[$(Pb, Sb, Sn^{2+})S$] and that of the *h* layer [$(Sn^{4+}, Fe)S_2$]. An approximate common multiple volume (pseudocell) could be determined in both cylindrite and franckeite using the lattice parameters given by Makovicky (1976) and Mozgova *et al.* (1976). The ratio of cations between the two layers in the pseudocells in the models are 2·58 and 1·37 for franckeite and cylindrite, respectively. Following Evans & Allmann (1968), we can express the crystallochemical formulae of franckeite and cylindrite as follows:

franckeite $2{\cdot}58[(Pb^{2+}, Sb^{3+}, Sn^{2+})S][(Sn^{4+}, Fe^{2+})S_2]$

cylindrite $1{\cdot}37[(Pb^{2+}, Sb^{3+}, Sn^{2+})S][(Sn^{4+}, Fe^{2+})S_2]$.

Concluding remarks

Makovicky (1976) and Williams & Hyde (1988*a*, *b*) previously presented b^*c^* transmission electron diffraction patterns of cylindrite and Williams & Hyde (1988*a*, *b*) also presented corresponding HRTEM images. Our results are somewhat different from these. The present TEM study indicates that the two layers in cylindrite have different stacking vectors. In addition, separate *h*- and *t*-layer electron diffraction patterns and their HRTEM images were obtained. Based on the common modulation of the two layers and the common CBED patterns, the relations between the two lattices and between the lattices and the modulations were determined. This study also indicates that there are two types of incommensurability in cylindrite and franckeite, the incommensurability between the two lattices in each structure and the incommensurability between the lattices and modulations. The alternative wave-structure models of cylindrite and franckeite are suggested, simulated and discussed in this paper, pointing out the modulations resulting from the mismatching relation between the two layers.

The authors thank Professors E. M. Huang and Z. S. Ma for providing the cylindrite and franckeite samples for this study. Discussions with Professors Y. M. Chu, K. K. Feng and F. H. Li are gratefully appreciated.

References

AALST, W. VAN, DEN HOLLANDER, J., PETERSE, W. J. A. M. & DE WOLFF, P. M. (1976) *Acta Cryst.* B**32**, 47-58.

EVANS, H. T. & ALLMANN, R. (1968). *Z. Kristallogr.* **127**, 73-79.

FRENZEL, A. (1893). *Neues Jahb. Mineral Geol. Palaeontol.* **2**, 125-128.

HIRSCH, P. B., HOWIE, A., NICHOLSON, R. B., PASHLEY, D. W. & WHELAN, M. J. (1977). *Electron Microscopy of Thin Crystals.* New York: Kireger.

HUANG, M., WU, G., CHEN, Y. & TANG, S. (1986). *Acta Geol. Sin.* **2**, 164-175.

JANNER, A. & JANSSEN, T. (1977). *Phys. Rev. B*, **15**, 643-658.

JANNER, A. & JANSSEN, T. (1980). *Acta Cryst.* A**36**, 399-408, 408-415.

KISSIN, S. A. & OWEUS, D. R. (1986). *Can. Mineral.* **24**, 45-50.

MAKOVICKY, E. (1971). *Neues Jahb. Mineral. Monatsh.* pp. 404-413.

MAKOVICKY, E. (1974). *Neues Jahrb. Mineral. Monatsh.* pp. 235-256.

MAKOVICKY, E. (1976). *Neues Jahrb. Mineral. Abh.* **126**, 304-306.

MOH, G. H. (1984). *Mineral. Abh.* **150**, 25-64.

MOH, G. H. (1986). *Neues Jahrb. Mineral. Abh.* **153**, 267-272.

MORITZ, H. (1933). *Neues Jahrb. Mineral. Beil.* **66**, Abt. A, 191-212.

MOZGOVA, N. N., BORODAYEV, YU. S. & SVESHNIKOVA, O. L. (1975). *Dokl. Akad. Nauk SSSR*, **220**, 107-110.

MOZGOVA, N. N., ORGANOVA, N. I. & GORSHKOV, A. I. (1976). *Dokl. Akad. Nauk SSSR* (*Engl. transl.*) **228**, 110-113.

RAMDOHR, P. (1960). *Die Erzmineralien und ihre Verwachanngen.* Berlin: Akademie Verlag.

WANG, S. (1988). *Inst. Phys. Conf. Ser.* No. 93, Vol. 2, p. 331.

WANG, S. (1989). *47th Annual Proc. Electron Microscopy Soc. of America*, edited by G. W. BAILEY, p. 420. San Francisco Press.

WILLIAMS, T. B. & HYDE, B. G. (1988*a*). *Acta Cryst.* B**44**, 467-474.

WILLIAMS, T. B. & HYDE, B. G. (1988*b*). *Phys. Chem. Miner.* **15**, 521-544.

WOLFF, P. M. DE (1974). *Acta Cryst.* A**30**, 777—785.

WOLFF, P. M. DE (1977). *Acta Cryst.* A**33**, 493-497.

YADA, K. (1979). *Can. Mineral.* pp. 679-691.

ZUSSMAN, J. (1954). *Mineral. Mag.* **30**, 498.

Acta Cryst. (1991). A**47**, 392-400

Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement

BY RICHARD A. ENGH AND ROBERT HUBER

*Max Planck Institut für Biochemie, D*8033 *Martinsried bei München, Germany*

(*Received* 14 *November* 1990; *accepted* 17 *January* 1991)

Abstract

Bond-length and bond-angle parameters are derived from a statistical survey of X-ray structures of small compounds from the Cambridge Structural Database. The side chains of the common amino acids and the polypeptide backbone were represented by appropriate chemical fragments taken from the Database. Average bond lengths and bond angles are determined from the resulting samples and the sample standard deviations provide information regarding the expected variability of the average values which

0108-7673/91/040392-09$03.00

can be parametrized as force constants. These parameters are ideally suited for the refinement of protein structures determined by X-ray crystallography since they are derived from X-ray structures, are accurate to within the deviations from target values suggested for X-ray structure refinement and use force constants which directly reflect the variability or uncertainty of the average values. Tests of refinement of the structures of BPTI and phycocyanin demonstrate the integrity of the parameters and comparisons of equivalent refinements with *XPLOR* parameters show improvement in *R* factors and geometry statistics.

Introduction

The determination and refinement of protein structures by X-ray diffraction requires structural information supplemental to the experimental X-ray data. This increases the ratio of observations (reflections, geometric information) to model parameters (coordinates and temperature factors). The structural information consists of a set of geometric parameters representing bond lengths, angles, planarity, dihedral angles and sometimes the anticipated deviations from these values for particular geometries (Jensen, 1985). Ideally, this geometric information should be as accurate as known, should reflect the quantities measured in the experiment (for example, a bond length is the average distance between the centers of scattering functions corrected for thermal motion) and should be weighted according to the reliability of the parameter. In practice, however, these conditions are usually not met. In fact, some of the parameters commonly used for refinement are not accurate to within the accuracy of the experimental data and many are not accurate to within the deviations from ideality considered acceptable for refined structures (Hendrickson, 1985). The greatest errors arise from the use of parameter lists with few atom types (Jack & Levitt, 1978; Brooks, Bruccoleri, Olafson, States, Swaminathan & Karplus, 1983; Brünger, Karplus & Petsko, 1989; van Gunsteren & Berendsen, 1987) most commonly seen with molecular dynamics parameters. Programs such as *PROLSQ* and *TNT* (Hendrickson & Konnert, 1980; Tronrud, Ten Eyck & Matthews, 1987) which derive parameter lists from coordinates of ideal fragments (Bowen, Donohue, Jenkin, Kennard, Wheatley & Whiffen, 1958; Vijayan, 1976; Allen, Kennard, Watson, Brammer, Orpen & Taylor, 1987) separately for each fragment are typically better.

The Cambridge Structural Database (Allen, Kennard & Taylor, 1983) is the appropriate source of geometrical information for the derivation of an updated set of parameters for structure refinement. The structures are themselves determined by X-ray crystallography (fewer than 1% by neutron diffraction) and are small enough to be fully determined by the diffraction data. Parameters derived from these structures then directly reflect average centers of electron density and are appropriate for protein crystallography. The Database has over 80 000 structures, providing ample population sizes for significant statistical analysis for most model fragments which occur in proteins. The organization of the Database facilitates rapid searching and statistical analysis.

Geometric restraints are usually applied such that the bond lengths are distributed about their ideal values with a standard deviation of less than 0·02 Å and bond angles about their ideal values with a standard deviation of about 2° (Hendrickson, 1985). This requirement provides a means of determining the weighting of the geometric restraints relative to diffraction data but does not reflect the accuracy of the restraints. In fact, some ideal values for identical bonds differ by more than 0·1 Å between different parameter sets and ideal angle values by more than 10°. This *exceeds* the estimated accuracy of a high-resolution protein structure (Jensen, 1985). Smaller errors in the parameters also cause systematic distortions of geometries with unforeseen consequences, particularly for statistical studies of protein geometric properties. Although some parameters may compensate for inaccuracies in other parameters, for example planarity restraints for aromatic ring structures with poor bond-angle parameters, and may lead to acceptable geometries, deviations from the 'ideal' values will be large.

The restraints are usually weighted either according to class (bonds, angles *etc.*) or by a harmonic force constant supposed to reflect the flexibility of the bond or angle. The former approach makes no allowance for greater variability of specific average bond lengths or angles. The latter approach may do so only indirectly: X-ray diffraction determines the *average* atomic positions and thus the *average* bond lengths and angles. The extent to which these *average* values vary as a function of the protein environment is not identical to the flexibility or the vibrational amplitudes of the bonds or angles. Since the restraints are most directly informational restraints, a better approach may be to determine the force constants from the observed variation in the Database. Restraints for bonds and angles which vary more widely should then have weaker force constants.

In this paper, we present the results of such a statistical analysis of geometric parameters taken from structures from the Cambridge Structural Database (CSD). We compare the parameters so derived with *EREF* (Jack & Levitt, 1978) and *XPLOR* param19x.pro (Brooks *et al.*, 1983; Brünger, Karplus & Petsko, 1989) parameters and describe the improvements. We have condensed these results for use by *XPLOR* by defining 14 new atom types not present in *XPLOR* topologies. As a test of the parameters,

we have re-refined the structures of C-phycocyanin of *Fremyella displosiphon* at 1·66 Å resolution (Duerring, Schmidt & Huber, 1991) and bovine pancreatic trypsin inhibitor at 1·2 Å resolution (Wlodawer, Walter, Huber & Sjöhn, 1984) using *XPLOR* with original and new parameters. Both structures show improvements in *R* factor and geometric energies (deviation from ideal values) with the new parameters. This demonstrates that the accuracy of the refinement parameters is an important consideration for final refinement of protein structures.

Methods

For each of the 20 commonly occurring amino acids, appropriate chemical fragments were selected from the Cambridge Structural Database in separate searches by name and by chemical connectivity using the program *QUEST**90. For larger residues, fragments representing particular groups were also selected, for example, all indole rings with a tetrahedral C atom substituted at the 3 position were selected for tryptophan. The chemical connectivity searches usually provided the largest samples.

The statistical properties of the bond and angle geometries of these selected structural fragments were calculated using the Cambridge Structural Database program *GSTAT**90. Flags NERR, NOO and NOD eliminated structures with known errors, overlapping chemical fragments and duplicate structures. The calculations were done for all structures with an *R* factor of less than 10% and then repeated for structures with an *R* factor of less than 6%. All bond lengths, angles and torsional angles were calculated. In addition, the planarity of certain groups such as the guanidine of arginine or carbonyl C atoms was investigated. Averages, standard deviations and standard errors were calculated, with and without elimination of values outside four standard deviations from the mean. There were usually one or no such outliers (the expected frequency in a normal distribution is ~0·01%).

These statistics deliver the values for the new geometric restraints for X-ray structure refinement. The standard deviation of the mean values provides an estimate of the accuracy of these values, depending primarily on the applicability of the choice of fragments to protein structure geometries. The standard deviation of a parameter in the sample provides its force constant. The reliability of these force constants can be estimated with the appropriate *F* distribution (Hamilton, 1964). The *F* test was also applied to test the consistency of samples with differing standard deviations; in some instances samples included unusual structures probably not representative of protein geometries.

A parameter set was created for the *XPLOR* program using the statistical information. This required the creation of 14 new atom types, shown in Table 1, in addition to the atom types in the standard *XPLOR* parametrization of param19x.pro (abbreviated here P19X). Most of these types distinguish between different types of 'bare' and ring C atoms; three additional types were required for the methionine S atom and proline and glycine CH_2 'extended' atoms. A new atom type was generally considered necessary when the average of a set of bond lengths or angles from two distinct chemical fragments differed by an amount greater than the standard deviation of the sample. The force constant was set to a value such that a thermal population at room temperature would be distributed, in the absence of other forces, with the same standard deviation about the mean as the statistical sample. These force constants were then scaled to provide consistency with the dihedral and improper dihedral force constants. Tables 2 and 3 show the bond and angle parameters, respectively.

Two structures, phycocyanin (Duerring, Schmidt & Huber, 1991) and BPTI (Wlodawer *et al.*, 1984), were re-refined with *XPLOR* using P19X and the new statistical parameters. To simplify the comparison, the P19X force constants were used for the CSD parameters (denoted CSD-X) as well. Thus, any differences between the final structures could unequivocally be attributed to the different ideal geometry values. The refinements proceeded from the

Table 1. *CSD parametrization atom types*

Atom type	Description
C	Carbonyl C atom of the peptide backbone
C5W*	Tryptophan C^{γ}
CW*	Tryptophan $C^{\delta 2}$, $C^{\epsilon 2}$
CF*	Phenylalanine C^{γ}
CY*	Tyrosine C^{γ}
CY2*	Tyrosine C^{ζ}
C5*	Histidine C^{γ}
CN*	Neutral carboxylic acid group C atom
CH1E	Tetrahedral C atom with one H atom
CH2E	Tetrahedral C atom with two H atoms (except CH2P, CH2G)
CH2P*	Proline C^{γ}, C^{δ}
CH2G*	Glycine C^{α}
CH3E	Tetrahedral C atom with three H atoms
CR1E	Aromatic ring C atom with one H atom (except CR1W, CRH, CRHH, CR1H)
CR1W*	Tryptophan $C^{\zeta 2}$, $C^{\eta 2}$
CRH*	Neutral histidine $C^{\epsilon 1}$
CRHH*	Charged histidine $C^{\epsilon 1}$
CR1H*	Charged histidine $C^{\delta 2}$
N	Peptide N atom of proline
NR	Unprotonated N atom in histidine
NP	Pyrrole N atom
NH1	Singly protonated N atom (His, Trp, peptide)
NH2	Doubly protonated N atom
NH3	Triply protonated N atom
NC2	Arginine $N^{\eta 1}$, $N^{\eta 2}$
O	Carbonyl O atom
OC	Carboxyl O atom
OH1	Hydroxyl O atom
S	S atom
SM*	Methionine S atom
SH1E	Singly protonated S atom

* Atom types marked with an asterisk are new (non-*XPLOR*) types.

Table 2. *Bond parameters*

Bond type	σ	Bond length (Å)
C5W–CW	0·018	1·433
CW–CW	0·017	1·409
C–CH1E	0·021	1·525
C5–CH2E	0·014	1·497
C5W–CH2E	0·031	1·498
CF–CH2E	0·023	1·502
CY–CH2E	0·022	1·512
C–CH2E	0·025	1·516
CN–CH2E	0·019	1·503
C–CH2G	0·018	1·516
C5W–CR1E	0·025	1·365
CW–CR1E	0·016	1·398
CW–CR1W	0·021	1·394
CF–CR1E	0·021	1·384
CY–CR1E	0·021	1·389
CY2–CR1E	0·024	1·378
C5–CR1H	0·011	1·354
C5–CR1E	0·011	1·356
C–N	0·016	1·341
C–NC2	0·018	1·326
C5–NH1	0·011	1·378
CW–NH1	0·011	1·370
C–NH1	0·014	1·329
C–NH2	0·021	1·328
C5–NR	0·017	1·371
C–O	0·020	1·231
CN–O	0·023	1·208
C–OC	0·019	1·249
CY2–OH1	0·021	1·376
C–OH1	0·022	1·304
CH1E–CH1E	0·027	1·540
CH1E–CH2E	0·020	1·530
CH1E–CH3E	0·033	1·521
CH1E–N	0·015	1·466
CH1E–NH1	0·019	1·458
CH1E–NH3	0·021	1·491
CH1E–OH1	0·016	1·433
CH2E–CH2E	0·030	1·520
CH2P–CH2E	0·050	1·492
CH2P–CH2P	0·034	1·503
CH2E–CH3E	0·039	1·513
CH2P–N	0·014	1·473
CH2G–NH1	0·016	1·451
CH2E–NH1	0·018	1·460
CH3E–NH1	0·018	1·460
CH2E–NH3	0·030	1·489
CH2E–OH1	0·020	1·417
CH2E–S	0·020	1·822
CH2E–SM	0·034	1·803
CH2E–SH1E	0·033	1·808
CH3E–SM	0·059	1·791
CR1E–CR1E	0·030	1·382
CR1E–CR1W	0·025	1·400
CR1W–CR1W	0·019	1·368
CR1E–NH1	0·021	1·374
CRH–NH1	0·020	1·345
CRHH–NH1	0·010	1·321
CR1H–NH1	0·011	1·374
CRH–NR	0·013	1·319

final refined structure of the two structures with *XPLOR* using conjugate gradients minimization and recalculation of phases with each step until the minimization failed to find further improvement in the total energy. The *R* factors and geometric parameters were compared.

The fit of the models to their electron-density maps were also compared. Based on the observation that errors in protein models can distort the electron-density map and disrupt the continuity of the density along bonds with equivalent *R* factors, we have defined a 'directed' real-space *R* factor:

$$R_b = \sum_{\text{bonds}} \left(\int_{\text{atom 1}}^{\text{atom 2}} |\rho_o - \rho_c| \, ds \right) \Big/ \sum_{\text{bonds}} \left(\int_{\text{atom 1}}^{\text{atom 2}} \rho_o \, ds \right)$$

where R_b denotes the 'bond *R* factor', the sum is over a selected set of bonds, the one-dimensional line integrals proceed along the line segment ('bond') between two atoms with differential element ds and ρ_o and ρ_c indicate the electron density (at the integration points along the bond) calculated from observed and calculated structure factors, respectively, using model phases. This is similar to the real-space *R* factor described by Brändén & Jones (1990), but differs in that the continuity of the density along bonds is emphasized, which can provide greater discrimination for the evaluation of structural errors (Engh, Sippl, Martin, Edwards & Huber, 1991).

Results

The greatest improvements in the geometric parameters we analyzed occurred for the aromatic amino acid residues. Fig. 1 shows the bond lengths for tryptophan derived from the Cambridge Database, plotted together with the standard deviation in the sample population and *EREF* and *XPLOR* param19x.pro (P19X) and *GROMOS* (van Gunsteren & Berendsen, 1987) parameters. It is readily seen that the statistical parameters deviate from the *EREF* and P19X parameters by amounts larger than the sample standard deviation, which is itself several times larger than the standard deviation of the mean. Similar results were seen for the bond-angle parameters, where in particular the P19X value of 122·5° for the $C^{\delta 1}$–C^{γ}–$C^{\delta 2}$ angle deviates from the statistical value of 106·2° by 16·3°. (This value in P19X is an artifact which results from the equivalence of the $C^{\delta 1}$, $C^{\epsilon 3}$ and $C^{\zeta 2}$ ring-atom types and the C^{γ}, $C^{\delta 2}$ and $C^{\epsilon 2}$ 'bare' C-atom types. See the derivation of new P19X parameters below for more information.) The sum of the interior angles of the five- and six-membered rings is 539·9 and 720·0°, respectively, as one would expect for the corresponding planar geometric figures. The corresponding *EREF* (P19X) interior-angle sums are 539·7 (555·5) and 716·0° (724·0°) for comparison.

Histidine could also be improved. The statistical geometries for histidine depended on its protonation state. The Database searches were carried out for all histidine (or appropriate imidazole analogs) fragments corresponding to each of the three natural protonation states: neutral with $N^{\delta 1}$ protonated and $N^{\epsilon 2}$ unprotonated (HISD), neutral with $N^{\delta 1}$ unprotonated and $N^{\epsilon 2}$ protonated (HISE) and charged with $N^{\delta 1}$ and $N^{\epsilon 2}$ both protonated (HISH). Only four occurrences of HISD were found with acceptable search criteria in the Database, so these parameters

Table 3. *Angle parameters*

Angle type	σ	Angle (°)	Angle type	σ	Angle (°)
C5W–CW–CW	1·2	107·2	CH3E–CH1E–CH3E	2·2	110·8
CW–C5W–CH2E	1·4	126·8	CH3E–CH1E–NH1	1·5	110·4
C5W–CW–CR1E	1·0	133·9	CH3E–CH1E–OH1	2·0	109·3
CW–CW–CR1E	1·0	118·8	C–CH2E–CH1E	1·0	112·6
CW–CW–CR1W	1·0	122·4	C5–CH2E–CH1E	1·0	113·8
CW–C5W–CR1E	1·6	106·3	CF–CH2E–CH1E	1·0	113·8
CW–CW–NH1	1·3	107·4	C5W–CH2E–CH1E	1·9	113·6
CH1E–C–N	1·5	116·9	CY–CH2E–CH1E	1·8	113·9
CH1E–C–NH1	2·0	116·2	C–CH2E–CH2E	1·7	112·6
CH1E–C–O	1·7	120·8	C–CH2G–NH1	2·9	112·5
CH1E–C–OC	2·5	117·0	C–CH2G–NH3	2·9	112·5
CH2E–C5–CR1E	1·3	129·1	CH1E–CH2E–CH1E	3·5	116·3
CH2E–C5–CR1H	1·3	131·2	CH1E–CH2E–CH2P	1·9	104·5
CH2E–CF–CR1E	1·7	120·7	CH1E–CH2E–CH2E	2·0	114·1
CH2E–C5W–CR1E	1·5	126·9	CH1E–CH2E–CH3E	2·1	113·8
CH2E–CY–CR1E	1·5	120·8	CH1E–CH2E–OH1	2·0	111·1
CH2E–C–N	2·1	118·2	CH1E–CH2E–S	2·3	114·4
CH2G–C–N	2·1	118·2	CH1E–CH2E–SH1E	2·3	114·4
CH2E–C5–NH1	1·5	122·7	CH2E–CH2E–CH2E	2·3	111·3
CH2E–C–NH1	2·1	116·5	CH2E–CH2P–CH2P	3·2	106·1
CH2G–C–NH1	2·1	116·4	CH2P–CH2P–N	1·5	103·2
CH2E–C–NH2	1·5	116·4	CH2E–CH2E–NH1	2·2	112·0
CH2E–C5–NR	1·5	121·6	CH2E–CH2E–NH3	3·2	111·9
CH2E–C–O	2·0	120·8	CH2E–CH2E–SM	3·0	112·7
CH2G–C–O	2·1	120·8	CY2–CR1E–CR1E	1·8	119·6
CH2E–C–OC	2·3	118·4	CW–CR1E–CR1E	1·3	118·6
CH2G–C–OC	2·3	118·4	CW–CR1W–CR1W	1·3	117·5
CR1E–CY2–CR1E	2·0	120·3	CF–CR1E–CR1E	1·7	120·7
CR1E–CY–CR1E	1·5	118·1	CY–CR1E–CR1E	1·5	121·2
CR1E–CF–CR1E	1·5	118·6	C5–CR1E–NH1	1·0	106·5
CR1W–CW–NH1	1·5	130·1	C5–CR1H–NH1	1·0	107·2
CR1E–C5–NH1	1·0	105·2	C5W–CR1E–NH1	1·3	110·2
CR1H–C5–NH1	1·0	106·1	C5–CR1E–NR	2·3	109·5
CR1E–CY2–OH1	3·0	119·9	CR1E–CR1E–CR1W	1·3	121·1
N–C–O	1·4	122·0	CR1W–CR1W–CR1E	1·3	121·5
NC2–C–NC2	1·8	119·7	CR1E–CR1E–CR1E	1·8	120·0
NC2–C–NH1	1·9	120·0	NH1–CRHH–NH1	1·0	108·4
NH1–C–O	1·6	123·0	NH1–CR1E–NR	1·3	111·7
NH2–C–O	1·0	122·6	C–N–CH1E	5·0	122·6
OC–C–OC	2·4	122·9	C–N–CH2P	4·1	125·0
C–CH1E–CH1E	2·2	109·1	CH1E–N–CH2P	1·4	112·0
C–CH1E–CH2E	1·9	110·1	C–NH1–CH1E	1·8	121·7
C–CH1E–CH3E	1·5	110·5	C–NH1–CH2G	1·7	120·6
C–CH1E–N	2·5	111·8	C–NH1–CH2E	1·5	124·2
C–CH1E–NH1	2·8	111·2	C–NH1–CH3E	1·7	120·6
C–CH1E–NH3	2·8	111·2	C5–NH1–CRHH	1·7	109·3
CH1E–CH1E–CH2E	1·7	110·4	C5–NH1–CRH	1·7	109·0
CH1E–CH1E–CH3E	1·7	110·5	CW–NH1–CR1E	1·8	108·9
CH1E–CH1E–NH1	1·7	111·5	CRHH–NH1–CR1H	1·0	109·0
CH1E–CH1E–OH1	1·5	109·6	CRH–NH1–CR1E	1·3	106·9
CH2E–CH1E–CH3E	3·0	110·7	C5–NR–CR1E	1·0	105·6
CH2E–CH1E–N	1·1	103·0	CR1E–NR–CR1E	3·0	107·0
CH2E–CH1E–NH1	1·7	110·5	CH2E–SM–CH3E	2·2	100·9
CH2E–CH1E–NH3	1·7	110·5	CH2E–S–S	1·8	103·8

remain uncertain, but were inferred approximately from the other structures for the parameter set.

Phenylalanine and tyrosine were in better agreement but particular bonds and angles were still outside the standard deviation of the sample and well outside the standard deviation of the mean. Particular problems occurred at the interior ring angle angle at the C^{γ} atom and the tyrosine C^{ζ}–O^{η} bond. The interior ring angles of both Phe and Tyr sum to 720·0° in contrast to 721·8° for *EREF* Phe and Tyr parameters and 720·0 and 717·0° for P19X parameters for Phe and Tyr, respectively.

The acidic amino acids glutamic acid and aspartic acid also have geometries which depend on the charge and thus on environment and pH. As with histidine, the Database searches were conducted for the appropriate charged and uncharged carboxylic acid groups. Since the pK_a for Glu and Asp is approximately 4·5 in a protein (Cantor & Schimmel, 1980) and depends on the environment, the protonation state for these residues may vary and should be assigned for accurate structure refinement. This assignment might be accomplished by considering the crystallization pH and the residue environment, either by qualitative considerations or by using electrostatic potential calculations (*e.g.* Gilson, Sharp & Honig, 1988) in combination with pK_a estimation techniques (*e.g.* Karshikoff, Engh, Bode & Atanasov, 1989). The

geometric parameters affected are the C–O bond lengths, which have an average of 1·23 Å in the charged residues and 1·21 and 1·30 Å for the double and single C–O bonds in the uncharged species, respectively. The C–C–O angles are similarly charge dependent.

The amide groups of glutamine and asparagine were represented by various model amide fragments. The O–C–N angle and C–O bonds could be improved. The planarity of the C atom was also investigated in both the carboxyl and the amide groups. In both cases, the average 'improper' dihedral angle measuring the planarity was nearly zero with a standard deviation of approximately 1·5°.

Arginine is rather well represented by *EREF* and P19X parameters, with the exception of the C^{δ}–N^{ε}–C^{ζ} angle parameter of P19X which deviates by ~3·5° from the statistical average. Lysine likewise required only small corrections. The neutral forms of these residues do not occur in the Database and with pK_a's of >12 and ~10·1 in a protein the neutral forms will occur only very rarely in proteins and can be neglected for refinement purposes.

The saturated aliphatic residues Leu, Ile, Val and Ala were in general well parametrized. All bond parameters were within one standard deviation from the mean (but not within one standard error of the mean) and only a few bond-angle parameters were a greater distance from the mean. The statistical bond-length variabilities of these residues were also substantially larger than those considered thus far. This is presumably because of the greater mutability of C–C single bonds in extended geometries compared to the greater relative constraint of ring structures and higher-order bonds. Only for these residues was the expected distribution of bond lengths at room temperature for P19X and *EREF* force constants of the same order as the distribution seen in the Database. In all other cases, the expected thermal distribution was two-three times as large as the deviation in the Database. This was not an effect of the relative sample sizes.

The thiol fragment representing cysteine occurred only twice with acceptable search criteria in the Database; all other cysteines had been modified at the S atom. Disulfide bridges, however, were more frequent, and our CSD cysteine geometry is derived also from these parameters. This affects only the C^{β}–S^{γ} bond, which was nearly identical in the two cysteine structures and in the cysteine disulfide bridge structures. Proline and glycine are described below in connection with the peptide backbone statistics. Residues serine, threonine and methionine were found with adequate statistics and some improvements over *XPLOR* and *EREF* were possible.

The geometries of the backbone were derived from the individual amino acid geometries as well as searches using chemical fragments representing the backbone. The Database includes a significant number of cyclic dipeptides which were eliminated from the search to avoid biasing the statistics with structures not found in proteins. The backbone geometries were found to be approximately equivalent for all residues except proline and glycine. The carbonyl bond of *EREF* and the O–C–N and C–N–C^{α} angles of both *EREF* and *XPLOR* were significantly altered. The C–N–C^{α} angle of glycine was approximately 1·2° smaller than the total average. The O–C–N angle of proline was approximately 1° smaller than the total average.

Marquart, Walter, Deisenhofer, Bode & Huber (1983) have noted that the average ω_1 (C^{α}–C–N–C^{α}) and ω_3 (O–C–N–C^{α}) dihedral angles in several proteins differ significantly from the ideal values of 180 and 0°, respectively, for a perfectly planar peptide group but are instead approximately 179 and −1·8°, respectively. Our peptide fragments have corresponding average values of 179 and −0·6°, values which both differ from the ideal planar values by more than two standard deviations of the mean. Elimination of

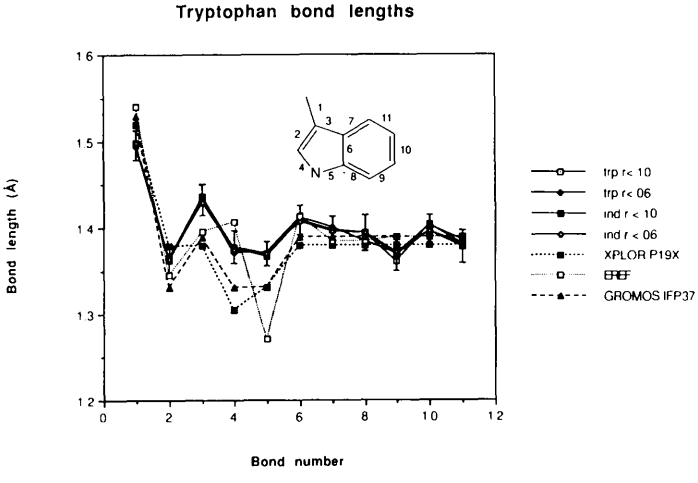

Fig. 1. Comparison of tryptophan bond lengths from four different Cambridge Database samples and *EREF*, P19X and *GROMOS* parameters. trp indicates tryptophan structures and ind indicates 3-methylindole fragment structures. *r* shows the maximum crystallographic *R* factor of the structure. The error bars indicate one standard deviation for the sample ind $r <$ 0·10, which is similar for all samples. The standard deviation of the mean is given by this value divided by $N^{1/2}$ where *N* is the number of the sample and is 31, 11, 75 and 33, respectively.

Table 4. *Tryptophan minimization*

		CSD-X	P19X
Bonds (Å)	Maximum	0·001	0·014
	r.m.s.	<0·001	0·007
Angles (°)	Maximum	0·07	9·72
	r.m.s.	0·04	3·29

glycine from the data set changes these values to 178 and −1·01°, now three-four standard deviations of the mean from ideal planarity. Also, the average 'improper' dihedral defined as the angle between the plane of atoms C–C^{α}–O and the plane of atoms C^{α}–O–N is 0·2 and 0·4° when glycine is eliminated, approximately four standard deviations of the mean greater than the ideal planar value of 0°. [This means that the central C atom lies, on average, on the *re* side of the plane defined by the atoms O–N–C^{α} in the nomenclature of Hanson (1966).] Peptides with glycyl carboxyl groups, on the other hand, show no significant deviation from planarity. The average value for ω_3 was suggested to depend on the helical content of a protein by Marquart *et al.* (1983). The results here suggest additionally that the effect probably arises from steric effects between the proximal C^{β} and O of the peptide group. The statistical average over a protein structure would then be affected both by the glycyl content of the protein and by the distribution of φ, ψ dihedral angles (secondary structure) in the protein.

As a test of the consistency of the parameters, we minimized tryptophan coordinates to convergence for P19X and the CSD parameters, respectively, considering only non-H-atom bond, angle, dihedral and improper dihedral potentials. Ideally self-consistent parameters will converge exactly to their target values under these conditions. Table 4 summarizes the results. It is noteworthy that the P19X parameters converge to a minimum r.m.s. deviation for angles of 3·29°, a value already higher than the target deviation for X-ray refinement. This is in contrast with the CSD value of 0·04°. This example serves to illustrate the self-consistency of this method of parametrization, even in the case of tryptophan where the population sizes were somewhat smaller than others. This is also the most extreme case, however, and does not reflect the overall magnitude of this effect.

Table 5. *Refinement comparisons*

BPTI *R* factor Resolution (Å)	P19X	CSD-X	Number of reflections
1·20–6·00	0·196	0·194	14428
1·20–1·25	0·287	0·286	1471
1·25–1·32	0·280	0·277	1580
1·32–1·40	0·254	0·251	1717
1·40–1·51	0·238	0·237	1806
1·51–1·66	0·228	0·224	1920
1·66–1·89	0·214	0·210	1937
1·89–2·36	0·198	0·195	1965
2·36–6·00	0·162	0·161	2032
Bond *R* factor	0·0487	0·0482	All bonds
Bond *R* factor	0·0426	0·0426	Backbone bonds
Geometry	Energy (kJ mol^{-1})		BPTI
Bond energies	190·7 (*P*)	166·1 (*C*)	*P* denotes calculation
Angle energies	456·3 (*P*)	348·8 (*C*)	of energies with P19X
Bond energies	204·7 (*C*)	232·7 (*P*)	*C* denotes calculation
Angle energies	392·1 (*C*)	478·9 (*P*)	with CSD-X
Phycocyanin *R* factor Resolution (Å)	P19X	CSD-X	Number of reflections
1·66–8·0	0·187	0·186	58464
Geometry	Energy (kJ mol^{-1})		
Bond energies	1761·0 (*P*)	1569·3 (*C*)	
Angle energies	3846·4 (*P*)	3265·7 (*C*)	See comment above
Bond energies	1896·1 (*C*)	2421·3 (*P*)	
Angle energies	3406·6 (*C*)	4561·7 (*P*)	

Table 5 and Fig. 2 show the results of the refinements for phycocyanin and BPTI. The refinements were done without manual intervention, temperature-factor refinement, occupancy refinement and with identical force constants to test the integrity of the

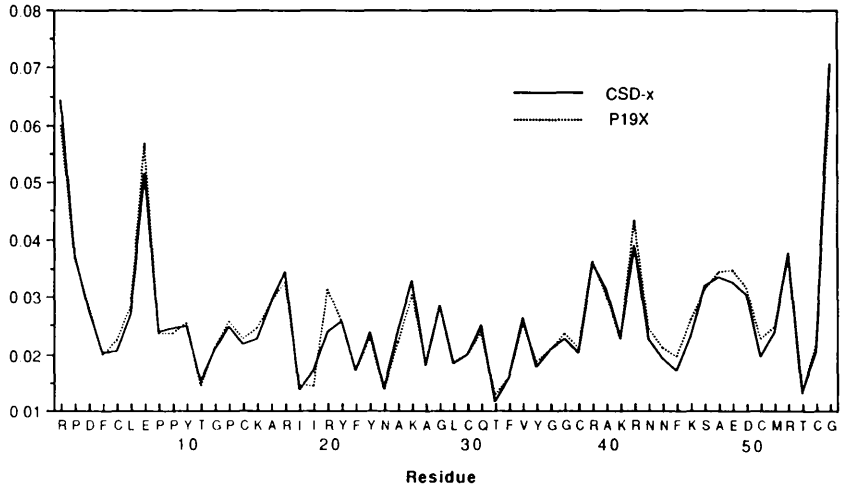

Fig. 2. The real-space 'bond' *R* factor (see text) showing the quality of fit between the two refined models with their respective electron-density maps. Some differences are probably due to minimization artifacts, but the overall improvement using the CSD-X parameters is significant. The improvements are largely confined to the side chains; the corresponding analysis over backbone bonds alone shows less improvement with the CSD-X parameters.

new ideal geometry values and minimize the possibility of external sources of divergent refinement. As a result, certain specific features of the CSD-X and P19X refinements may arise from 'random' divergence during the minimization. However, the statistical results show general improvement, both in the *R* factor and in geometry statistics.

For both structures, the bond and angle energies are lower for the CSD-X refined structures than for the P19X structures, indicating that the CSD-X parameters are indeed more accurate. Further evidence is provided by the calculation of the energies of the CSD-X refined structures using P19X parameters and *vice versa*. The energies are much higher for both bonds and angles when calculating the energies of the CSD-X refined structures with P19X parameters. However, the bond energies are only somewhat higher and the angle energies are actually *lower* when calculating the energies of the P19X refined structures with CSD-X parameters. The force constants of these two sets are identical, as was the relative weighting of the energy terms. This means that the influence of the diffraction term on the refinement moves the structure away from the P19X parameters and toward the CSD-X parameters, so much so that the angles of the P19X refined structure satisfy CSD-X ideal values better than their own restraints.

The *R* factor of BPTI shows some general overall improvement, also consistently in smaller resolution shells (Table 5). The *R* factor of phycocyanin is also very slightly improved. Fig. 2 shows the real-space 'bond' *R* factor for BPTI, comparing the CSD-X and P19X structures. Here also there is overall improvement, although most differences are small and some P19X residues have a better fit. Some of the differences may be minimization artifacts. Others are due to differences between the refined coordinates, such as that at residue 20. (This is an arginine and the side-chain target angles are from 1 to 4° larger in CSD-X.) Still others are due to small differences in the electron-density maps, such as in the residue range 40–50 where the coordinates in the two models are nearly identical but improvement in the 'bond' *R* factor may be seen.

Discussion

We have argued in this work that the parameters used for X-ray structure refinement should be as accurate as possible, first to ensure the integrity of refined protein structures and secondly to provide improvements in *R* factors and geometry statistics. These considerations are particularly important for high-resolution structures. The parameters which we have formulated for use with *XPLOR* offer improvements especially over parameters used in simulated annealing or molecular dynamics refinement (Brooks *et al.*, 1983; Brünger, Kuriyan & Karplus, 1987; Brünger, 1988; Brünger, Karplus & Petsko, 1989; van Gunsteren & Berendsen, 1987; Fujinaga, Gros & van Gunsteren, 1989) and refinement using similar energy functions (Jack & Levitt, 1978). They update and extend other parameters (Hendrickson & Konnert, 1980; Tronrud, Ten Eyck & Matthews, 1987; Bowen *et al.*, 1958; Vijayan, 1976; Allen, Kennard, Watson, Brammer, Orpen & Taylor, 1987) by parametrizing according to the charge state of histidine, glutamic acid and aspartic acid.

This parametrization may be further extended in several ways. Periodic updates will deliver improved average values and especially the statistical reliability of the standard deviations determined for the parameters. The Cambridge Database currently adds more than 7000 structures per year to its archives; these are certain to include new fragments appropriate for this analysis. With increasing data, it should be possible to parametrize bonds and angles according to more specific criteria, such as introducing parameters as a function of torsion angles. Another obvious extension is the parametrization of the many small-molecule substrates, unusual amino acids, carbohydrate structures *etc.*

We have also introduced in this work the idea that the expected variation in the parameters should determine the force constants rather than 'physical' force constants or uniform weighting. This is justified by the argument that energy-function refinement is a form of supplementing X-ray diffraction refinement with additional information; the anticipated variation may be regarded as additional information which specifies the reliability of the bond or angle target value. On the other hand, the variation parameters are less reliably established by small populations and it is not certain that the variation seen in small molecules is applicable to protein structures where tertiary structure features may influence these values. It is not within the scope of this study finally to resolve these questions; further studies of high-resolution protein structure and improved statistics of ever larger databases are required for this.

This work is one of a number of studies which use structural databases to determine energy parameters associated with potentials of mean force (McQuarrie, 1976). Sippl and co-workers have used databases of protein structures to determine potentials of mean force to predict the likelihood of specific secondary structures (Sippl, 1990) and to evaluate the quality of models of protein structure (Hendlich *et al.*, 1990). Similar methods could be applied to protein refinement by generating potentials of mean force from a variety of databases. This would generalize the potentials over those suggested in this work and provide an easily extensible method to include a wide range of structural information. Bürgi & Dubler-Steudle (1988) have used structural data as a source of information for parametrizing potential-energy

surfaces and free energies of activation. Dauber & Hagler (1980) used crystal structures to parametrize the 'non-bonded' interactions or packing, nuclear repulsion and hydrogen bonding. Potentials or closely related probability distributions have been determined from databases also for packing calculations as a function of residue pairs (Singh & Thornton, 1990; Narayana & Argos, 1984; Gregoret & Cohen, 1990; Ponder & Richards, 1987) and for hydrogen bonding (Baker & Hubbard, 1984; Taylor & Kennard, 1984; Ippolito, Alexander & Christianson, 1990). Many other applications have also been published which testify to the growing recognition of the wealth and variety of information available from structural databases.

References

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

ALLEN, F. H., KENNARD, O., WATSON, D. G., BRAMMER, L., ORPEN, A. G. & TAYLOR, R. (1987). *J. Chem. Soc. Perkin Trans.* 2, pp. S1-S19.

BAKER, E. N. & HUBBARD, R. E. (1984). *Prog. Biophys. Mol. Biol.* **44**, 97-179.

BOWEN, H. J. M., DONOHUE, J., JENKIN, D. G., KENNARD, O., WHEATLEY, P. J. & WHIFFEN, D. H. (1958). *Molecules and Ions*, edited by A. D. MITCHELL & L. C. CROSS. London: The Chemical Society.

BRÄNDÉN, C.-I. & JONES, T. A. (1990). *Nature (London)*, **343**, 687-689.

BROOKS, B. R., BRUCCOLERI, R. E., OLAFSON, B. D., STATES, D. J., SWAMINATHAN, S. & KARPLUS, M. (1983). *J. Comput. Chem.* **4**, 187-217.

BRÜNGER, A. T. (1988). *J. Mol. Biol.* **203**, 803-816.

BRÜNGER, A. T., KARPLUS, M. & PETSKO, G. A. (1989). *Acta Cryst.* A**45**, 60-61.

BRÜNGER, A. T., KURIYAN, J. & KARPLUS, M. (1987). *Science*, **235**, 458-460.

BÜRGI, H.-B. & DUBLER-STEUDLE, K. C. (1988). *J. Am. Chem. Soc.* **110**, 7291-7299.

CANTOR, C. R. & SCHIMMEL, P. R. (1980). *Biophysical Chemistry.* Part I: *The Conformation of Biological Macromolecules.* San Francisco: W. H. Freeman.

DAUBER, P. & HAGLER, A. T. (1980). *Acc. Chem. Res.* **13**, 105-112.

DUERRING, M., SCHMIDT, G. B. & HUBER, R. (1991). *J. Mol. Biol.* **217**, 577-592.

ENGH, R. A., SIPPL, M. J., MARTIN, P., EDWARDS, B. & HUBER, R. (1991). In preparation.

FUJINAGA, M., GROS, P. & VAN GUNSTEREN, W. F. (1989). *J. Appl. Cryst.* **22**, 1-8.

GILSON, M. K., SHARP, K. A. & HONIG, B. H. (1988). *J. Comput. Chem.* **9**, 327-335.

GREGORET, L. M. & COHEN, F. E. (1990). *J. Mol. Biol.* **211**, 959-974.

GUNSTEREN, W. F. VAN & BERENDSEN, H. J. C. (1987). *GROMOS.* Groningen Molecular Simulation Library, BIOMOS b.v., Groningen, The Netherlands.

HAMILTON, W. C. (1964). *Statistics in Physical Science.* New York: Ronald Press.

HANSON, K. R. (1966). *J. Am. Chem. Soc.* **88**, 2731-2742.

HENDLICH, M., LACKNER, P., WEITCKUS, S., FLÖCKNER, H., FROSCHAUER, R., GOTTSBACHER, K., CASARI, G. & SIPPL, M. J. (1990). *J. Mol Biol.* **216**, 167-180.

HENDRICKSON, W. A. (1985). *Methods Enzymol.* **115**, 252-270.

HENDRICKSON, W. A. & KONNERT, J. H. (1980). *Computing in Crystallography*, edited by R. DIAMOND, S. RAMASESHAN & K. VENKATESAN, pp. 13.01-13.23. Bangalore: Indian Academy of Sciences.

IPPOLITO, J. A., ALEXANDER, R. S. & CHRISTIANSON, D. W. (1990). *J. Mol. Biol.* **215**, 457-471.

JACK, A. & LEVITT, M. (1978). *Acta Cryst.* A**34**, 931-935.

JENSEN, L. H. (1985). *Methods Enzymol.* **115**, 227-234.

KARSHIKOFF, A. D., ENGH, R., BODE, W. & ATANASOV, B. (1989). *Eur. Biophys. J.* **17**, 787-297.

MCQUARRIE, D. A. (1976). *Statistical Mechanics.* New York: Harper & Row.

MARQUART, M., WALTER. J., DEISENHOFER, J., BODE, W. & HUBER, R. (1983). *Acta Cryst.* B**39**, 480-490.

NARAYANA, S. V. L. & ARGOS, P. (1984). *Int. J. Pept. Protein Res.* **24**, 25-39.

PONDER, J. W. & RICHARDS, F. M. (1987). *J. Mol. Biol.* **193**, 775-791.

SINGH, J. & THORNTON, J. M. (1990). *J. Mol. Biol.* **211**, 595-615.

SIPPL, M. J. (1990). *J. Mol. Biol.* **213**, 859-883.

TAYLOR, R. & KENNARD, O. (1984). *Acc. Chem. Res.* **17**, 320-326.

TRONRUD, D. E., TEN EYCK, L. F. & MATTHEWS, B. W. (1987). *Acta Cryst.* A**43**, 489-501.

VIJAYAN, M. (1976). *CRC Handbook of Biochemistry and Molecular Biology*, 3rd ed., *Proteins*, Vol. II, edited by G. D. FASMAN, pp. 742-759. Cleveland: CRC Press.

WLODAWER, A., WALTER, J., HUBER, R. & SJÖLIN, L. (1984). *J. Mol. Biol.* **180**, 301-329.

Acta Cryst. (1991). A**47**, 400-404

The Phase Problem and its Relation to the Spin-Glass Problem

BY RAJESWARI VENKATESAN

Raman Research Institute, Bangalore 560080, *India*

(*Received* 5 *September* 1990; *accepted* 18 *February* 1991)

Abstract

The maximum-entropy method of image reconstruction is discussed in the context of the crystallographic phase problem. Entropy is the function to be maximized in a space of phases which has dimension equal to the number of structure-factor constraints. The function $\int[1-\exp(-\rho/\rho_0)]\,dV$ is proposed as a suitable one. An analogy between the phase problem and the spin-glass problem of condensed-matter

0108-7673/91/040400-05$03.00