at this case and determined that the problem could be traced to one particular PDB file (1w2p) for which the 'Geometry' tool of the PDB portal reported unusual C—N bonds (in dual-conformation fragments), resulting in a very wide spread of values for this structure [1.330 (32) Å overall, 1.331 (38) Å for chain *B*]. When structure 1w2p, as an outlier, was eliminated from the sample (a reduction of C—N cases from 4496 to 3583), the sample mean remained unchanged but the standard deviation was halved [1.329 (8) Å].

The above results attest to the robustness of our approach and suggest that the conclusions regarding sample averages will also be valid for differently chosen random samples. Indeed, when the mean values of Table 2 corresponding to all residues in the retrieved structures (no exclusion of disordered regions) were derived from a smaller set of structures (1ejg, 1r6j and 2b97 not included), all the mean values remained the same within 0.001 Å and the standard deviations within 0.002 Å, except for the σ value for the C—N bond, which changed from 0.013 to 0.018 Å on the expansion of the data set. This problem was traced to the presence of unrealistic C—N distances calculated for structure 2b97 (from 1.166 to 1.509 Å). When four extreme outliers outside the 1.20–1.46 Å interval (about ±10 'normal' standard deviations from the mean) were eliminated, the sample standard deviation dropped to 0.015 Å, with the mean value unchanged. To minimize such problems with unrealistic geometrical parameters, we based our final analysis of the highest resolution PDB structures only on well ordered regions (no multiple conformations, $B_{\rm iso} < 40$ Å^2). As seen in Table 2, the elimination of disordered regions does not produce a uniform pattern in the results. In many cases, the sample standard deviations for individual structures are reduced by as much as 50% or more. However, there are also cases (for structures with little disorder as defined by our criteria) where the sample standard deviations are practically unchanged or even increase (N—C^α for structure 1us0, C^α—C for structure 1r6j). There are also cases where the standard deviation remains high even on exclusion of disorder (C—N for structure 2b97), indicating that suspicious geometry outliers exist in the well ordered part of the model. The mean values are usually not affected to a significant extent by the inclusion of disordered regions, but exceptions are observed for some structures, where those changes can be as high as 0.010 Å (N—C^α for structure 1hje). Since the largest changes are accompanied by a very significant increase of the sample standard deviation, it is obvious that they represent inclusion of statistical outliers in the mean-value calculations. This emphasizes the notion that for proper statistical analyses of individual geometrical parameters, disordered regions should be excluded from the calculations or down-weighted according to their site occupancy and *B* factors. However, when large pools of structures are used for averaging and the purpose is to estimate the trends in the data, it is possible to use all the available data without disorder screens. This is illustrated by the mean bond-length values in Table 2 calculated over all the highest resolution structures, where the mean values are practically unchanged and the standard deviations show only a moderate increase (20% on average).

3.10. Comparison of peptide geometry estimated from PDB and CSD data

A comparison of the main-chain bond distances evaluated from highest resolution protein structures with those compiled by Engh and Huber (Table 2) and those evaluated from $R1 \leq 0.075$ structures of the current CSD database (Table 1) reveals that the values obtained from the currently available structural depositories are in good agreement, the only exception being the C═O and C^α—C bonds, which in protein structures are somewhat longer than in small molecules. This may be interpreted as reflecting the systematic involvement of the main-chain carbonyl groups in proteins in similar hydrogen-bonding interactions. The small-molecule lengths of these bonds are the same in the EH set and in the current evaluation, although the increased sample size and quality dictate a lower standard deviation. The small-molecule mean values of the N—C^α and C—N bond lengths from the previous and current CSD analyses differ by 0.003–0.004 Å and it is interesting to note that the current values are practically identical to those derived from the PDB data.

4. Discussion

From a superficial confrontation of the EH stereochemical standards with the vastly expanded database (CSD) from

which they were originally generated, as well as with the holdings in the PDB, which to a large degree bear their influence, we conclude that the situation is not uniform; *i.e.* while some of the original geometrical targets have withstood the test of time, some others might need small but clear adjustments. Although the current applications of refinement restraints do not use individualized weights for various instances of the same parameter class (*e.g.* various bond types), new refinement algorithms should consider the individualization of restraint weights. For some parameter adjustments, the indications from the CSD and the PDB are not consistent. One such example is provided by the main-chain C=O bond, which in proteins is 0.004 Å longer than in small molecules. A question thus arises: which indication should be used for optimization of the target values? The CSD data are more precise and are generally not 'contaminated' by restraining. On the other hand, only protein structures can give accurate information about protein geometry, taking into account that the molecular parameters may be (and most likely are) influenced by the specific nature of protein conformation and interactions with the environment. Those influences could be larger for some parameters (*e.g.* torsion angles) and smaller for others (bond distances), but even in the latter case they could be detectable, especially when correlated with significant noncovalent interactions (hydrogen bonding, ionic interactions). Since the accelerating accumulation of protein structures in the PDB provides not only quantitative but also qualitative improvement of the data, we are of the opinion that at the present moment sufficiently accurate information is on hand to justify an attempt to correct the restraint targets by using ultrahigh-resolution protein structures refined with loose restraints.

Tight restraints are necessary at low resolution and in areas poorly defined by diffraction. With increasing resolution, the weights should be progressively less tight for well defined parts of the model. However, one should not be tempted to 'improve' the *R* factor by violating the rigors of stereochemistry. As a practical guideline, we recommend to aim at an r.m.s.d. for bonds of 0.020 Å for models at 1.5 Å or lower resolution. The analysis presented above allowed us to observe that the current practice of using equally tight restraints for all geometrical parameters of the same kind often leads to the undesirable situation where the well behaved protein fragments are over-restrained while the flexible or disordered parts are refined with very poor geometry. At very high resolution, even the very tight restraints that are necessary to keep the flexible fragments under good stereochemical control are not able to override the experimental information reflected in the geometry of well ordered parts and it is normal in such situations for the model bond distances to deviate from the idealized targets by 0.02–0.03 Å. As a practical solution, at ultrahigh resolution (higher than 0.8 Å) restraints could be limited to side chains and fragments of main chains with multiple conformations (Dauter *et al.*, 1992). However, at medium resolution it would be beneficial to couple the weights of individual restraints with the occupancies and *B* factors of the participating atoms. This is in keeping with earlier observations [*e.g.* Fig. 4(*a*) in Cruickshank (1999) or Fig. 2(*c*) in Parisini *et al.* (1999)] that in fragments displaying high *B* factors the geometrical parameters tend to reproduce the restraint values, whereas without restraints their geometry 'explodes' to an unacceptable level, since the X-ray terms alone cannot successfully refine such fragments. However, the situation with the well behaving parts is different. The analysis of the ultrahigh-resolution structures suggests that tight restraints may bias some parameters that truly differ from the library standards owing to, for example, an unusual chemical environment. At high resolution, restraints are required for badly behaving fragments, but can be applied with less weight to well ordered fragments with low *B* factors. Such atypical features would not be believable at lower resolution and in disordered or mobile fragments. At high resolution, the identification and justification of such fragments can be achieved by the use of local rather than global validators, *e.g.* by inspection of the real-space *R* factor or identification of features in difference Fourier maps. In the final refinement, the geometrical restraints could be relaxed or otherwise individualized for these fragments. The individualization of restraints in genuinely distorted fragments may lead to better modeling of particularly important parts of protein structures, such as the active or binding sites of enzymes, where the unusual features often have a functional significance.

This project was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. The research of MJ was supported by a Faculty Scholar fellowship from the National Cancer Institute and by a subsidy from the Foundation for Polish Science.

References

Addlagatta, A., Krzywda, S., Czapinska, H., Otlewski, J. & Jaskolski, M. (2001). *Acta Cryst.* D**57**, 649–663.
Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bönisch, H., Schmidt, C. L., Bianco, P. & Ladenstein, R. (2005). *Acta Cryst.* D**61**, 990–1004.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.
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.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.
Dauter, Z., Botos, I., LaRonde-LeBlanc, N. & Wlodawer, A. (2005). *Acta Cryst.* D**61**, 967–975.
Dauter, Z., Dauter, M. & Dodson, E. (2002). *Acta Cryst.* D**58**, 494–506.
Dauter, Z., Sieker, L. C. & Wilson, K. S. (1992). *Acta Cryst.* B**48**, 42–59.
Dodson, E., Kleywegt, G. J. & Wilson, K. (1996). *Acta Cryst.* D**52**, 228–234.

Dunlop, K. V., Irvin, R. T. & Hazes, B. (2005). *Acta Cryst.* D**61**, 80–87.
Edison, A. S. (2001). *Nature Struct. Biol.* **8**, 201–202.
Engh, R. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Engh, R. A. & Huber, R. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 382–392. Dordrecht: Kluwer Academic Publishers.
Esposito, L., Vitagliano, L., Sica, F., Sorrentino, G., Zagari, A. & Mazzarella, L. (2000). *J. Mol. Biol.* **297**, 713–732.
Esposito, L., Vitagliano, L., Zagari, A. & Mazzarella, L. (2000). *Protein Eng.* **13**, 825–828.
EU 3-D Validation Network (1998). *J. Mol. Biol.* **276**, 417–436.
Evans, P. R. (2007). *Acta Cryst.* D**63**, 58–61.
Hakanpää, J., Linder, M., Popov, A., Schmidt, A. & Rouvinen, J. (2006). *Acta Cryst.* D**62**, 356–367.
Hansen, N. K. & Coppens, P. (1978). *Acta Cryst.* A**34**, 909–921.
Hendrickson, W. A. (1985). *Methods Enzymol.* **115**, 252–270.
Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990). *EMBO J.* **9**, 1665–1672.
Hooft, R. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.
Howard, E. I., Sanishvili, R., Cachau, R. E., Mitschler, A., Chevrier, B., Barth, P., Lamour, V., Van Zandt, M., Sibley, E., Bon, C., Moras, D., Schneider, T. R., Joachimiak, A. & Podjarny, A. (2004). *Proteins*, **55**, 792–804.
Jabs, J., Weiss, M. S. & Hilgenfeld, R. (1999). *J. Mol. Biol.* **286**, 291–304.
Jelsch, C., Guillot, B., Lagoutte, A. & Lecomte, C. (2005). *J. Appl. Cryst.* **38**, 38–54.
Jelsch, C., Teeter, M. M., Lamzin, V., Pichon-Pesme, V., Blessing, R. H. & Lecomte, C. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 3171–3176.
Kang, B. S., Devedjiev, Y., Derewenda, U. & Derewenda, Z. S. (2004). *J. Mol. Biol.* **338**, 483–493.
Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. & Shore, V. C. (1960). *Nature (London)*, **185**, 422–427.
Kleywegt, G. J. & Jones, T. A. (1995). *Structure*, **3**, 535–540.
Ko, T.-P., Robinson, H., Gao, Y. G., Cheng, C. H., DeVries, A. L. & Wang, A. H. (2003). *Biophys. J.* **84**, 1228–1237.
Kuhn, P., Knapp, M., Soltis, S. M., Ganshaw, G., Thoene, M. & Bott, R. (1998). *Biochemistry*, **37**, 13446–13452.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
MacArthur, M. W. & Thornton, J. M. (1996). *J. Mol. Biol.* **264**, 1180–1195.
Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. (2006). *Acta Cryst.* D**62**, 859–866.
Morris, R. J. & Bricogne, G. (2003). *Acta Cryst.* D**59**, 615–617.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Parisini, E., Capozzi, F., Lubini, P., Lamzin, V., Luchinat, C. & Sheldrick, G. M. (1999). *Acta Cryst.* D**55**, 1773–1784.
Patterson, W. R., Anderson, D. H., DeGrado, W. F., Cascio, D. & Eisenberg, D. (1999). *Protein Sci.* **8**, 1410–1422.
Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G. & North, A. C. T. (1960). *Nature (London)*, **185**, 416–421.
Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Sheldrick, G. M. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 401–411. Dordrecht: Kluwer Academic Publishers.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.
Terwilliger, T. C. (2003). *Methods Enzymol.* **374**, 22–37.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
Wang, B.-C., Chen, C.-J., Liu, Z. J., Wu, C. K., Schubot, F. D., Rosenbaum, G., Vysotski, E. S., Lee, J., Dailey, H. A., Ferrara, J., Schiffer, M., Pokkulur, P. R., Joachimiack, A., Zhang, R., Howard, A., Chrzas, J., Robbins, A. H. & Rose, J. P. (2000). Am. Crystallogr. Assoc. Meet. Abstr., p. 66.
Wlodawer, A. & Hendrickson, W. A. (1982). *Acta Cryst.* A**38**, 239–247.
Yeates, T. O. (1990). *Acta Cryst.* A**46**, 625–626.