Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Direct methods and protein crystallography at low resolution

**Christopher J. Gilmore**

Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland

Correspondence e-mail: chris@chem.gla.ac.uk

Received 4 April 2000
Accepted 27 June 2000

The tools of modern direct methods are examined and their limitations for solving protein structures discussed. Direct methods need atomic resolution data (1.1–1.2 Å) for structures of around 1000 atoms if no heavy atom is present. For low-resolution data, alternative approaches are necessary and these include maximum entropy, symbolic addition, Sayre's equation, group scattering factors and electron microscopy.

## 1. The tools of direct methods

Direct methods have evolved from the early 1950s to become the method of choice for solving small-molecule crystal structures from diffraction data. In this context, 'small' extends from ten atoms in the asymmetric unit to a small protein with 1000 or more atoms. In this section, the tools that direct methods use and their limitations are examined. This is necessarily brief; for a full description, see Giacovazzo (1998), Woolfson & Fan (1995) or Fortier (1997).

### 1.1. Normalization

Intensity data are first normalized to give normalized structure factors or $E$ magnitudes,

$$|E_{\mathbf{h}}|^2 = \frac{k\,|F_{\mathbf{h}}^{\rm obs}|^2}{\varepsilon_{\mathbf{h}} \sum_{j=1}^{N} f_j^2 \exp(2B \sin^2\theta/\lambda^2)}, \qquad (1)$$

where $k$ is a scale factor that puts the observed intensities $|F_{\mathbf{h}}|^2$ on an absolute scale, $\varepsilon_{\mathbf{h}}$ is the statistical weight for reflection $\mathbf{h}$ and $f_j$ is the atomic scattering factor for atom $j$. There are $N$ atoms in the unit cell, with an overall isotropic temperature factor $B$. $B$ and $k$ need to be determined and this is carried out using Wilson's method (Wilson, 1949). This assumes that the atoms in the unit cell are uniformly and randomly distributed and such an assumption forms the basis of Wilson statistics. Obviously, in proteins and other biological macromolecules this is not the case; at the very least, we have an ordered protein and a disordered solvent volume that really requires a different treatment. Nonetheless, it is possible to obtain useful values of $k$ and $B$ by this method.

The distribution of $E$ magnitudes depends on whether the space group is centrosymmetric or not and does not depend on structural complexity. In the non-centrosymmetric case, ~37% of the normalized structure factors are expected to be >1.0, 1.8% >2.0 and only 0.01% >3.0.

### 1.2. Triplets

Each structure-factor magnitude $|F_{\mathbf{h}}|$ has an asscoiated phase angle $\varphi_{\mathbf{h}}$ which we wish to determine. Triplets are the

fundamental phase relationship in direct methods and they take the form

$$\Phi_3 = \varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}} \simeq 0. \quad (2)$$

It is obvious that the indices of the three reflections sum to zero. Associated with each triplet is a concentration parameter $\kappa_{\mathbf{h},\mathbf{k}}$

$$\kappa_{\mathbf{h},\mathbf{k}} = \frac{2|E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}|}{N^{1/2}}, \quad (3)$$

where $N$ is the number of atoms assumed equal in the unit cell, excluding H atoms. Relationship (2) implies a probabilistic origin and the Cochran distribution (Cochran, 1955) gives us the required formula,

$$P(\Phi_3|\kappa_{\mathbf{h},\mathbf{k}}) = \frac{1}{2\pi I_0(\kappa_{\mathbf{h},\mathbf{k}})}\exp(\kappa_{\mathbf{h},\mathbf{k}}\cos\Phi_3). \quad (4)$$

$I_0$ is a zeroth-order Bessel function of the first kind. The expression $2\pi I_0(\kappa_{\mathbf{h},\mathbf{k}})$ is a normalizing term. Fig. 1 shows how Bessel functions appropriate to direct methods behave as a function of their argument. The Cochran distribution assumes the viability of Wilson statistics. Fig. 2 shows how the probability distribution (4) varies with the concentration parameter. It can be seen that the mode of the distribution is always zero and that as $\kappa_{\mathbf{h},\mathbf{k}}$ decreases the information content of the Cochran distribution also decreases, until at $\kappa_{\mathbf{h},\mathbf{k}} = 1$ very little useful information can be obtained concerning the value of the triplet. $\kappa_{\mathbf{h},\mathbf{k}}$ decreases as $1/N^{1/2}$. If the three $E$ magnitudes in the triplet have values of 2.5 then this corresponds to a limit of ~1000 atoms in the unit cell.

### 1.3. Quartets

Quartets are the logical extension of triplets and involve four phases instead of three,

$$\Phi_4 = \varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{\mathbf{l}} + \varphi_{-\mathbf{h}-\mathbf{k}-\mathbf{l}}. \quad (5)$$

The distribution (Schenk, 1973; Hauptman, 1975) is more complex than that of the triplet. Defining the principal terms

$$R_1 = |E_{\mathbf{h}}|,\ R_2 = |E_{\mathbf{k}}|,\ R_3 = |E_{\mathbf{l}}|,\ R_4 = |E_{-\mathbf{h}-\mathbf{k}-\mathbf{l}}| \quad (6)$$

and the unique cross-terms

$$R_{12} = |E_{\mathbf{h}+\mathbf{k}}|,\ R_{23} = |E_{\mathbf{k}+\mathbf{l}}|,\ R_{31} = |E_{\mathbf{l}+\mathbf{h}}|, \quad (7)$$

the required distribution is

$$P(\Phi_4|\,R_1, R_2, R_3, R_4, R_{12}, R_{23}, R_{31}) = \frac{1}{L}\exp(-2B\cos\Phi_4)I_0(2N^{-1/2}R_{12}X_{12}) \times I_0(2N^{-1/2}R_{23}X_{13})I_0(2N^{-1/2}R_{31}X_{31}), \quad (8)$$

where $L$ is a normalizing term (usually determined numerically),

$$\begin{aligned} B &= (2/N)R_1R_2R_3R_4, \\ X_{12} &= [R_1^2R_2^2 + R_3^2R_4^2 + 2R_1R_2R_3R_4\cos\Phi_4]^{1/2}, \\ X_{23} &= [R_2^2R_3^2 + R_1^2R_4^2 + 2R_1R_2R_3R_4\cos\Phi_4]^{1/2}, \\ X_{31} &= [R_3^2R_1^2 + R_2^2R_4^2 + 2R_1R_2R_3R_4\cos\Phi_4]^{1/2}. \end{aligned} \quad (9)$$

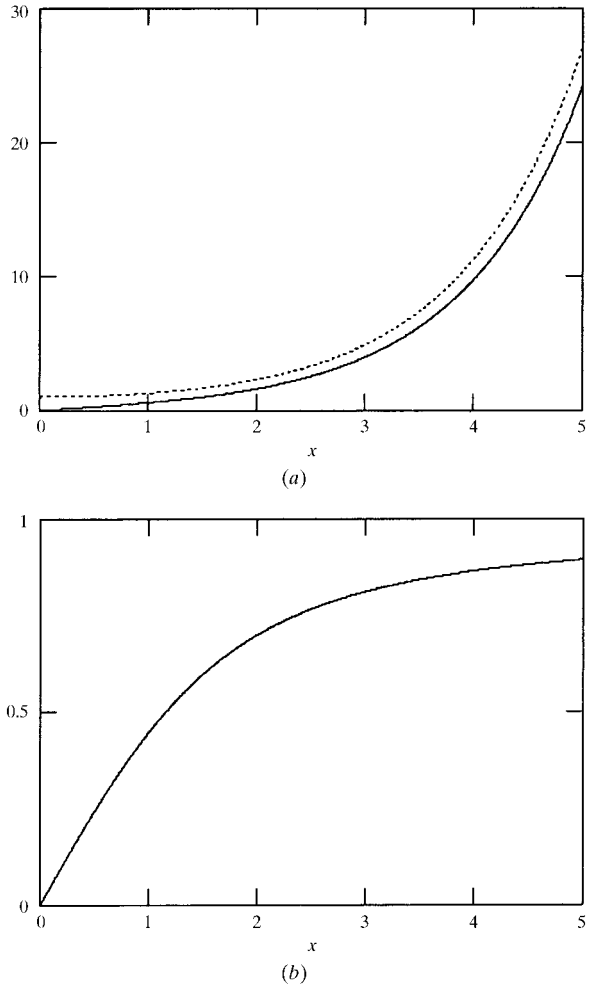


**Figure 1**
The variation of Bessel functions (*a*) $I_0(x)$ (dotted line) and $I_1(x)$ (full line) and (*b*) $I_1(x)/I_0(x)$ as a function of $x$ in the range 0–5.

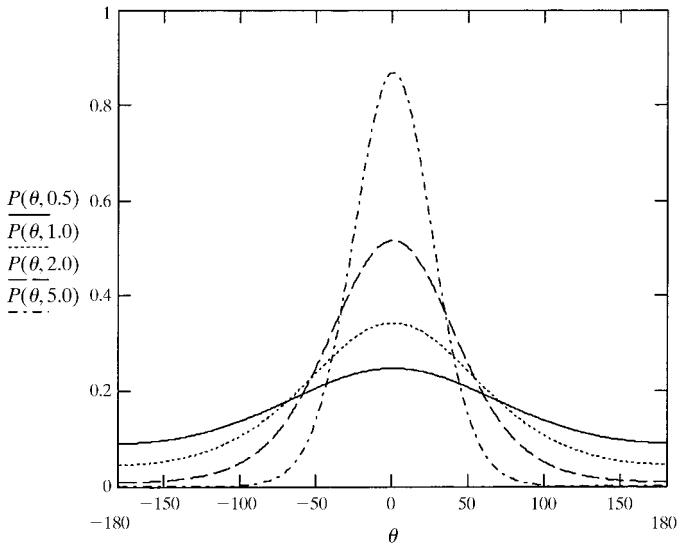


**Figure 2**
The Cochran distribution as a function of the concentration parameter $\kappa_{\mathbf{h},\mathbf{k}}$.

Three sorts of quartet can be identified as follows.

(i) Positive quartets, in which the principal and cross-terms are large. These are strongly correlated with triplets which makes them difficult to use.

(ii) Negative quartets, where the principal terms are large and the cross-terms are small.

(iii) Enantiomorph-sensitive quartets, where the principal terms are large and the cross-terms have an intermediate value. Both the negative and enantiomorph-sensitive quartets are largely independent of triplets.

Fig. 3 shows typical quartet distributions for these three cases using a structure with 200 equal atoms in the unit cell, all four principal terms set to 3.0 and with varying cross-terms. Note that the reliability of quartets is a function of $1/N$ and not $1/N^{1/2}$ as in the triplet case. This makes the use of these invariants in protein crystallography rather questionable.

### 1.4. The tangent formula

The tangent formula (Karle & Hauptman, 1956) is a key formula in direct methods that lets us refine phase values and determine new ones. Consider the situation in which we have a series of triplets with a common reflection $\varphi_{\mathbf{h}}$. They can be written

$$\begin{aligned}\varphi_{\mathbf{h}} &= \varphi_{\mathbf{h}2} - \varphi_{\mathbf{h}-\mathbf{h}2}\\ \varphi_{\mathbf{h}} &= \varphi_{\mathbf{h}3} - \varphi_{\mathbf{h}-\mathbf{h}3}\\ \varphi_{\mathbf{h}} &= \varphi_{\mathbf{h}4} - \varphi_{\mathbf{h}-\mathbf{h}4}\ \textit{etc.}\end{aligned} \tag{10}$$

Consider also a situation in which all the phases on the RHS of (10) are known at least approximately; the tangent formula then gives us an estimate for $\varphi_{\mathbf{h}}$ which in its simplest form is

$$\tan\varphi_{\mathbf{h}} = \frac{\sum_{\mathbf{k}} |E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}|\ \sin(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})}{\sum_{\mathbf{k}} |E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}|\ \cos(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})}. \tag{11}$$

It can be extended to include quartets of all types and various weighting schemes which help impose stability on a formula that can be prone to instabilities.

### 1.5. Symbolic addition

The triplets (and quartets) form a set of linear equations relating phase values. The technique of symbolic addition (Karle & Karle, 1966) assigns algebraic symbols to a small number (typically 4–8) phases that have large associated $E$ magnitudes and which interact strongly through phase relationships. Triplets and quartets are used to determine 50–100 new phases as functions of these symbols; this is the process of symbolic addition. The symbols are converted into numerical values using relationships between them made manifest by the symbolic addition procedure or by giving unassigned symbols permuted values in the range 0–$2\pi$. The phases are then extended and refined using either tangent refinement or the Sayre equation (see §1.7).

Symbolic addition is not much used currently for solving small molecules; it has been superseded by methods that are much easier to automate. It does, however, have the virtues of stability when used when used with macromolecules at low resolution; this is explored further in §3.2.

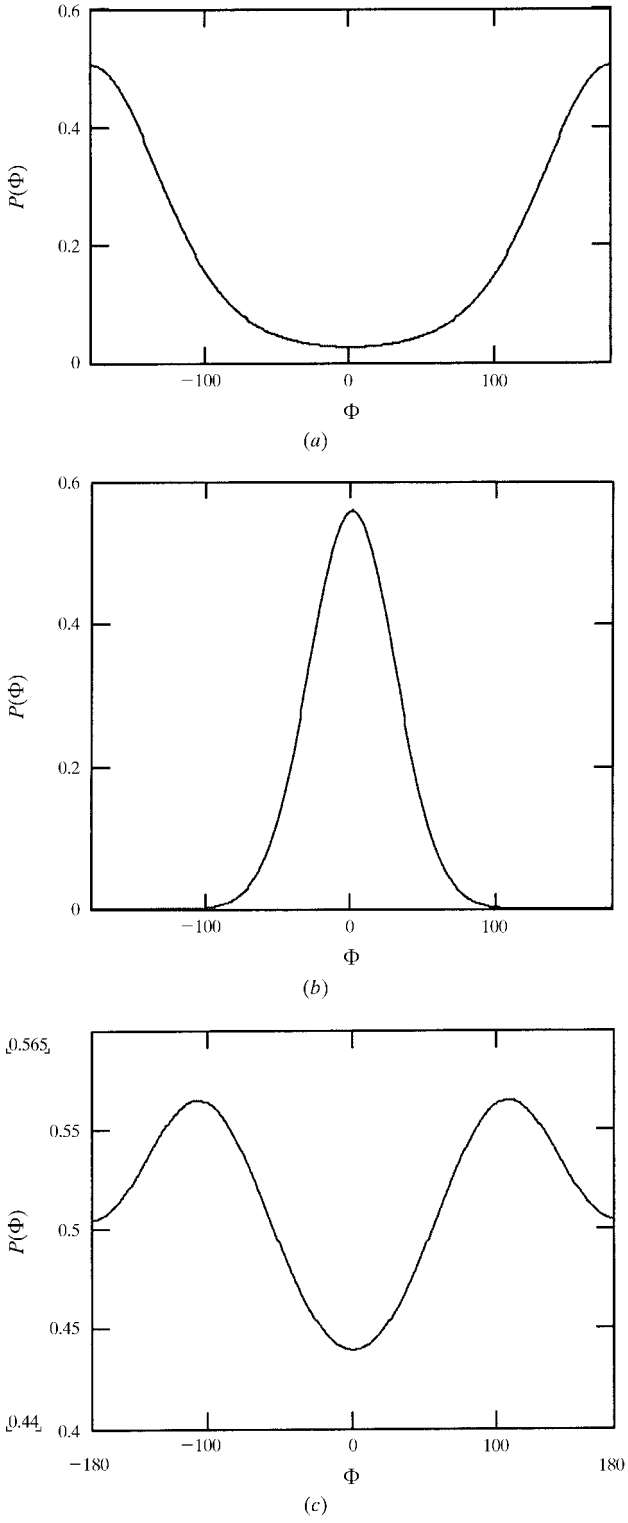


**Figure 3**
Quartet distributions for a crystal structure with $N = 200$. The principal terms are given by $R_1 = R_2 = R_3 = R_4 = 3.0$. (*a*) A positive quartet with cross-terms $R_{12} = R_{23} = R_{31} = 3.0$, (*b*) a negative quartet with cross-terms $R_{12} = R_{23} = R_{31} = 0.25$, (*c*) an enantiomorph quartet with cross-terms $R_{12} = R_{23} = R_{31} = 0.9$.

### 1.6. The minimal principle

The mode of the Cochran distribution for triplets is always zero. However, the mean can be computed as

$$\langle \cos \Phi_3 \rangle = \int_0^{\pi} \cos \Phi_3 P(\Phi_3|\kappa_{\mathbf{h},\mathbf{k}}) = \frac{I_1(\kappa_{\mathbf{h},\mathbf{k}})}{I_0(\kappa_{\mathbf{h},\mathbf{k}})}. \quad (12)$$

This expression gives rise to the minimal function (DeTitta *et al.*, 1994)

$$R(\Phi_3) = \sum_{\mathbf{h},\mathbf{k}} \kappa_{\mathbf{h},\mathbf{k}} \left[ \cos T_{\mathbf{h},\mathbf{k}} - \frac{I_1(\kappa_{\mathbf{h},\mathbf{k}})}{I_0(\kappa_{\mathbf{h},\mathbf{k}})} \right]^2 \Big/ \sum_{\mathbf{h},\mathbf{k}} \kappa_{\mathbf{h},\mathbf{k}}, \quad (13)$$

where $\cos T_{\mathbf{h},\mathbf{k}}$ is the value of the triplet computed from known phases. The function $R(\Phi_3)$ serves two purposes: (i) as a formula to refine and estimate new phases by minimizing the difference between the estimated value and the mean of the cosine of the triplet and (ii) as the minimal principle which uses (13) to define the best phase set, *i.e.* as a figure of merit.

### 1.7. The Sayre equation

The Sayre equation (Sayre, 1952) is algebraic rather than probabilistic in origin and is derived from the expression for the electron density and its square,

$$F_{\mathbf{h}} = (\theta/V) \sum_{\mathbf{k}} F_{\mathbf{k}} F_{\mathbf{h}-\mathbf{k}}. \quad (14)$$

In terms of $E$ magnitudes this takes the form of the Sayre–Hughes (Hughes, 1953) equation,

$$E_{\mathbf{h}} = N^{1/2} \langle E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}} \rangle. \quad (15)$$

The Sayre equation can be used in the same way as the tangent formula, but has a more general validity and is not constrained to use only large structure-factor magnitudes.

### 1.8. Figures of merit

In general, direct methods are multi-solutional: they give rise to multiple phase sets and we need to select those which are most likely to give useful structural information. Figures of merit serve this purpose and are used to rank phase sets. There are numerous such indicators, including the following. (i) The minimal function (13). An optimal phase set will have a minimum value of $R(\Phi_3)$. (ii) The negative quartet figure of merit (DeTitta *et al.*, 1975),

$$\mathrm{NQEST} = \frac{\sum_{\text{negative}} B \cos(\varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{\mathbf{l}} + \varphi_{-\mathbf{h}-\mathbf{k}-\mathbf{l}})}{\sum_{\text{negative}} B}, \quad (16)$$

where the summation spans all those quartets assumed negative using (8). An optimal phase set should have a minimum value of NQEST.

Usually, several figures of merit are calculated for a given phase set and these are combined to given an overall figure called a CFOM.

### 1.9. Correlation coefficients

Let $E_o$ be the observed $E$ magnitude and $E_c$ the calculated value from, for example, a variant of the tangent formula; let $w$ be the associated weight. For a set of such magnitudes we can then compute the correlation coefficient CC, which takes many forms. A useful expression from Read (1986) is

$$\mathrm{CC} = \Big( \sum w E_o^2 E_c^2 \sum w - \sum w E_o^2 \sum w E_c^2 \Big) / \Big( \Big\{ \Big[ \sum w E_o^4 \sum w - \Big( \sum w E_o^2 \Big)^2 \Big] \times \Big[ \sum w E_c^4 \sum w - \Big( \sum w E_c^2 \Big)^2 \Big] \Big\}^{1/2} \Big). \quad (17)$$

Correlation coefficients lie between $-1 \leq \mathrm{CC} \leq 1.0$. They can be used as figures of merit.

### 1.10. *E* maps

So far our discussions have involved reciprocal-space quantities; the transform into real space is carried out using $E$ magnitudes *via* $E$ maps,

$$\rho(\mathbf{x}) \simeq \frac{1}{V} \sum_{\mathbf{h}} |E_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}}) \exp(-2\pi i \mathbf{h} \cdot \mathbf{x}). \quad (18)$$

The use of $E$ magnitudes and the limits we shall impose on the reflections entering the summation in (16) mean that the electron density is only approximate (at the very least, there are serious series-termination errors), but hopefully is sufficient to reveal structural features so that model building can begin.

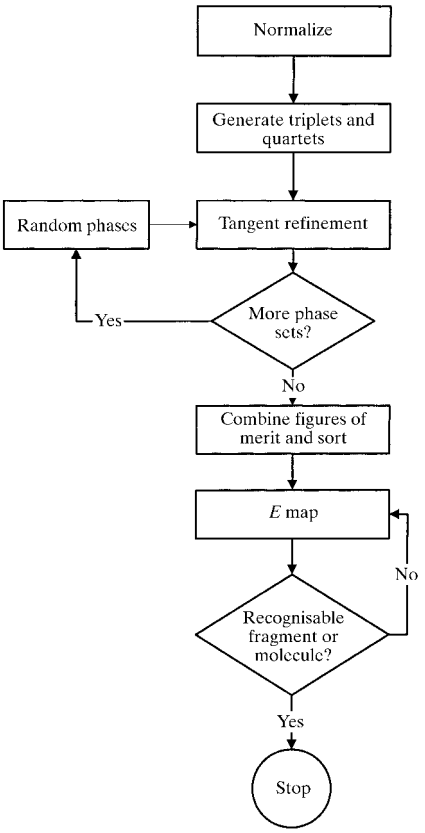


**Figure 4**
A flow chart for traditional direct methods as used to solve small-molecule structures.

### 1.11. Simplifying the problem

The problem of direct phasing can be simplified by the following heuristic rules.

(i) Only the top 8–10 $N_a$ need to be phased, where $N_a$ is the nunber of atoms in the asymmetric unit.

(ii) In centrosymmetric space groups, all the phases are centric with phases restricted to 0. Non-centrosymmetric space groups often have centrosymmetric projections giving rise to centric reflections which have restricted phase choices, *e.g.* 0, $\pi$, $\pm\pi/2$.

(iii) We can tolerate relatively large ($\sim$40°) random errors, but smaller systematic errors.

## 2. Using the tools to solve crystal structures

There are numerous procedures for solving structures *via* direct methods. A typical, though somewhat simplified, sequence is as follows.

(i) The data are normalized using Wilson's method to give $E$ magnitudes. These are sorted in descending order.

(ii) Triplets are generated for the top $10N_a$ reflections. Quartets (usually just the negative ones) are also optionally generated.

(iii) The top 8–$10N_a$ reflections are given random phases.

(iv) The phases are refined to convergence using the tangent formula in one of its many variants.

(v) Figures of merit are calculated for this phase set and combined together to give an overall figure of merit CFOM.

(vi) Steps (iii)–(v) are repeated 24–1000 times depending on the difficulty and complexity of the structure.

(vii) The phase sets are sorted on CFOM.

(viii) An $E$ map is computed for the best set and the peaks picked. We then use our knowledge of molecular dimensions and conformations to extract a trial structure. This is the first point at which chemical knowledge is used actively, *i.e.* the direct-methods procedure is model-free until this point.

(ix) The structure is completed and refined in the usual way.

(x) If no identifiable fragment can be found then the next ranked phase set from step (vii) is used and steps (viii) and (ix) are repeated. This can be performed for the top ten or more phase sets.

This is shown diagrammatically in Fig. 4.

### 2.1. What is needed for this method to work?

The procedure is usually routine if the following criteria are met.

(i) Atomicity. We need intensity data to a resolution of 1.1–1.2 Å.

(ii) Completeness. The data must be complete to this resolution.

(iii) Accuracy. Accurate data are required.

(iv) Complexity. The number of non-H atoms in the asymmetric unit should be <200.

Clearly, none of these criteria apply to most protein data sets, where a resolution of 2 Å is common, where low-angle data may be missing, where accuracy is limited by poor crystalline specimens and where the number of atoms in the asymmetric unit is several thousand.

This latter problem can be overcome using atomicity as a stronger constraint and this gives rise to the computer

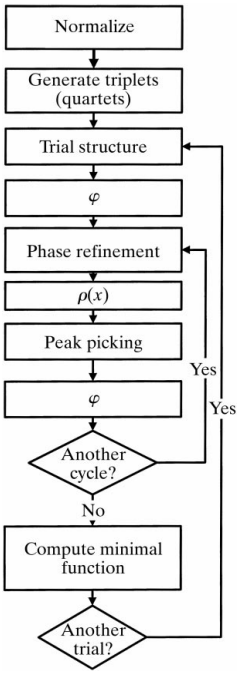


**Figure 5**
A flow chart for the *SnB* program as applied to small proteins with atomic resolution data. (From Weeks & Miller, 1997.)

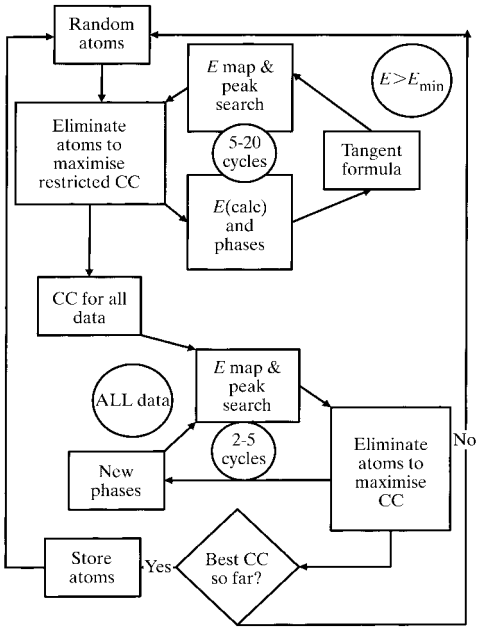


**Figure 6**
A flow chart for the *Half-Bake* computer program as applied to small proteins with atomic resolution data. (From Sheldrick, 1997.)

programs *Shake-and-Bake* (*SnB*; Weeks & Miller, 1997) and *Half-Bake* (Sheldrick, 1997).

### 2.2. *Shake-and-Bake* (*SnB*)

*SnB* starts in a conventional way by normalizing the data and generating triplet (and optionally negative quartet) invariants. To assign initial phases, an extension of the random-phase procedure into reciprocal space is made; trial structures are generated by placing random atoms in the unit cell with distance constraints, *i.e.* atoms may not be closer than 1.5 Å. No angle constraints are applied. A Fourier transform gives phase values which, because of the distance constraint, tend to have lower errors than a simple random-phasing algorithm. Note that an imposition of atomicity is being invoked from the very beginning in this procedure.

The random phases are now refined using either tangent methods or a grid search based on the minimal function in which each phase is modified by a phase shift that minimizes $R(\Phi_3)$. A new map is generated from these refined phases and this is subjected to a peak-search procedure (again we have atomicity) in which $N$ peaks are selected (for an $N$-atom problem) subject to the same distance constraint that was used in the initial phase generation. The new peaks give new phases, which are then refined in a cyclical fashion. At convergence, $R(\Phi_3)$ is stored, a new phase set is generated and the procedure is repeated.

As phase sets accumulate, one looks for a set which has a much lower value of $R(\Phi_3)$ than the others. This is usually an indication of phase correctness and the atoms corresponding to this solution form the starting point of a traditional completion.

The procedure is shown in flow-chart form in Fig. 5.

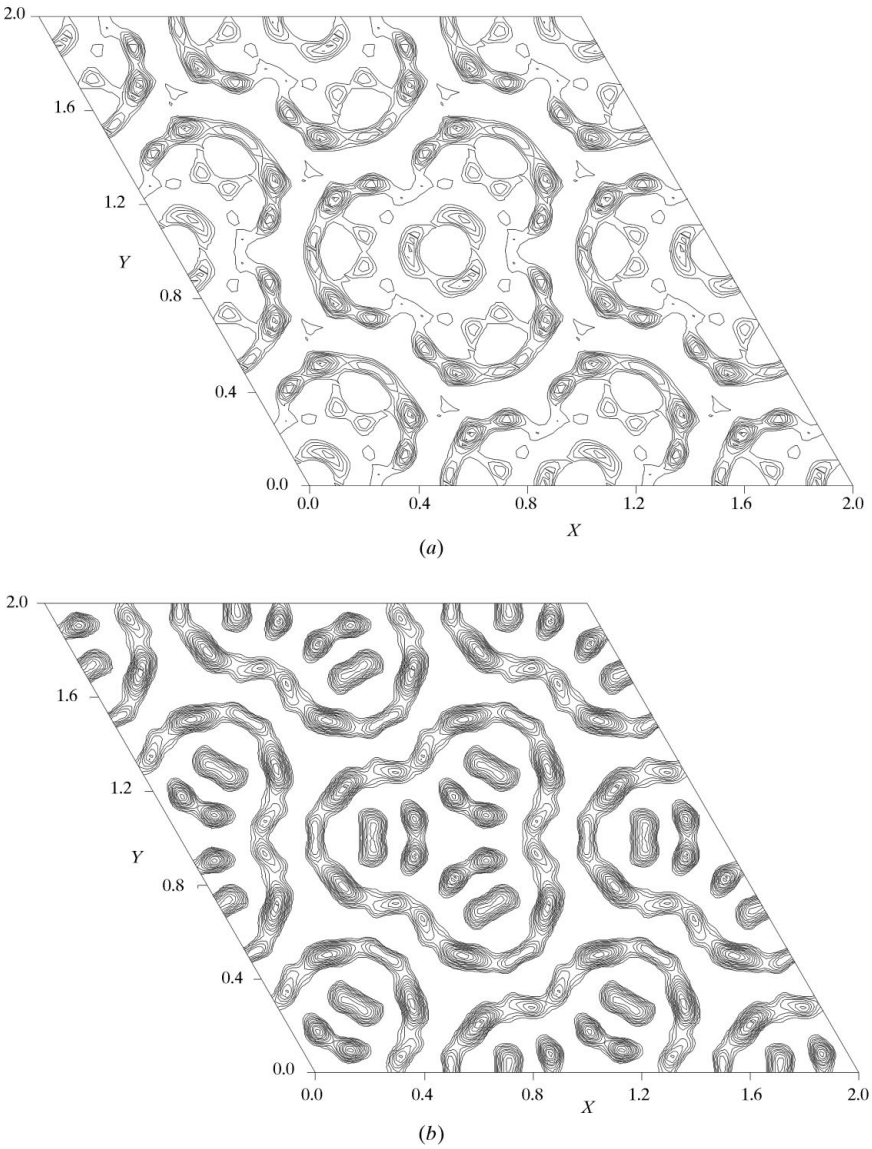


**Figure 7**
Potential maps for Omp F porin: (*a*) the true map using the image derived phases of Sass *et al.* (1989), (*b*) the best map derived from ME phasing. The mean phase error is 9° and the correlation coefficient between (*a*) and (*b*) is 0.94. (From Gilmore *et al.*, 1996.)

### 2.3. *Half-Bake*

*Half-Bake* uses the ideas of *SnB* in a different way but still requires and imposes atomicity. Instead of the minimal function, correlation coefficients (17) and a restricted coefficient

$$\sum_{E_o > E_{\min}} E_c^2 (E_o^2 - 1), \qquad (19)$$

(where $E_{\min}$ is typically 1.3–1.5) are used as indicators of phase correctness. The tangent formula is used in phase refinement. Fig. 6 shows the flow chart for this procedure.

Both *SnB* and *Half-Bake* have solved structures with $N > 1000$ and have also become valuable tools in deriving heavy-atom substructures in proteins. In this case, the resolution limit can be substantially relaxed because even at 2 Å atoms such as Se are clearly resolved and the necessary atomicity is still present.

## 3. Solving protein structures at low resolution using direct methods

For reasons that should now be clear, there is no general solution of the phase problem at low resolution, but the following direct-methods (*i.e.* model-free) techniques have been explored: (i) maximum entropy, (ii) globular scattering factors, (iii) symbolic addition and (iv) electron microscopy and electron crystallography.

Other techniques such as sphere packings (Andersson, 1999; Andersson & Hovmöller, 1996) are outside the scope of this paper and other methods are fully described elsewhere in this issue.

### 3.1. Maximum entropy

The maximum-entropy (ME) formalism was first applied to the phase problem by Bricogne (1984) and subsequently incorporated in a more general Bayesian statistical approach applied to macromolecules. For a review, see Gilmore (1986). The ME method is not constrained to the use of Wilson statistics and is stable irrespective of data resolution; it is thus better able to deal with model-free *ab initio* structure determination at low resolution. Associated with the Bricogne formalism is likelihood as a figure of merit and this is also a resolution-independent indicator of phase correctness of great power.

For an example of the ME method applied to low-resolution electron diffraction data from membrane proteins, see Gilmore *et al.* (1996). In this work, two protein structures were solved in projection: Omp F porin and halorhodopsin.

**3.1.1. Omp F porin**. The structure of Omp F porin from the outer membrane of *Escherichia coli* (MW = 36 500 Da) was originally determined using images at 3.2 Å resolution by Sass *et al.* (1989). Their data were obtained at 100 kV from glucose-embedded samples on a liquid-helium-cooled superconducting cryomicroscope. Most of the diffracted power from these images was contained within a 6 Å limit and so *ab initio* phasing was carried out to this same limit. There were 42 unique reflections; the plane group of the projection is $p31m$ with a unit cell of side $a = 72$ Å. The true map using the image-derived phases of Sass is shown in Fig. 7(*a*). The best map derived from ME phasing is shown in Fig. 7(*b*). At this resolution, the preferred map has a basis set mean absolute phase error of only 9°. With only minor details there is an essential correspondence with this map and that computed with all correct angles from the image data; the correlation coefficient is 0.94.

**3.1.2. Halorhodopsin**. Electron-diffraction amplitudes and electron-micrograph-derived crystallographic phases from halorhodopsin to 6 Å resolution were reported by Havelka *et al.* (1993) from frozen hydrated samples. The centrosymmetric tetragonal plane group is $p4gm$ with unit-cell parameter $a = 102$ Å. Within the 6 Å resolution limit, this corresponds to 76 unique reflections. The true map is shown in Fig. 8(*a*). Using the ME method, 16 reflections to 9 Å were phased with only one incorrect indication; the corresponding map is shown in Fig. 8(*b*). The correlation coefficient between these two maps is 0.82.

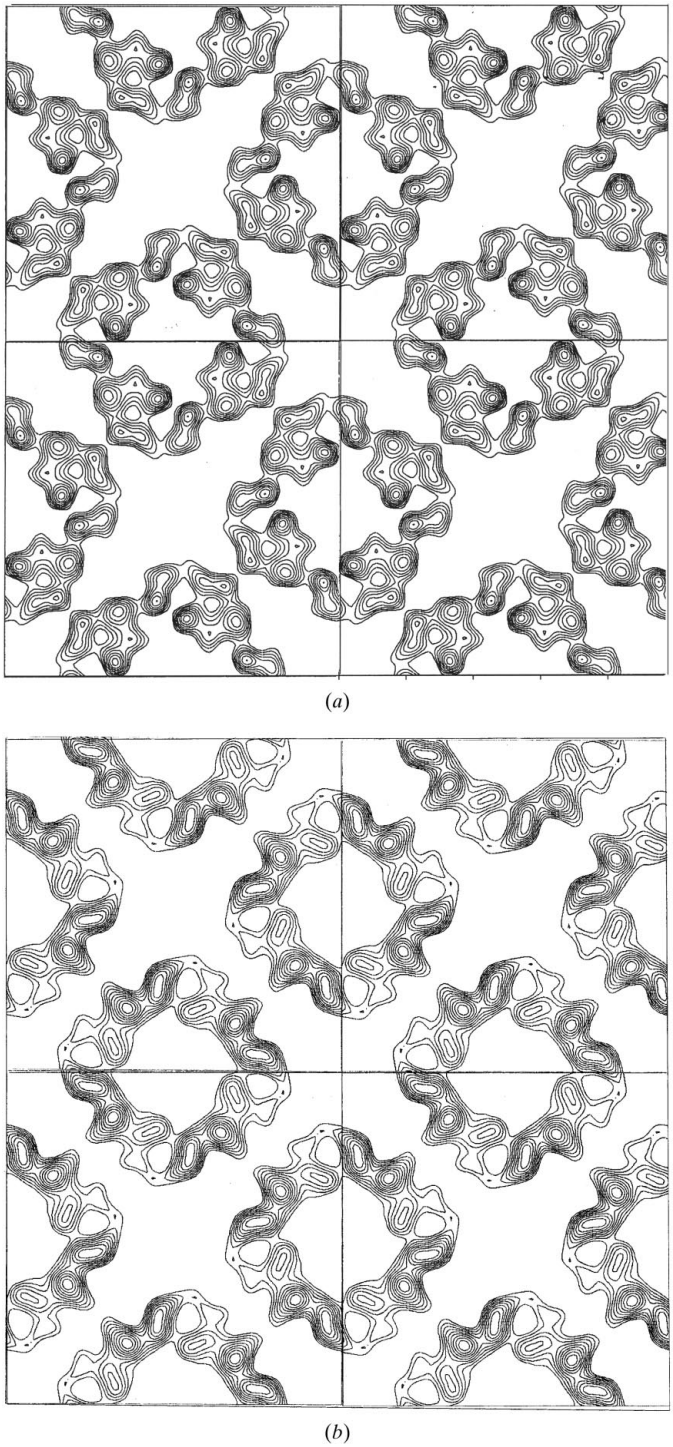

**Figure 8**
Potential maps for halorhodopsin: (*a*) the true map, (*b*) the best ME map with only one incorrect phase indication and a correlation coefficient between (*a*) of 0.82. (From Gilmore *et al.*, 1996.)

### 3.2. Globular scattering factors

Harker (1953) discussed the problem of normalizing data *via* the Wilson method when its resolution was less than atomic. Wilson statistics only hold if the resolution is less than the shortest interatomic distance in the crystal. If this is not the case, then the expression

$$\langle I \rangle_s = \sum_{j=1}^{N} f_j^2 \tag{20}$$

used by Wilson (where $s = \sin\theta/\lambda$) has to be replaced by

$$\langle I \rangle_s = \sum_{g} F_g^2, \tag{21}$$

where $F_g$ is a *globular scattering factor*. For a sphere,

$$F_g(s) = \sum_{i}^{N} f_i(s) \frac{\sin 2\pi s r_i}{2\pi s r_i} \tag{22}$$

and for $G$ globs in the unit cell,

$$F_{\mathbf{h}}^{\rm calc} = \sum_{g=1}^{G} F_g \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_g). \quad (23)$$

Clearly, this reduces a cell with $N$ atoms to one containing $G$ globs. The associated phase relationships will reflect this by showing a large increase in the concentration parameter. This idea has been used extensively by Dorset (see, for example, Dorset & McCourt, 1999) in conjunction with symbolic addition to solve a variety of structures at 10–20 Å resolution.

### 3.3. Globular structure factors and symbolic addition: beef liver catalase

As an example of this (and of the ME method) see Dorset & Gilmore (1999), which examines beef liver catalase in projection at 9 Å using room-temperature electron-diffraction data. The plane group is *pgg*, with unit-cell parameters $a = 69.7$, $b = 177$ Å. Both the ME formalism and symbolic addition coupled with the Sayre–Hughes equation were used. In addition to using likelihood, the Luzzati figure of merit $\langle \Delta\rho^4 \rangle_{\rm min}$ was also employed, where $\Delta\rho = \rho - \overline{\rho}$ (Luzzati *et al.*, 1972). Note that the minimum value of this figure of merit corresponds to maps with a minimum dynamic range and maximum flatness (rather like entropy); this seems intuitively reasonable under low-resolution conditions.

The results of a symbolic addition calculation in which the best map was selected *via* $\langle \Delta\rho^4 \rangle_{\rm min}$ are shown in Fig. 9(*a*), which has a resolution of ~9 Å. At first sight, Fig. 9(*b*), derived from ME calculations, shows no resemblance to Fig. 9(*a*), but it is a Babinet solution. Babinet solutions are those in which all the phase angles are shifted by $\pi$, *i.e.* $\varphi_{\mathbf{h}} \rightarrow \pi + \varphi_{\mathbf{h}}$, and in real space the maps are characterized as the inverse of the non-reversed one. Babinet solutions are not uncommon when phasing at low resolution in a model-free environment and care needs to be exercised. The Babinet of Fig. 9(*b*) is shown in Fig. 9(*c*) and the correspondence between this and Fig. 9(*a*) is obvious. Finally, the symbolic addition–Luzzati method is combined with the Babinet in Fig. 9(*c*) to give Fig. 9(*d*). For comparison, an image-derived solution at 23 Å using data from Akey & Edelstein (1983) is shown in Fig. 10.

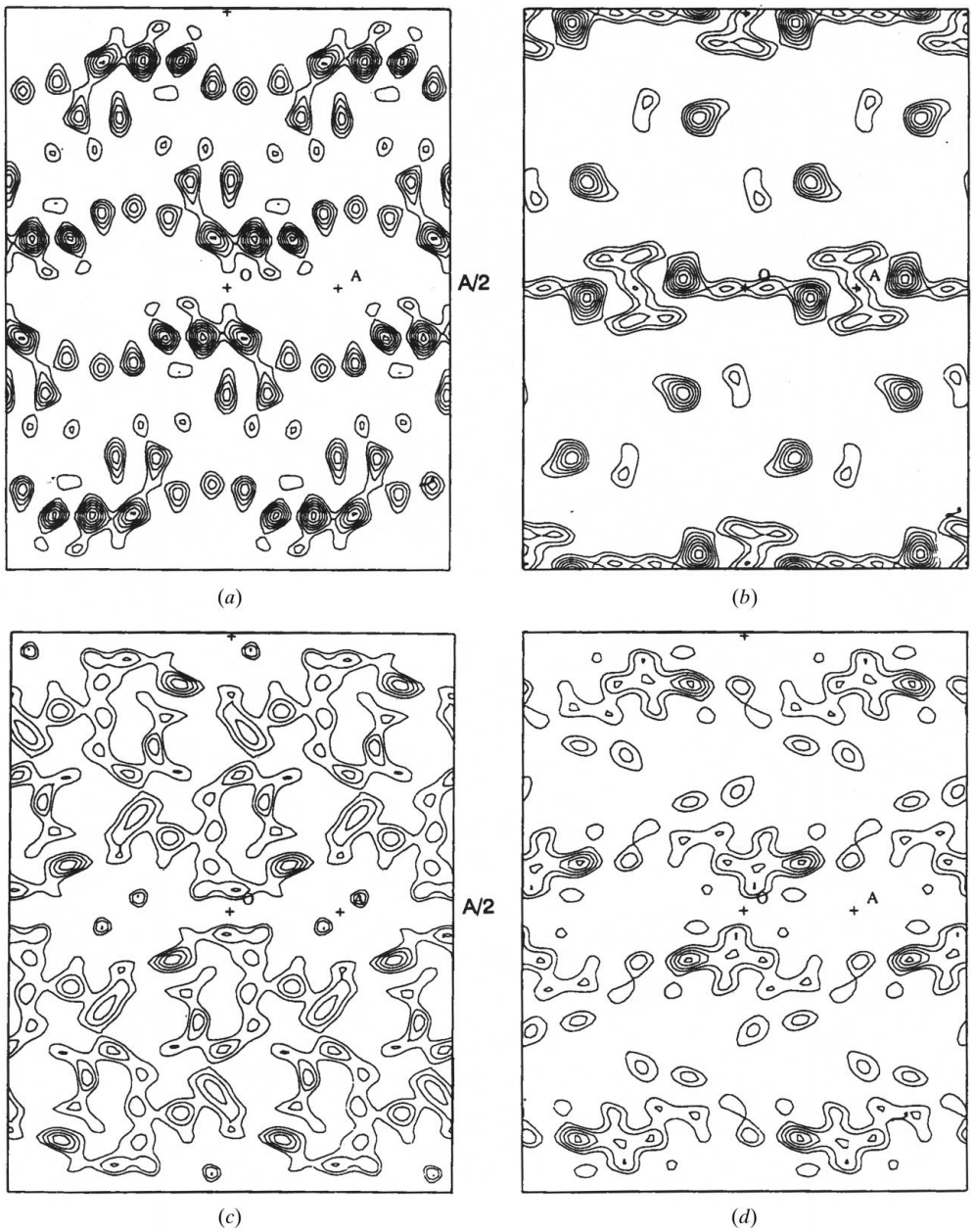


**Figure 9**
Potential maps for beef liver catalase at approximately 9 Å resolution: (*a*) using symbolic addition and selecting the map with the lowest value of $\langle \Delta\rho^4 \rangle$, (*b*) derived from maximum-entropy calculations and selecting the map with the lowest value of $\langle \Delta\rho^4 \rangle$, (*c*) the Babinet map of (*b*), (*d*) based on a subset of (*c*) using the same reflections as in (*a*). The crystallographic *b* axis is horizontal. (From Dorset & Gilmore, 1999.)

### 3.4. The electron microscope and electron crystallography

The electron microscope is an invaluable tool in low-resolution imaging of biological macro-

molecules. It is the source of two sorts of data for crystallographic purposes.

(i) Phased reflections, where the phase information comes from the Fourier transform of electron-microscope images after suitable filtering. Usually the phases so derived correspond to intensities that have a significantly lower resolution than the diffraction data and there are some significant sources of error in image data arising from radiation damage, curvilinear paracrystalline distortion and transfer-function uncertainties, but they are invaluable.

(ii) The electron diffraction data, which are less problematic and have a higher resolution than those in (i). Clearly, to achieve optimum resolution we need to extend the image-derived phases to phase the diffraction intensities and this is a problem that direct methods can address.

Two sorts of situation arise in phase extension as follows.

(i) The image phase set is sufficiently large and well distributed in reciprocal space to permit an unambiguous phase-extension procedure without recourse to multi-solution methods, *i.e.* those that involve phase permutation.

(ii) When the basis set is small or in some way inadequate we have a *branching problem*. This problem arises when phases are selected without exploring the relevant phase space in sufficient detail, so that what appears to be an unambiguous phase choice is no such thing. The methods outlined in §3 can be employed here; for a survey of electron crystallography and these problems in an ME context, see Gilmore (1996).

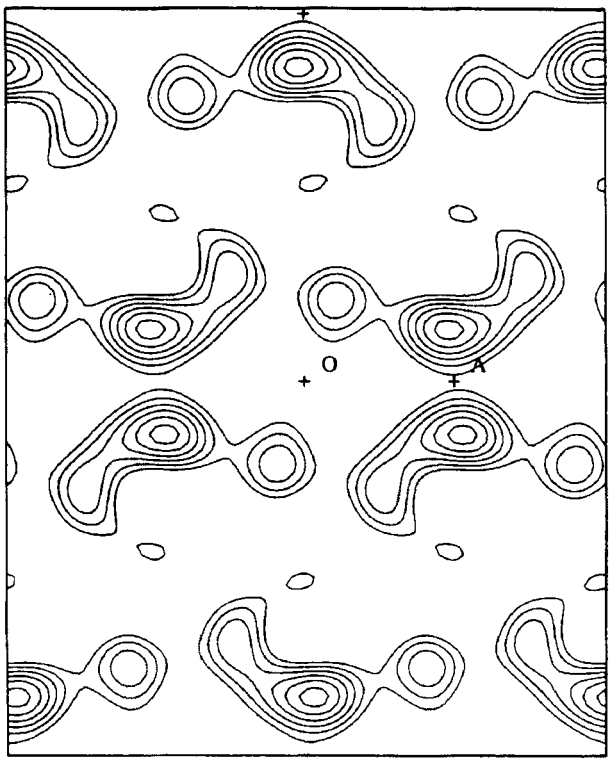


**Figure 10**
Potential map for beef liver catalase from image-derived phases at 23 Å by Akey & Edelstein (1983). (From Dorset & Gilmore, 1999.)

### 3.5. The use of very low resolution reflections

Traditional wisdom dictates that very low angle reflections in protein crystallography are of minimal value and their use can prevent a successful structure solution. This is effectively refuted by Andersson (1999) and by the results presented above where the very low order reflections played a key role. To summarize his arguments: the solvent contribution to a given reflection depends on the difference between the electron density of the solvent and that of the protein. At very low resolution, the Babinet principle means that the phases of the solvent are shifted by $\pi$ relative to the protein (see Fig. 11). The correlation coefficient is close to 100% to 15 Å resolution, but becomes effectively zero at less than 3 Å. This means that no bulk-solvent correction is needed when using only low-angle reflections. When mixing data at low and high resolution then the magnitude of the solvent vector depends on the solvent–protein contrast and we can write the total structure factor $|F_{\mathbf{h}}|^{\rm total}$ as

$$|F_{\mathbf{h}}|^{\rm total} = |F_{\mathbf{h}}|^{p} - k_s|F_{\mathbf{h}}|^{p} \exp(-B_s \sin^2\theta/\lambda^2), \qquad (24)$$

where $k_s$ measures the density ratio of solvent and protein and $B_s$ is the solvent temperature factor. Thus, properly handled, there is no reason to exclude low-order data from *ab intio* structure determination.

I wish to acknowledge invaluable and stimulating discussions with Klas Andersson and Doug Dorset and support from Eastman-Kodak (Rochester), EPSRC and BBSRC.

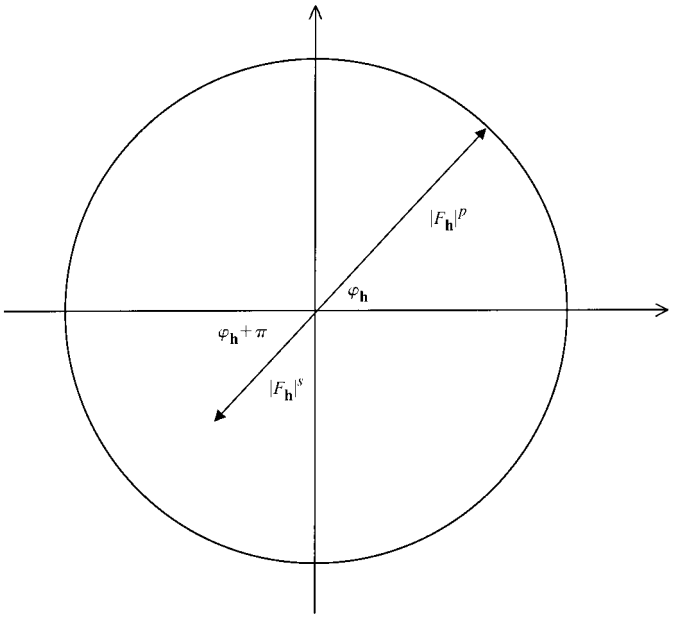


**Figure 11**
Solvent and Babinet effects at very low resolution. $|F_{\mathbf{h}}|^{p}$ is the contribution of the protein to the total structure factor $|F_{\mathbf{h}}|^{ps}$ with a phase angle $\varphi_{\mathbf{h}}$ and $|F_{\mathbf{h}}|^{s}$ is the solvent contribution with a phase, by Babinet's principle, of $\varphi_{\mathbf{h}} + \pi$. (Taken from Andersson, 1999.)

## References


Akey, C. W. & Edelstein, S. J. (1983). *J. Mol. Biol.* **163**, 575–612.
Andersson, K. M. (1999). *J. Appl. Cryst.* **32**, 530–535.
Andersson, K. M. & Hovmöller, S. (1996). *Acta Cryst.* D**52**, 1174–1180.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.
DeTitta, G. T., Langs, D. A., Edmonds, J. W. & Duax, W. L. (1975). *Acta Cryst.* A**31**, 472–479.
DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.
Dorset, D. L. & Gilmore, C. J. (1999). *Acta Cryst.* A**55**, 448–456.
Dorset, D. L. & McCourt, M. P. (1999). *Z. Kristallogr.* **214**, 652–658.
Fortier, S. (1997). Editor. *Direct Methods for Solving Macromolecular Structures.* Dortrecht: Kluwer.
Giacovazzo, C. (1998). *Direct Phasing in Crystallography. Fundamentals and Applications.* Oxford University Press.
Gilmore, C. J. (1986). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 317–321. Dortrecht: Kluwer.
Gilmore, C. J. (1996). *Acta Cryst.* A**52**, 561–589.
Gilmore, C. J., Nicholson, W. V. & Dorset, D. L. (1996). *Acta Cryst.* A**52**, 937–946.
Harker, D. (1953). *Acta Cryst.* **6**, 731–736.
Hauptman, H. A. (1975). *Acta Cryst.* A**31**, 680–687.
Havelka, W. A., Henderson, R., Heymann, J. A. W. & Oesterhelt, D. (1993). *J. Mol. Biol.* **234**, 837–846.
Hughes, E. W. (1953). *Acta Cryst.* **6**, 871.
Karle, J. & Hauptman, H. A. (1956). *Acta Cryst.* **9**, 635–651.
Karle, J. & Karle, I. L. (1966). *Acta Cryst.* **21**, 849–859.
Luzzati, V., Tardieu, A. & Taupin, D. (1972). *J. Mol. Biol.* **64**, 269–286.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Sass, H. J., Büldt, G., Beckmann, E., Zemlin, F., Van Heel, M., Zeitler, E., Rosenbusch, J. P., Dorset, D. L. & Massalski, A. (1989). *J. Mol. Biol.* **209**, 171–175.
Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.
Schenk, H. (1973). *Acta Cryst.* A**29**, 77–82.
Sheldrick, G. M. (1997). *Proceedings of the CCP4 Study Weekend. Recent Advances in Phasing*, edited by K. S. Wilson, G. Davies, A. W. Ashton & S. Bailey, pp. 147–157. Warrington: Daresbury Laboratory.
Weeks, C. M. & Miller, R. (1997). *Proceedings of the CCP4 Study Weekend. Recent Advances in Phasing*, edited by K. S. Wilson, G. Davies, A. W. Ashton & S. Bailey, pp. 139–146. Warrington: Daresbury Laboratory.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Woolfson, M. M. & Fan, H.-F. (1995). *Physical and Non-Physical Methods of Solving Crystal Structures.* Cambridge University Press.