J. Appl. Cryst. (1995). **28**, 338–340

Dictionary of Protein Stereochemistry

BY VICTOR S. LAMZIN, ZBIGNIEW DAUTER AND KEITH S. WILSON

European Molecular Biology Laboratory (EMBL), c/o DESY, Notkestrasse 85, 22603 *Hamburg, Germany*

(*Received* 31 *January* 1994; *accepted* 1 *November* 1994)

Abstract

A computer-readable dictionary for protein structure refinement is presented. The dictionary is based on the set of interatomic bonds and bond angles previously derived from small-molecule structures by Engh & Huber [*Acta Cryst.* (1991), A**47**, 392–400]. The internal inconsistency of the set is discussed. Some preliminary results derived from several protein structures refined at atomic resolution are compared with this set.

1. Introduction

X-ray crystallography is the most powerful technique for determining the three-dimensional structure of molecules. For small molecules, where diffraction extends to atomic resolution (about 1.0 Å or further), the spatial coordinates of every atom in the structure can usually be refined with high accuracy. However, if diffraction does not extend to atomic resolution, there are not enough observations to model the structure solely from the X-ray data. For a protein crystal, diffracting to about 2.0 Å resolution (or lower), the number of observations is comparable to the number of parameters to be determined. This requires additional information to be introduced on the basis of *a priori* knowledge of protein stereochemistry. Such information is usually provided as a set of interatomic distances and/or angles, planarity, van der Waals radii *etc.* as restraints and is derived from structures of small molecules and peptides. The most recent library is the set of bond distances and angles with their standard deviations generated from the Cambridge Structural Database (Allen, Kennard & Taylor, 1983) by Engh & Huber (1991), henceforth referred to as EH.

We describe here the use of the EH parameters in the form of a computer-readable library for the *PROTIN/PROLSQ* package (Konnert & Hendrickson, 1980; CCP4, 1994), one of those most widely used to refine protein structures.

2. Methods

The EH parameters are not completely self-consistent. For example, in the planar phenyl ring of the tyrosine side chain, the angles do not sum exactly to 360° (there are small rounding errors: the sum of three bond angles around the CG atom is 359.7° and around CZ is 360.1°). Thus, the EH set does not exactly represent the ideal structures of the amino acid residues. The simplest way to overcome this problem is to use the existing *PROTIN/PROLSQ* dictionary and match it by least-squares minimization of (1) for each residue. Conditions for planar groups should be introduced as constraints.

$$\mathcal{F} = \sum_i w_i(a_i - a_{\rm EH})^2 + \lambda \sum_i v_i(\theta_i - \theta_{\rm EH})^2, \qquad (1)$$

where a_i and $a_{\rm EH}$ are the current and EH target distances, θ_i and $\theta_{\rm EH}$ are the current and EH target angles and the weights w_i and v_i are $1/\sigma^2(a_i)$ and $1/\sigma^2(\theta_i)$, respectively. There is, however, a complication in defining the parameter λ, which is the relative weight for angles and bonds. This complication can be avoided by the minimization, instead, of

$$\mathcal{F} = \sum_i w_i(a_i - a_{\rm EH})^2, \qquad (2)$$

where a_i and $a_{\rm EH}$ are current and EH distances between 1–2 and 1–3 bonded atoms, with corresponding weights w_i. The angle distance a, which we define as a distance between 1–3 bonded atoms, is easily derived using the cosine law if the two bond lengths b and c, as well as the angle θ, are known (Fig. 1).

If the errors in position for atoms A, B and C follow an isotropic three-dimensional Gaussian distribution, (3) and (4) relate the variances in atomic position $\sigma^2(A)$, $\sigma^2(B)$ and $\sigma^2(C)$ to the variances in bond distances $\sigma^2(b)$ and $\sigma^2(c)$. The variance of the angle can be expressed by (5).

$$\sigma^2(b) = \sigma^2(A) + \sigma^2(B), \qquad (3)$$

$$\sigma^2(c) = \sigma^2(A) + \sigma^2(C), \qquad (4)$$

$$\sigma^2(\theta) = [\sigma^2(C)/c^2] + [\sigma^2(A)a^2/b^2c^2] + [\sigma^2(B)/b^2]. \qquad (5)$$

In theory, the system of three linear equations (3)–(5) allows determination of variances in positions for atoms B and C and, therefore, a variance for the angle distance to use in function (2). The solutions for σ^2 must be positive. However, in practice, when we tried to evaluate σ^2 for an angle distance, 70% of the solutions were negative. Thus, for most bond and angle distances in

the EH set, either the lengths and standard deviations for a and b (see Fig. 1) are not consistent with the value and standard deviation for the angle θ, or the standard deviations given do not represent a Gaussian distribution.

An upper estimate of the standard deviation of an angle distance can be obtained with the assumption that the central atom A is fixed and $\sigma^2(A)$ is zero. Then, $\sigma(a)$ is simply equal to $[\sigma^2(b) + \sigma^2(c)]^2$. This simplified assumption was used in minimization of function (2).

3. Results and discussion

The results of regularization for 20 amino acids are briefly summarized in Table 1. There is a problem with proline, arising from the fact that the main-chain parameters for this residue differ from those for other residues in the EH set. This indicates that proline requires a separate definition for its main chain. Apart from proline, all the dictionary entries are geometrized with only a marginal round-off error of 0.001 Å, which corresponds to the precision quoted by Engh & Huber (1991). A preliminary version of our dictionary for the *PROTIN/PROLSQ* package was deposited in the public domain of CCP4 (1994) about a year ago. An updated version has now been submitted.

Recently, there has been a parallel attempt to implement the EH parameters in a computer-readable form (Priestle, 1994). However, the dictionary generated suffers from some discrepancies from the EH set. These concern C–CA–CB angles for all groups as well as angle distances for some side chains. This introduces an additional root-mean-square (r.m.s.) error of about 0.01 Å, comparable to the r.m.s. deviation in distances usually obtained for highly refined protein models.

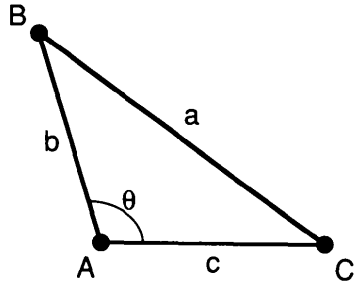

Fig. 1. A bond angle.

Table 1. *R.m.s. deviation* (Å) *of dictionary parameters compared to values suggested by Engh & Huber* (1991)

Main-chain parameters were the same for every residue.

Residue	Bond distances	Angle distances
Ala	0.0002	0.0002
Arg	0.0005	0.0007
Asn	0.0004	0.0004
Asp	0.0003	0.0005
Cys	0.0001	0.0004
Gln	0.0004	0.0000
Glu	0.0004	0.0005
Gly	0.0000	0.0001
His	0.0004	0.0012
Ile	0.0004	0.0007
Leu	0.0003	0.0006
Lys	0.0002	0.0010
Met	0.0003	0.0004
Phe	0.0004	0.0012
Pro	0.0041	0.0189
Ser	0.0002	0.0003
Thr	0.0002	0.0003
Trp	0.0005	0.0010
Tyr	0.0006	0.0011
Val	0.0004	0.0005

The EH set is based on the analysis of small molecules. Structural parameters of protein residues may differ from these. Recently, we began the determination of several protein crystal structures at atomic resolution (approaching 1.0 Å). Synchrotron radiation and an imaging-plate scanner provided high-quality X-ray data and details of the full analyses will be published elsewhere. Data to atomic resolution allow comprehensive refinement of protein structures without, or with only marginal, restraints. Moreover, full/block-matrix least-squares refinement, *e.g.* with *SHELXL*93 (Sheldrick, 1993), provides proper estimation of coordinate variance. Accurate and statistically meaningful stereochemical

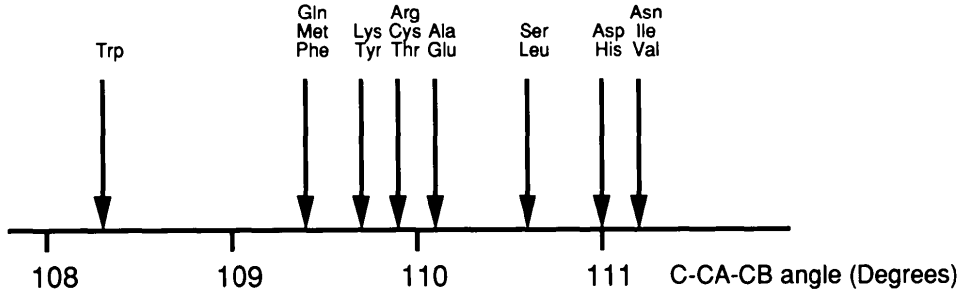

Fig. 2. Schematic representation of the mean value for the C–CA–CB angle in different side chains as derived from four protein structures refined at atomic resolution. The standard deviation of the mean value for a residue type ranges between 0.2° for Thr with 77 observations and 0.7° for Trp with 9 observations.

parameters can be derived when enough structures have been refined. As an example, based on four such structures whose refinement is now complete, we looked at how the C–CA–CB angle varies with side-chain type. The distribution obtained is shown in Fig. 2. This result is, of course, a preliminary one, but clearly differs from the EH library, where the C–CA–CB angle is specified as 110.5° for Ala, 110.1° for Asp, Asn, Arg, Gln, Glu, Cys, Lys, His, Leu, Met, Phe, Ser, Trp and Tyr and 109.1° for Ile, Thr and Val.

References

ALLEN, F. H., KENNARD, O. & TAYLOR, R. (1983). *Acc. Chem. Res.* **16**, 146–153.

CCP4 (1994). *Acta Cryst.* D**50**, 760–763.

ENGH, R. A. & HUBER R. (1991). *Acta Cryst.* A**47**, 392–400.

KONNERT, J. H. & HENDRICKSON, W. A. (1980). *Acta Cryst.* A**36**, 344–350.

PRIESTLE, J. P. (1994). *Structure*, **2**, 911–913.

SHELDRICK, G. M. (1993). *SHELXL*93. *Program for Crystal Structure Refinement*. Univ. of Göttingen, Germany.