Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Bulk-solvent correction in direct-methods phasing

D. Y. Guo,* Robert H. Blessing and David A. Langs

Hauptman-Woodward Institute, 73 High Street, Buffalo, New York 14203, USA

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

Received 15 October 1999
Accepted 10 January 2000

It is shown that for crystals of large proteins at low diffraction resolution, with $N \simeq 10\,000$ independent non-H protein atoms and $d_{\min} \simeq 8$ Å, a simple bulk-solvent correction yields the Sayre equation in its classical form, $F_{\mathbf{h}} = q \sum_{\mathbf{k}} F_{\mathbf{k}} F_{\mathbf{h}-\mathbf{k}}$. In the low-resolution protein case, the proportionality factor becomes $q = 1/[(\langle\rho_P\rangle - \rho_S)V]$, where V is the unit-cell volume, ρ_S is the assumed constant electron density in the solvent regions of the crystal and $\langle\rho_P\rangle$ is the average electron density in the protein regions. The classical form of the tangent formula follows from the bulk-solvent-corrected Sayre equation and its validity at low resolution is verified in empirical calculations.

1. Introduction

Bulk solvent occupies anywhere from ~25 to ~75% of the unit-cell volume in protein crystals (Matthews, 1968, 1985). A typical protein molecule has a partial specific volume close to $\overline{v} = 0.74$ mm^3 mg^{-1} (Matthews, 1968; Edelstein & Schachman, 1973; Kratky *et al.*, 1973; Westbrook, 1985) and an empirical formula close to C_4H_6NO, so that the average electron density in the protein part of a typical protein crystal is $\langle\rho_P\rangle \simeq 0.44$ e Å^{-3} and the volume per non-H protein atom is $V_{\text{non-H}} \simeq 17.2$ Å^3. If the bulk solvent resembles liquid water, with mass density $\rho_{H_2O} = 1.00$ mg mm^{-3}, the average electron density in the solvent part of a typical protein crystal would be $\langle\rho_S\rangle \simeq 0.33$ e Å^{-3} and the volume per water molecule would be $V_{H_2O} \simeq 29.9$ Å^3 . Buffer, salt, preservative or crystallizing agent dissolved in the bulk solvent would increase $\langle\rho_S\rangle$ somewhat.

For the most part, the bulk solvent has a disordered liquid-like structure, so that the the amplitude of scattering by the bulk solvent falls off very steeply with increasing scattering angle. The general flatness of the bulk-solvent electron-density distribution is used to advantage in the solvent-flattening technique in protein crystal structure determination (Wang, 1985), but in structure refinement it has been a common expedient to deal with bulk-solvent scattering by simply omitting reflections for which $d_{hkl} = \lambda/(2\sin\theta_{hkl}) \gtrsim$ 6–8 Å. Better, more recent, refinement practice, however, is to include all observed reflections and apply to the model-calculated structure factors a bulk-solvent correction based on Babinet's principle, which is described below. A typical bulk-solvent-corrected form is

$$F_{\text{Total}} = F_{\text{Prot}}[1 - k_{\text{Solv}} \exp(-2\pi^2 \langle u^2_{\text{Solv}}\rangle / d^2_{hkl})],$$

in which $k_{\text{Solv}} = \langle\rho_S\rangle/\langle\rho_P\rangle$, $\langle u^2_{\text{Solv}}\rangle = B_{\text{Solv}}/(8\pi^2)$, $0.75 \lesssim k_{\text{Solv}} \lesssim 1.0$, $200 \lesssim B_{\text{Solv}} \lesssim 400$ Å^2 and $2.5 \lesssim \langle u^2_{\text{Solv}}\rangle \lesssim 5$ Å^2 (Moews &

Kretsinger, 1975; Jiang & Brünger, 1994; Urzhumtsev & Podjarny, 1995; Kostrewa, 1997; Tronrud, 1997; Badger, 1997).

In the context of efforts to develop *ab initio* direct methods of phasing for protein crystals that yield only low- to medium-resolution diffraction data, the presence of more-or-less distinct higher density protein and lower density solvent regions presents an apparent violation of the starting hypothesis of probabilistic direct methods theory, *viz*. the hypothesis of uniform random distributions of independent atoms. In this paper, we describe an algebraic analysis that leads to a bulk-solvent-compensated (BSC) Sayre equation, triplet-phase relationship and tangent formula and we report empirical tests of these BSC results with low-resolution data for two protein crystals. In a subsequent paper, we shall describe corresponding results of a BSC probabilistic analysis for three-phase structure invariants.

2. BSC crystal structure factors

Assume that protein crystals may be described as containing liquid-like bulk-solvent regions of approximately constant relatively low electron density ρ_S between protein molecular regions of relatively high but fluctuating electron density $\rho_P(\mathbf{r})$. Then, the unit-cell volume is partitioned into protein and solvent subvolumes,

$$V = V_P + V_S,$$

and the unit-cell electron-density distribution,

$$\rho(\mathbf{r}) = \begin{cases} \rho_P(\mathbf{r}) & \text{if } \mathbf{r} \subseteq V_P \\ \rho_S & \text{if } \mathbf{r} \subseteq V_S \end{cases},$$

and the crystal structure factors,

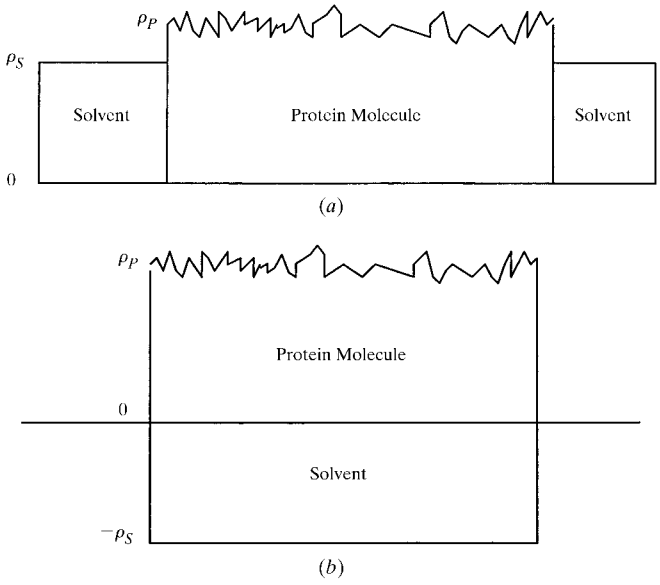

Figure 1
Schematic illustration of the background contrast principle as applied to the Babinet-complementary protein molecular regions and liquid-like bulk-solvent regions in a protein crystal. Low-resolution scattering by the electron densities ρ_P inside and ρ_S outside the protein is equivalent to scattering by the density difference $(\rho_P - \rho_S)$ inside and zero density outside.

$$F_{\mathbf{h}} = \mathcal{F}_V[\rho(\mathbf{r})] = \mathcal{F}_{V_P}[\rho_P(\mathbf{r})] + \mathcal{F}_{V_S}[\rho_S],$$

are similarly partitioned.

In the last equation, $\mathcal{F}$ denotes the Fourier transform operation, according to which

$$\begin{cases} F_{\mathbf{h}} = \mathcal{F}[\rho(\mathbf{r})] = \int\limits_V \rho(\mathbf{r}) \exp(+2\pi i \mathbf{h}\cdot\mathbf{r})\, \mathrm{d}V \\ \rho(\mathbf{r}) = \mathcal{F}^{-1}(F_{\mathbf{h}}) = (1/V)\sum\limits_{\mathbf{h}} F_{\mathbf{h}} \exp(-2\pi i \mathbf{h}\cdot\mathbf{r}) \end{cases},$$

where $\mathbf{r} = x\mathbf{a} + y\mathbf{b} + z\mathbf{c}$, $\mathbf{h} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$, $|\mathbf{h}| = 1/d_{hkl} = 2(\sin\theta_{hkl})/\lambda$ and $\mathbf{h}\cdot\mathbf{r} = hx + ky + lz$. The complex exponential function in the Fourier transform operator has the property that, for integration over the whole unit-cell volume V,

$$\int\limits_V \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\, \mathrm{d}V = \int\limits_V \cos(2\pi \mathbf{h}\cdot\mathbf{r})\, \mathrm{d}V + i\int\limits_V \sin(2\pi \mathbf{h}\cdot\mathbf{r})\, \mathrm{d}V = 0, \quad \forall \mathbf{h} \neq 0.$$

Therefore, for constant ρ_S integrated over the subvolumes V_S and V_P,

$$\int\limits_{V_S} \rho_S \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\, \mathrm{d}V = -\int\limits_{V_P} \rho_S \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\, \mathrm{d}V, \quad \forall \mathbf{h} \neq 0. \quad (1)$$

Then, for the protein and solvent partitioned structure factors,

$$F_{\mathbf{h}} = \int\limits_V \rho(\mathbf{r}) \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\, \mathrm{d}V = \int\limits_{V_P} \rho_P(\mathbf{r}) \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\, \mathrm{d}V + \int\limits_{V_S} \rho_S \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\, \mathrm{d}V, \quad \forall \mathbf{h},$$

it follows that, at low resolution such that ρ_S is practically constant,

$$F_{\mathbf{h}} = \int\limits_{V_P} [\rho_P(\mathbf{r}) - \rho_S] \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\, \mathrm{d}V, \quad \forall \mathbf{h} \neq 0. \quad (2)$$

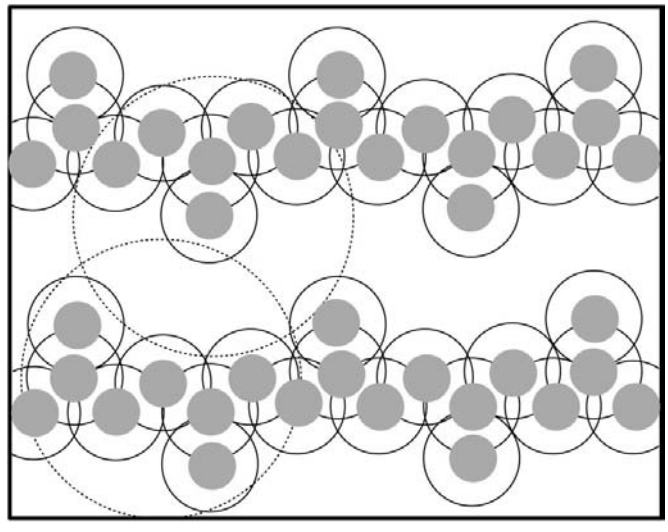

Figure 2
Schematic illustration of bulk-solvent-compensating overlapping spheres of negative scattering density centered on each non-H protein atom. The filled circles represent protein atoms in a pair of neighboring chain segments. The solid-outlined circles surrounding the filled circles represent atom-centered spheres of volume equal to the average volume per protein atom with $V_{\text{non-H}} \simeq 17.2$ Å^3 and $R_{\text{non-H}} \simeq 1.60$ Å. The dashed-outlined circles represent the atom-centered bulk-solvent-compensating spheres with $R_S \simeq 5$ Å and $V_S \simeq 525$ Å^3.

In other words, at low resolution the protein and the bulk-solvent regions constitute Babinet-complimentary scattering masks which scatter with opposite phase (see, for example, Strong, 1958). This background contrast principle, according to which the scattering by the total electron-density distribution is equivalent to scattering by protein minus solvent difference density in the protein regions and zero density in the solvent regions, is illustrated schematically in Fig. 1.

3. BSC Sayre equation, triplet relationship and tangent formula

Historically, the triplet relationship and the tangent formula were established through Sayre's theoretical analysis (Sayre, 1952; see also Fan, 1998) of hypothetical squared structures $\rho^2(\mathbf{r})$ corresponding to resolved-equal-atoms structures $\rho(\mathbf{r})$. To incorporate a bulk-solvent correction into Sayre's analysis, we apply (1) and (2) to the squared-structure factor $Q_\mathbf{h} = \mathcal{F}[\rho^2(\mathbf{r})]$ and obtain

$$
\begin{aligned}
Q_\mathbf{h} &= \int_V \rho^2(\mathbf{r}) \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\,\mathrm{d}V, \quad \forall \mathbf{h},\\
Q_\mathbf{h} &= \int_{V_P} [\rho_P^2(\mathbf{r}) - \rho_S^2] \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\,\mathrm{d}V, \quad \forall \mathbf{h} \neq 0,\\
&= \int_{V_P} [\rho_P(\mathbf{r}) + \rho_S][\rho_P(\mathbf{r}) - \rho_S] \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\,\mathrm{d}V.
\end{aligned}
$$

Then, defining a dimensionless $\Delta = \Delta(\mathbf{r})$ according to

$$
\begin{aligned}
\rho_P(\mathbf{r}) + \rho_S &= (\langle\rho_P\rangle + \rho_S)(1+\Delta),\\
\Delta &= [\rho_P(\mathbf{r}) - \langle\rho_P\rangle]/(\langle\rho_P\rangle + \rho_S)
\end{aligned}
$$

and expecting $|\Delta| << 1$ in low-resolution density images, since local fluctuations about the average protein density should be small compared with the sum of the average protein and solvent densities, we obtain

$$
\begin{aligned}
Q_\mathbf{h} &= (\langle\rho_P\rangle + \rho_S) \int_{V_P} (1+\Delta)[\rho_P(\mathbf{r}) - \rho_S] \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\,\mathrm{d}V\\
&\simeq (\langle\rho_P\rangle + \rho_S) \int_{V_P} [\rho_P(\mathbf{r}) - \rho_S] \exp(2\pi i \mathbf{h}\cdot\mathbf{r})\,\mathrm{d}V\\
&\simeq (\langle\rho_P\rangle + \rho_S) F_\mathbf{h}, \quad \forall \mathbf{h} \neq 0. \qquad (3)
\end{aligned}
$$

Thus, for low-resolution protein-plus-bulk-solvent structures, just as for resolved-equal-atom structures, the squared-structure structure factor is simply proportional to the crystal structure factor.

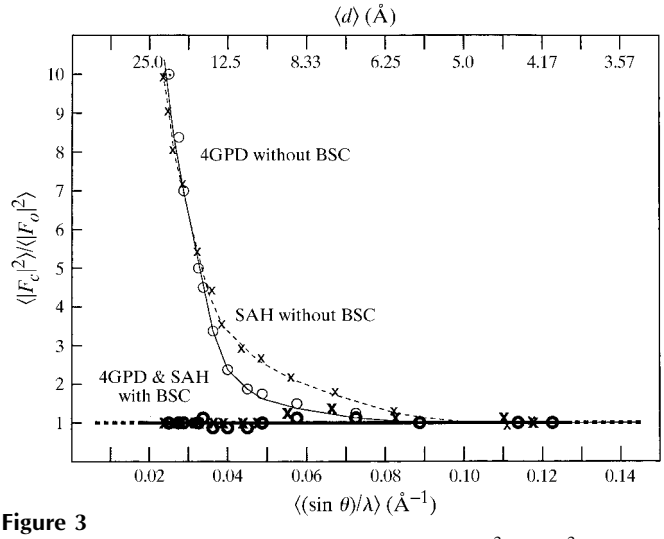

Figure 3
Resolution dependence of the intensity ratios $\langle|F_c|^2\rangle/\langle|F_o|\rangle^2$ with and without bulk-solvent correction.

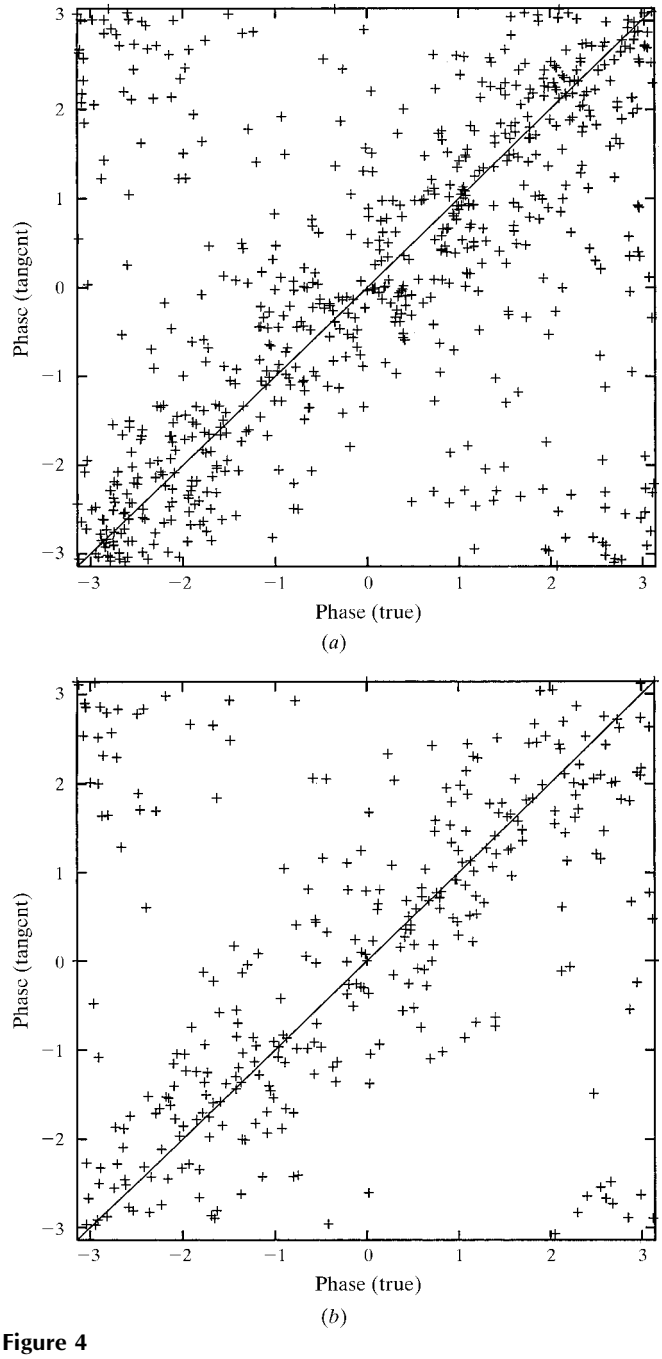

Figure 4
Scatter plot of tangent-formula-estimated *versus* refined-structure-calculated phases. The calculated phases are based on the protein atoms and the atom-centered bulk-solvent correction. Resolution range $\infty > d \geq 8$ Å, calculation-completed experimental $|F|$ data for which $|E| \geq 1.0$. (*a*) 669 GPD phases; (*b*) 409 SAH phases. See also Table 2.

At the same time, the Fourier transforms convolution theorem gives the squared-structure structure factor to be

$$\begin{aligned} Q_{\mathbf{h}} &= \mathcal{F}[\rho^2(\mathbf{r})] = \mathcal{F}[\rho(\mathbf{r})] \otimes \mathcal{F}[\rho(\mathbf{r})] = F_{\mathbf{h}} \otimes F_{\mathbf{h}} \\ Q_{\mathbf{h}} &= (1/V) \sum_{\mathbf{k}} F_{\mathbf{k}} F_{\mathbf{h}-\mathbf{k}}, \end{aligned} \tag{4}$$

so that the bulk-solvent-compensated Sayre equation is

$$\begin{aligned} F_{\mathbf{h}} &\simeq q \sum_{\mathbf{k}} F_{\mathbf{k}} F_{\mathbf{h}-\mathbf{k}}, \\ q &= 1/[(\langle \rho_P \rangle + \rho_S)V]. \end{aligned} \tag{5}$$

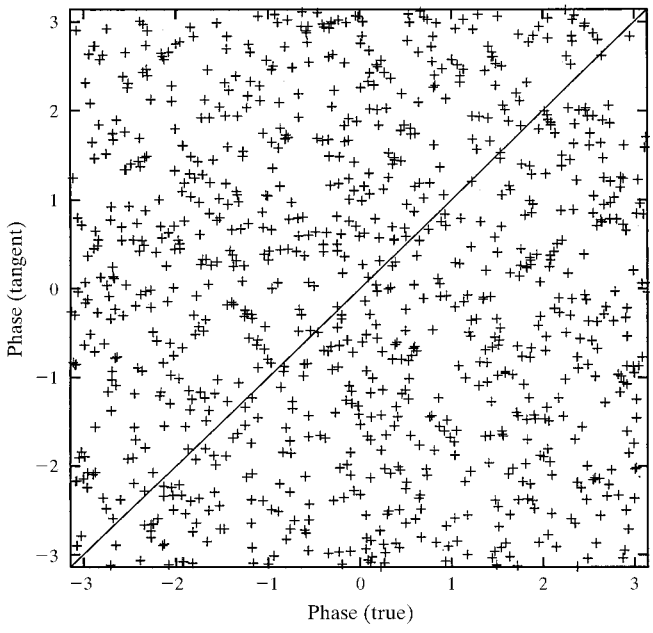

Figure 5
As Figs. 4(*a*) and 4(*b*), but with $8 > d \geq 6$ Å. (*a*) 949 GPD phases; (*b*) 565 SAH phases. See also Table 2.

Apart from the different physical meaning of its proportionality constant, (5) has precisely the same form as the classical Sayre equation. It therefore follows that in favorable cases, for the largest $|F_{\mathbf{h}}|$ and the largest $|F_{\mathbf{k}} F_{\mathbf{h}-\mathbf{k}}|$ leading terms in the Sayre equation sum, phases can be estimated *via* the triplet or three-phase invariant relationship,

$$\varphi_{\mathbf{h}} \simeq \varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}}. \tag{6}$$

More importantly, rewriting (5) as

$$\begin{aligned} |F_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}}) &\simeq q \sum_{\mathbf{k}} |F_{\mathbf{k}}||F_{\mathbf{h}-\mathbf{k}}| \exp[i(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})] = A + iB, \\ A &= q \sum_{\mathbf{k}} |F_{\mathbf{k}}||F_{\mathbf{h}-\mathbf{k}}| \cos(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}}), \\ B &= q \sum_{\mathbf{k}} |F_{\mathbf{k}}||F_{\mathbf{h}-\mathbf{k}}| \sin(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}}), \end{aligned}$$

gives the tangent formula

$$\tan \varphi_{\mathbf{h}} \simeq \frac{B}{A} = \frac{\sum_{\mathbf{k}} |F_{\mathbf{k}}||F_{\mathbf{h}-\mathbf{k}}| \sin(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})}{\sum_{\mathbf{k}} |F_{\mathbf{k}}||F_{\mathbf{h}-\mathbf{k}}| \cos(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})} \tag{7}$$

also in its classical form, even for the case of low-resolution protein-plus-bulk-solvent structures, in which the unit-cell distributions of atomic positions are non-uniform.

4. Atomic sum (and globbic sum) BSC crystal structure factors

For a unit cell containing N_P protein atoms plus approximately constant bulk-solvent electron density ρ_S, applying (1) and (2) yields

$$\begin{aligned} F_{\mathbf{h}} &= \left[\sum_{j=1}^{N_P} f_j \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j) \right] + \int_{V_S} \rho_S \exp(2\pi i \mathbf{h} \cdot \mathbf{r}) \, \mathrm{d}V \\ F_{\mathbf{h}} &= \left[\sum_{j=1}^{N_P} f_j \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j) \right] - \int_{V_P} \rho_S \exp(2\pi i \mathbf{h} \cdot \mathbf{r}) \, \mathrm{d}V \\ F_{\mathbf{h}} &= \sum_{j=1}^{N_P} (f_j - f_S) \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j), \end{aligned} \tag{8}$$

where

$$f_S = \rho_S \int_{V_P/N_P} \exp(2\pi i \mathbf{h} \cdot \mathbf{r}) \, \mathrm{d}V \tag{9}$$

is a bulk-solvent correction subtracted from each atomic scattering factor. As indicated in Fig. 2 and described further below, the correction corresponds to a bulk-solvent compensating sphere of negative scattering density with volume V_P/N_P centered on each protein atom.

Similarly, for low- to medium-resolution structure-factor calculations that employ globbic scattering factors g_j for spherically averaged polyatomic globs or groups of atoms, such as main-chain peptide groups and amino-acid side-chain groups (Guo *et al.*, 1999),

$$F_{\mathbf{h}} = \sum_{j=1}^{N_g} (g_j - g_S) \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j) \tag{10}$$

and

Table 1
Crystal data for glyceraldehyde-3-phosphate dehydrogenase (GDP) and for *S*-adenosylhomocysteine hydrolase (SAH).

	GPD	SAH
Reference	Murthy *et al.* (1980)	Turner *et al.* (1998)
PDB access code	4gpd	1a7a
Space group	*P*1	*C*222
Unit-cell parameters (Å, °)		
a	83.019	92.014
b	80.956	168.019
c	82.545	137.857
α	110.848	90
β	71.473	90
γ	116.858	90
$V_{\rm cell}$ (Å^3)	454260	2131300
Z	1	8
Crystal chemical asymmetric unit (protein atoms†)	$C_{6364}H_{10072}N_{1676}O_{1928}S_{60}$	$C_{4289}H_{6767}N_{1168}O_{1276}S_{20}Se_{30}P_4$
M_r (kDa)	142.8	98.3
Matthews coefficient $V_m = V_{\rm cell}/(ZM_r)$ (Å^3 Da^{-1})	3.18	2.71
Solvent volume ($V_S/V_{\rm cell}$) (%)	61.3	54.6
Radiation	Cu $K\alpha$, λ = 1.5418 Å	Se *K*-edge remote, λ = 0.95 Å
No. unique measured $\|F\|$	37665 (Laue group $\bar{1}$)	26591 (Laue group *mmm*)
$d_{\rm min}$ (Å)	2.5	2.8
$d_{\rm max}$ (Å)	45.9	47.8
Completeness (%)	61.8	99.5

† The SAH crystal chemical unit includes an enzyme dimer complexed with two bound NADH cofactor molecules and two bound adenosine-analog inhibitor molecules. The cofactor and inhibitor atoms were included in our calculations as protein (as distinct from solvent) atoms.

$$g_S = \rho_S \int_{V_P/N_g} \exp(2\pi i\mathbf{h}\cdot\mathbf{r})\,{\rm d}V, \qquad (11)$$

where $N_g < N_P$ is the number of globs per unit cell.

5. BSC atomic scattering factors

In the present work, the bulk-solvent correction subtracted from each protein atom scattering factor in (8) was approximated by the Fourier transform of a protein-atom-centered uniform-density sphere (Guinier, 1994; Patterson, 1967) with parametrically adjustable radius. Thus, (9) becomes

$$f_S = \rho_S \int_{V_P/N_P} \exp(2\pi i\mathbf{h}\cdot\mathbf{r})\,{\rm d}V \simeq Z_S(4/3)\pi R_S^3\Phi(2\pi|\mathbf{h}|R_S), \qquad (12)$$

where

$$\Phi(u) = [3(\sin u - u\cos u)]/u^3. \qquad (13)$$

Assuming bulk-solvent water,

$$Z_S = \rho_S\frac{V_P}{N_P} = \frac{Z_{\rm H_2O}}{V_{\rm H_2O}}V_{\text{non-H}} \simeq \frac{10\,{\rm e}}{29.9\ {\rm \AA}^3}\times 17.2\ {\rm \AA}^3, \qquad (14)$$

and from the empirical calculations described below, $R_S \simeq 5$ Å. This large radius results in a dilute electron density of -0.011 e Å^{-3} within each protein-atom-centered bulk-solvent-compensating uniform-density sphere. As indicated in Fig. 2, overlap of the large-radius spheres centered on neighboring protein atoms (on average, ∼30 neighboring atoms) raises the level of the bulk-solvent-compensating electron density in the protein region to a level commensurate with the average electron density in the bulk-solvent region.

6. Empirical BSC tests

Equations (8) and (12)–(14) have been tested against the measured structure-factor amplitudes and the refined atomic positional, mean-square displacement and site-occupancy parameters for crystals of glyceraldehyde-3-phosphate dehydrogenase (GPD; 10 028 independent non-H protein atoms and 61% solvent volume) and *S*-adenosylhomocysteine hydrolase (SAH; 6 787 independent non-H protein atoms and 55% solvent volume). Crystal data are given in Table 1; structure-factor agreement before and after applying the bulk-solvent correction is summarized in Table 2.

The radii of the bulk-solvent compensating spheres defined by (12)–(14) were fitted by simplex refinements to minimize the normalized mean absolute deviation

$$r = \frac{\sum_{i=1}^{n}\left|\langle|F_o|^2\rangle_i - \langle|F_c|^2\rangle_i\right|}{\sum_{i=1}^{n}\langle|F_o|^2\rangle_i},$$

where *i* indexes the n = 14 resolution subsets listed in Table 2, and where, writing (8) in more explicit terms,

$$F_c = k^{-1}\sum_{j=1}^{N_P} p_j\{f_j\exp[-B_j(\sin\theta_{\mathbf{h}})^2/\lambda^2] - f_S\}\exp(2\pi i\mathbf{h}\cdot\mathbf{r}_j), \qquad (15)$$

in which k is the absolute scaling factor for the $|F_{\rm o}|$ data measured on a relative scale, the $p_j \leq 1$ are atomic site-occupation factors, the $B_j = 8\pi^2\langle u_j^2\rangle$ are atomic mean-square displacement parameters and f_S is the bulk-solvent compensation given by (12)–(14). The simplex refinements converged to bulk-solvent-compensating sphere radii $R_S \simeq 5.51$ and 4.99 Å with fit residuals $r(|F|^2) = 6.8$ and 8.4% for the GPD and SAH structures, respectively.

Table 2 and Fig. 3 show that with no solvent correction the protein-only structure model substantially overestimates the average structure-factor amplitudes for $d \gtrsim 4$ Å resolution, increasingly so with decreasing resolution, by as much as a factor of ten or more for $d \gtrsim 15$ Å. The simple bulk-solvent correction *via* (15) and (12)–(14) leaves only a 'ripple' of overestimation near 5 Å and underestimation near 9 Å in the $|F_c|^2/|F_o|^2$ *versus* resolution statistics.

Table 2
Tests of the bulk-solvent correction *via* equations (14), (12) and (13).

		$\langle\|F_c\|^2\rangle/\langle\|F_o\|\rangle^2$	
d_{min} (Å)	N_{meas}	Protein alone	Protein plus bulk solvent
GPD			
2.5	37665	1.00	1.00
3.0	25877	0.97	0.95
4.0	11506	1.04	1.01
5.0	5644	1.31	1.16
6.0	3135	1.57	1.15
7.0	1885	1.74	1.04
8.0	1204	1.95	0.93
9.0	793	2.38	0.88
10.0	564	3.20	0.93
11.0	415	4.44	0.97
12.0	307	5.77	1.00
13.0	230	6.98	1.01
14.0	175	8.56	1.07
15.0	135	10.33	1.12
SAH			
2.8	26591	1.00	1.00
3.0	21675	1.02	1.01
4.0	9264	1.32	1.24
5.0	4830	1.94	1.46
6.0	2824	2.36	1.32
7.0	1803	2.60	1.10
8.0	1214	2.91	0.98
9.0	873	3.49	0.94
10.0	642	4.47	0.96
11.0	489	5.74	0.99
12.0	372	7.23	1.01
13.0	301	8.16	1.01
14.0	237	9.27	1.00
15.0	195	10.05	1.02

7. Low-resolution tests of the tangent formula

To test the applicability of the tangent formula, phases $\varphi_{\mathbf{h}}$ were estimated *via* (7) from phases $\varphi_{\mathbf{k}}$ and $\varphi_{\mathbf{h}-\mathbf{k}}$ calculated *via* (15) and (12)–(14) for the GPD and SAH structures, and phase errors were evaluated as the unweighted average absolute phase differences

$$\langle|\Delta\varphi|\rangle = \langle|\varphi_{est} - \varphi_{calc}|\rangle.$$

As indicated in (7), the tangent-formula sums were $|F|$ weighted, but as indicated in Tables 3 and 4, the reflection triplets for the tangent-formula calculations were selected based on $|E|$ thresholds, where the $|E|$ values were from local-average-normalized intensities,

$$|E_{\mathbf{h}}| = [(|F_{\mathbf{h}}|^2/\varepsilon_{\mathbf{h}})/(\langle|F|^2/\varepsilon\rangle_{|\mathbf{h}|})]^{1/2}.$$

Table 3 includes results for only those triplets composed of reflections present in the (incomplete) measured data sets. Table 4 and Figs. 4 and 5 present results based on 100% complete sets of calculated and of calculation-completed measured data.

Table 3 shows that for the low-resolution data with $d \geq 6$ Å, the tangent formula produces from the calculated phases of the measured reflections estimated phases that are substantially better than random, with average phase errors of ∼65° for the more complete SAH data and ∼75° for the less complete GPD data. For triplets composed of only reflections from the $8 > d \geq 6$ Å resolution shell, however, the estimated phases are no better than random, with average phase errors of ∼90° for both structures. That the low-resolution tangent-formula phase estimates are on average some 15–25° better than random is noteworthy in view of the large size of the structures under consideration, with ∼10 000 independent non-H atoms in the GPD structure and ∼7 000 in the SAH structure.

Table 3
Average absolute differences between tangent-formula estimated and protein-plus-bulk-solvent model calculated phases for reflections present in the (incomplete) experimental data sets.

Resolution range (Å)	$\|E\|$ threshold	Completeness (%)	N_{data}	$N_{triplets}$	$\langle\|\Delta\varphi\|\rangle$ (°)
GPD					
$\infty > d \geq 8$	$\|E\| \geq 1.0$	58	390	23706	78
$\infty > d \geq 6$		70	1131	254452	75
$8 > d \geq 6$		75	741	40422	92
$\infty > d \geq 8$	$\|E\| \geq 0.6$	63	819	238820	71
$\infty > d \geq 6$		71	2165	1915710	72
$8 > d \geq 6$		76	1346	247128	87
SAS					
$\infty > d \geq 8$	$\|E\| \geq 1.0$	97	398	147162	62
$\infty > d \geq 6$		99	961	966745	61
$8 > d \geq 6$		100	563	110665	85
$\infty > d \geq 8$	$\|E\| \geq 0.6$	98	761	1041719	66
$\infty > d \geq 6$		99	1831	6787048	61
$8 > d \geq 6$		100	1070	775883	87

Table 4
As Table 3, but for 100% complete sets of either calculated data or calculation-completed experimental data.

Resolution range (Å)	$\|E\|$ threshold	N_{data}	$N_{triplets}$	Calc. $\|F\|$s $\langle\|\Delta\varphi\|\rangle$ (°)	Calc.-Comp. Obs. $\|F\|$s $\langle\|\Delta\varphi\|\rangle$ (°)
GPD					
$\infty > d \geq 8$	$\|E\| \geq 1.0$	669	145310	44	45
$\infty > d \geq 6$		1618	878611	52	51
$8 > d \geq 6$		949	94572	96	94
$\infty > d \geq 8$	$\|E\| \geq 0.6$	1290	1064436	46	48
$\infty > d \geq 6$		3062	6061003	52	54
$8 > d \geq 6$		1772	614382	92	92
SAS					
$\infty > d \geq 8$	$\|E\| \geq 1.0$	409	160133	40	41
$\infty > d \geq 6$		974	1005215	45	45
$\infty > d \geq 6$		565	111206	83	84
$\infty > d \geq 8$	$\|E\| \geq 0.6$	777	1102172	47	50
$\infty > d \geq 6$		1849	6966369	48	49
$\infty > d \geq 6$		1072	778304	85	87

Compared with Table 3, Table 4 shows quite dramatic improvements in the low-resolution tangent-formula phase estimates when the data sets are completed by supplying calculated data for the reflections missing from the measured data sets. The low-resolution ($d \geq 8$ Å) average phase errors drop from values of 60–80° for the incomplete data sets to values of 40–45° for the complete data sets. Still, for triplets composed of only reflections from the $8 > d \geq 6$ Å resolution shell, the estimated phases are essentially random, with average phase errors ∼90° even for the completed shell. The

powerful effects of even a very few key low-resolution reflections are highlighted in Tables 3 and 4 by the SAH data, for which supplying only 11 missing reflections to complete the 97% complete $d \geq 8$ Å subset of $|E| \geq 1.0$ measured data, or only 13 reflections to complete the 99% complete $d \geq 6$ Å subset reduced the average phase error from ~60 to ~40°.

8. Concluding comments

The tendency of the tangent-formula-estimated phases toward the model-calculated phases in the $d \geq 8$ Å low-resolution range is shown clearly in the diagonally dominant scatter plots presented in Fig. 4, while the essentially random character of the estimates for triplets restricted to the $8 > d \geq 6$ Å resolution shell is apparent in the plots presented in Fig. 5. Including higher resolution data to $d_{\min} = 2.5$ Å also yields random scatter plots (not shown) that resemble Fig. 5. The high-resolution 'breakdown' of the tangent-formula phase estimation in the absence of $d \geq 8$ Å low-resolution data presumably corresponds to violation of the hypothesis of essentially constant ρ_S and flattened low-resolution ρ_P. This hypothesis is the basis for the expectation $|\Delta| << 1$, which led to (3) and thence to the bulk-solvent-corrected low-resolution Sayre equation (5) and tangent formula (7).

Anticipating results from our forthcoming paper on probabilistic analysis of bulk-solvent-compensated low-resolution phase relationships, we note that the low-resolution success and high-resolution failure of the tangent formula also has a probabilistic explanation. Since the variance of the probability distribution of three-phase structure invariants is proportional to $N^{1/2}$, where N is the number of atoms per primitive unit cell, phase estimation becomes unreliable, essentially random, for large protein structures. At low resolution, however, the *effective* number of atoms becomes much smaller, because all the solvent atoms effectively disappear as shown by (1) and (2) and because the effective scattering units are globs or groups of protein atoms as indicated by (10) and (11). Thus, $N_{\rm eff} < N_P$, where $N_{\rm eff}$ is the effective number of scattering units and N_P is the number of non-H protein atoms per primitive unit cell, and the reliability of low-resolution phase estimation improves (see also Dorset & Jap, 1998). We have observed that, as an empirical rough rule of thumb, $N_{\rm eff} \simeq kN_P/d_{\min}$, with $k \simeq 1$ Å.

We thank Herbert Hauptman and Douglas Dorset for helpful discussions of the low-resolution phase problem and Lynne Howell and co-workers for the SAH data and refined structural model. We are grateful for research support from USDHHS PHS NIH grant No. GM46733.

References

Badger, J. (1997). *Methods Enzymol.* **277**, 344–352.
Dorset, D. L. & Jap, B. K. (1998). *Acta Cryst.* D**54**, 615–621.
Edelstein, S. J. & Schachman, H. K. (1973). *Methods Enzymol.* **27**, 82–98.
Fan, H. F. (1998). *Direct Methods for Solving Macromolecular Structures*, NATO ASI Series Volume, Series C: *Mathematical and Physical Sciences*, Vol. 507, edited by S. Fortier, pp. 79–85. Dordrecht, The Netherlands: Kluwer Academic Publishers.
Guinier, A. (1994). *X-ray Diffraction in Crystals, Imperfect Crystals and Amorphous Bodies*, p. 359. New York: Dover Publications, Inc.
Guo, D. Y., Blessing, R. H., Langs, D. A. & Smith, G. D. (1999). *Acta Cryst.* D**55**, 230–237.
Jiang, J.-S. & Brünger, A. T. (1994). *J. Mol. Biol.* **243**, 100–115.
Kostrewa, D. (1997). *CCP4 Newslett.* **34**, 9–22. http://www.dl.ac.uk/CCP/CCP4/newsletter34/bsdk_text.html
Kratky, O., Leopold, H. & Stabinger, H. (1973). *Methods Enzymol.* **27**, 98–110.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Matthews, B. W. (1985). *Methods Enzymol.* **114**, 176–187.
Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.* **91**, 201–228.
Murthy, M. R. N., Garavito, R. M., Johnson, J. E., Rossmann, M. G. (1980). *J. Mol. Biol.* **138**, 859.
Patterson, A. L. (1967). *International Tables for X-ray Crystallography*, Vol. II, p. 72. Birmingham: The Kynoch Press.
Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.
Strong, J. (1958). *Concepts of Classical Optics*, p. 200. San Francisco: W. H. Freeman & Co., Inc.
Tronrud, D. E. (1997). *Methods Enzymol.* **277**, 306–319.
Turner, M. A., Yuan, C.-S., Borchardt, R. T., Hershfeld, M. S., Smith, G. D. & Howell, P. L. (1998). *Nature Struct. Biol.* **5**, 369–376.
Urzhumtsev, A. G. & Podjarny, A. D. (1995). *CCP4 Newslett.* **31**, 12–16.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
Westbrook, E. M. (1985). *Methods Enzymol.* **114**, 187–196.