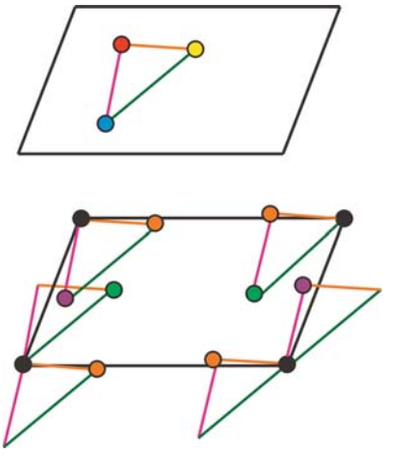

Fig. 12.2.2.2. The vector superposition method. The Patterson map of Fig. 12.2.1.1 can be regarded as the superposition of the structure (and its inverse), with each of its atoms placed alternately at the origin. By shifting each peak of the Patterson function to the origin and calculating the correlation of all remaining peaks with the unshifted map, it is possible to deconvolute the Patterson function.

*i.e.* there are three Harker sections: $u = \frac{1}{2}$, $v = \frac{1}{2}$ and $w = \frac{1}{2}$. Peaks occurring on the Harker sections must reduce to a self-consistent set of coordinates $(x, y, z)$, allowing reconstruction of the atomic positions.

If we have two isomorphous (see below) data sets $F_{PH}$ and $F_P$, then the difference in the two Patterson functions,

$$P_{PH} - P_P = \int [F_{PH}^2(\mathbf{S}) - F_P^2(\mathbf{S})] \exp\{-2\pi i \mathbf{r} \cdot \mathbf{S}\} \, \mathrm{d}^3\mathbf{S},$$

will deliver information about the heavy-atom structure. Such a difference function gives rise to non-negligible peaks arising from interference between the $F_H$ and $F_P$ terms, however (Perutz, 1956). Rossmann (1960) showed that these interference terms could be reduced through calculation of the modified Patterson function

$$P_H = \int [F_{PH}(\mathbf{S}) - F_P(\mathbf{S})]^2 \exp\{-2\pi i \mathbf{r} \cdot \mathbf{S}\} \, \mathrm{d}^3\mathbf{S}.$$

In the case of a single-site derivative, peaks should occur only at the Harker vectors corresponding to the heavy-atom position. Even so, there is a choice of positions for the heavy atom: *e.g.*, in the $P2_12_12_1$ case, coordinates $(\pm x + \xi, \pm y + \nu, \pm z + \zeta)$, where $\xi$, $\nu$ and $\zeta$ can each take the value 0 or $1/2$, will all give rise to the same Harker vectors. This in itself is not a problem, relating to equivalent choices of origin and of handedness, but has important ramifications for multisite derivatives or multiple isomorphous replacement (see below).

If there is more than one site, then there will be two sets of peaks: one set corresponding to the Harker sections (self-vector set) and one set corresponding to the difference vectors between different heavy-atom sites (the cross-vector set). In this case, the choice of one heavy-atom position $(x_{H1}, y_{H1}, z_{H1})$ determines the origin and the handedness to which all other peaks *must* correspond. Thus, in the $P2_12_12_1$ example, only one cross vector will occur for

$$(x_{h1} \pm x_{h2} + \xi, y_{h1} \pm y_{h2} + \nu, z_{h1} \pm z_{h2} + \zeta).$$

An alternative to the Harker-vector approach is Patterson-vector superposition (Sheldrick *et al.*, 1993; Richardson & Jacobson, 1987). The Patterson map contains several images of the structure that have been shifted by interatomic vectors (Fig. 12.2.2.2). If this structure is relatively simple (as is to be hoped for in a 'normal' heavy-atom derivative), then it should be possible to deconvolute the superimposed structures by vector shifts (Buerger, 1959).

### 12.2.3. The difference Fourier

Once the heavy-atom positions have been found, they can be used to calculate approximate phases and Fourier maps. Ideally, difference Fourier maps calculated with phases from a single site should reveal the other positions determined from the Harker search procedure. This ensures that all heavy-atom positions correspond to a single origin and hand. Similarly, phases calculated from derivative *H*1 should reveal the heavy-atom structure for derivative *H*2. Merging and refinement of all phase information will result in a phase set that can be used to solve the structure.

### 12.2.4. Reality

#### 12.2.4.1. *Treatment of errors*

Until now, we have dealt with cases involving perfect data. Although this ideal may now be attainable using MAD techniques, this is not necessarily the usual laboratory situation. In the first place, it is necessary to scale the derivative data $F_{PH}$ to the native $F_P$. One of the most common scaling procedures is based on the expected statistical dependence of intensity on resolution (Wilson, 1949). This may not be particularly accurate when only low-resolution data are available, in which case a scaling through equating the Patterson origin peaks of native and derivative sets may provide better results (Rogers, 1965).

A model to account for errors in the data, determination of heavy-atom positions *etc.* was proposed by Blow & Crick (1959), in which all errors are associated with $|F_{PH}|_{\mathrm{obs}}$ (Fig. 12.2.4.1); a more detailed treatment has been provided by Terwilliger & Eisenberg

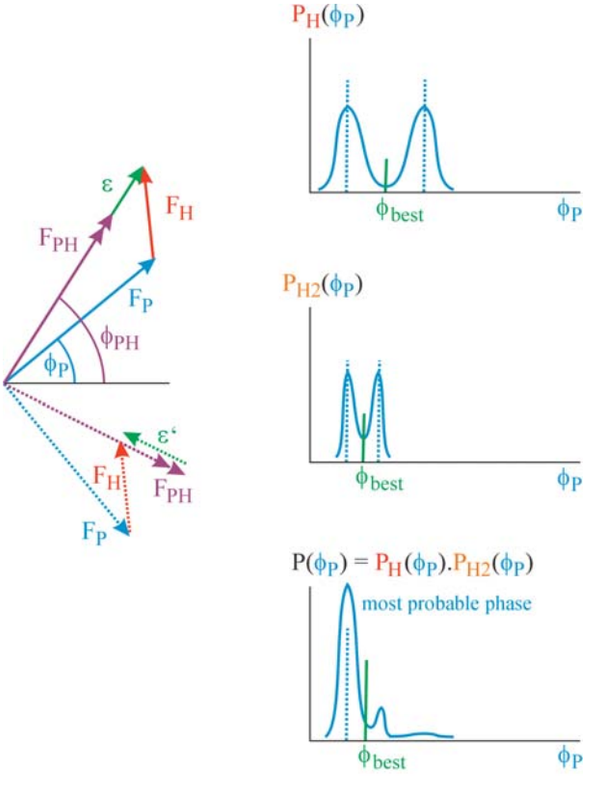


Fig. 12.2.4.1. The treatment of phase errors. The calculated heavy-atom structure results in a calculated value for both the phase and magnitude of $F_H$ (red). According to the value of $\varphi_P$, the triangle $F_P$–$F_H$–$F_{PH}$ will fail to close by an amount $\varepsilon$, the lack of closure (green). This gives rise to a phase distribution which is bimodal for a single derivative. The combined probability from a series of derivatives has a most probable phase (the maximum) and a best phase (the centroid of the distribution), for which the overall phase error is minimum.



Copyright © 2006 International Union of Crystallography

(1987). Owing to errors, the triangle formed by $F_P$, $F_{PH}$ and $F_H$ fails to close. The *lack of closure error* $\varepsilon$ is a function of the calculated phase angle $\varphi_P$:

$$\varepsilon(\varphi_P) = |F_{PH}|_{\rm obs} - |F_{PH}|_{\rm calc}.$$

Once an initial set of heavy-atom positions has been found, it is necessary to refine their parameters ($x$, $y$, $z$, occupancy and thermal parameters). This can be achieved through the minimization of

$$\sum_{\mathbf{S}} \varepsilon^2/E,$$

where $E$ is the estimated error ($\simeq \langle(|F_{PH}|_{\rm obs} - |F_{PH}|_{\rm calc})^2\rangle$) (Rossmann, 1960; Terwilliger & Eisenberg, 1983). This procedure is safest for noncentrosymmetric reflections ($\varphi$ restricted to 0 or $\pi$) if enough are present. Phase refinement is generally monitored by three factors:

$$R_{\rm Cullis} = \sum\||F_{PH} + F_P| - |F_H|_{\rm calc}|/\sum|F_{PH} - F_P|$$

for noncentrosymmetric reflections only; acceptable values are between 0.4 and 0.6;

$$R_{\rm Kraut} = \sum\||F_{PH}|_{\rm obs} - |F_{PH}|_{\rm calc}|/\sum|F_{PH}|_{\rm obs},$$

which is useful for monitoring convergence; and the

$$\text{phasing power} = \sum|F_H|_{\rm calc}/\sum\||F_{PH}|_{\rm obs} - |F_{PH}|_{\rm calc}|,$$

which should be greater than 1 (if less than 1, then the phase triangle cannot be closed *via* $F_H$).

The resulting phase probability is given by

$$P(\varphi_P) = \exp\{-\varepsilon^2(\varphi_P)/2E^2\}.$$

The phases have a minimum error when the *best phase* $\varphi_{\rm best}$, *i.e.* the centroid of the phase distribution,

$$\varphi_{\rm best} = \int \varphi_P P(\varphi_P)\ {\rm d}\varphi_P,$$

is used instead of the most probable phase. The quality of the phases is indicated by the *figure of merit m*, where

$$m = \int P(\varphi_P)\exp(i\varphi_P)\ {\rm d}\varphi_P \Big/ \int P(\varphi_P)\ {\rm d}\varphi_P.$$

A value of 1 for $m$ indicates no phase error, a value of 0.5 represents a phase error of about 60°, while a value of 0 means that all phases are equally probable.

The *best Fourier* is calculated from

$$\rho_{\rm best}(\mathbf{r}) = (1/V)\sum_{\mathbf{S}} m|F_P(\mathbf{S})|\exp\{i\varphi_{P\rm best}(\mathbf{S})\},$$

where the electron density should have minimal errors.

#### 12.2.4.2. *Automated search procedures*

If the derivative shows a high degree of substitution, then the Harker sections become more difficult to interpret. Furthermore, Terwilliger *et al.* (1987) have shown that the intrinsic noise in the difference Patterson map increases with increasing heavy-atom substitution. It is at this stage that automated procedures are invaluable.

One such automated procedure is implemented in *PROTEIN* (Steigemann, 1991). The unit cell is scanned for possible heavy-atom sites; for each search point ($x$, $y$, $z$), all possible Harker vectors are calculated, and the difference-Patterson-map values at these points are summed or multiplied. As the origin peak dominates the Patterson function, this region is set to zero. The resulting correlation map should contain peaks at all possible heavy-atom positions. The peak list can then be used to find a set of consistent heavy-atom locations through a subsequent search for difference vectors (cross vectors) between putative sites. It should be possible to locate all major and minor heavy-atom sites through repetition of this procedure. A similar strategy is adopted in the program *HEAVY* (Terwilliger *et al.*, 1987), but sets of heavy-atom sites are ranked according to the probability that the peaks are not random. The program *SOLVE* (Terwilliger & Berendzen, 1999) takes this process a stage further, where potential heavy-atom structures are solved and refined to generate an (interpretable) electron density in an automated fashion.

The search method can also be applied in reciprocal space, where the Fourier transform of the trial heavy-atom structure is calculated, and the resulting $F_{H\rm calc}$ is compared to the measured differences between derivative and native structure-factor amplitudes (Rossmann *et al.*, 1986). In the programme *XtalView* (McRee, 1998), the correlation coefficient between $|F_H|$ and $|F_{PH} - F_P|$ is calculated, whilst a correlation between $F_H^2$ and $(F_{PH} - F_P)^2$ is used by Badger & Athay (1998). Dumas (1994*b*) calculates the correlation between $|F_{H\rm calc}|^2$ and $|F_{H\rm estimated}|^2$, based on the estimated lack of isomorphism.

Vagin & Teplyakov (1998) have reported a heavy-atom search based on a reciprocal-space translation function. In this case, low-resolution peaks are not removed but weighted down using a Gaussian function. Potential solutions are ranked not only according to their translation-function height, but also through their phasing power, which appears to be a stronger selection criterion.

All these searches are based upon the sequential identification of heavy-atom sites and their incorporation in a heavy-atom partial structure. Problems arise when bogus sites influence the search for further heavy-atom positions. In an attempt to overcome this problem, the heavy-atom search has been reprogrammed using a genetic algorithm, with the Patterson minimum function as a selection criterion (Chang & Lewis, 1994). This approach has the potential to reveal all heavy-atom positions in one calculation, and tests on model data have shown it to be faster than traditional sequential searches.

### 12.2.5. Special complications

#### 12.2.5.1. *Lack of isomorphism*

This problem is by far the most common in protein crystallography. An isomorphous derivative is one in which the crystalline arrangement has not been disturbed by derivatization. An early study of Crick & Magdoff (1956) proposed a rule of thumb that a change in any of the cell dimensions by more than around 5% would result in a lack of isomorphism that would defeat any attempt to locate the heavy-atom positions or extract useful phase information. Lack of isomorphism can, however, be more subtle; sometimes a natural variation in the native crystal form may occur, resulting in poor merging statistics of data obtained from different crystals. Coupling this variation with commonly observed structural changes upon heavy-atom binding can provide a considerable barrier to obtaining satisfactory phases. Dumas (1994*a*) has provided a theoretical consideration of this problem.

One practical approach is to collect native and derivative data sets from the same crystal, a technique that has been successful in the structure determination of cyclohydrolase (Nar *et al.*, 1995), proteosome (Löwe *et al.*, 1995) and a number of other proteins. Nonisomorphism can be used, however. In the structure solution of carbamoyl sarcosine hydrolase (Romao *et al.*, 1992), derivatives fell into two (related) crystalline classes. By judicious use of two 'native' crystal forms, heavy-atom positions could be obtained in each of the two classes. Phasing and resultant averaging between the two classes provided an interpretable electron density. In the case of ascorbate oxidase (Messerschmidt *et al.*, 1989), multiple isomorphous replacement failed to provide an interpretable density. It was possible, however, to place the initial density into a second



references