Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Statistical analysis of crystallographic data obtained from squid ganglion DFPase at 0.85 Å resolution

Juergen Koepke,[a]* Eileen I. Scharff,[b]† Christian Lücke,[b]‡ Heinz Rüterjans[b] and Günter Fritzsch[a]

[a]Max-Planck-Institute of Biophysics, Department of Molecular Membrane Biology, Marie-Curie-Strasse 15, D-60439 Frankfurt/Main, Germany, and [b]Institute of Biophysical Chemistry, Johann Wolfgang Goethe-University, Marie-Curie-Strasse 9, D-60439 Frankfurt/Main, Germany

† Present address: BASF Aktiengesellschaft, GVX/P-C006, D-67052 Ludwigshafen, Germany.
‡ Present address: Max Planck Research Unit for Enzymology of Protein Folding, Weinbergweg 22, D-06120 Halle, Germany.

Correspondence e-mail: koepke@mpibp-frankfurt.mpg.de

Received 4 June 2003
Accepted 21 July 2003

PDB Reference: squid ganglion DFPase, 1pjx, r1pjxsf.

The X-ray crystal structure of squid-type diisopropylfluorophosphatase (DFPase) has been refined to a resolution of 0.85 Å and a crystallographic R value of 9.4%. Crystal annealing improved both the mosaicity and resolution of the crystals considerably. The overall structure of this protein represents a six-bladed β-propeller with two calcium ions bound in a central water-filled tunnel. 496 water, two glycerol and two MES buffer molecules and 18 PEG fragments of different lengths could be refined in the solvent region. 45 of the 314 residues have been refined with alternative orientations. H atoms have been omitted from disordered residues. For the residues of the inner β-strands, H atoms are visible in a normal $F_o - F_c$ difference map of a hydrogen-deficient structure model. The 208 most reliable residues, without disorder or reduced occupancy in their side chains, were finally refined without restraints. A subsequent full-matrix refinement cycle for the positional parameters yielded estimated standard deviations (e.s.d.s) by matrix inversion. The thus calculated bond lengths and bond angles and their e.s.d.s were used to obtain averaged bond lengths and bond angles, which were compared with the restraints applied in the preceding refinement cycles. The lengths and angles of the hydrogen bonds inside the antiparallel β-sheets of the DFPase structure were compared with data averaged over 11 high-resolution protein structures. Torsion angles were averaged according to angle types used as restraints in *X-PLOR* and *CNS* and subsequently compared with values obtained from 46 high-resolution structures. Side-chain torsion angles were also classified into rotamer types according to the Penultimate Rotamer Library. Moreover, precise dimensions for both Ca^{2+}-coordination polyhedra could be obtained and the coordination of one Ca^{2+} ion by an imidazole N atom was confirmed. This statistical analysis thus provides a first step towards a set of restraints that are founded completely on macromolecular data; however, 10–20 additional protein data sets of comparable accuracy and size will be required to obtain a larger statistical base, especially for side-chain analysis.

1. Introduction

In a previous study (Scharff, Koepke *et al.*, 2001), we presented the structure of squid ganglion DFPase at a resolution of 1.8 Å. Furthermore, we have described the ability of this enzyme to hydrolyze organophosphorus triesters by cleaving their P—F bond and proposed a catalytic mechanism based on nine different point mutations close to the catalytic site.

Recently, details of the crystallization and handling of DFPase crystals that diffract to atomic resolution were presented (Koepke *et al.*, 2002). Data collection, refinement of data to 0.85 Å resolution and preliminary results were

described. The mean bond lengths of 11 backbone and $C^{\alpha}-C^{\beta}$ bond types labelled according to Engh & Huber (1991) were compared with their respective restraints. At that time, only insignificant differences were found compared with the restraints and data from another high-resolution protein structure (Longhi *et al.*, 1998).

The 25 β-strands of the squid ganglion DFPase are folded into six twisted antiparallel β-sheets surrounding a central tunnel and forming a sixfold β-propeller structure (Fig. 1). The innermost β-strands are almost parallel to the tunnel axis, which corresponds to a pseudo-sixfold axis. The outer β-strands, however, are oriented nearly perpendicular to the tunnel axis. A low-affinity and a high-affinity calcium-binding site have been found in the tunnel. One end of the tunnel is blocked by a short helical turn. At the other end, the low-affinity calcium (Ca1) was found to be part of the active site. The high-affinity calcium (Ca2) is completely embedded in the centre of the tunnel and is believed to stabilize the propeller fold. It is octahedrally coordinated by two protein O atoms, three water molecules and, most remarkably, by an N atom from the side chain of a histidine residue. The ultrahigh resolution obtained in this study was required to unambiguously confirm this unusual coordination between Ca2 and His274 $N^{\delta1}$.

Since protein structures of such ultrahigh resolution well below 1 Å are still rare, the data obtained from DFPase have been used for the first time in a detailed statistical analysis of all kinds of geometrical data, *i.e.* bond lengths, bond angles, torsion angles, lengths and angles of hydrogen bonds and coordination lengths of the metal ions. This wealth of data can now be compared with the corresponding data found in the literature, in order to test the respective restraints against which protein structures with lower resolution are usually refined. Therefore, in Appendix *A* we demonstrate the important role that accurate restraints play for the crystallographic refinement of biological macromolecule structures. Until now, restraints used in crystallographic refinement programs have been based only on small-molecule structures that cannot fully reflect the special conditions inside a protein. At present, a sufficient amount of ultrahigh-resolution protein data has already been collected to obtain protein-based restraints, which are presumably better suited for the aims of biocrystallography. To this end, we demonstrate here the high accuracy that is provided by data obtained from only a single ultrahigh-resolution structure.

2. Materials and methods

2.1. X-ray data collection and processing

Crystals of the squid ganglion DFPase were grown by the hanging-drop vapour-diffusion technique. Crystallization in the presence of 12%(*w*/*v*) PEG 6000 and 0.1 *M* MES buffer has been described previously (Scharff, Lücke *et al.*, 2001). To minimize the mosaicity and to increase the resolution limit, the crystals were annealed several times (Samygina *et al.*, 2000). Data were collected at the EMBL Hamburg Outstation (beamline BW7B) to a resolution of 0.82 Å and processed with *MOSFLM* (Leslie, 1992) to a resolution of 0.835 Å. The data from four different crystals were required to obtain a complete data set, since the highest resolution reflections decayed after about 100 images. A total of 210 frames from three crystals diffracting to the highest resolution were used. For all other data, the resolution was cut to either 0.86 or 0.91 Å. Low-resolution data to 1.5 Å were added from all four crystals used, leading to an overall completeness of 93.8% and 77.8% completeness for the highest resolution shell (0.835–0.85 Å) (Table 1). Finally, structure factors were calculated by employing *TRUNCATE* from the *CCP*4 package (Collaborative Computational Project, Number 4, 1994).

Table 1
Summary of crystallographic data.

Unit-cell parameters	
a (Å)	43.1
b (Å)	81.8
c (Å)	86.5
V (Å^3)	305128
Space group	$P2_12_12_1$
Molecules per AU	1
Molecular weight (Da)	35079
Amino acids	314
V_M (Å^3 Da^{-1})	2.188
Solvent content (%)	43.8
B_{Wilson}† (Å^2)	5.9
Highest resolution (Å)	0.835
Unique reflections	264417
Overall completeness‡§ (%)	93.8 (7.7)
Completeness, highest resolution shell‡§ (%)	77.8 (1.8)
Atoms refined	5410
Solvent molecules	
Water	496
Glycerol	2
MES buffer	2
PEG fragments	18
R_{sym}¶ (%)	6.5
R_{cryst}‡ (%)	11.1 (9.4)††
R_{free}‡§§ (%)	12.8 (11.1)††

† B_{Wilson} is an average temperature factor. ‡ For all reflections to a resolution of 0.85 Å. § $I/\sigma(I)$ is given in parentheses. ¶ $R_{sym} = \sum_{hkl}\sum_i |I_i - \langle I\rangle|/\sum\langle I\rangle$, where I_i is the intensity of the *i*th measurement of reflection *hkl* and $\langle I\rangle$ is the average intensity of a reflection. †† *R* values for $F_o > 4\sigma(F_o)$ is given in parentheses. §§ R_{free} is calculated from 1% of the measured unique data that were not used during refinement.

2.2. Refinement

At first, refinement was performed using *REFMAC*5 (Murshudov *et al.*, 1999), starting from the model deposited in the Protein Data Bank under accession code 1e1a. Subsequently, the model was refined against diffraction intensities, with H atoms defined according to standard geometry using a parallelized version of *SHELX*97 (Diederichs, 2000; Sheldrick & Schneider, 1997). In addition, anisotropic temperature factors and a diffuse solvent correction (Moews & Kretsinger, 1975) were applied at a maximum resolution of 0.85 Å. As a consequence, more details appeared in the solvent region and between each refinement round $2F_o - F_c$ and $F_o - F_c$ electron-density maps were inspected using the graphics program *XTALVIEW* (McRee, 1999). New water molecules were added by identification of peaks >3σ in the $F_o - F_c$ difference density map with a geometry suitable for hydrogen

bonding, eventually resulting in a total of 496 water molecules (Table 1). Based on the size and electron density of some peaks accounting for water molecules in the $2F_o - F_c$ maps, it became necessary to reduce their occupancy in several cases. In 18 cases, it was possible to introduce PEG fragments of different length into the refinement process at positions where water molecules were connected by positive difference density. We have found eight ethylene glycol molecules (EDO), two di(hydroxyethyl)ether molecules (PEG), two triethylene glycol molecules (PGE), three 1,2-dimethoxyethane molecules (DXE), two 2-methoxyethanol molecules (MXE) and one 1-ethoxy-2-(2-methoxyethoxy)ethane (ME2) molecule. For two such areas, a glycerol molecule (GOL) was the better fit and finally two MES buffer molecules [2-(*n*-morpholino)-ethanesulfonic acid] could be identified and refined in the solvent region (Fig. 2). The above molecule names and three-letter abbreviations in parentheses are identical with the heterogroup names listed in PDB entry 1pjx. 45 residues and 17 water molecules have been refined with alternative orientations. Backbone atoms were refined in two orientations for the following residues: 1–3, 216–217 and 311–313. In residues with alternative orientations or reduced occupancy, no H atoms have been added to the model.

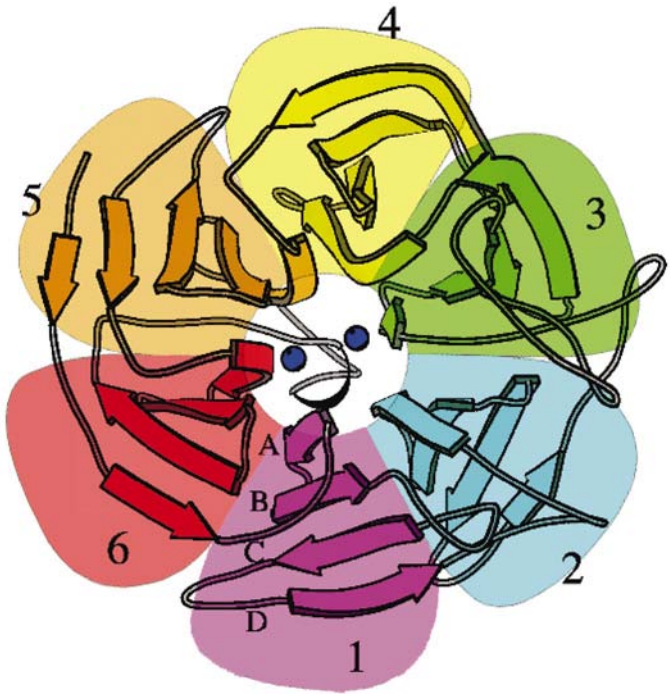

Figure 1
Overall structure of the squid-type DFPase. View of the molecule down the pseudo-sixfold axis. The shaded propeller blades are labelled 1 to 6 and are colour coded from magenta to red, as are the four antiparallel β-strands belonging to each blade, which are themselves named *A* to *D* from the inside to the outside (molecule drawing from Scharff, Koepke *et al.*, 2001).

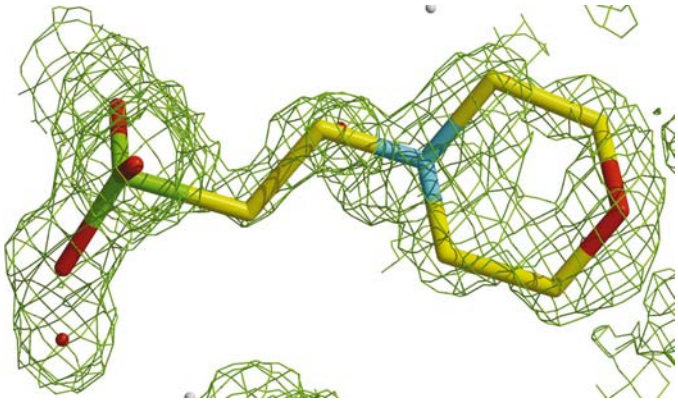

Figure 2
One of the two MES buffer molecules identified in the solvent region. The $2F_o - F_c$ map shown in green is contoured at 0.75σ. The occupancy of the MES buffer molecule was set to 0.8 to meet the temperature factors of neighbouring water molecules. Figures containing electron-density information are extracted from *XTALVIEW* (McRee, 1999) and have been generated with *Raster3D* (Bacon & Anderson, 1988; Merritt & Murphy, 1994).

With H atoms defined according to standard geometry in the riding positions, $R_{\rm free}$ improved by 1.3% to a value of 12.9%, while $R_{\rm cryst}$ improved by 1.2% to 11.2%. In a final round of refinement, restraints for the 208 most reliable residues were removed, resulting in a marginally improved $R_{\rm cryst}$ of 11.1% and indicating that no major changes had occurred. The subsequent blocked full-matrix least-squares cycle for the positional parameters improved $R_{\rm free}$ to 12.8%. The matrix inversion yielded estimated standard deviations (e.s.d.s) for the refined parameters, from which the e.s.d.s for bond lengths and bond angles could be calculated. For the 47 (53) different bond-length and 86 (97) bond-angle restraints used in *SHELX*97, weighted mean values and weighted standard deviations were calculated from the data of 208 unrestrained residues. The values in parentheses take into account restraints for different protonation states of histidines. Accurate distances could also be calculated for the two Ca^{2+}-coordination polyhedra.

3. Results and discussion

3.1. Quality of the electron-density map supporting the Ca^{2+}-ion coordination by an N atom

Ten years ago, Nayal & Di Cera (1994) found only O atoms coordinating Ca^{2+} ions in the PDB (Bernstein *et al.*, 1977). To date, 3170 different Ca^{2+}-binding sites have been identified in the Metalloprotein Database (Castagnetto *et al.*, 2002), 13 of which contain a histidine side-chain N atom as a ligand, corresponding to nine structurally different sites. These sites belong to eight different protein structures: squid-type DFPase (Scharff, Koepke *et al.*, 2001), yeast frequenin (Ames *et al.*, 2000), sex-hormone-binding globulin (Grishkovskaya *et al.*, 2000), Rop (Willis *et al.*, 2000), Pnb esterase (Spiller *et al.*, 1999), gingipain R (Eichinger *et al.*, 1999), concanavalin A (Bouckaert *et al.*, 2000) and exo-amylase (Morishita *et al.*, 1997), with the Ca^{2+} coordination distributed about equally between $N^{\delta1}$ and $N^{\varepsilon2}$ of the respective histidines.

We recently elucidated the orientation of the imidazole ring of His274 by the size of the contoured atomic electron-density volumes (Koepke *et al.*, 2002). However, there remained some

uncertainty about the protonation of the N atoms of this histidine, because of the bond lengths around $C^{\varepsilon1}$. In Fig. 3 the protonation of $N^{\varepsilon2}$ is clearly visible, while $N^{\delta1}$ is unprotonated and coordinates Ca2. The coordination of the Ca^{2+} ion by $N^{\delta1}$ produces an aromatic imidazolium ion with the positive charge delocalized between the two N atoms. This resonance stabilization thus apparently induces identical $C^{\varepsilon1}-N^{\delta1}$ and $C^{\varepsilon1}-N^{\varepsilon2}$ bond lengths, analogous to a doubly protonated imidazole ring. Moreover, such a charge delocalization might impart additional stability to the Ca^{2+}-binding site.

3.2. Hydrogen bonds between antiparallel β-strands

The H atoms between the carbonyl O atoms and the backbone N atoms are clearly visible for the two innermost antiparallel β-strands *A* and *B* of each blade, as indicated in Fig. 4. Since the hydrogen bonds are rather long, the protons stay close to the donor. Only in the shortest of the four hydrogen bonds has the proton moved slightly to the centre of the bond. Moreover, all four protons that account for secondary hydrogen bonds from the neighbouring C^{α} atom to the same backbone O atom are also visible.

In Table 2, the hydrogen bonds found in DFPase are compared with data published by Fabiola *et al.* (1997). Since the DFPase structure contains only antiparallel β-sheets as secondary-structure elements, except for a small α-helical turn at the C-terminus, all hydrogen bonds found inside these β-sheets belong to only two different hydrogen-bond types. The location and length of the secondary-structure elements in DFPase were classified according to Kabsch & Sander (1983). The primary hydrogen bonds of the type N—H···O═C could be drawn from the *SHELX* listing (Sheldrick & Schneider,

Table 2
Hydrogen bonds in antiparallel β-sheets.

Hydrogen-bond type	DFPase mean (μ) (Å)	t†	11 protein structures mean (m)‡ (Å)
N—H···O═C			
H···O (Å)	2.10 (0.14)§	**11.1**	1.94
N···O (Å)	2.91 (0.13)	2.0	2.89
N—H···O (°)	160.0 (9.9)	1.7	158
n	96		49
C^{α}—H···O═C			
H···O (Å)	2.49 (0.14)	**6.3**	2.37
C^{α}···O (Å)	3.31 (0.14)	1.3	3.27
C^{α}—H···O (°)	141.9 (5.2)	2.4	141
n	53		49

† $t = |\mu - m|/\sigma_{\mu}$. ‡ Fabiola *et al.* (1997). § The values in parentheses represent the sample standard deviation (σ).

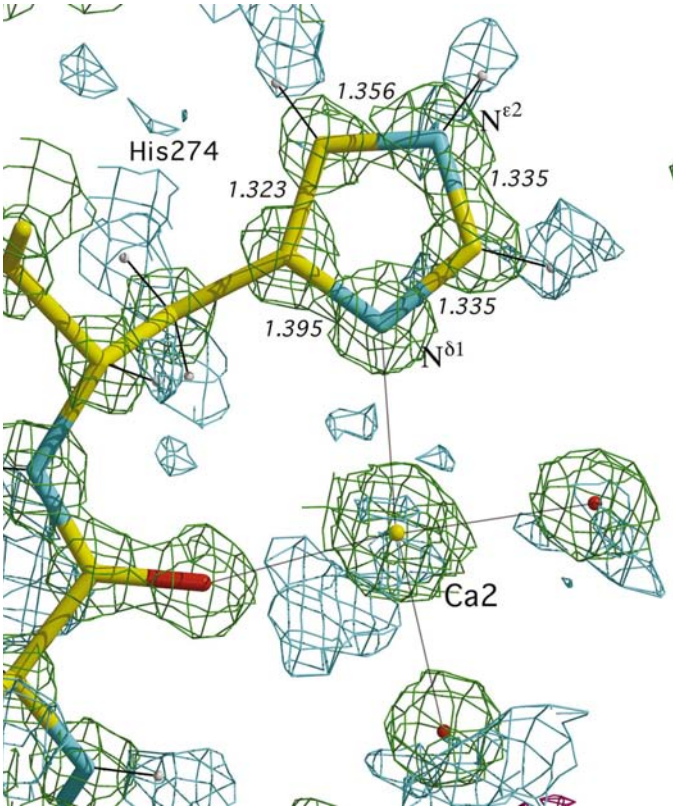

Figure 3
Electron density around His274 and Ca2. The $2F_o - F_c$ map shown in green is contoured at 3σ. C, N and O atoms as well as the Ca^{2+} ion can be distinguished by their different sizes. The $F_o - F_c$ difference map of a hydrogen-deficient model is contoured in cyan at 2σ. The densities of the H atoms attached to the His274 imidazole ring are clearly visible. The bond lengths in the ring are given in Å and the coordination of Ca2 is indicated by thin grey lines.

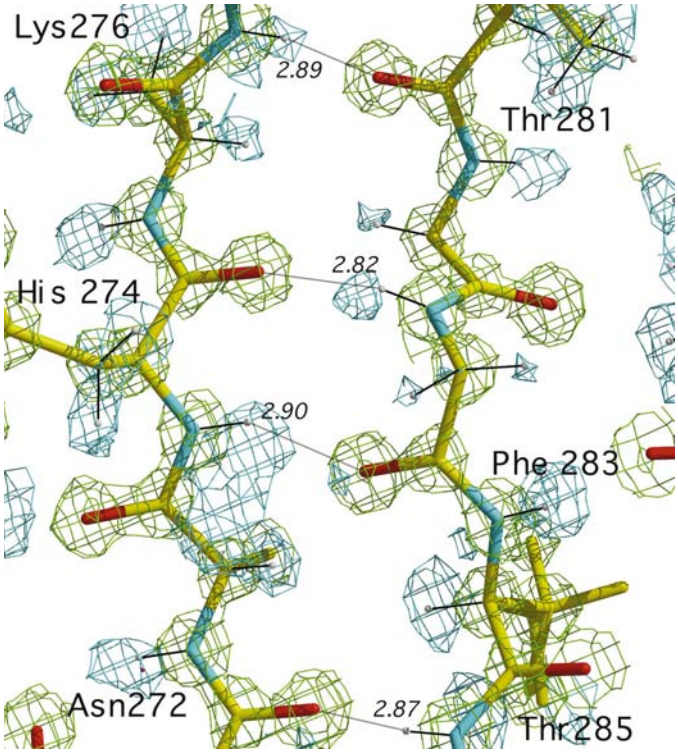

Figure 4
Hydrogen bonds between two antiparallel β-strands. The $2F_o - F_c$ electron density shown in green is contoured at 3σ, while the $F_o - F_c$ difference density calculated from a model with missing H atoms is contoured in cyan at 2σ. Putative hydrogen bonds are indicated by thin grey lines, with the distances between the respective N and O atoms given in Å.

Table 3
Comparison of mean bond lengths and restraints.

Residue	Bond type†	Bonds	Mean‡ (Å)	σ_μ§ (Å)	Restraint¶ (Å)	*F*-test††	χ^2_1
	C—O	334	1.234 (14)	0.001	1.229 (19)	1.52	3.90
Trp	C—O	6	1.237 (12)	0.005	1.229 (19)	*2.36*	1.54
	CH1E—NH1	255	1.452 (14)	0.001	1.459 (20)	1.19	2.50
His	C^α—N	6	1.443 (16)	0.007	1.459 (20)	*2.27*	2.55
	C—NH1	292	1.331 (15)	0.001	1.336 (23)	2.16	2.44
Met	C—N	7	1.330 (9)	0.003	1.336 (23)	2.03	2.03
	CH1E—CH1E	104	1.538 (19)	0.002	1.542 (23)	1.01	3.73
Ile	C^α—C^β	19	1.542 (22)	0.005	1.544 (23)	**3.07**	**19.63**
	Gaussian fit		1.542 (20)	0.005		**5.76**	1.11
	C—CH2E	60	1.514 (25)	0.003	1.516 (25)	1.68	*8.37*
	Gaussian fit		1.515 (26)	0.003		*2.90*	1.71
Gln	C^γ—C^δ	12	1.517 (23)	0.007	1.506 (23)	*2.60*	0.78
	CH2G—NH1	31	1.444 (14)	0.003	1.456 (15)	**3.62**	1.06
Gly	C^α—N	31					
	C—NH2	27	1.322 (26)	0.005	1.324 (25)	1.06	0.64
Gln	C^δ—$N^{\varepsilon 2}$	12	1.313 (21)	0.006	1.324 (25)	**5.63**	**30.43**
	Gaussian fit		1.315 (32)	0.009		**9.38**	0.58
	CH2E—CH3E	19	1.501 (43)	0.010	1.513 (39)	**3.45**	1.54
Ile	$C^{\gamma 1}$—$C^{\delta 1}$	19	1.501 (43)	0.010	1.500 (69)	1.84	1.77
	C—NC2	11	1.318 (24)	0.007	1.326 (13)	**4.77**	0.80
Arg	C^ζ—$N^{\eta 1}$	6	1.315 (15)	0.006	1.326 (13)	**3.97**	0.36
Arg	C^ζ—$N^{\eta 2}$	5	1.320 (33)	0.014	1.326 (13)	**3.54**	1.15
	C5—CR1E	6	1.335 (14)	0.006	1.356 (11)	**3.17**	0.34
His	C^γ—$C^{\delta 2}$	6	1.335 (14)	0.006	1.354 (9)	**3.27**	0.35
HisD	C^γ—$C^{\delta 2}$	2	1.341 (21)	0.015	1.353 (17)	1.94	0.56
HisE	C^γ—$C^{\delta 2}$	4	1.333 (13)	0.006	1.353 (14)	*2.48*	1.11
	C5—NH1	6	1.386 (15)	0.006	1.378 (11)	*2.19*	0.57
His	C^γ—$N^{\delta 1}$	6	1.386 (15)	0.006	1.380 (10)	*2.18*	0.41
HisD	C^γ—$N^{\delta 1}$	2	1.379 (4)	0.003	1.369 (15)	1.03	0.72
HisE	C^γ—$N^{\delta 1}$	4	1.389 (17)	0.009	1.383 (22)	1.01	0.29

† The bond types in bold are labelled according to Engh & Huber (1991). ‡ Weighted mean, $\mu = \sum_i(\omega_i x_i)/\sum_i(\omega_i)$, and weighted sample standard deviation, $\sigma = \{n/(n-1)\sum_i[\omega_i(\mu - x_i)^2]/\sum_i(\omega_i)\}$, with weight $\omega_i = 1/\sigma_i^2$ (Bevington, 1969). The values in parentheses refer to the last digits of the mean and represent the sample standard deviation. § Mean standard deviation, $\sigma_\mu = (\sigma^2/n)^{1/2}$. ¶ Engh & Huber (1991, 2001). †† $F = \chi_1^2/\chi_2^2$, with χ_1^2 calculated with the mean values and their standard deviations, while χ_2^2 was calculated using the restraints.

1997), while the secondary hydrogen bonds were only taken into account at the C^α atom of the neighbouring amino acid one position down the sequence when its $C^\alpha\cdots O$ distance was <3.5 Å and additionally the corresponding $C^\alpha—H\cdots O$ angle was >130°, as defined in Fabiola *et al.* (1997). The function *t* in column 3 of Table 2 can be used as a test value, indicating whether the data found for DFPase deviate significantly from the values given by Fabiola *et al.* (1997). Values of low confidence (in bold type) do not belong to the same distribution (see Appendix *B* for definitions). As documented in Table 2, the $H\cdots O$ distances of the two hydrogen-bond types are significantly different, probably owing to the donor-to-hydrogen distances used in our study. We did not refine the hydrogen positions, but used values preset in the *SHELX*96 program (Sheldrick & Schneider, 1997), positioning the H atoms 0.89 Å away from the N and 0.98 Å from the C^α atom, in a geometrical orientation defined by the other atoms bound to the donor. All other mean values drawn from the hydrogen bonds in DFPase lie below the confidence level and are therefore comparable to the values found in the literature.

3.3. Bond lengths and bond angles *versus* restraints

In Tables 3 and 4, standard values of statistical analysis are listed in cases where great differences to the respective restraints occur: the weighted means (μ), the respective weighted sample standard deviations (σ) and the mean standard deviations (σ_μ) were calculated from the 208 most reliable residues of the refined DFPase structure. According to Engh & Huber (1991), the standard deviation of the mean provides an estimate of the accuracy of this value, while the standard deviation of a parameter in the sample provides its force constant, since the force constants of stereochemical restraints are directly proportional to $1/\sigma^2$. Therefore, we have listed in parentheses the sample standard deviation and as an additional column the mean standard deviation, which was calculated as indicated in Appendix *B* and can be used for an immediate test according to the function $t = |\mu - m|/\sigma_\mu$.

To test the significance of deviations from the Engh & Huber (1991, 2001) parameters listed in the sixth column, *F*-test values were calculated as quotients of two χ^2 values. Engh & Huber (1991) employed the *F* distribution to investigate the reliability of the force constants and to *F*-test the consistency of samples with differing standard deviations. The *F*-test values are listed in column 7 of Tables 3 and 4. Significant values must exceed 2.18 or 3.06, corresponding to a 95% (italics) or 99% (bold) confidence level, respectively. In this work, the χ^2 values were calculated by distributing the measured values from -3σ to $+3\sigma$ into 21 bins, followed by a comparison with a Gaussian distribution, as described in Appendix *B*. The first of the two χ^2 values, listed as last column in Tables 3 and 4, was calculated with a Gaussian distribution derived from the weighted means of the samples and their standard deviations. A second χ^2 value was calculated analogously with a Gaussian distribution, derived from the respective restraint and its standard deviation. In a strict sense, the *F*-test values are only valid when the distribution of the samples is truly Gaussian, which can be tested by the first χ^2 value.

The restraints in column 6 are taken from Engh & Huber (1991) only when type-based mean values are listed, as indicated in column 2 by the corresponding bond-type names in bold. Since in a more recent publication (Engh & Huber, 2001) type-based mean values are no longer included, we also compared atom-based mean values listed for different amino acids (labelled with the corresponding amino-acid name in column 1) with the mean values from that publication. We use the terms 'type-based' and 'atom-based' in analogy to Brünger (1992). The χ^2 values in the last columns of Tables 3 and 4 indicate significant deviations from a Gaussian distribution

Table 4
Comparison of mean bond angles and restraints.

Residue	Angle type†	Angles	Mean‡ (°)	σ_μ§ (°)	Restraint¶ (°)	*F*-test††	χ_1^2
	C—CH1E—NH1	254	109.8 (2.1)	0.1	111.0 (2.7)	1.71	2.89
Thr	C—C$^\alpha$—N	18	110.5 (1.9)	0.5	111.0 (2.7)	*2.23*	0.19
	CH1E—C—NH1	258	116.4 (1.6)	0.1	117.2 (2.2)	1.82	1.89
Met	C$^\alpha$—C—N	7	116.5 (1.1)	0.4	117.2 (2.2)	**3.50**	3.43
HisD	C$^\alpha$—C—N	2	114.6 (2.9)	2.0	117.2 (2.2)	*2.29*	1.30
Trp	C$^\alpha$—C—N	6	116.9 (0.9)	0.4	117.2 (2.2)	2.16	1.23
	O—C—CH1E	276	120.7 (1.4)	0.1	120.1 (2.1)	*2.28*	2.73
Cys	O—C—C$^\alpha$	8	120.9 (1.4)	0.5	120.1 (2.1)	2.13	1.05
	O—C—NH1	287	122.9 (1.3)	0.1	122.7 (1.6)	1.48	1.41
Ser	O—C—N	7	122.1 (2.3)	0.9	122.7 (1.6)	*2.29*	1.47
	C—CH1E—CH2E	180	110.7 (2.5)	0.2	110.4 (2.0)	1.01	1.32
Asp	C—C$^\alpha$—C$^\beta$	19	112.7 (3.2)	0.7	110.4 (2.0)	**3.54**	1.01
HisD	C—C$^\alpha$—C$^\beta$	2	107.5 (0.1)	0.1	110.4 (2.0)	*2.21*	0.73
Tyr	C—C$^\alpha$—C$^\beta$	8	109.1 (2.4)	0.8	110.4 (2.0)	*2.64*	0.65
	CH2E—CH1E—NH1	160	110.8 (1.6)	0.1	110.6 (1.8)	1.01	1.23
Ser	C$^\beta$—C$^\alpha$—N	6	110.6 (2.1)	0.8	110.5 (1.5)	*2.33*	0.69
Tyr	C$^\beta$—C$^\alpha$—N	8	111.1 (2.5)	0.9	110.6 (1.8)	*2.79*	0.96
	CH1E—CH2E—C	34	113.3 (2.6)	0.5	113.4 (2.2)	1.01	0.79
Asn	C$^\alpha$—C$^\beta$—C$^\gamma$	15	113.3 (2.0)	0.5	113.4 (2.2)	*2.45*	1.24
	CH1E—CH2E—CH1E	12	114.4 (0.8)	0.2	115.3 (2.3)	*2.65*	2.08
Leu	C$^\alpha$—C$^\beta$—C$^\gamma$	12					
	OC—C—OC	34	124.0 (1.6)	0.3	123.3 (1.9)	**4.36**	6.30
Asp	O$^{\delta1}$—C$^\gamma$—O$^{\delta2}$	19	124.1 (1.3)	0.3	123.3 (1.9)	1.08	0.66
Glu	O$^{\varepsilon1}$—C$^\delta$—O$^{\varepsilon2}$	15	123.7 (2.2)	0.6	123.3 (1.2)	1.14	0.82
	CH2E—NH1—C	6	126.5 (2.3)	0.9	123.6 (1.4)	**8.78**	0.77
Arg	C$^\delta$—N$^\varepsilon$—C$^\zeta$	6					
	CH2E—C5—CR1E	6	131.5 (1.2)	0.5	129.1 (1.3)	**3.86**	0.94
His	C$^\beta$—C$^\gamma$—C$^{\delta2}$	6	131.5 (1.2)	0.5	131.4 (1.2)	1.13	0.98
HisD	C$^\beta$—C$^\gamma$—C$^{\delta2}$	2	131.7 (0.5)	0.4	130.8 (3.1)	1.65	0.56
HisE	C$^\beta$—C$^\gamma$—C$^{\delta2}$	4	131.4 (1.4)	0.7	129.7 (1.6)	*2.58*	0.80
	C5—CR1E—NH1	4	107.6 (1.4)	0.7	106.7 (1.2)	**5.41**	1.47
HisE	C$^\gamma$—C$^{\delta2}$—N$^{\varepsilon2}$	4					
	CH1E—CH1E—CH2E	18	111.6 (1.8)	0.4	111.0 (1.9)	*2.78*	2.45
Ile	C$^\alpha$—C$^\beta$—C$^{\gamma2}$	18					
	C5W—CW—CW	6	106.7 (1.0)	0.4	107.3 (0.8)	**4.04**	1.17
Trp	C$^\gamma$—C$^{\delta2}$—C$^{\varepsilon2}$	6					
	C5W—CW—CR1E	6	134.4 (1.5)	0.6	133.9 (0.9)	**3.73**	0.27
Trp	C$^\gamma$—C$^{\delta2}$—C$^{\varepsilon3}$	6					
	CR1E—NH1—CW	6	108.4 (1.0)	0.4	109.0 (0.9)	*2.88*	0.66
Trp	C$^{\delta1}$—N$^{\varepsilon1}$—C$^{\varepsilon2}$	6					

† The bond types in bold are labelled according to Engh & Huber (1991). ‡ Weighted mean, $\mu = \sum_i(\omega_i x_i)/\sum_i(\omega_i)$, and weighted sample standard deviation, $\sigma = \{n/(n-1)\sum_i[\omega_i(\mu - x_i)^2]/\sum_i(\omega_i)\}$, with weight $\omega_i = 1/\sigma_i^2$ (Bevington, 1969). The values in parentheses refer to the last digits of the mean and represent the sample standard deviation. § Mean standard deviation, $\sigma_\mu = (\sigma^2/n)^{1/2}$. ¶ Engh & Huber (1991, 2001). †† $F = \chi_1^2/\chi_2^2$, with χ_1^2 calculated with the mean values and their standard deviations, while χ_2^2 was calculated using the restraints.

whenever 7.6 or 10.1 are exceeded (corresponding to confidence levels of 95 or 99%, which are marked in bold or italic type, respectively). In Table 3, 15 *F*-test values exceed the tabulated 5% limit and nine values exceed the 1% limit for an *F*-distribution with $n - 2$ degrees of freedom (Bronshtein & Semendyayev, 1985), two of which are not Gaussian, while 21 values exceed the 5% and eight values even the 1% limit in Table 4.

Interestingly, each of the two deviations from a Gaussian distribution in Table 3 can be overcome with a Gaussian function fitted to the sample distribution. The lines labelled 'Gaussian fit' in the second column of Table 3 are shown whenever these fits have reduced the χ^2 value. In each case, the respective *F*-test value was increased and the peak positions of the fit were insignificantly shifted. The difference might arise from outliers ($>|3\sigma|$) which are not regarded in the sample distribution. Aside from that, we observed another abnormality only for mean values with high population ($n > 150$), which is not reflected in the χ^2 values but might be worth mentioning in this context. All these mean value distributions have a positive deviation close to their peak position, which can be described by a second very sharp Gaussian function. This sharp deviation seems to broaden when refinement without restraints progresses, but it does not vanish when the convergence of refinement is reached. Presumably, these deviations are remainders from the restrained refinement that was performed before the restraints were switched off.

The first three bond types in Table 3 belong to backbone bonds and according to their *F*-test values, two residues of these types show a significant difference in the DFPase: the C—O bond of the tryptophans and the N—C$^\alpha$ bond of the histidines. The differences observed for the C—N bond of methionines are on the borderline. Two other type-based mean values (CH1E—CH1E, C—CH2E) have no significant deviation from their restraints, but each shows a significant difference for an atom-based type of a single amino acid: (i) the C$^\alpha$—C$^\beta$ bond of the isoleucines and (ii) the C$^\gamma$—C$^\delta$ bond of the glutamines. All these five type-based mean values at the top of Table 3 have values in good agreement with their restraints owing to their relatively high populations, but cannot reflect the differences occurring for certain amino acids that belong to these respective types. On the other hand, the type-based mean values CH2G—NH1, CH2E—CH3E, C—NC2, C5—CR1E and C5—NH1 have been *F*-tested not to belong to the same parent distribution as the restraints; however, these values are based on small populations. Hence, these values need to be put on a larger base to obtain better statistics. The differences in the C5—CR1E and C5—NH1 restraints can be accounted for by the protonation state of the histidines. Subsequently, only the HisE entry belonging to the C5—CR1E type remains significantly different from the restraints, while the corresponding histidine bonds for the C5—NH1 type average to two new mean values of 1.389 (17) and 1.379 (4) Å, respectively, with new *F*-test values close to unity.

Table 5
Coordination distances in the Ca^{2+} polyhedrons.

(*a*) Individual bonds.

Bond	Length† (Å)	Bond	Length† (Å)
Ca1—Glu21 $O^{\varepsilon 2}$	2.348 (5)	Ca2—Asp232 $O^{\delta 2}$	2.212 (6)
Ca1—Asn120 $O^{\delta 1}$	2.348 (5)	Ca2—Leu273 O	2.263 (4)
Ca1—Asn175 $O^{\delta 1}$	2.397 (5)	Ca2—His274 $N^{\delta 1}$	2.373 (6)
Ca1—Asp229 $O^{\delta 1}$	2.357 (5)	Ca2—Wat	2.246 (5)
Ca1—Wat	2.349 (5)	Ca2—Wat	2.253 (5)
Ca1—Wat	2.476 (5)	Ca2—Wat	2.300 (5)
Ca1—Wat	2.515 (6)		

(*b*) Bond type.

	Bonds	Mean†‡ (Å)	σ_μ§ (Å)	Restraint¶ (Å)	*F*-test††
Ca—OC	6	2.321 (65)	0.027	2.320 (20)	1.10

† The numbers in parentheses refer to the last digits of the mean and represent the e.s.d.s or sample standard deviations, respectively. ‡ Weighted mean, $\mu = \sum_i(\omega_i x_i)/\sum_i(\omega_i)$, and weighted sample standard deviation, $\sigma = \{n/(n-1)\sum_i[\omega_i(\mu - x_i)^2]/\sum_i(\omega_i)\}$, with weight $\omega_i = 1/\sigma_i^2$ (Bevington, 1969). The values in parentheses refer to the last digits of the mean and represent the sample standard deviation. § Mean standard deviation, $\sigma_\mu = (\sigma^2/n)^{1/2}$. ¶ From *REFMAC*5 (Murshudov *et al.*, 1999). †† $F = \chi_1^2/\chi_2^2$, with χ_1^2 calculated with the mean values and their standard deviation, while χ_2^2 was calculated using the restraints.

Table 6
Averaged torsion angles.

	DFPase		Average†	
Angle type‡	θ_{min} (°)	σ (°)	θ_{min} (°)	σ (°)
NH1—CH1E	−105.5, 62.9	26.6	−89.2, 60.0	29.9
NH1—CH2G	−87.7, 93.4	28.8	90.9, 92.2	30.8
NH1—CH1P	−66.0	11.0	−65.8	10.4
CH1E—C	−13.4, 139.3	22.5	−27.0, 138.1	24.8
CH2G—C	8.7, 178.3	25.6	−7.0, 179.2	28.3
C—NH1	178.6	7.3	179.3	6.2
CH1E—CH2E	−64.7, 66.1, 183.7	7.1	−65.8, 64.7, 179.6	11.6
CH1E—CH1E	−58.0, 63.7, 180.8	5.5	−60.4, 62.7, 178.3	8.0
CH2E—CH2E	−65.5, 69.3, 179.7	10.6	−67.9, 69.4, 179.5	15.7
CH2E—C	−23.4, 152.1	38.9	−12.3, 165.1	41.0
CH2E—C5	−79.1, 100.3	30.1	−83.0, 97.6	34.4
CH2E—CF	−96.1, 91.0	15.8	−70.4, 83.2	26.0
CH2E—C5W	−126.4, 90.6	10.0	−77.3, 82.7	33.8
CH2E—CY	−99.8, 119.0	28.7	−76.1, 84.1	22.8
CH2E—SM	−88.2, 67.6	16.0	−67.7, 72.4, 181.4	19.0
CH2E—NH1	−86.1, 76.9, 178.6	19.9	−89.5, 91.8, 181.0	16.4
CH1P—CH2E	−24.4, 27.2	8.3	−24.0, 25.7	8.0
CH2E—CH2P	−36.3, 33.3	8.4	−33.4, 34.2	8.9
CH2P—CH2P	−28.9, 30.5	7.4	−30.6, 27.2	7.8
CH2P—N	−13.8, 14.0	6.5	−12.0, 15.6	6.5

† Priestle (2003). ‡ Angle types are labelled according to Engh & Huber (1991).

The statements made above regarding significant deviations of the bond-length mean values from the restraints also hold for the bond-angle mean values listed in Table 4. Again, the first type-based mean values with high sample populations show no major discrepancies from their corresponding restraints, but for individual amino acids they display several atom-based mean values with significant differences according to the *F*-test values. For example, in Table 4 the C—CH1E—CH2E type shows significant differences for Asp, HisD and Tyr. This bond-angle type was evaluated by Lamzin *et al.* (1995) for four protein structures refined at atomic resolution. We can confirm their threonine value, which was based in Lamzin *et al.* (1995) on 77 observations; however, in particular for less populated amino acids, we find other values. The values further down in Table 4 do not have high populations, but also show discrepancies in their type-based mean values. The mean value for CH2E—C5—CR1E can be split into two values by differentiating between the various histidine protonation states, thus generating smaller *F*-test values. Nevertheless, the value for HisE shows a significant deviation (95% confidence) in analogy with the $C^{\gamma}—C^{\delta 2}$ atom-based mean value mentioned above. On the other hand, the large discrepancy between the CH2E—C5—CR1E type-based mean value and the restraint published by Engh & Huber (1991) disappears when the value for a double-protonated histidine (His) reported in the more recent publication by Engh & Huber (2001) is used instead for comparison. Another type-based mean value for a bond angle involving the same $C^{\gamma}—C^{\delta 2}$ bond, C5—CR1E—NH1, also shows a significant *F*-test value for the $N^{\varepsilon 2}$-protonated histidines. We therefore propose that the deviations observed for HisE are real, in contrast to the deviation of the CH2E—C5—CR1E mean value, even though the number of samples on which this statement is based is rather small.

3.4. Distances of Ca^{2+}-ion coordination

The coordination distances involving the two Ca^{2+} polyhedra are listed in Table 5. A mean value is calculated from the Ca^{2+}—Glu, Ca^{2+}—Asn, Ca^{2+}—Asp and Ca^{2+}—Leu oxygen coordinations, which are all restrained to the same value in *REFMAC*5 (Murshudov *et al.*, 1999). Despite the different coordination numbers with seven (Ca1) and six (Ca2) ligands and despite the obvious elongations in the case of the sevenfold coordination at Ca1, the mean value agrees very well with the restraint, as confirmed by the low *F*-test value of 1.10. The two different coordination numbers are the reason for the relatively high sample standard deviation of this mean value. When the data are split according to the coordination number, the sample standard deviations are reduced and averaged to new mean values of 2.362 (24) and 2.247 (37) Å for sevenfold and sixfold coordination, respectively. Therefore, we assume that the restraint is only a rough estimate, which takes into account several different coordination numbers at once.

In *REFMAC*5 the Ca—His coordination is restrained to 2.325 (20) Å, a value considerably shorter than the measured coordination distance of 2.373 (6) Å (Table 4). Merging the N—Ca distances data found in the Metalloprotein Database (Castagnetto *et al.*, 2002), except for the distance from yeast frequenin (Ames *et al.*, 2000) whose coordinates were derived by modelling, the value becomes even shorter. The mean value of 2.2 Å with a standard deviation of 0.2 Å demonstrates that very few or no N—Ca distances in these structures were restrained, probably owing to the unavailability of a reliable restraint. A better choice for the application of a Ca—N coordination restraint in protein-structure refinements is the measured high-resolution coordination distance presented here.

Table 7
Penultimate rotamers (Lowell *et al.*, 2000) (four and three torsion angles).

	DFPase		Library
Type	Rotamers	%	%†
Arg			
ttp 180	1	17	3 (3)
mmm 180	1	17	1 (2)
ttm −85	1	17	3 (3)
mtp 180	1	17	5 (3)
ptp 85	1	17	18 (19)
Rest	1	15	<1 (1)
Lys			
mtpt	1	6	3 (2)
mttt	6	38	20 (14)
mttm	2	13	5 (5)
mtmt	1	6	3 (2)
tttp	2	13	4 (5)
mmtm	1	6	1 (1)
mmtt	1	6	6 (5)
tptt	1	6	3 (1)
Rest	1	6	19 (20)
Met			
mmm	2	50	19 (16)
tpp	1	25	5 (2)
ptp	1	25	2 (3)
Rest			14 (16)
Glu			
mm −40	7	47	13 (7)
mt **−10**	4	27	33 (29)
tt **0**	1	7	24 (42)
pt **−20**	1	7	5 (9)
Rest	2	12	9 (8)
Gln			
mm −40	4	34	15 (13)
pt **20**	1	8	4 (5)
mt **−30**	3	25	35 (26)
tt **0**	3	25	16 (29)
mm 100	1	8	3 (1)
Rest			12 (14)
Pro			
exo	7	35	43 (28)
endo	9	45	44 (54)
cis, *endo*	1	5	6 (1)
Rest	3	15	7 (16)

† Values in parentheses are for side chains with a β-fold at the respective backbone.

3.5. Torsion angles and rotamers

Recently, Priestle (2003) has analyzed the distributions of dihedral angles obtained from 46 high-resolution structures (<1.2 Å resolution) found in the PDB (Bernstein *et al.*, 1977) and compared them with restraints used for torsion angles in the programs *X-PLOR* (Brünger, 1992) and *CNS* (Brünger *et al.*, 1998). We have summed up the corresponding averaged torsion angles for DFPase and compared them in Table 6 with the values found by Priestle (2003). Obviously, when comparing respective values, the standard deviations (σ) of the torsion angles are in general large enough to prevent the observable differences from becoming significant. The only exception, the first angle in the CH2E—C5W angle type, is based on only one observation in DFPase, whereas the standard deviations were derived from all torsion angles of the respective type. Therefore, the difference for this torsion angle reflects only a normal deviation inside a larger ensemble, but not a significant difference from the average. For the CH2E—SM angle type we found only two different maxima instead of three as in Priestle (2003), probably owing to our limited number of observations. Nevertheless, because of the large standard deviations of the torsion angles, it appears that their force constants become small and their influence as restraints is very limited.

Another concept for incorporating the valuable information inherent in torsion angles into the refinement is worth considering, although it has not yet found entrance into any refinement program. Torsion angles of amino-acid side chains in proteins can only occupy a limited number of conformations in space, owing to energy minima that are related to van der Waals constraints and avoidance of clashes. These possible conformations are usually called rotamers of the respective amino acid. The first systematical classification of the rotamers for all 20 amino acids was presented by Ponder & Richards (1987). Dunbrack & Karplus (1993) introduced a backbone-dependence in their rotamer library for side-chain prediction. We have calculated torsion angles from our coordinates and classified them according to the Penultimate Rotamer Library, published by Lovell *et al.* (2000). The rotamers from four and three torsion angles are shown in Table 7, while Table 8 contains the rotamers of amino acids with two and one torsion angles. The symbols for the listed rotamers follow the nomenclature of Lovell *et al.* (2000). Letters are used in the symbols when the orientations of the respective torsion angles cluster close to the values 180°, +60° or −60°, where *t* stands for *trans*, *i.e.* 180°, *p* for *gauche*$^+$ or +60° and *m* for *gauche*$^-$ or −60°. In well determined cases, numbers are used in the last torsion-angle position for values different from the three standard angles by more than 5 or 10°. In addition, very flat distributions 180° wide are shown in bold.

The small number of rotamers which could not be classified and were thus accounted for in Tables 7 and 8 by the line called 'rest' indirectly confirms that most rotamers in DFPase fit well to the proposed types. In most cases there is only a single 'rest'. Only for proline and glutamate is the 'rest' larger (3 and 2, respectively), while phenylalanine, glutamine and all amino acids with only one torsion angle have no unclassifiable 'rest'. In the third column, the distribution of each amino acid side chain into the different rotamer types is given in percent for better comparison with the percentage of rotamers in the library distribution (fourth column), which was calculated from 240 structures of the PDB with a resolution better than 1.7 Å. The agreement between these distributions is poor for arginine, lysine and methionine. Still, some of these values show the same tendency, even though the number of samples is too low for reasonable statistics. Since for proline the number of rotamers is limited to three different possibilities, in this case the consensus with the library is quite high.

The agreement is generally better for rotamers of amino acids with two or one torsion angle. The highest agreement, of course, is found for the bottom amino acids of Table 8, where in particular isoleucine, leucine, tyrosine and phenylalanine fit well when the values in parentheses in the fourth column, which represent the percent distribution for rotamers of side chains with a β-folded backbone, are taken into account. Hence, the classification of side-chain torsion angles in the

Table 8
Penultimate rotamers (Lowell *et al.*, 2000) (two and one torsion angle).

	DFPase		Library
Type	Rotamers	%	%†
Asp			
m −20	6	32	51 (38)
p 30	6	32	9 (5)
t 0	5	26	21 (44)
p −10	1	5	10 (2)
Rest	1	5	4 (5)
Asn			
m −20	6	40	39 (28)
m 120	3	20	4 (3)
t 30	2	13	15 (18)
t −20	3	20	12 (21)
m −80	1	7	8 (9)
Rest			6 (12)
Ile			
pt	3	17	13 (13)
mt	9	50	60 (58)
mm	5	28	15 (16)
tt	1	5	8 (8)
Rest			1 (2)
Leu			
tp	5	42	29 (36)
mt	6	50	59 (46)
mp	1	8	2 (5)
Rest			7 (7)
His			
m 80	2	32	13 (10)
m −70	1	17	29 (30)
p −80	1	17	9 (6)
t 60	1	17	16 (17)
m 170	1	17	7 (3)
Rest			6 (8)
Trp			
m 95	3	60	32 (43)
t −105	1	20	16 (10)
Rest	1	20	6 (2)
Tyr			
m −85	4	50	43 (50)
m −**30**	1	13	9 (4)
t 80	2	24	34 (25)
p 90	1	13	13 (21)
Rest			2 (1)
Phe			
m −85	8	53	44 (51)
t 40	3	20	33 (18)
p 90	4	27	13 (24)
Rest			2 (1)
Thr			
m	3	19	43 (55)
t	1	6	7 (13)
p	12	75	49 (31)
Rest			1 (1)
Val			
t	13	76	73 (72)
p	1	6	6 (8)
m	3	18	20 (20)
Rest			1 (1)
Ser			
p	4	67	48 (36)
m	2	33	29 (29)
Rest			2 (0)
Cys			
t	3	37	26 (45)
m	3	37	50 (32)
p	2	26	23 (23)
Rest			1 (0)

† Values in parentheses are for side chains with a β-fold at the respective backbone.

Penultimate Rotamer Library also holds for atomic resolution structures and it is therefore a valuable tool for predicting side-chain orientations and could even aid the refinement process in future refinement programs. This may include a certain weakness for longer side chains, which should be improved by future investigations producing a larger statistical basis.

4. Conclusions

In the current study, with data derived from only a single atomic resolution data set, significant differences to the restraints derived by Engh & Huber (1991) were detected for bond lengths and bond angles affecting single amino acids. On the other hand, mean values derived from a larger number of samples and merged over several amino acids, such as backbone data for example, show lesser deviations from the restraints. To decide whether this conclusion is sound or not, it is necessary to consider the quality of the data obtained.

When the number of samples is small, the standard deviations calculated for the averaged bond lengths and bond angles remain within the range of values derived from small-molecule structures of the Cambridge Structural Database (CSD), which are commonly used as restraints. However, when the number of samples is larger ($n > 100$), the standard deviations calculated were clearly better than those of the restraints. This number of samples can only be reached when for example all backbone bonds or angles are combined. Data for single amino acids do not reach this extent. To improve the statistics for single amino acids, averaging over data of several ultrahigh-resolution data sets is required. Presumably, 10–20 atomic resolution structures of the same size as DFPase should be sufficient to create such a database which provides enough information to reach a comparable data quality as we obtained with the combined backbone data. To date, 135 protein and peptide entries with a resolution better than 1.2 Å and 48 with a resolution below 1.0 Å are listed in the PDB (Bernstein *et al.*, 1977). Hence, it now becomes feasible to construct a protein-based library of stereochemical parameters based on a sufficiently large database of atomic resolution structures.

The quality of data obtained from the hydrogen bonds in the antiparallel β-sheets of DFPase appears to be of high quality and agrees with the data published for primary and secondary hydrogen bonds. Hence, one could envisage using these data as additional restraints in low-resolution data sets where the number of observations is not high enough.

Rotamers are another class of data which are already used to facilitate homology modelling (Bower *et al.*, 1997), structure prediction (Schrauber *et al.*, 1993) and structure determination (Jones *et al.*, 1991). It is conceivable to use rotamer classes with high population as a guideline for the spatial orientation of side chains. With current computing power, it becomes feasible to test all possible orientations and to select only the best solution. In future structure-refinement programs, the use of rotamers might overcome the weak torsion-angle restraints

which, in our opinion, do not utilize the full information contained in these data.

At 0.8 Å resolution Schmidt & Lamzin (2002) drew an additional border, below which subatomic details of individual atoms become available. Multipole refinement (Guillot *et al.*, 2001) can account for deformations of the electron clouds owing to bonding orbitals between the individual atoms. This technique might provide interesting new results in a future study using all DFPase data obtained to the highest resolution of 0.82 Å.

APPENDIX *A*
Crystallographic target function

The crystallographic target function $\chi^2_{\rm Xray}$ is minimized in least-squares refinements to reduce discrepancies between the model and the N observed structure amplitudes,

$$\chi^2_{\rm Xray} = \sum_{\mathbf{h}}^{N} \frac{1}{\sigma_F^2(\mathbf{h})} [|F_o(\mathbf{h})| - |F_c(\mathbf{h})|]^2,$$

where $\mathbf{h} = (h, k, l)$ is the reciprocal-lattice vector, $|F_o|$ and $|F_c|$ are the observed and calculated structural amplitudes, respectively, and σ_F is the uncertainty of the measured structural amplitudes. Terms accounting for stereochemical restraints can be introduced into the target function when R differences between ideal values r^s and observations r^m (as for bond lengths, bond angles, torsion angles *etc.*), derived from the model and weighted with their inverse variances σ_r^2, are additionally minimized,

$$\chi^2_{\rm Rst} = \sum_{j}^{R} \frac{1}{\sigma_r^2(j)} [r^s - r^m(j)]^2.$$

The different restraint types are accounted for in n different restraint targets, $\chi^2_{\rm Rst}$ and an overall target function, X^2, can be defined as the sum of these individual restraint targets and the crystallographic target function,

$$X^2 = \sum_{i}^{n+1} \chi_i^2 = \chi^2_{\rm Xray} + \sum_{i}^{n} \chi^2_{\rm Rst}.$$

Precise ideal values, commonly termed restraints and used in protein crystallography to increase the number of observations, were derived by Engh & Huber (1991) from small-molecule crystallographic data stored in the Cambridge Structural Database. Moreover, in a more recent contribution by Engh & Huber (2001) to *International Tables for Crystallography Vol. F*, individual mean values for the 20 different amino acids are listed, calculated from a more recent version of the same database. In a macromolecular structure refinement at low resolution, up to half of the terms used to calculate the overall target function X^2 arise from restraints and may therefore generate a huge bias in the minima search in χ^2-space. Accurate restraints are thus essential for refinement at low resolution.

APPENDIX *B*
Error analysis

The best experimental estimate of the parent standard deviation σ is given by the sample standard deviation s of a distribution function fitting the data. Provided the measurements of unequal uncertainties follow a Gaussian distribution, the most probable value for the mean μ is the weighted average

$$\mu = \frac{\sum_i \omega_i x_i}{\sum_i \omega_i}$$

and the standard deviation can be estimated from the weighted variance of the data (Bevington, 1969),

$$\sigma \simeq s = \left[\frac{n}{(n-1)} \frac{\sum_i \omega_i (\mu - x_i)^2}{\sum_i \omega_i} \right]^{1/2},$$

with the weight $\omega_i = 1/\sigma_i^2$. Finally, the standard deviation of the mean σ_μ is

$$\sigma_\mu = (\sigma^2/n)^{1/2}.$$

With these data, tests can be made to verify a hypothesis concerning the distribution around a mean value. A test function t can be defined as

$$t = |\mu - m|/\sigma_\mu.$$

This follows the Student's distribution (Fisher & Yates, 1953) and is used to test whether the distribution underlying the mean value μ deviates significantly from a distribution with the mean value m when the standard deviation σ_m is unknown. For a greater number of samples ($n > 30$), $t > 3$ provides a confidence level of 99.8% for this deviation.

To calculate χ^2 values, the measured samples are distributed into n bins to obtain a sample distribution D. For each bin, the distributed values D_i are compared with values derived from a Gaussian distribution G, *e.g.* calculated from the corresponding mean and its standard deviation,

$$\chi^2 = \sum_{i=1}^{n} \frac{(G_i - D_i)^2}{\sigma^2(D_i)} \simeq \sum_{i=1}^{n} \frac{(G_i - D_i)^2}{D_i}.$$

Since the standard deviation of the sample distribution is not known, its value is approximated by the mean of all D_i, which is strictly valid only for a Poisson distribution. The χ^2 value has to be normalized to the number of degrees of freedom,

$$\chi^2_\nu = \chi^2/(n - \nu),$$

where ν is the number of refined parameters.

A χ^2 test can be performed if the distribution D of the samples is normal. This test yields a criterion of whether the hypothesis is true or has to be rejected (*e.g.* according to a confidence level of 95 or 99%) when tabulated limits of a χ^2 distribution with $n - \nu$ degrees of freedom are exceeded (Bronshtein & Semendyayev, 1985). Analogously, χ^2 values of two different distributions can be tested using the F test if they belong to the same parent distribution. In this case, the F-test value proves this hypothesis to be 95 or 99% probable when

the tabulated 5 or 1% limits, respectively, for an F-distribution with $n - \nu$ degrees of freedom (Bronshtein & Semendyayev, 1985) are exceeded. The F-test value can be calculated as the quotient of the two respective χ^2 values,

$$F = \chi_1^2/\chi_2^2.$$

If $F < 1$, then the reciprocal $1/F$ is used.

We thank A. N. Popov for introducing us to the flash-cooling method, and V. Lamzin and A. Schmidt for fruitful discussions about atomic resolution. We acknowledge help by the staff of EMBL Hamburg Outstation and the 'European Community Access to Research Infrastructure Action of the Improving Human Potential Programme to the EMBL Hamburg Outstation, contract No. HPRI-CT-1999-00017'.

References

Ames, J. B., Hendricks, K. B., Strahl, T., Huttner, I. G., Hamasaki, N. & Thorner, J. (2000). *Biochemistry*, **39**, 12149–12161.
Bacon, D. J. & Anderson, W. F. (1988). *J. Mol. Graph.* **6**, 219–220.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Bevington, P. R. (1969). *Data Reduction and Error Analysis for the Physical Sciences.* New York: McGraw–Hill.
Bouckaert, J., Dewallef, Y., Poortmans, F., Wyns, L. & Loris, R. (2000). *J. Biol. Chem.* **275**, 19778–19787.
Bower, M. J., Cohen, F. E. & Dunbrack, R. L. (1997). *J. Mol. Biol.* **267**, 1268–1282.
Bronshtein, I. N. & Semendyayev, K. A. (1985). *Handbook of Mathematics.* New York: Van Nostrand Reinhold Co.
Brünger, A. T. (1992). *X-PLOR, Version* 3.1: *A System for X-ray Crystallography and NMR.* New Haven, USA: Yale University Press.
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.
Castagnetto, J. M., Hennessy, S. W., Roberts, V. A., Getzoff, E. D., Tainer, J. A. & Pique, M. E. (2002). *Nucleic Acids Res.* **30**, 379–382.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Diederichs, K. (2000). *J. Appl. Cryst.* **33**, 1154–1161.
Dunbrack, R. L. Jr & Karplus, M. (1993). *J. Mol. Biol.* **230**, 543–574.
Eichinger, A., Beisel, H.-G., Jacob, U., Huber, R., Medrano, F.-J., Banbula, A., Potempa, J., Travis, J. & Bode, W. (1999). *EMBO J.* **18**, 5453–5462.
Engh, R. A. & 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. 382–416. Dordrecht: Kluwer Academic Publishers.
Fabiola, G. F., Krishnaswamy, S., Nagarajan, V. & Pattabhi, V. (1997). *Acta Cryst.* D**53**, 316–320.
Fisher, R. A. & Yates, F. (1953). *Statistical Tables for Biological, Agricultural and Medical Research.* New York: Oliver & Bond.
Grishkovskaya, I., Avvakumov, G. V., Sklenar, G., Dales, D., Hammond, G. L. & Muller, Y. A. (2000). *EMBO J.* **19**, 504–512.
Guillot, B., Viry, L., Guillot, R., Lecomte, C. & Jelsch, C. (2001). *J. Appl. Cryst.* **34**, 214–223.
Hoskin, F. G. C., Kirkish, M. & Steinmann, K. (1984). *Fundam. Appl. Toxicol.* **4**, 165–172.
Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.
Koepke, J., Scharff, E. I., Lücke, C., Rüterjans, H. & Fritzsch, G. (2002). *Acta Cryst.* D**58**, 1757–1759.
Lamzin, V. S., Dauter, Z. & Wilson, K. S. (1995). *J. Appl. Cryst.* **28**, 338–340.
Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EAMCB Newsl. Protein Crystallogr.* **26**, 27–33.
Longhi, S., Czjzek, M. & Cambillau, C. (1998). *Curr. Opin. Struct. Biol.* **8**, 730–737.
Lovell, S. C., Word, M. J., Richardson, J. S. & Richardson, D. C. (2000). *Proteins*, **40**, 389–408.
McRee, D. E. (1999). *Practical Protein Crystallography.* San Diego: Academic Press.
Merritt, E. A. & Murphy, M. E. P. (1994). *Acta Cryst.* D**50**, 869–873.
Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.* **91**, 201–228.
Morishita, Y., Hasegawa, K., Matsuura, Y., Katsube, Y., Kubota, M. & Sakai, S. (1997). *J. Mol. Biol.* **267**, 661–672.
Murshudov, G. N., Lebedev, A., Vagin, A. A., Wilson, K. S. & Dodson, E. J. (1999). *Acta Cryst.* D**55**, 247–255.
Nayal, M. & Di Cera, E. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 817–821.
Ponder, J. W. & Richards, F. M. (1987). *J. Mol. Biol.* **193**, 775–791.
Priestle, J. P. (2003). *J. Appl. Cryst.* **36**, 34–42.
Samygina, V. R., Antonyuk, S. V., Lamzin, V. S. & Popov, A. N. (2000). *Acta Cryst.* D**56**, 595–603.
Scharff, E. I., Koepke, J., Fritzsch, G., Lücke, C. & Rüterjans, H. (2001). *Structure*, **9**, 493–502.
Scharff, E. I., Lücke, C., Fritzsch, G., Koepke, J., Hartleib, J., Dierl, S. & Rüterjans, H. (2001). *Acta Cryst.* D**57**, 148–149.
Schmidt, A. & Lamzin, V. S. (2002). *Curr. Opin. Struct. Biol.* **12**, 698–703.
Schrauber, H., Eisenhaber, F. & Argos, P. (1993). *J. Mol. Biol.* **230**, 592–612.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.
Spiller, B., Gershenson, A., Arnold, F. H. & Stevens, R. C. (1999). *Proc. Natl Acad. Sci. USA*, **96**, 12305–12310.
Willis, M. A., Bishop, B., Regan, L. & Brunger, A. T. (2000). *Structure*, **8**, 1319–1328.