Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On 'globbicity' of low-resolution protein structures

D. Y. Guo,[a]* Robert H. Blessing,[a] David A. Langs[a] and G. David Smith[a,b]

[a]Hauptman–Woodward Medical Research Institute, Inc., 73 High Street, Buffalo, New York 14203, USA, and [b]Roswell Park Cancer Institute, Elm and Carlton Street, Buffalo, New York 14263, USA

Correspondence e-mail: guo@hwi.buffalo.edu

Received 30 October 1997
Accepted 10 June 1998

Using Harker's [Harker (1953). *Acta Cryst.* **6**, 731–736] idea of spherically averaged polyatomic groups or 'globs' as the units of structure suitable for analyzing low-resolution diffraction data from protein crystals, 'globbic' scattering factors have been calculated for main-chain peptide units and amino-acid side-chain groups to 3 Å resolution *via* Debye's [Debye (1915). *Ann. Phys. (Leipzig)*, **46**, 809–823] scattering formula. It is shown that the scattering factors are insensitive to intra-globbic conformational variation and can be approximated fairly well by a single-Gaussian formula, *i.e.* $f_g(s) = Z_g \exp(-1.7 Z_g s^2)$, where $s = (\sin\theta)/\lambda$ and Z_g is the total electron count for the atoms of the glob. Phase errors due to the globbic approximation and their effect on electron-density maps at 3.5 Å resolution have been assessed *via* calculations for the crambin structure; this analysis indicates that the globbic scattering factors will be useful in efforts to develop procedures for direct-methods phasing of diffraction data to ~3.5 Å resolution from protein crystals.

1. Introduction

As is well known, crystals of biological macromolecules rarely yield X-ray diffraction data that extend to atomic resolution.[1] Therefore, three-dimensional atomic structures of biomolecular crystals are modelled by adjusting main-chain and side-chain conformation angles for functional groups of atoms that have more or less rigid known structures, and atomic refinements of group-modelled structures with diffraction data that do not extend to atomic resolution are carried out with stereochemical restraints imposed to assure that bond lengths, valence angles, conformation angles and non-covalent interaction distances remain within tolerable limits.

Lack of high-resolution data hinders not only atomic modelling of macromolecular structures, but also *ab initio* phasing of low- to moderate-resolution data sets from native crystals by probabilistic direct methods. In cases in which atomic resolution ($d_{\min} < 1.2$ Å) data have been obtained, recently developed and still developing direct methods (see, for example, Hauptman, 1997) have been remarkably successful in phasing native protein structures with as many as ~1000 independent non-H atoms. Two examples are a scorpion protein toxin (Smith *et al.*, 1997) and triclinic lysozyme (Sheldrick, 1997; Deacon, 1997). The new methods have also been used successfully with lower resolution ($d_{\min} \simeq 3$ Å) single-wavelength anomalous scattering data to determine *de novo* multi-selenium anomalous scattering substructures in several bioengineered Se-Met proteins. In such cases, even data sets with $d_{\min}$ as large as 3 or 4 Å provide atomic reso-

[1] The threshold for atomic resolution is operationally defined (Sheldrick *et al.*, 1993) to be $d_{\min} = \lambda/(2\sin\theta_{\max}) \leq 1.2$ Å, just slightly less than the 1.24 Å average C=O bond length, the expected shortest distance between non-H atoms in biomolecules.

lution of the heavy-atom substructure. Two examples are a 35 kDa Se_8 protein (Smith *et al.*, 1998) and a 95 kDa Se_{32} dimeric asymmetric subunit of a tetrameric protein (Turner *et al.*, 1997). These and other noteworthy successes notwithstanding, *ab initio* phasing in the common cases of normal (non-anomalous) scattering data from native protein crystals to diffraction resolution limits in the 2–4 Å range remains problematic and needs further research.

2. From atomicity to 'globbicity'

One of the early theoretical analyses of the problem of the limited diffraction resolution attainable with biomolecular crystals was that of Harker (1953), who estimated the average intensity of low-resolution X-ray scattering, in spherical shells of reciprocal space, from large-unit-cell single crystals composed of groups of unresolved atoms. Harker called the polyatomic groups 'globs' and showed that the X-ray scattering factor for a rotationally averaged or spherical glob can be approximated as

$$f_g(s) = \sum_{a=1}^{n} f_a(s) \frac{\sin(4\pi s r_a)}{4\pi s r_a}, \qquad (1)$$

where $2s = 2(\sin\theta)/\lambda = 1/d$ is the reciprocal-space radius, n is the number of atoms in the glob, f_a is the atomic X-ray scattering factor of atom a and $r_a = |\mathbf{r}_a|$ is the length of the position vector of atom a referred to an origin at the glob scattering centroid.

Harker's result is closely related to an earlier result by Debye (Debye, 1915; cited in Guinier, 1994, and Warren, 1990) who, considering X-ray scattering from gases, liquids, amorphous solids or microcrystalline powders, derived a formula for the average intensity of X-ray scattering from a statistically large sample of randomly oriented identical n-atom molecules or microcrystallites or, equivalently, from a single rotationally averaged rigid group of n atoms. Debye's treatment gives the scattering factor for a spherically averaged rigid group of atoms as

$$f_g(s) = \left[\sum_{a=1}^{n} \sum_{b=1}^{n} f_a(s) f_b(s) \frac{\sin(4\pi s r_{ab})}{4\pi s r_{ab}} \right]^{1/2}, \qquad (2)$$

where $r_{ab} = |\mathbf{r}_b - \mathbf{r}_a|$ is the distance between atoms a and b.

Harker's formula (1) involves a single sum over n one-centered terms and Debye's formula (2) involves the square root of a double sum over n^2, in general, two-centered terms. Although the approximate Harker formula has a computationally simpler form, the exact Debye formula has the computational advantage that it obviates any ambiguity concerning definition of the glob scattering centroid when calculating a glob scattering factor.

For crystal structure analyses with diffraction data that do not extend to atomic resolution, the concepts atoms, atomic and atomicity can be usefully replaced by the concepts globs, globbic and globbicity. Depending on the diffraction resolution limit in a given case, appropriate globbic scattering factors calculated *via* (1) or (2) can be used to calculate globbic crystal structure factors. With $\mathbf{r}_a$ and $\mathbf{r}_g$ denoting atomic and globbic position vectors with respect to the unit-cell origin, and W_a and W_g denoting atomic and globbic Debye–Waller factors, we obtain the following relationships. For N atoms per unit cell,

$$\begin{aligned} F_{\mathbf{h}} &= \sum_{a=1}^{N} f_a(s) W_a(\mathbf{h}) \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_a) \\ &= |F_{\mathbf{h}}(\text{atoms})| \exp[i\varphi_{\mathbf{h}}(\text{atoms})], \end{aligned} \qquad (3)$$

or, for $M < N$ globs per unit cell with, on average, $N/M = \langle n \rangle$ atoms per glob,

$$\begin{aligned} F_{\mathbf{h}} &\simeq \sum_{g=1}^{M} f_g(s) W_g(\mathbf{h}) \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_g) \\ &= |F_{\mathbf{h}}(\text{globs})| \exp[i\varphi_{\mathbf{h}}(\text{globs})]. \end{aligned} \qquad (4)$$

Then, given at least approximate phases $\varphi_{\mathbf{h}}$ for a sufficiently large subset of the largest amplitudes $|F_{\mathbf{h}}|$ measured within the diffraction resolution limit $|\mathbf{h}|_{\max} = 2s_{\max} = 2(\sin\theta_{\max})/\lambda = 1/d_{\min}$, the unit-cell electron density distribution is given by

$$\rho(\mathbf{r}) \simeq V^{-1} \sum_{|\mathbf{h}|_{\min}}^{|\mathbf{h}|_{\max}} |F_{\mathbf{h}}| \exp[i(\varphi_{\mathbf{h}} - 2\pi \mathbf{h} \cdot \mathbf{r})]. \qquad (5)$$

In the context of protein crystallography, just as the 1.24 Å average peptide C′═O bond length defines a natural atomic resolution threshold at ∼1.2 Å, the 3.80 Å average transplanar peptide $C^{\alpha(i)} \cdots C^{\alpha(i+1)}$ repeat distance defines a natural globbic resolution limit at ∼3.5 Å. At this resolution, main-chain peptide units $-C^{\alpha}_{1/2}-C'(=O)-N-C^{\alpha}_{1/2}-$ and amino-acid side-chain groups $-C^{\beta}-R$ are natural globs (*cf.* Leherte *et al.*, 1994), and we have calculated globbic scattering factors to 3 Å resolution for the natural globs using both Harker's formula (1) and Debye's formula (2).

2.1. Globbic scattering factors for amino-acid residues

The Harker and Debye formulae give globbic scattering factors for spherically averaged main-chain peptide and amino-acid side-chain globs which differ very little to 3.5 Å resolution. The differences are so slight that in a preliminary publication of our work on globbic scattering factors we inadvertently mislabeled our results: in the captions of Tables 1 and 2 in our earlier paper (Guo *et al.*, 1995) 'Debye group scattering factor' should be replaced by 'Harker group scattering factor'.

2.2. Four-Gaussian, nine-coefficient fitting

We have now carried out new calculations of both the Harker and Debye globbic scattering factors, and have fitted to them the nine coefficients of the four-Gaussian fitting functions (Cromer & Waber, 1965),

$$g(s) = \sum_{j=1}^{4} a_j \exp(-b_j s^2) + c. \qquad (6)$$

The fitting was performed by means of a simplex refinement to minimize the normalized root-mean-square error of fit,

Table 1
Glob nomenclature and residual errors of fit (7) for the four-Gaussian, nine-coefficient functions (6) fitted to globbic scattering factors from the Harker (1) and Debye (2) scattering formulae.

Chemical name	Three-letter symbol	One-letter symbol	Number of atoms	$R_{\rm Debye}$	$R_{\rm Harker}$
Peptide	$C^{\alpha}C'(O)N$	X	4	0.0002	0.0000
Peptide	$C^{\alpha}(C^{\beta})C'(O)N$	X′	5	0.0004	0.0000
Cysteine	Cys	C	2	0.0000	0.0000
Serine	Ser	S	2	0.0000	0.0000
Valine	Val	V	3	0.0001	0.0000
Threonine	Thr	T	3	0.0001	0.0000
Proline	Pro	P	3	0.0001	0.0000
Isoleucine	Ile	I	4	0.0006	0.0000
Leucine	Leu	L	4	0.0003	0.0000
Methionine	Met	M	4	0.0001	0.0000
Asparagine	Asn	N	4	0.0002	0.0000
Aspartate	Asp	D	4	0.0002	0.0000
Glutamine	Gln	Q	5	0.0004	0.0000
Glutamate	Glu	E	5	0.0004	0.0000
Lysine	Lys	K	5	0.0016	0.0002
Histidine	His	H	6	0.0003	0.0000
Phenylalanine	Phe	F	7	0.0003	0.0000
Arginine	Arg	R	7	0.0004	0.0001
Tyrosine	Tyr	Y	8	0.0003	0.0000
Tryptophan	Trp	W	10	0.0009	0.0002

$$R = \left\{ \frac{\sum_{j=1}^{n} [f_g(s_j) - g(s_j)]^2}{\sum_{j=1}^{n} f_g^2(s_j)} \right\}^{1/2}, \qquad (7)$$

where $f_g(s)$ is given by (1) or (2), $g(s)$ is given by (6), $s_j = j\Delta s$, $\Delta s = 0.01$ Å and $j = 1, 2, ..., n$ with $n = 17$ so that $0 \leq s_j \leq 0.17$ Å^{-1} and $\infty > d_j \geq 2.94$ Å. Results of the new calculations are tabulated in Tables 1, 2 and 3.

In the new calculations, the five-atom peptide main-chain glob $-C_{1/2}^{\alpha(i)}-C'(=O)-N-C_{1/2}^{\alpha(i+1)}-$ with two half-weight C^{α} atoms was approximated by the chemically equivalent four-atom glob $-C^{\alpha}-C'(=O)-N-$ with one full-weight C^{α}, and the amino-acid side-chain globs $-C^{\beta}-R$ corresponded to the most favored side-chain conformations. Only non-H atoms were included in the calculations and, for the calculations with the Harker formula (1), the glob scattering centroid was taken to be the Z_a-weighted geometric centroid of the coordinates (Prince *et al.*, 1992) of the n atoms of the glob, where Z_a is the atomic number of atom a.

The globbic scattering factors reported in our earlier publication were fitted over the range $0 \leq (\sin\theta)/\lambda \leq 0.2$ Å^{-1}, corresponding to $\infty > d \geq 2.5$ Å, but in the present work, the range was reduced to $0 \leq (\sin\theta)/\lambda \leq 0.17$ Å^{-1}, corresponding to $\infty > d \geq 2.94$ Å. This was done because we have found that what we have called natural globs for protein crystal structure analysis are suited to diffraction resolution limits in the range $4 \gtrsim d_{\rm min} \gtrsim 3$ Å, but for $d_{\rm min} < 3$ Å the natural globs would need to be resolved into smaller pieces.

Since our earlier work showed that the natural protein globs all have scattering-factor curves of approximately Gaussian shape (Fig. 1 in Guo *et al.*, 1995), the simplex fitting procedure used in the present work was modified to emphasize the fit of the first exponential coefficient in (6). The modified simplex strategy was: first, with fixed $a_1 = Z_g$ (where Z_g is the total electron count for the neutral non-H atoms of the glob), starting from $b_1 = 0$ and with fixed $a_2 = a_3 = a_4 = b_2 = b_3 = b_4 = c = 0$, refine b_1 to convergence. Then, allow a_1, b_1 and c to refine with the other parameters fixed at zero. Finally, allow all nine parameters to refine with starting values of $a_2 = a_3 = a_4 = 0.001$ and $b_2 = 2b_1$, $b_3 = b_1/2$ and $b_4 = -b_1$. The simplex cycles consisted of a three-step refinement of each parameter in the order b_1, a_1, b_2, a_2, b_3, a_3, b_4, a_4 and c. For each refined parameter the initial simplex step length was 0.001; for each succeeding step the step length was multiplied by 1.001 if the residual (7) decreased, or divided by 1.001 if the residual increased. A residual as small as 4×10^{-5} or a step length as small as 4×10^{-7} was taken to indicate convergence. Floating-point operations were performed in double precision.

Due to physical and numerical correlations among the fitted coefficients (*e.g.* the $s = 0$ boundary condition, $a_1 + a_2 + a_3 + a_4 + c = Z_g$), the close similarity of our results from the Harker and Debye formulae is not readily apparent by comparison of corresponding coefficients in Tables 2 and 3, but the close agreement of the Harker and Debye scattering-factor curves is evident in plots such as the one shown in Fig. 1. To 3.5 Å resolution, the Harker and Debye scattering factors for the natural protein globs differ by at most 5%, with the Harker values, which involve sums over only n terms, tending to be smaller than the Debye values, which involve sums over n^2 terms. In principle, since the Harker formula (1) is more approximate and the Debye formula (2) is more exact, the Debye coefficients from Table 2 are to be preferred to the Harker coefficients from Table 3.

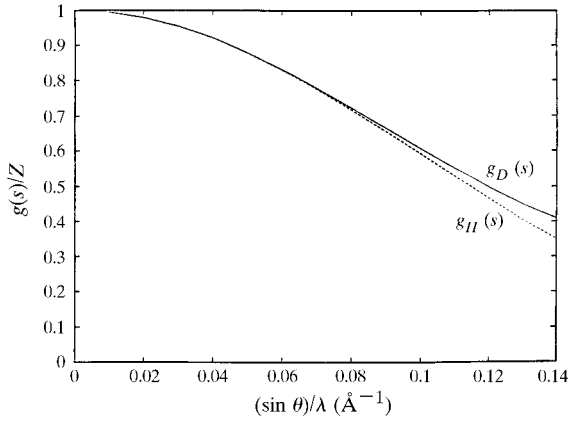

Figure 1
Comparison of Harker (1) and Debye (2) globbic scattering-factor curves to $d_{\rm min} = 3.5$ Å resolution for the trans-planar $-C^{\alpha}-C'(=O)-N-$ peptide unit.

Table 2
Coefficients for the nine-coefficient four-Gaussian functions (6) fitted to the Debye globbic scattering factors (2) for the trans-peptide main-chain group and amino-acid side-chain groups in their most favored conformations.

Values were fitted over the range $0 \leq (\sin\theta)/\lambda \leq 0.17$ Å^{-1} corresponding to $\infty > d \geq 2.94$ Å.

	a_1	b_1	a_2	b_2	a_3	b_3	a_4	b_4	c
X	27.02376	73.13853	−10.33178	108.44619	8.51072	72.45635	5.38563	−16.24203	−3.60115
C	12.79681	57.09464	−0.25531	169.48871	0.11055	70.68722	0.01219	−85.05642	9.33401
S	10.02265	36.70575	−0.00303	202.35540	0.00163	−6.82776	0.00200	101.23940	3.97533
V	14.71802	73.67335	−3.13719	113.58062	2.29951	72.74583	0.93182	−11.65331	3.18416
T	15.57804	77.26596	−4.82360	116.98499	3.90685	76.14320	1.37829	−11.27015	3.95618
P	15.32862	65.28794	−2.19129	105.90038	1.59645	64.47120	0.86731	−18.82024	2.39551
I	20.48284	125.09532	−22.85164	153.93403	20.50559	124.77207	5.03650	−3.80551	0.81176
L	24.95235	80.48446	−10.89699	113.69264	9.19889	79.91213	6.48695	−13.51836	−5.75023
M	22.45396	100.48688	1.55682	121.81561	−1.35469	100.64265	−1.65130	−6.56737	12.99719
N	27.00883	72.98491	−10.21644	108.34578	8.40013	72.30120	5.31272	−16.32248	−3.51799
D	28.15225	68.43158	−9.24152	104.48541	7.48048	67.78325	4.65902	−18.74271	−3.05766
Q	27.49114	120.27427	−24.37188	151.38734	22.58133	119.88089	6.69116	−5.08679	0.58865
E	27.86725	128.23825	−28.79278	158.61767	25.97153	127.88369	4.02047	−2.83659	4.91648
K	160.92175	47.82762	93.05266	131.86146	−180.33085	86.76572	−104.59072	8.78196	61.99539
H	30.88960	86.42441	0.05177	86.43046	0.01960	86.43704	−0.00008	−253.21383	7.03817
F	36.15148	138.36408	−30.41439	171.99004	28.41935	137.98209	3.38549	−2.58312	4.44102
R	92.21626	19.00075	18.75058	302.61505	−167.23244	5.03549	−2.40948	5.29983	103.82188
Y	37.71098	62.02162	27.47199	166.18083	−21.05330	51.30931	−4.73311	41.52935	10.61023
W	133.87148	250.24599	−92.11440	287.85313	18.98229	8.27610	−21.98182	−5.60775	22.18250

2.3. Conformational independence

Our purpose in developing globbic scattering factors is to use them for low-resolution inter-globbic conformational analysis to phase protein crystal structures. Preliminary to this, we have carried out several tests for intra-globbic conformational dependence of the globbic scattering factors at 3–4 Å resolution.

We first tested for conformational dependence by calculating Debye scattering-factor curves for globs corresponding to the low-energy extended and folded conformations observed in proteins for five-atom $-C^{\beta}-C^{\gamma}-C^{\delta}(=O^{\varepsilon 1})-O^{\varepsilon 2}$ glutamate side chains. As shown in Fig. 2, which also shows the scattering-factor curve for a free C atom for comparison, the curves for the two glutamate conformers are practically indistinguishable to 3 Å resolution.

In a second test for conformational dependence we calculated scattering-factor curves for various conformations of a five-atom $-C^{\alpha}(-C^{\beta})-C'(=O)-N-$ glob corresponding to a main-chain alanyl residue. Without regard to conformational energies, we varied the $N-C'-C^{\alpha}-C^{\beta}$ conformation angle depicted in Fig. 3 through the values $\psi = 0, 10, 20, ..., 180^{\circ}$ and computed the Debye scattering-factor curve for each conformation. The resulting 19 curves (not shown here) are practically indistinguishable from one another or from the glutamate curves of Fig. 2.

The small numerical differences we have observed among scattering factors for various five-atom protein globs, including α-helical and β-sheet conformations for alanyl main-chain globs, extended- and folded-conformation glutamate side-chain globs and most favored conformations of glutamine $-C^{\beta}-C^{\gamma}-C^{\delta}(=O^{\varepsilon 1})-N^{\varepsilon 2}$ and lysine $-C^{\beta}-C^{\gamma}-C^{\delta}-C^{\varepsilon}-N^{\zeta}$ side-chain globs, are summarized in Table 4. Differences among the five-atom glob scattering factors are in general too small to be discernable on plots such as that shown in Fig. 2,

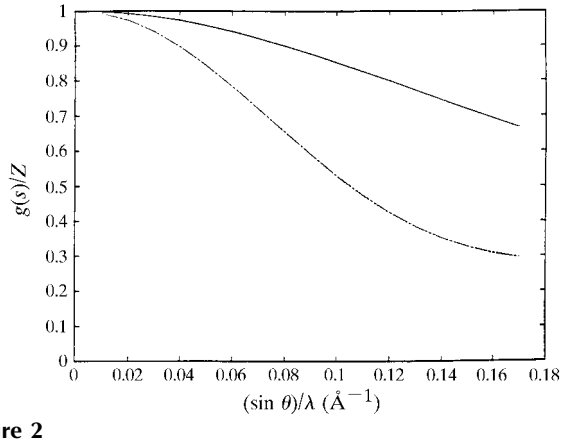

Figure 2
Debye globbic scattering-factor curves (dashed and dotted curves) for extended and folded glutamate side chains. In this diagram, the two curves are indistinguishably superimposed. For comparison, the scattering-factor curve (solid curve) for a single C atom is also illustrated.

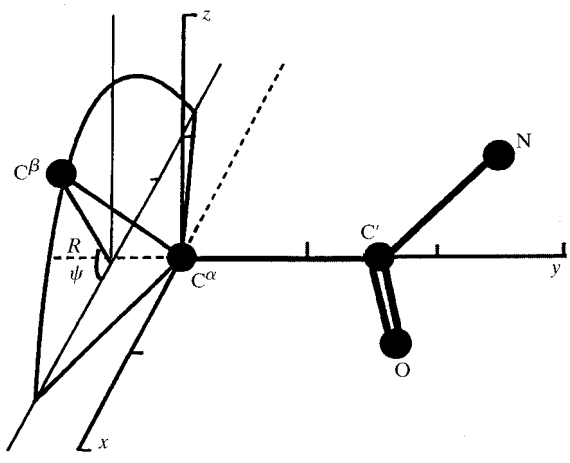

Figure 3
Conformational ψ-rotamers for a $-C^{\alpha}(-C^{\beta})-C'(=O)-N-$ alanyl peptide residue. The C^{α}, C′, O and N atoms lie in the *xy* plane.

Table 3
Coefficients for the nine-coefficient four-Gaussian functions (6) fitted to the Harker globbic scattering factors (1) for the trans-peptide main-chain group and amino-acid side-chain groups in their most favored conformations.

Values were fitted over the range $0 \leq (\sin\theta)/\lambda \leq 0.17$ Å^{-1} corresponding to $\infty > d \geq 2.94$ Å.

	a_1	b_1	a_2	b_2	a_3	b_3	a_4	b_4	c
X	28.22272	50.04199	−0.14686	158.20162	0.07998	62.49410	0.00360	−105.33758	−1.16790
C	16.13514	42.28708	0.08977	155.13994	−0.03267	51.73549	−0.00210	−106.36913	5.80778
S	12.43156	28.78123	0.05415	199.82450	−0.00557	−79.85930	0.00526	−41.40541	1.51386
V	18.36280	47.54203	0.08888	182.10863	−0.03393	76.04687	−0.00337	−92.29040	−0.41694
T	20.68012	44.48178	0.03307	155.40871	−0.01123	52.31097	−0.00154	−83.13023	−0.70363
P	18.42035	45.56694	0.10992	173.90218	−0.03413	46.68166	−0.00390	−94.97349	−0.49478
I	26.25343	62.01296	−0.14967	149.16783	0.05914	57.81843	0.00190	−126.64079	−2.16914
L	25.13822	57.38466	−0.13142	138.26076	0.03621	55.20963	0.00048	−160.84073	−1.04789
M	29.41166	80.40095	−0.25868	215.11364	0.01482	−30.80091	0.00784	−73.22785	4.82070
N	28.18671	50.06531	−0.14068	157.70925	0.07842	62.17868	0.00365	−104.77434	−1.13637
D	29.33235	47.75647	−0.09731	157.84300	0.04980	59.57077	0.00238	−104.95420	−1.29068
Q	34.51337	69.21179	−0.46114	177.75335	0.17016	54.66029	0.04512	−70.29239	−1.27712
E	35.34334	68.12250	−0.36445	176.59415	0.10735	56.42072	0.02816	−73.61735	−1.11819
K	32.10410	128.36703	−17.84178	162.68882	16.26063	127.94771	7.13098	−6.43653	−6.66786
H	38.76504	67.62402	0.23262	169.08270	−0.10768	62.42881	−0.01217	−88.53369	−0.89226
F	46.09927	82.44782	−0.60074	170.65129	0.18135	82.93029	0.05600	−69.12550	−3.74291
R	43.41026	168.24143	−23.15171	209.71546	21.56797	167.75229	−0.00272	−150.98724	3.15302
Y	51.01288	113.22057	−0.68939	291.68460	0.43510	1.72185	−0.11208	−71.08480	−0.65429
W	71.04781	128.71277	−33.55700	163.28495	30.47379	128.31185	19.96383	−8.00807	−26.95079

and we can be confident that at ∼3.5 Å resolution the natural globbic scattering factors for proteins show no significant dependence on intra-globbic conformation.

2.4. Single-Gaussian approximation

Our experience with the four-Gaussian, nine-coefficient fitting, and the evidence for conformational independence of the globbic scattering factors, led us to try fitting single-Gaussian exponential coefficients according to

$$g(s) = Z_g \exp(-b_g s^2). \quad (8)$$

The fitted b_g values are given in Table 5 along with the error-of-fit R values (7), which show that the single-Gaussian approximation is not bad. The one-parameter errors of fit exceed $R = 10\%$ only for a few of the largest side-chain globs, and for the most important single case, the main-chain peptide glob, $R < 3\%$. Table 5 also shows that, as might be expected, the b_g values tend to increase with increasing Z_g. As shown in Fig. 4, the single parameter $\langle b_g\rangle/\langle Z_g\rangle = 1.7$ from the data in Table 5 provides the very simple empirical relationship

$$g(s) = Z_g \exp(-1.7 Z_g s^2), \quad (9)$$

which can be used to approximate globbic scattering factors for biochemical groups, such as uncommon amino acids, not already tabulated.

3. Phase errors due to the globbic approximation

As we hope to develop techniques for low-resolution phasing utilizing the globbic scattering factors, we have carried out calculations of the amplitude-weighted average phase errors,

$$\langle|\Delta\varphi|\rangle = \frac{\sum_{\mathbf{h}} |F_{\mathbf{h}}||\varphi_{\mathbf{h}}(\text{globs}) - \varphi_{\mathbf{h}}(\text{atoms})|}{\sum_{\mathbf{h}} |F_{\mathbf{h}}|}, \quad (10)$$

that result from the globbic approximation to a known structure.

For these tests, we used diffraction data for crambin crystals from the data set measured at $T = 130$ K to $d = 0.83$ Å resolution by Hope (1988) and the structure model fitted to these data by Teeter *et al.* (1993). For three different globbic structure models, phase errors (10) were calculated for resolution shells of 100 reflections for the 850 lowest resolution measured reflections, for which $40.96 \geq d \geq 2.94$ Å. The three models were: (i) all main-chain peptide globs, all amino-acid side-chain globs and all modelled water molecule globs (simply O atoms), (ii) main- and side-chain globs only, water molecules omitted, and (iii) main-chain globs only, side chains and water molecules omitted. For the structure-factor calculations to obtain the globbic phases, the globs were positioned at the Z_a-weighted geometric centroids of the corresponding

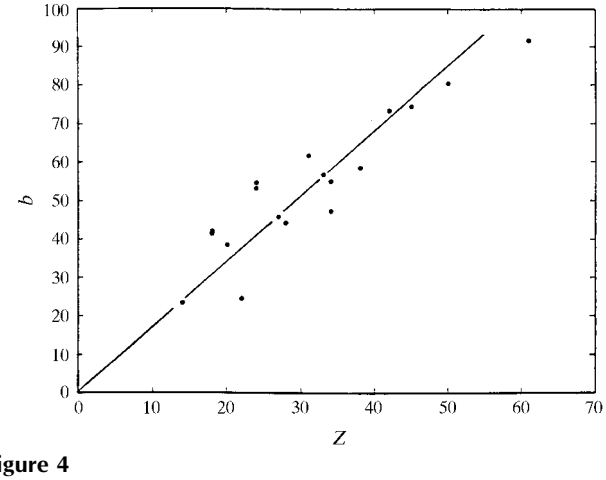

Figure 4
A plot of the single-Gaussian exponential coefficient, b_g, against the number of electrons, Z_g, in each of the 20 amino-acid side-chain globs and the peptide unit main-chain glob.

Table 4
Numerical Debye globbic scattering-factor values $f_g(s)/Z_g$ for five-atom globs with similar values of Z_g.

$(\sin\theta)/\lambda$ (Å^{-1})	d (Å)	Glu extended	Glu folded	$-C^{\alpha}(-C^{\beta})C'(=O)N-$		Gln	Lys
Z = 31		Z = 34	Z = 34	α-helix Z = 33	β-sheet Z = 33	Z = 33	Z = 31
0.01	50.00	0.99	0.99	0.99	0.99	0.99	0.99
0.02	25.00	0.97	0.97	0.97	0.97	0.97	0.96
0.03	16.67	0.94	0.94	0.94	0.94	0.94	0.91
0.04	12.50	0.90	0.90	0.90	0.90	0.89	0.85
0.05	10.00	0.84	0.84	0.84	0.84	0.84	0.78
0.06	8.33	0.78	0.78	0.78	0.78	0.78	0.71
0.07	7.14	0.72	0.72	0.72	0.72	0.72	0.64
0.08	6.25	0.65	0.65	0.65	0.65	0.65	0.58
0.09	5.56	0.59	0.59	0.58	0.58	0.58	0.53
0.10	5.00	0.53	0.53	0.52	0.52	0.53	0.49
0.11	4.55	0.47	0.47	0.46	0.46	0.47	0.45
0.12	4.17	0.42	0.42	0.41	0.41	0.42	0.42
0.13	3.85	0.39	0.38	0.37	0.37	0.38	0.39
0.14	3.57	0.35	0.35	0.34	0.34	0.35	0.3.6
0.15	3.33	0.33	0.33	0.32	0.32	0.32	0.33
0.16	3.13	0.31	0.31	0.30	0.30	0.30	0.30
0.17	2.94	0.30	0.30	0.29	0.29	0.29	0.28

Table 5
Exponential coefficients b_g and residual errors of fit (7) for the single-Gaussian approximation (8) to the Debye globbic scattering factors (2).

Glob		Number of atoms	Z_g	b_g	R
$C^{\alpha}C'(O)NC^{\alpha}$	X	4	27	45.85765	0.031
Cys	C	2	22	24.53398	0.028
Ser	S	2	14	23.31156	0.012
Val	V	3	18	41.71702	0.033
Thr	T	3	20	38.47072	0.035
Pro	P	3	18	41.25170	0.028
Ile	I	4	24	54.53361	0.060
Leu	L	4	24	53.17353	0.037
Met	M	4	34	47.31583	0.070
Asn	N	4	27	45.85304	0.034
Asp	D	4	28	44.23273	0.032
Gln	Q	5	33	56.76736	0.061
Glu	E	5	34	54.96528	0.063
Lys	K	5	31	61.68775	0.106
His	H	6	38	58.51054	0.044
Phe	F	7	42	73.33941	0.066
Arg	R	7	45	74.47762	0.125
Tyr	Y	8	50	80.45150	0.097
Trp	W	10	61	91.61687	0.106

groups of non-H atoms, each glob was assigned a mean-square displacement parameter equal to the average for the atoms of the corresponding group, and globs corresponding to disordered groups were assigned the appropriate fractional site-occupation parameters. The fully refined structure model was used to the fullest extent possible in order to test only the effect of the globbic approximation.

Results are presented in Fig. 5, with the three models represented by (i) dotted, (ii) dashed and (iii) full curves, respectively. At high resolution, for $d_{\min} < 3$ Å, the globbic phase errors rapidly approach 90° random phase errors, but with decreasing resolution the globbic phase errors decrease rapidly. For $d_{\min} \gtrsim 3.5$ Å, the phase errors are reduced to and below a quite acceptable 45°, even for the crude model (iii) which includes only the main-chain globs and excludes all side-chain and water structure. The main-chain-only model (iii) exhibits a broad local maximum phase error $\langle\Delta\varphi\rangle \simeq 40°$ at $(\sin\theta)/\lambda \simeq 0.1$ Å^{-1}, $d \simeq 5$ Å. This maximum spans a spacing interval $7 \gtrsim d \gtrsim 4$ Å that corresponds to important repeat spacings in the omitted side-chain structure: 3.8 Å $C^{\alpha i}\cdots C^{\alpha(i+1)}$ to 5.4 Å β-helix and 6.9 Å β-sheet $C^{\alpha(i-1)}\cdots C^{\alpha(i+1)}$.

4. Electron-density maps from globbic phases

Satisfactorily small phase errors should produce satisfactorily interpretable electron-density maps, and to test this point the experimental crambin $|F_o|$ data with $d \leq 3.5$ Å were phased with several different, somewhat more approximate (and therefore probably somewhat more realistic) models than the models described above for the tests of phase error due to the globbic approximation alone. The map-phasing models were: (*a*) a simplified atomic model with all water molecules and the

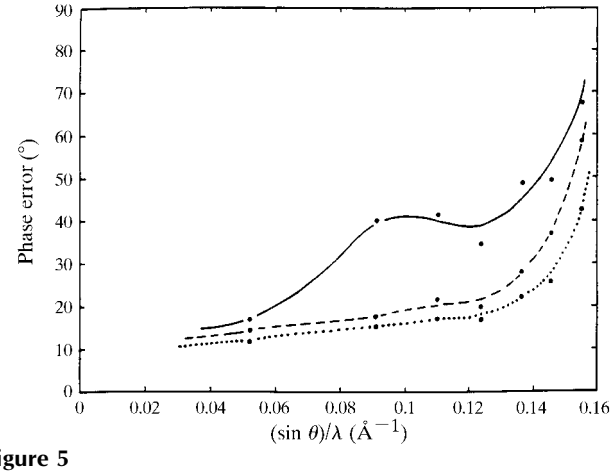

Figure 5
Resolution distribution of $|F_o|$-weighted average globbic phase errors, equation (10), for crambin. The dotted curve corresponds to phasing by (i) all main- and side-chain globs and solvent molecules, the dashed curve corresponds to phasing by (ii) main- and side- chain globs without solvent, and the solid curve corresponds to phasing by (iii) polypeptide main-chain globs only, with all side-chains and solvent omitted.

secondary conformations of disordered main- and side-chain groups omitted, (*b*) a globbic model representing all main- and side-chain globs in their primary conformations, (*c*) a globbic polyalanine model, *i.e.* main-chain globs with all side chains truncated to an alanyl side chain (simply a C atom), and (*d*) a globbic polyglycine model, *i.e.* main-chain peptide globs only. For these models, the globs were positioned at the corresponding unit-weighted geometric centroids of group atomic coordinates and all globs were assigned unit occupation parameters and the same mean-square displacement parameter, $B = 4.5$ Å^2, which is the unit-cell average for the refined atomic model. Any reflections for which $|F_c| < 0.1|F_o|$ were omitted from further consideration. Amplitude agreement factors

$$R(|F|) = \frac{\sum_{\mathbf{h}} \big||F_{\mathbf{h}}(\text{globs})| - |F_{\mathbf{h}}(\text{atoms})|\big|}{\sum_h |F_{\mathbf{h}}(\text{globs})|}, \qquad (11)$$

as well as mean phase errors (10) are given in Table 6, and electron-density maps are shown in Fig. 6.

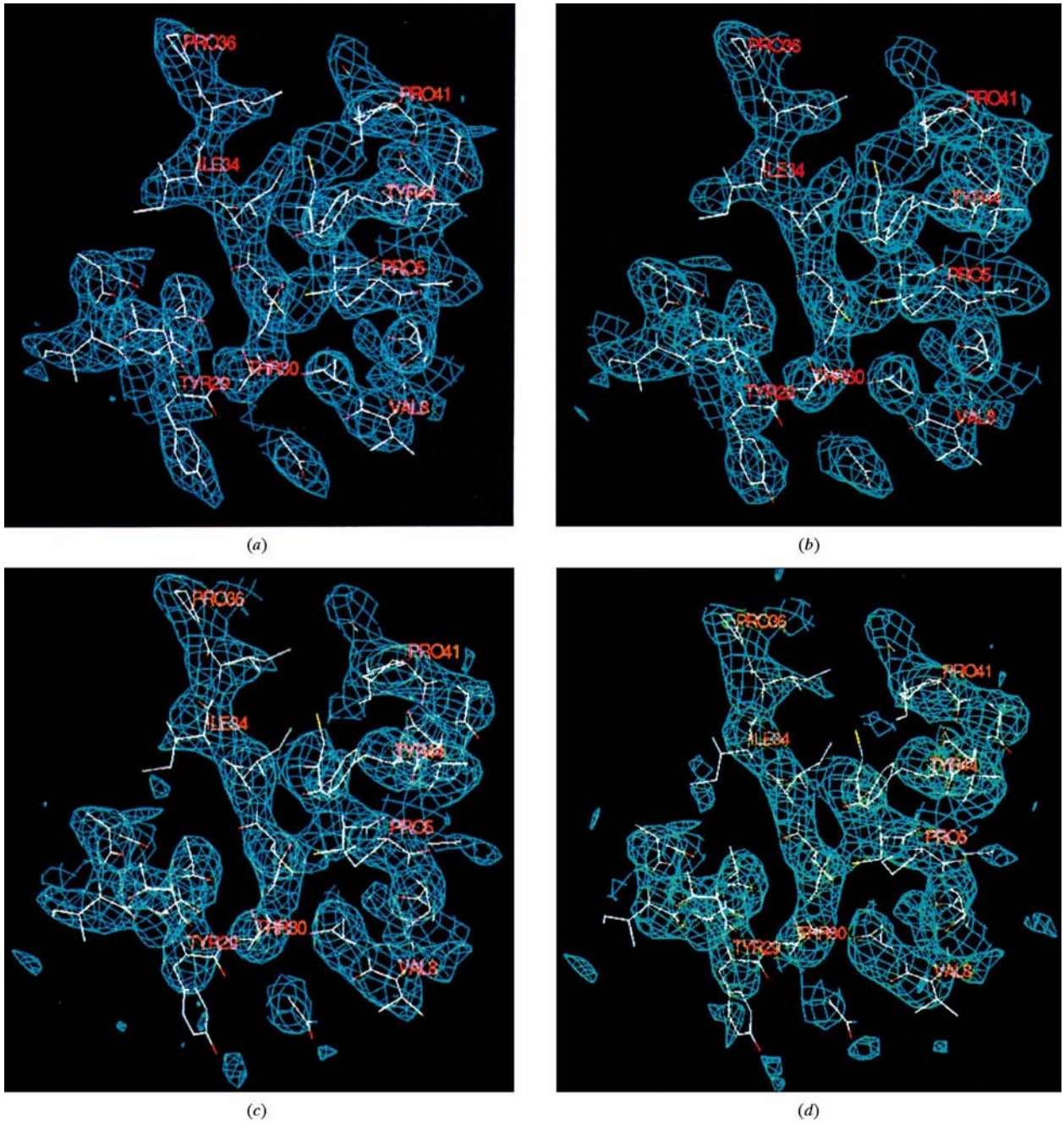

Figure 6
Crambin electron-density maps to 3.5 Å resolution plotted as 1σ, $\rho = \langle(\rho - \langle\rho\rangle)^2\rangle^{1/2}$, isodensity surfaces. In order to exclude, as far as possible, density from symmetry-equivalent molecules, density farther than 2.0 Å from any protein atom in the independent molecule is not plotted. The four maps are phased by: (*a*) a simplified atomic model representing main- and side-chain primary conformers, with all solvent molecules and minor conformers omitted, (*b*) the globbic model corresponding to the simplified atomic model, (*c*) a globbic polyalanine model and (*d*) a globbic polyglycine model.

Table 6
Figures of merit for globbic modelling of the crambin structure.

Model	Number of reflections	$R(\|F\|)$	$\langle\|\Delta\varphi\|\rangle$
(*a*) Atomic model	472	35.4	0.0
(*b*) Globbic model	460	40.0	24.6
(*c*) Globbic polyalanine	459	50.7	37.0
(*d*) Globbic polyglycine	455	53.4	42.2

The maps shown in Fig. 6 all correspond to 3.5 Å resolution and are contoured at 1σ, *i.e.* at $\rho = \langle(\rho - \langle\rho\rangle)^2\rangle^{1/2}$. The map phased by the simplified atomic model (Fig. 6*a*) showed one break in main-chain density between Tyr29 and Thr30. Not surprisingly, this break also occurred in the maps phased by the globbic models (Figs. 6*b*, 6*c* and 6*d*), and there was additionally a second break between Ala38 and Thr39. As the globbic models were abbreviated, first to polyalanine and then to polyglycine, side-chain density tended, of course, to disappear, but even in the globbic polyglycine map, many residues showed a bulge in the side-chain direction. Given that the maps correspond to models that range from a quite precise atomic main-plus-side-chain model to a quite crude main-chain-only globbic model, with amplitude and phase agreement statistics that range over $0.35 < R(|F|) < 0.55$ and $0 < \langle|\Delta\varphi|\rangle < 42^\circ$, we think the maps are remarkably similar, and we are encouraged to continue our research into finding globbic methods for *ab initio* phasing at ordinary protein diffraction resolution.

We are grateful for support of our research by USDHHS PHS NIH grant No. GM46733.

References

Cromer, D. T. & Waber, J. T. (1965). *Acta Cryst.* **18**, 104–109.
Deacon, A. (1997). Personal communication.
Debye, P. (1915). *Ann. Phys. (Leipzig)*, **46**, 809–823.
Guinier, A. (1994). *X-ray Diffraction in Crystals, Imperfect Crystals, and Amorphous Bodies*, pp. 49–50. New York: Dover Publications, Inc.
Guo, D. Y., Smith, G. D., Griffin, J. F. & Langs, D. A. (1995). *Acta Cryst.* A**51**, 945–947.
Harker, D. (1953). *Acta Cryst.* **6**, 731–736.
Hauptman, H. A. (1997). *Curr. Opin. Struct. Biol.* In the press.
Hope, H. (1988). *Acta Cryst.* B**44**, 22–26.
Leherte, L., Fortier, S., Glasgow, J. & Allen, F. H. (1994). *Acta Cryst.* D**50**, 155–166.
Prince, E., Finger, L. W. & Konnert, J. H. (1992). *International Tables for Crystallography*, Vol. C, edited by A. J. C. Wilson, pp. 609–617. Dordrecht: Kluwer Academic Press.
Sheldrick, G. M. (1997). Personal communication.
Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.
Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A., Housset, D., Langs, D. A. & Miller, R. (1997). *Acta Cryst.* D**53**, 551–557.
Smith, G. D., Nagar, B., Rini, J. M., Hauptman, H. A. & Blessing, R. H. (1998). *Acta Cryst.* D**54**, 799–804.
Teeter, M. M., Roe, S. M. & Heo, N. H. (1993). *J. Mol. Biol.* **230**, 292–311.
Turner, M. A., Yuan, C.-S., Hershfield, M., Borchardt, R. T., Smith, G. D. & Howell, P. L. (1997). Personal communication.
Warren, B. E. (1990). *X-ray Diffraction*, pp. 116–117. New York: Dover Publications, Inc.