electronic reprint

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

A highly automated heavy-atom search procedure for macromolecular structures

Ralf W. Grosse-Kunstleve and Axel T. Brunger

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

A highly automated heavy-atom search procedure for macromolecular structures

Ralf W. Grosse-Kunstleve and Axel T. Brunger*

Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520-8114, USA

Correspondence e-mail: brunger@laplace.csb.yale.edu

Received 12 April 1999
Accepted 4 June 1999

A new highly automated heavy-atom search procedure combines a fast Fourier transform translation function, Patterson superposition functions and Patterson correlation refinement. The search procedure can be applied to various native and difference Patterson maps and their statistically weighted averages. The procedure was tested with diffraction data for several crystal structures with up to 30 heavy-atom sites in the asymmetric unit and with minimum Bragg spacings ranging from 3 to 4 Å. In all cases, the correct sites were found with modest computing time.

1. Introduction

Experimental phasing techniques are in most cases essential for the solution of new macromolecular crystal structures. For a single or multiple isomorphous replacement (SIR or MIR) experiment, suitable heavy-atom derivatives have to be obtained by trial-and-error. In contrast, single or multiple anomalous-dispersion (SAD or MAD) experiments with selenomethionyl (SeMet) proteins (Hendrickson, 1991) provide a relatively straightforward solution to the phase problem for protein crystal structures. This method has become increasingly popular, owing to both significant advances in recombinant protein expression and the ever-increasing availability of tunable synchrotron beamlines.

The most difficult step in all heavy-atom techniques is the determination of the heavy-atom sites from the experimental diffraction data. If only a few (two or three) heavy-atom sites are to be determined, the sites can often be found by manual interpretation of the Harker sections in the Patterson function. However, if there are more than a few sites, the manual interpretation becomes very tedious and often fails. This is a consequence of the quadratic growth of the number of peaks in the Patterson map as a function of the number of sites. Furthermore, thermal motion and disorder of the sites is accentuated in Patterson maps, as the width of the distribution around Patterson peaks is twice that of the corresponding electron-density peaks. The larger the number of Patterson peaks, the more these peaks will overlap, thereby hampering the interpretation. Another possible cause of incorrect interpretation of a heavy-atom Patterson map is the intrinsic noise resulting from the approximation of the heavy-atom Patterson map by using the amplitude differences of observed structure factors (Terwilliger *et al.*, 1987). The noise level is roughly proportional to the number of heavy-atom sites.

Heavy-atom site-determination procedures can be classified into three groups.

(i) Direct methods of phase determination (Sheldrick *et al.*, 1993; DeTitta *et al.*, 1994; Weeks *et al.*, 1994; Weeks & Miller, 1999).

(ii) Patterson methods working in direct space (Sheldrick, 1997; Sheldrick *et al.*, 1993; Terwilliger *et al.*, 1987; Terwilliger, 1998) also using non-crystallographic symmetry (Tong & Rossmann, 1993) and using a combination of Patterson superposition methods with a genetic algorithm (Chang & Lewis, 1994).

(iii) Reciprocal-space Patterson searches (McRee, 1993; Dumas, 1994; Vagin & Teplyakov, 1998).

The direct-space Patterson search methods can be viewed as 'reversing' a convolution process. In contrast, the reciprocal-space Patterson searches employ a direct 'forward' computation, where structure factors are computed from a trial set of heavy-atom sites and compared with the observed structure-factor differences or MAD F_A structure factors (Karle, 1989; Hendrickson, 1991; Terwilliger, 1994).

The method presented in this paper combines direct-space and reciprocal-space Patterson searches and Patterson correlation (PC) refinement (Brunger, 1991). We show that the reciprocal-space search is intrinsically more accurate than the direct-space search, especially for a large number of heavy-atom sites. However, prior to the development of the fast translation function (FTF; Navaza & Vernoslova, 1995), the reciprocal-space search was not practical owing to the high computational cost associated with more conventional translation functions (Fujinaga & Read, 1987; Lipson & Cochran, 1957). Therefore, one of the key features of our heavy-atom search method is an enhanced implementation of the FTF, which is 300 to 500 times faster than conventional translation functions. The FTF was added as Fortran source code to the *Crystallography & NMR System* (*CNS*; Brunger *et al.*, 1998). The FTF is invoked from the high-level *CNS* language which provides the basis for the implementation of the heavy-atom search procedure.

2. Overview of the heavy-atom search procedure

The heavy-atom search procedure consists of four stages which are described in more detail below and in Fig. 1. In the first stage, the observed diffraction intensities are filtered by various cutoff criteria and then used to compute native structure factors (F) or difference structure factors (ΔF). If MIR, MAD or MIRAS data are available, two or more sets of difference structure factors can be averaged and optionally weighted by empirical diffraction ratios. The second stage consists of a Patterson search by either a single-atom translation function or a symmetry-minimum function, or a combination of both. A given number (typically 100) of highest peaks in the resulting Patterson search map are sorted and subsequently used as initial trial sites. The third stage consists of a sequence of alternating reciprocal-space or direct-space Patterson searches and PC refinements starting with each of the initial trial sites. This stage produces a large a number of potential solutions. The final stage consists of sorting these solutions ranked by the value of the target function of the PC refinement. If the correct solution is found, it is normally characterized by the highest value of the target function and a significant separation from incorrect solutions.

3. Patterson map ($F_{\rm patt}$) computation

The suppression of aberrant reflections can be essential for the success of a heavy-atom search (Sheldrick, 1997). In our procedure, an amplitude-based σ cutoff is applied to all diffraction data sets, *i.e.* reflections with amplitudes

$$|F| < f_{\rm cut}\sigma_F \qquad (1)$$

are rejected, where σ_F is the estimated standard deviation of the observed structure-factor amplitude and the factor $f_{\rm cut}$ is usually set to a value between 1 and 5.

For isomorphous replacement data, the ΔF values and the corresponding $\sigma_{\Delta F}$ are defined as

$$\Delta F_H = |F_{H,1}| - |F_{H,2}|,$$
$$\sigma_{\Delta F_H} = (\sigma^2_{F_{H,1}} + \sigma^2_{F_{H,2}})^{1/2}, \qquad (2)$$

where $F_{H,1}$ and $F_{H,2}$ refer to the structure factors of the native and derivative crystal, respectively. For anomalous dispersion data, the same formula is applied to compute dispersive differences between measurements at two wavelengths. The corresponding formulas for anomalous differences of measurements made at or close to an absorption edge are given by

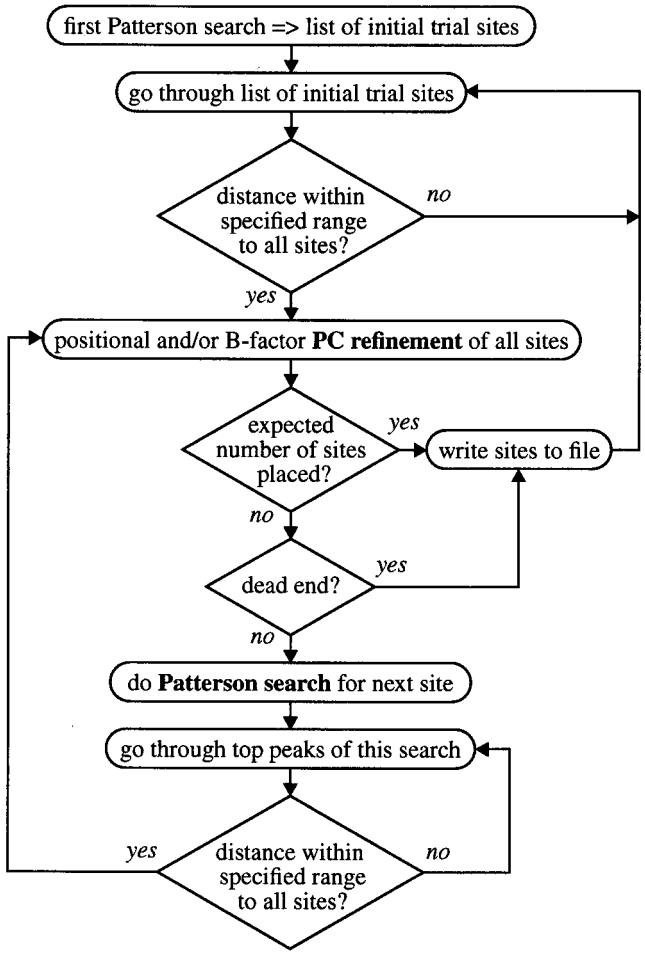

Figure 1
Flow chart of the heavy-atom search algorithm.

$$\Delta F_H = |F_{+H}| - |F_{-H}|,$$
$$\sigma_{\Delta F_H} = (\sigma^2_{F_{+H}} + \sigma^2_{F_{-H}})^{1/2}. \quad (3)$$

Our procedure also uses a σ cutoff for the ΔF differences (Hendrickson, 1988), *i.e.* all reflections with

$$|\Delta F| < d_{\rm cut}\sigma_{\Delta F} \quad (4)$$

are rejected. The factor $d_{\rm cut}$ is usually set to a value between 0.5 and 1. Finally, an outlier cutoff is used to reject all reflections with

$$|\Delta F| > c_{\rm rms}{\rm rms}(\Delta F) \quad (5)$$

(Hendrickson, 1988). The factor $c_{\rm rms}$ is usually set to a value between 3.5 and 5, and rms(ΔF) is defined as

$${\rm rms}(\Delta F) = \left[\left(\sum_H \Delta F_H^2\right)/N_{\rm Ref}\right]^{1/2}, \quad (6)$$

where the sum extends over all $N_{\rm Ref}$ reflections which satisfied all previous filtering steps.

Overall k scaling and B scaling is used to compensate for systematic errors caused by differences between crystals and data-collection conditions. The scaling and the temperature factors are obtained by least-squares minimization of

$$\sum_H [|F_{H,{\rm scaled}}| - |F_H| k \exp(-\alpha)]^2. \quad (7)$$

For isotropic scaling,

$$\alpha = B/4d_H^2, \quad (8)$$

where d_H is the d spacing corresponding to the reflection H, and for anisotropic scaling

$$\alpha = 0.25(B_{11}h^2a^{*2} + B_{22}k^2b^{*2} + B_{33}l^2c^{*2} + 2B_{12}hka^*b^* + 2B_{13}hla^*c^* + 2B_{23}klb^*c^*), \quad (9)$$

where h, k, l are the reciprocal-space indices, a^*, b^*, c^* are the lengths of the reciprocal-cell edges and B_{ij} are the six temperature factors to be refined. The selection criteria and scaling are iteratively performed until convergence is achieved, since scaling affects the selection of the cutoff criteria and *vice versa*.

If only one Patterson map is used, structure factors for the heavy-atom search ($F_{\rm patt}$) are simply assigned by $F_{\rm patt} = \Delta F$. Otherwise, $F_{\rm patt}$ is taken as the statistically weighted sum of the ΔF values for each reflection, *i.e.* the Patterson maps are effectively averaged. For reflections with some of the ΔF values missing (unobserved or removed owing to the cutoff criteria) the corresponding $F_{\rm patt}$ are set to zero. The statistical weighting is empirically carried out by grouping the reflections into resolution shells. For each resolution shell, the diffraction ratio

$$\Delta F_{\rm weighted} = [\Delta F^2_{\rm old}/\langle(n/\varepsilon)\Delta F^2_{\rm old}\rangle]^{1/2} \quad (10)$$

is computed, where $n = 2$ for acentric reflections and $n = 1$ for centric reflections. The factor ε corrects for the difference in expected intensity for different reciprocal-lattice zones (Read, 1986; Steward & Karle, 1976). The statistically weighted average of a dispersive and an anomalous Patterson map produces a better approximation of the 'true' heavy-atom Patterson map (Drenth, 1994).

It is also possible to use MAD F_A structure factors (Karle, 1989; Hendrickson, 1991; Terwilliger, 1994). Of course, in this case no differences have to be computed, but similar outlier-cutoff, averaging and weighting procedures can be applied.

Table 1
CPU times for a conventional translation function (CTF) and the fast translation function (FTF) for several test cases with different unit-cell sizes and symmetries.

Space group	Unit-cell dimensions (Å)	$d_{\rm min}$ (Å)	Time CTF (s)	Time FTF (s)	Factor
$P2_12_12_1$	a = 65.5, b = 72.2, c = 45.0	4	245	0.8	306
$C222_1$	a = 42.1, b = 97.1, c = 91.9	3	1700	8	210
$C222_1$	a = 64.1, b = 102.0, c = 187.0	4	3000	13	230
$C222$	a = 91.9, b = 168.0, c = 137.8	4.5	7850	17	460
$P4_332$	a = 272.8	6	1129644	2400	470

4. Determination of initial trial sites

4.1. Reciprocal-space method: single-atom fast translation function

For the determination of initial trial sites, a single heavy-atom site is translated throughout an asymmetric unit and the standard linear correlation coefficient of $F^2_{\rm patt}$ and $F^2_{\rm calc}(t)$ (referred to as F2F2) is computed for each position t. Other target expressions can be used, including the correlation coefficients between $F_{\rm patt}$ and $F_{\rm calc}(t)$ (referred to as F1F1), $E^2_{\rm patt}$ and $E^2_{\rm calc}$ (referred to as E2E2), where the E are normalized structure factors, and $E_{\rm patt}$ and $E_{\rm calc}(t)$ (referred to as E1E1). The F2F2 target function is preferred because it allows one to use a new and enhanced implementation of the fast translation function (FT; Navaza & Vernoslova, 1995; see *Appendix*). Table 1 shows a comparison of the computing times needed for a conventional translation function (Fujinaga & Read, 1987; Lipson & Cochran, 1957) and the FTF for several test cases with different unit-cell sizes and symmetries. The FTF is between 300 and 500 times faster than the conventional translation function. Thus, the FTF makes the automated reciprocal-space heavy-atom search procedure practical even for large numbers of heavy-atom sites.

4.2. Direct space-method: symmetry minimum function

The symmetry-minimum function (SMF; Simpson *et al.*, 1965; Pavelcik, 1986; Estermann, 1995) makes maximal use of the information contained in the Harker regions. The computation of an SMF requires a Patterson map and a table of the unique Harker vectors and their weights (Simpson *et al.*, 1965).

Theoretically, the structure factor F_{000} has to be included in the calculation of the Patterson map for the SMF. Unfortunately, F_{000} is often very difficult to estimate, especially for

difference Patterson maps. To test the effect of omitting F_{000}, SMFs were computed with and without including F_{000}, using simulated data. The results (not shown) indicated that the F_{000} term has no influence on the outcome of a SMF. This is a consequence of the relatively large number of reflections for a typical macromolecular crystal structure. We therefore conclude that F_{000} can be neglected for macromolecular problems.

4.3. Combination of FTF and SMF

Both the FTF and the SMF can be viewed as likelihood maps showing the likelihood that a particular grid point is close to a heavy-atom site. Ideally, both maps would be identical. However, owing to noise, systematic errors resulting from the use of difference data and the different methods of generating the likelihood maps, the correlation between both maps is typically only in the range 0.4–0.7 (data not shown). In order to obtain a likelihood map with a better signal-to-noise ratio, both maps are combined. The SMF is normalized by linear scaling such that the minimum value is zero and the maximum value is one. The FTF map is then multiplied with this normalized SMF map.

4.4. Peak search and special position check

In general, the symmetry of both the FTF and the SMF, and of their combination, is that of the subgroup $\mathbf{L}(\mathbf{G})$ of the Euclidean normalizer of the crystal space group $\mathbf{G}$ (Koch & Fischer, 1983). The list of initial trial sites is determined by a peak search in an asymmetric unit of $\mathbf{L}(\mathbf{G})$ of the likelihood map. This means redundancies arising from the $\mathbf{L}(\mathbf{G})$ symmetry are removed. A grid point is considered to be a peak if the corresponding density in the map is at least as high as that of its six nearest neighbors.

By default, sites at or close to a special position are rejected. For each of the initial trial sites, the shortest distance to all its symmetry mates is computed. If this distance is less than a given cutoff distance, typically 3.5 Å, the site is rejected. Optionally, sites on special positions can be used. Once an initial trial site is accepted, PC refinement of coordinates or B factors is carried out. The PC refinement target is either the F2F2 or the E2E2 correlation coefficient. Note that this refinement target can be different from that used in the translation search.

4.5. Special treatment of space group *P*1

The procedure for the determination of the initial trial sites as explained above is valid for all space groups except for space group $P1$. In $P1$, the first atom is arbitrarily fixed at the origin and the initial trial sites are determined by a Patterson search for the second site, as explained in §5.

5. Determination of additional sites

For each of the refined initial trial sites, additional sites are determined by alternating Patterson searches and PC refinements (Fig. 1). The peaks obtained from the Patterson search are sorted by peak height. The highest peak is selected which has distances to its symmetry mates and all pre-existing sites larger than the given cutoff distance. Once a new site is found, positional and B-factor PC refinement is carried out with the two-site structure. If only two sites are to be found, the refined coordinates and B factors are written to a file. The algorithm then continues in a similar fashion with the next entry from the list of initial trial sites. If two or more sites were already placed, a dead-end elimination test is performed. The correlation coefficient before placing and refining the last new site is compared with the correlation coefficient after the addition of the new site. If the target value did not increase by a certain amount, typically 0.01, the search for that particular initial trial site is deemed to have reached a dead end and no additional sites are placed. Otherwise, another Patterson search is carried out until the expected number of sites are found.

5.1. Patterson search for additional sites

5.1.1. Reciprocal-space method. The reciprocal-space search for an additional site is similar to the search for the initial trial sites discussed before, except that the previously placed sites are kept fixed and are included in the structure-factor ($F_{\rm calc}$) calculation (Navaza & Vernoslova, 1995).

5.1.2. Direct-space method. Additional sites can also be found by an image-seeking minimum function (IMF; Simpson *et al.*, 1965; Estermann, 1995). Computing an IMF map is equivalent to a deconvolution of the Patterson map using knowledge of the already placed heavy-atom sites. The symmetry of an IMF map is identical to that of the corresponding FTF map. The redundancies arising from this symmetry are removed as before.

In theory, an IMF map based on one known site would show the superimposed images of the correct site configuration and its inverse image. An IMF map based on two or more sites would show the image of the correct configuration. However, because of coincidental overlap of peaks in the Patterson map, thermal motion of the sites and noise in the data, the IMF maps often do not provide much information for macromolecular crystal structures.

5.1.3. Combination of FTF and IMF. Similar to the FTF and SMF, the IMF can be viewed as a likelihood map which provides the likelihood of each grid point being close to an additional heavy-atom site. Ideally, the FTF and IMF maps would be identical. However, the correlation coefficient between both maps is typically only in the range 0.1–0.4 and decreases with the number of existing sites (data not shown). This is not only because of noise, systematic errors resulting from the use of difference data and the different methods of generating the likelihood maps, but also reflects a shortcoming of the IMF compared with the FTF. The IMF algorithm assumes that all Patterson peaks are resolved. Peaks which coincidentally overlap are incorrectly weighted. Therefore, a combination of FTF and IMF is only useful for searches for the first few additional sites, when Patterson peaks of previously placed sites are not likely to be overlapped.

Table 2
Summary of test cases.

PDB code	Space group	Unit-cell dimensions (Å)	Data type, N_{maps}†	d_{min} (Å)	N_{total}‡	N_{search}§	Reference
1ytt	$P2_12_12_1$	$a = 65.5, b = 72.2, c = 45.0$	Yb MAD, 1 ΔF	4	4	4	Burling *et al.* (1996)
1kwa	$C222_1$	$a = 42.1, b = 97.1, c = 91.9$	Se MAD, 2 ΔF	4	12	8	Daniels *et al.* (1998)
			Se MAD, 2 ΔF	3			
3bct	$C222_1$	$a = 64.1, b = 102.0, c = 187.0$	Se MAD, 1 ΔF	4	15	10	Huber *et al.* (1997)
			Se MAD, 1 F_A	4			
1ecf	$C222_1$	$a = 117.1, b = 157.5, c = 106.7$	Se MAD, 1 ΔF	4	21	14	Muchmore *et al.* (1998)
1auv	$P322_1$	$a = 76.4, b = 180.9$	Se MAD, 1 ΔF	3	22	15	Esser *et al.* (1998)
1a7a	$C222$	$a = 91.9, b = 168.0, c = 137.8$	Se MAD, 1 ΔF	4	30	20	Turner *et al.* (1998)

† Number of maps (wavelengths) used. ‡ Total number of heavy-atom sites. § Number of sites searched for.

The decreasing accuracy of the IMF is empirically taken into account by a weighting factor W_{IMF}. In the search for the first additional site, the IMF is normalized by linear scaling and shifting such that the minimum value is zero and the maximum value is one, and each grid point of the IMF is then used as a multiplicative weight for the FTF map. In the search for the second additional site, the IMF is scaled and shifted such that the minimum value is $1/W_{IMF}$, for the third site it is $2/W_{IMF}$ and so on. The maximum value is always one. If the minimum value is close to one, the IMF is no longer used.

The exact description of the symmetry of both the FTF and the IMF, and of their combination, can be quite complex as it depends on the inherent symmetry of the existing sites. However, the most important redundancies arising from symmetry are easy to remove. For example, in the search for a second site, redundancies owing to a center of inversion (space group $P1$) or a mirror plane (for example in $P2_1$ or $P6$) are removed. More involved redundancies are ignored, as their treatment would require complex algorithms and the effect on the efficiency of the search algorithm is in general insignificant.

5.2. Generalized dead-end elimination

Experience shows that for structures with many heavy-atom sites (>20) the correlation coefficient sometimes does not significantly increase when adding a correct new site. Therefore, the parameter N_{dead} was introduced. It specifies the total number of times a decrease or no change in the correlation coefficient is tolerated upon addition of a new site. The main drawback of increasing N_{dead} is that the total run time for the heavy-atom search increases as more time is spent on incorrect solutions.

6. Test results

The six test cases listed in Table 2 will be referred to by their corresponding PDB accession codes. For two of the test structures, 1kwa and 3bct, two different test series were calculated. For 1kwa, two minimum Bragg spacings were used: d_{min} = 3 and 4 Å. These test series will be referred to by 1kwa/3 Å and 1kwa/4 Å, respectively. For 3bct, one test series was calculated with ΔF data and a second one with F_A data. These test series will be referred to as 3bct/ΔF and 3bct/F_A, respectively.

For all MAD data sets except 1kwa, it was possible to use only the diffraction data for the wavelength with the largest anomalous signal to locate the heavy-atom sites. The 1kwa MAD data were affected by oxidation of the Se atoms (Daniels *et al.*, 1998), which severely perturbed the anomalous signal. When diffraction data collected at a single wavelength were used, only partially correct solutions were obtained. In this case, a significant improvement in the signal-to-noise ratio was obtained by averaging the anomalous difference Patterson maps computed with the anomalous-peak and the high-energy remote diffraction data.

Optimal parameters (Table 3) for the heavy-atom search method were found empirically by using the test cases. A resolution range of d_{min} = 15–4 Å appears to be a reasonable default. Using lower resolution diffraction data did not improve the search in any of the test cases and caused failure of the 1a7a test case. For 1auv, no solution was found with a high-resolution cutoff of 4 Å, but a cutoff of 3 Å was successful. For 1kwa, only a partially correct solution was found at 4 Å resolution, but more correct sites were found when using higher resolution diffraction data.

The exact number of ordered heavy-atom sites is often unknown. Test calculations suggest the use of a conservative initial estimate of the number of sites rather than risking the introduction of too many incorrect sites. The remaining sites could always be determined with difference Fourier techniques (see below).

6.1. Comparison of potential solutions

To evaluate the performance of a heavy-atom search protocol, we compared the heavy-atom configurations obtained in the tests with the correct solution. For unknown structures, it is also useful to compare the heavy-atom configurations with each other; for example, the configuration with the highest correlation coefficient with all other configurations. Unfortunately, comparison of heavy-atom configurations is not straightforward. The configurations can be shifted with respect to each other by allowed origin shifts or one solution can be the inverse image of the other. The possible transformations for a given space group can be obtained from the corresponding Euclidean normalizer (Koch

Table 3
Common (default) parameters for all test cases.

Maximum Bragg spacing	d_{max} = 15 Å
Target for PC refinement	F2F2
PC refinement parameters	Positional refinement with shift damping and 15 *B*-factor refinement steps
Averaging of Patterson maps	Maps are weighted by diffraction ratios which are computed for ten equal volume reciprocal-space shells
Amplitude-based σ cutoff (equation 1)	f_{cut} = 1
σ cutoff for ΔF data (equation 4)	d_{cut} = 0.5
R.m.s. outlier cutoff (equation 5)	$c_{r.m.s.}$ = 4
Overall *k* scaling and *B* scaling (equation 9)	Both *k* scaling and anisotropic *B* scaling with four scaling iterations
Grid resolution for translation-search maps	1/3 of high-resolution limit
Number of initial trial sites	100
W_{IMF} for 'Reciprocal + Direct' runs	3
Required minimum distance from new site to its symmetry mates and all previously placed sites	3.5 Å
Special positions	Not allowed
Expected increase in correlation coefficient for dead-end test	0.01

& Fischer, 1983). If there are only discrete allowed origin shifts (coincidentally this is the case for all structures in Table 2) the comparison is relatively simple, but for space groups with polar axes there are continous allowed origin shifts which render the comparison much more difficult.

6.2. Comparison of search methods

For each test case, three different heavy-atom search protocols were carried out using direct-space, reciprocal-space and combined reciprocal-space and direct-space searches. Fig. 2 shows the results for 1kwa/3 Å. For brevity, the detailed results for the other test cases are not shown.

The SMF/IMF direct-space search ('Direct space' in Fig. 2) is generally only moderately successful in finding correct sites, especially for structures with more than about ten sites. The reciprocal-space search ('Reciprocal space' in Fig. 2) generally finds more sites which are also more likely to be correct. The combination of reciprocal-space and direct-space methods ('Reciprocal + Direct' in Fig. 2) is generally more successful than either method alone and the efficiency of the search is improved relative to the pure reciprocal-space search. The trial numbers for the correct solutions are lower and/or there are more solutions or partial solutions.

Although the direct-space search is the least powerful method, it is able to find the correct solution for the high-quality data sets 1ytt and 3bct/F_A, and partially correct solutions for 1kwa/3 Å, 1ecf and 1a7a. For 1ecf and 1a7a, the ratio of correct and incorrect sites is sufficiently large to allow one to eliminate the incorrectly placed sites and determine the missing sites by difference Fourier techniques (not shown). For 1kwa/3 Å, one can detect and remove the two incorrectly placed sites and then find all but one of the seven missing sites by difference Fourier techniques. For 1kwa/4 Å, the solution with the highest correlation coefficient contains only one correct site out of eight. For 3bct/ΔF and 1auv, there are no correct sites in the solutions with the highest correlation coefficient. In contrast, the reciprocal-space search produces the correct sites for all cases except 1kwa/4 Å, where it produces a partially correct solution which is sufficient to locate the remaining sites by difference Fourier techniques. The combination of reciprocal-space and direct-space methods also finds the correct solution for all cases except 1kwa/4 Å. It is the most powerful search method because it finds more correct or partially correct solutions and they appear sooner during the search (Fig. 2). The main reason for the increased efficiency is the use of the SMF for the initial sites. The very first trial for 1kwa/3 Å produces the correct solution when using only the SMF for the determination of initial trial sites and the FTF for the determination of additional sites. The main benefit of using the IMF is that more solutions or partial solutions are found.

The use of F_A structure factors increases the efficiency of the SMF/IMF direct-space search for 3bct. However, the effect on the efficiency of the FTF reciprocal-space search and the combined reciprocal-space and direct-space search is only moderate (Fig. 3). Since the computation of F_A structure factors requires at least two wavelengths (Hendrickson, 1991; Terwilliger, 1994), in most cases the desired solution is probably more easily obtained using ΔF data from the peak wavelength and using the combined reciprocal-space and direct-space search.

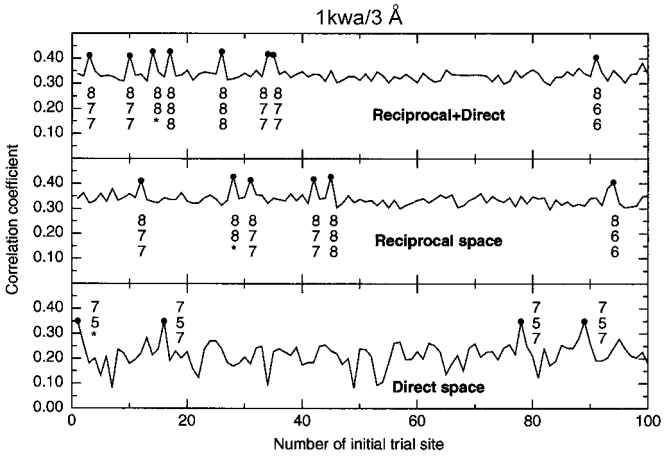

Figure 2
Comparison of direct-space, reciprocal-space and combined reciprocal and direct search methods. Correlation coefficients which are 2 rms (cc) above the mean are marked with a dot, where rms (cc) = $[\sum_{i=1}^{N_{trials}}(cc_i - \langle cc\rangle^2)/N_{trials}]^{1/2}$. Outstanding correlation coefficients are marked with a dot and annotated with a triplet of numbers: top, number of sites found in the trial; middle, number of sites which are correct within 1.5 Å; bottom, number of sites which are identical within 1.5 Å to sites of the corresponding top solution. For the top solution, the bottom number is replaced by a star.

6.3. Results obtained with the combined reciprocal and direct search method

The results for the combination of reciprocal-space and direct-space methods are shown in Fig. 3. In most cases, there are only a few outstanding correlation coefficients and they correspond to correct solutions with at most one wrong site. The exception is 1kwa/4 Å, where the top solution contains only five correct sites out of eight. The solution ranked second is the same as the top solution. However, the next two solutions have no sites in common with the top solution, although they have only slightly lower correlation coefficients. This is a clear indication that the top solution may be unreliable. In general, if solutions with similar correlation coefficients have no sites in common, it should be assumed that the correct solution was not found or that only a partial solution was found.

For 3bct/ΔF and 3bct/F_A, it is difficult to judge from the correlation coefficient alone whether or not the correct solution was found. When the solutions are compared, it appears that similar correlation coefficients correspond to the same set of sites, producing confidence in the top solution.

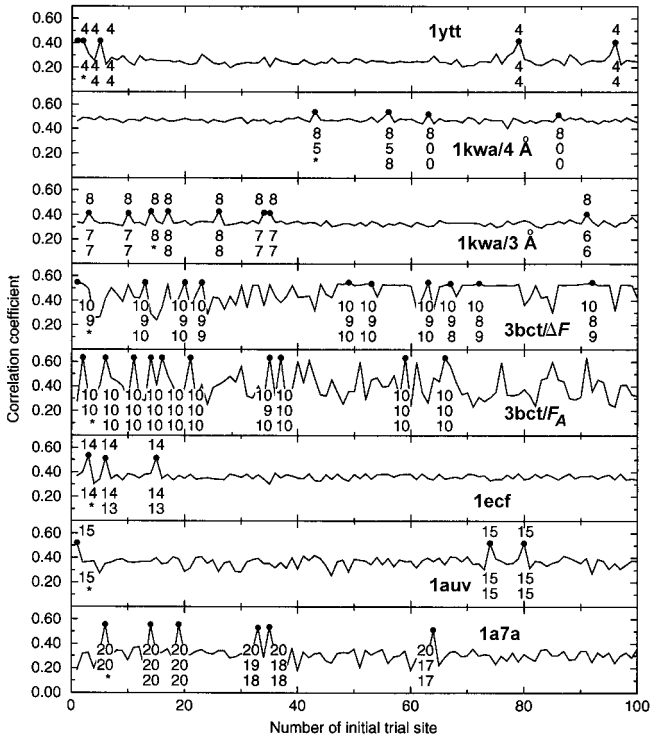

Figure 3
Results for all test cases with the Reciprocal + Direct search method. For all cases except 3bct/ΔF and 3bct/F_A, correlation coefficients which are 2 rms(cc) above the mean are marked with a dot, where rms(cc) = $[\sum_{i=1}^{N_{\rm trials}}(\mathrm{cc}_i - \langle \mathrm{cc}\rangle^2)/N_{\rm trials}]^{1/2}$. For 3bct/$\Delta F$ and 3bct/F_A, there are no correlation coefficients above 2 rms(cc) and the ten highest correlation coefficients are marked instead. The outstanding correlation coefficients are also annotated with a triplet of numbers: top, number of sites found in the trial; middle, number of sites which are correct within 1.5 Å; bottom, number of sites which are identical within 1.5 Å to sites of the corresponding top solution. For the top solution, the bottom number is replaced by a star.

Table 4
Run times (h:min) for heavy-atom searches on a DEC Alpha EV56/533 MHz processor.

Reference code	Direct space	Reciprocal space	Reciprocal + Direct
1ytt	0:09	0:11	0:15
1kwa/4 Å	1:13	2:29	2:45
1kwa/3 Å	0:34	1:05	1:12
3bct/ΔF	0:50	4:55	4:50
3bct/F_A	1:29	5:31	5:36
1ecf	1:33	11:34	12:21
1auv	1:10	6:47	7:02
1a7a	1:32	9:26	10:27

The run times for each test series are shown in Table 4. On average, the SMF/IMF direct-space search is almost five times faster than the reciprocal-space search. However, the longer run time of the reciprocal-space search is highly correlated with the success rate of the method. It is more successful in finding the correct solution, and the additional run time is well spent. The combination of the direct-space and the reciprocal-space search method is on average only 6% slower than the reciprocal-space search alone. This slight run time increase is rewarded by a higher success rate.

6.4. Completion of the heavy-atom model

In many cases, a subset of the heavy-atom sites is sufficient to obtain an interpretable electron-density map. However, to obtain the best experimental map for subsequent model building, it is desirable to place as many heavy-atom sites as possible. Therefore, incorrectly placed sites must be eliminated and missing sites must be located. Incorrectly placed sites can often be identified by unusually high refined B factors. It can also be useful to inspect the anomalous difference Fourier map computed with the diffraction data with the largest anomalous signal and the phases from the heavy-atom refinement. An anomalous difference map shows strong peaks at the refined sites, but incorrectly placed sites are likely to have relatively small peaks. There are often additional peaks at missing sites. A more powerful method for the location of missing sites is the computation of a double-difference or log-likelihood gradient map (Bricogne, 1984). Positive peaks in this map may indicate missing sites. Negative peaks may indicate incorrectly placed sites.

3bct/ΔF is the only test case where the top solution of the heavy-atom search contained an incorrect site. Upon refinement of the coordinates and B factors for the ten sites (Burling *et al.*, 1996), the B factors for nine sites were distributed between 10 and 36 Å^2. The B factor of the ninth site was exceptionally high (57.88 Å^2). This observation is corroborated by the peak heights in the anomalous difference map. The peak corresponding to the ninth site is significantly smaller (10.95σ) than the other peaks (19.30–44.65σ). The gradient map shows an outstanding peak at 21.45σ, with the second highest peak being much lower (7.72σ). The ninth site was therefore removed and the highest peak from the gradient map was added as a new site. After heavy-atom refinement of the ten sites and application of solvent flipping (Abrahams &

Leslie, 1996) as implemented in *CNS* (Brunger *et al.*, 1998), the electron-density map was readily interpretable.

Fig. 4 shows the anomalous difference and gradient maps for 1a7a after refinement of the 20 sites from the top solution of the heavy-atom search. The ten highest peaks in the gradient map correspond to the positions of the ten missing sites. This example demonstrates that only two-thirds of the sites are sufficient to determine the remaining sites by difference Fourier techniques.

7. Practical considerations

Based on our experience, the heavy-atom search procedure is highly successful when using the parameters listed in Table 3, a high-resolution cutoff of $d_{\min} = 4$ Å and a search for two-thirds of the number of expected sites. If there is an outstanding correlation coefficient or a group of outstanding correlation coefficients of similar solutions, one should continue with phase refinement and difference Fourier techniques. If no clear solution was found with the 4 Å high-resolution cutoff, a 3 Å cutoff should be tried. For structures with a large number of expected heavy-atom sites, it could also be useful to increase the value of $N_{\rm dead}$ to one or two.

Our test calculations suggest that for structures with less than about 30 expected heavy-atom sites, given reasonably good diffraction data, the correct solution is very likely to emerge. It is likely that the algorithm will also work for structures with more than 30 expected heavy-atom sites, although the success rate for increasingly complex structures will be related to the quality of the diffraction data.

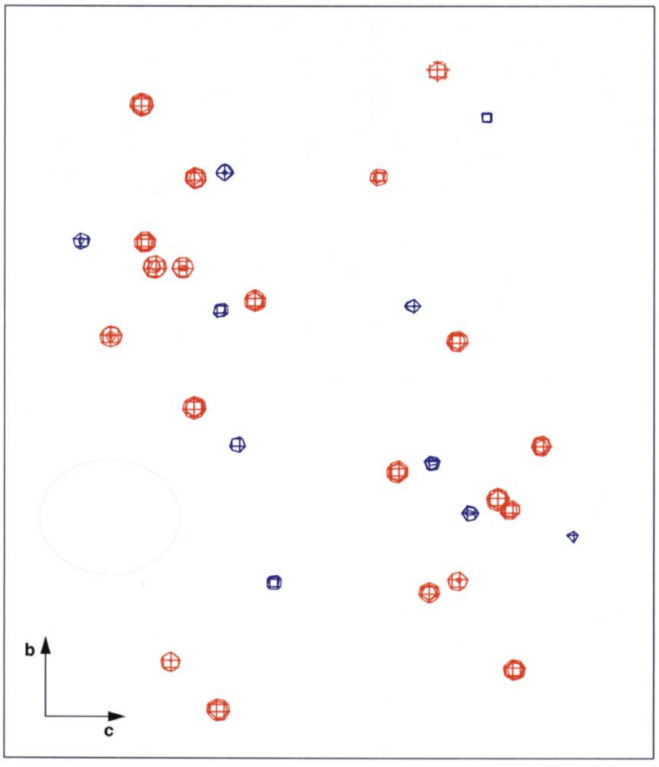

Figure 4
Anomalous difference (red) and log-likelihood gradient map (blue) for 1a7a. The maps were computed after MAD refinement using the 20 sites found by the heavy-atom search procedure. The contour level for both maps is 11σ. The anomalous difference map shows peaks at the 20 refined sites. The gradient map shows peaks for the ten missing sites.

8. Conclusions

In our laboratory, a large number of SeMet-MAD structures were recently solved using the automated heavy-atom search algorithm. One example is the test case 1kwa. Five selenium sites were found in anomalous difference Fourier maps with the SIR phases obtained from a mercury derivative. Three additional sites were found with the heavy-atom search algorithm (Daniels *et al.*, 1998). Another example is the structure of the ATP-dependent oligomerization domain of the *N*-ethylmaleimide-sensitive factor complexed with ATP1 (Yu *et al.*, 1998). All nine selenium sites except for a disordered N-terminal methionine were found within minutes of data collection. The structure of the core of the synaptic SNARE complex (Sutton *et al.*, 1998) was a challenging case because it contained 46 ordered SeMet residues. Given the limited crystal quality, multiple MAD experiments were required, each using a different combination of native and SeMet-labeled proteins. One of the combinations with a disperse set of sites allowed the automatic determination of 15 ordered selenium sites. The remaining sites were found by difference Fourier techniques. The structure of the small G protein Rab3A complexed with the effector domain of rabphilin-3A (Ostermeier & Brunger, 1999) has two Zn atoms in the asymmetric unit. Zinc has a significant anomalous signal at the wavelengths used for SeMet MAD phasing and was therefore included in the interpretation of the heavy-atom search solutions and the heavy-atom refinement. The best solution from the heavy-atom search with six sites was used to initiate the heavy-atom refinement. Combination of refinement and difference Fourier techniques then produced all nine ordered selenium sites and the two zinc sites.

The test cases and the application to many unknown structures clearly demonstrate that the combination of reciprocal-space and direct-space Patterson searches and PC refinements is a fast and powerful tool for the detection of heavy-atom sites. It is an important step towards a fully automated procedure which leads directly from SeMet-MAD data to atomic coordinates. In this context, it should be noted that the heavy-atom search algorithm is very well suited for parallel computers. Even searches for many sites in complex structures could be performed in a matter of a few minutes on a massively parallel computer.

APPENDIX
Improved implementation of the fast translation function

The fast translation function requires the computation and fast Fourier transform (FFT) of three three-dimensional intermediate arrays. In the following, it is sufficient to consider a

condensed form of the most time-consuming portion of equation (15) in Navaza & Vernoslova (1995),

$$\sum_{s}\sum_{s'}\sum_{s''}\sum_{s'''} V_{s,s',s'',s'''} \exp(-2\pi i \mathbf{A}_{s,s',s'',s'''}). \qquad (11)$$

Each sum runs over all symmetry operations of the space group and has to be computed for each reflection. The integer vector $\mathbf{A}_{s,s',s'',s'''}$ can be viewed as the address of a point in the three-dimensional intermediate array to which the complex value $V_{s,s',s'',s'''}$ is added.

(11) involves reciprocal vectors up to four times the diffraction data resolution. In general, if reciprocal vectors are to be stored on a grid prior to an FFT, the resolution of that grid must be twice that of the highest angle reciprocal vector. Therefore, if the high-resolution cutoff is 4 Å, the resolution of the intermediate array must be 0.5 Å. This can easily result in excessive memory requirements. For example, for 1a7a and a 4 Å cutoff, the grid sizes of the intermediate array are 184, 337 and 276 in the *a*, *b* and *c* directions, respectively. If the Hermitian symmetry of the reciprocal-space coefficients is used, one of the grid sizes can be reduced by a factor of two, but the array still has more than 8.5 million elements. Single-precision (4 byte) floating-point arithmetic is insufficient. Depending on the space-group symmetry, the round-off errors in the summation of the coefficients can completely invalidate the results. Therefore, double-precision (8 byte) arithmetic must be used, and consequently the size of the intermediate array for 1a7a is about 64 Mbyte.

The situation is significantly better for the other two intermediate arrays, which correspond to equation (14) in Navaza & Vernoslova (1995). The condensed form of the most time-consuming sub-equation is

$$\sum_{s}\sum_{s'} V_{s,s'} \exp(-2\pi i \mathbf{A}_{s,s'}). \qquad (12)$$

This equation only involves reciprocal-space vectors up to twice the data resolution. Therefore, the memory requirement is only one-eighth of the array corresponding to equation (15) in Navaza & Vernoslova (1995).

The simplest and fastest approach for carrying out the summations of (11) and (12) is to allocate a three-dimensional array large enough to hold all complex values *V*. This approach is fast because each integer vector **A** can be directly mapped to a location in the three-dimensional array. However, the total memory requirement of a heavy-atom search based on this approach can easily exceed 500 Mbyte.

Depending on the space group, typically only 1–2% of the intermediate three-dimensional arrays have non-zero values. Therefore, one can use one-dimensional tables of pairs (**A**, $\sum V$) instead of three-dimensional arrays, where $\sum V$ is the accumulated sum of the *V*. This idea is the basis of the implementation of the FTF in *AMoRe* (Navaza, 1994). However, for each addition in the summations corresponding to (11) and (12), the one-dimensional array has to be searched for the index **A**. This is a much more time-consuming operation than a direct access to a three-dimensional array and consequently the summations are slowed significantly.

We chose a compromise between the fast three-dimensional arrays and the much more memory-efficient one-dimensional tables. Naturally, this leads to a two-dimensional array. The first two components a_x and a_y of a vector $\mathbf{A} = (a_x, a_y, a_z)$ serve as an address to a two-dimensional array. Each element of this array is a table of pairs (a_z, $\sum V$). Initially, each table is empty. During the summation, space for the indices a_z and the accumulated values $\sum V$ is dynamically allocated as needed. Each individual table is much shorter than the single table in the one-dimensional approach and therefore the table lookup operations are significantly faster. The memory overhead for the two-dimensional array is negligible compared with other storage requirements.

These considerations lead to two implementations of the FTF in *CNS*. In both implementations, integer expansion factors are computed by which the translation-search grid sizes must be multiplied in order to accommodate all the coefficients in (11) and (12). With the resolution of the FTF maps typically set to one-third of the high-resolution cutoff for the data, the factors for (11) are usually three and the factors for (12) are usually two.

In the first implementation, the large three-dimensional intermediate arrays for the summations are kept in core memory. After the summation for one array is finished, it is Fourier transformed in the first dimension. The size of the array is then reduced by the corresponding expansion factor. If, for example, the expansion factor for the first transform direction is three, only every third grid point is retained. The other values are not needed. After the reduction, the second dimension is transformed and the size of the array is reduced again. Finally, the third dimension is transformed and reduced. This re-sampling between the transform steps significantly reduces the run time.

The second implementation uses two-dimensional arrays of tables of pairs (a_z, $\sum V$) in combination with re-sampling between the Fourier transformations for each direction. An additional benefit of this algorithm is that it is immediately obvious which columns and planes of the (virtual) three-dimensional array contain only coefficients equal to zero and thus do not need to be transformed. This simple and inexpensive book-keeping allows for further reduction of the run time. Although the summation step of the second implementation is always slower than that of the first implementation, the second implementation is faster for most structures. The exceptions are large high-symmetry (hexagonal or cubic) structures.

For the tests reported in this paper, we always used the second implementation. It is worth noting that the run time needed for the PC refinements is typically about the same as that for the FTFs. Therefore, for the vast majority of cases, any further reduction of the run time for the FTF (for example by using symmetrized FFT algorithms) would not have a significant impact on the total run time for the heavy-atom search.

We are grateful to P. D. Adams for valuable discussions, J. Navaza for helping us understand the fast translation function, M. A. Estermann for help in the implementation of the

minimum functions, Elke Koch for helping us understand the symmetries of the search maps and N. Ban, R. B. Sutton and L. M. Rice for critical reading of the manuscript. We wish to thank W. I. Weis, M. Turner, J. Krahn and L. Esser for providing us with their SeMet-MAD data sets.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Brunger, A. T. (1991). *Acta Cryst.* A**47**, 195–204.
Brunger, A. T., Adams, P. D., Clore, G. M., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, N., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Burling, F. T., Weis, W. I., Flaherty, K. M. & Brunger, A. T. (1996). *Science*, **271**, 72–77.
Chang, G. & Lewis, M. (1994). *Acta Cryst.* D**50**, 667–674.
Daniels, D. L., Cohen, A. R., Anderson, J. M. & Brunger, A. T. (1998). *Nature Struct. Biol.* **5**(4), 317–325.
DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.
Drenth, J. (1994). *Principles of Protein X-ray Crystallography.* New York: Springer–Verlag.
Dumas, P. (1994). *Acta Cryst.* A**50**, 537–546.
Esser, L., Wang, C. R., Hosaka, M., Smagula, C. S., Sudhof, T. C. & Deisenhofer, J. (1998). *EMBO J.* **17**(4), 977–984.
Estermann, M. A. (1995). *Nucl. Instrum. Methods Phys. Res. A*, **354**, 126–133.
Fujinaga, M. & Read, R. J. (1987). *J. Appl. Cryst.* **20**, 517–521.
Hendrickson, W. A. (1988). *Proteins*, **4**(2), 77–88.
Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). *Cell*, **90**(5), 871–882.
Karle, J. (1989). *Acta Cryst.* A**45**, 303–307.
Koch, E. & Fischer, W. (1983). *International Tables for Crystallography*, Vol. A, ch. 15. Dordrecht: Kluwer Academic Publishers.
Lipson, H. & Cochran, W. (1957). *The Determination of Crystal Structures*, p. 235. London: Bell.
McRee, D. E. (1993). *Practical Protein Crystallography.* San Diego: Academic Press.
Muchmore, C. R., Krahn, J. M., Kim, J. H., Zalkin, H. & Smith, J. L. (1998). *Protein Sci.* **7**(1), 39–51.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Navaza, J. & Vernoslova, E. (1995). *Acta Cryst.* A**51**, 445–449.
Ostermeier, C. & Brunger, A. T. (1999). *Cell*, **96**, 363–374.
Pavelcik, F. (1986). *J. Appl. Cryst.* **19**, 488–491.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Sheldrick, G. M. (1997). *Methods Enzymol.* **276**, 628–641.
Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.
Simpson, P. G., Dobrott, R. D. & Lipscomb, W. N. (1965). *Acta Cryst.* **18**, 169–179.
Steward, J. M. & Karle, J. (1976). *Acta Cryst.* A**32**, 1005–1007.
Sutton, R. B., Fasshauer, D., Jahn, R. & Brunger, A. T. (1998). *Nature (London)*, **395**, 347–353.
Terwilliger, T. C. (1994). *Acta Cryst.* D**50**, 11–16.
Terwilliger, T. C. (1998). *Solve – Automated Structure Solution for MIR and MAD*, http://www.solve.lanl.gov/.
Terwilliger, T. C., Kim, S.-H. & Eisenberg, D. (1987). *Acta Cryst.* A**43**, 1–5.
Tong, T. & Rossmann, M. G. (1993). *J. Appl. Cryst.* **26**, 15–21.
Turner, M. A., Yuan, C. S., Borchardt, R. T., Hershfield, M. S., Smith, G. D. & Howell, P. L. (1998). *Nature Struct. Biol.* **5**(5), 369–376.
Vagin, A. & Teplyakov, A. (1998). *Acta Cryst.* D**54**, 400–402.
Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.
Weeks, C. M. & Miller, R. (1999). *J. Appl. Cryst.* **32**, 120–124.
Yu, R. C., Hanson, P. I., Jahn, R. & Brunger, A. T. (1998). *Nature Struct. Biol.* **5**, 803–811.