International Tables for Crystallography (2006). Vol. F, Section 14.2.2, pp. 303–307.

anomalous scattering factors at any time during phasing. These quantities are incorporated into heavy-atom atomic 'occupancies' and refined along with other parameters. Of course, the partial structure of anomalous scatterers must be known, and its refinement is concurrent with phasing. This may be a principal advantage of the pseudo-MIR approach, because the anomalous-scatterer parameter refinement may be more reliable when incorporated into phasing than when done against $|{}^0F_A|$ estimates. Greater weight is given to the data set selected as 'native' in refinement of the 'heavy-atom' parameters in some implementations of the pseudo-MIR approach, although others treat data at all wavelengths equivalently (Terwilliger & Berendzen, 1997). The amplitudes $|{}^0F_A|$ are not a by-product of the pseudo-MIR approach.

14.2.1.7. *Determination of the anomalous-scatterer partial structure*

Determination of the partial structure of anomalous scatterers is a prerequisite for MAD-phased electron density, regardless of the phasing technique. As described above, the optimal quantities for solving and refining the partial structure of anomalous scatterers are the normal structure amplitudes $|{}^0F_A|$. Frequently $|{}^0F_A|$ values are not extracted from the MAD measurements, and the largest Bijvoet or dispersive differences are used instead. This involves the approximation of representing structure amplitudes ($|{}^0F_A|$) as the subset of larger differences ($||F^+| - |F^-||$ or $||F_{\lambda^1}| - |F_{\lambda^2}||$). The approximation is accurate for only a small fraction of reflections because there is little correlation between φ_A and φ_T. However, it suffices for a suitably strong signal and a suitably small number of sites. Patterson methods are quite successful in locating anomalous scatterers when the number of sites is small. However, the aim of MAD is to solve the macromolecule structure from one MAD data set using any number of anomalous scatterer sites. For larger numbers of sites, statistical direct methods may be employed.

The correct enantiomorph for the anomalous-scatterer partial structure also must be determined (φ_A *versus* $-\varphi_A$) in order to obtain an electron-density image of the macromolecule. However, it cannot be determined directly from MAD data. The correct enantiomorph is chosen by comparison of electron-density maps based on both enantiomorphs of the partial structure. Unlike the situation for pure MIR, the density based on the incorrect enantiomorph of the anomalous-scatterer partial structure is not the mirror image of that based on the correct enantiomorph and contains no image of the macromolecule. The correct map is distinguished by features such as a clear solvent boundary, positive correlation of redundant densities and a macromolecule-like density histogram. If the anomalous-scattering centres form a centric array, then the two enantiomorphs are identical and both maps are correct.

14.2.1.8. *General anomalous-scatterer labels for biological macromolecules*

MAD requires a suitable anomalous scatterer, of which none are generally present in naturally occurring proteins or nucleic acids. However, selenomethionine (SeMet) substituted for the natural amino acid methionine (Met) is a general anomalous-scattering label for proteins (Hendrickson, 1985), and is the anomalous scatterer most frequently used in MAD. The *K* edge of Se is the most accessible for MAD experiments ($\lambda = 0.98$ Å).

The SeMet label is especially general and convenient because it is introduced by biological substitution of SeMet for methionine. This is achieved by blocking methionine biosynthesis and substituting SeMet for Met in the growth medium of the cells in which the protein is produced. Production of SeMet protein in bacteria is generally straightforward (Hendrickson *et al.*, 1990; Doublié, 1997) and has also been accomplished in eukaryotic cells (Lustbader *et al.*, 1995; Bellizzi *et al.*, 1999).

Methionine is a particularly attractive target for anomalous-scatterer labelling because the side chain is usually buried in the hydrophobic core of globular proteins where it is relatively better ordered than are surface side chains. The labelling experiment provides direct evidence for isostructuralism of Met and SeMet proteins. All proteins in the biological expression system have SeMet substituted for Met at levels approaching 100%. The cells are viable, therefore the proteins are functional and isostructural with their unlabelled counterparts to the extent required by function.

The natural abundance of methionine in soluble proteins is approximately one in fifty amino acids, providing a typical MAD phasing signal of 4–6% of $|F|$ [equations (14.2.1.14) and (14.2.1.15)]. Typical extreme values for the anomalous scattering factors are $f'_{\rm min} \sim -10$ e and $f''_{\rm max} \sim 6$ e (Fig. 14.2.1.1). SeMet is more sensitive to oxidation than is Met, and care must be taken to maintain a homogeneous oxidation state. Generally, the reduced state is maintained by addition of disulfide reducing agents to SeMet protein and crystals. However, the oxidized forms of Se have sharper *K*-edge features and f' and f'' values of greater magnitude than does the reduced form (Smith & Thompson, 1998). This property has been exploited to enhance anomalous signals by intentional oxidation of SeMet protein (Sharff *et al.*, 2000). SeMet is also a useful isomorphous-replacement label with a signal of $\sim$10% of $|F|$. Prior knowledge of the sites of labelling is extremely useful during initial fitting of a protein sequence to electron density. Also, noncrystallographic symmetry operators can usually be defined more reliably from Se positions in SeMet protein than by heavy-atom positions in MIR due to the uniformity and completeness of labelling (Tesmer *et al.*, 1996).

An analogous general label is available for nucleic acids in the form of brominated bases, particularly 5-bromouridine, which is isostructural with thymidine. The *K* edge of Br corresponds to a wavelength of 0.92 Å, which is quite favourable for data collection.

14.2.2. Automated MAD and MIR structure solution

(T. C. Terwilliger and J. Berendzen)

14.2.2.1. *Introduction*

In favourable cases, structure solution by X-ray crystallography using the MAD or MIR methods can be a straightforward, though often lengthy, process. The recently developed *Solve* software (Terwilliger & Berendzen, 1999*b*) is designed to fully automate this class of structure solution. The overall approach is to link together all the analysis steps that a crystallographer would normally carry out into a seamless procedure, and in the process to convert each decision-making step into an optimization problem.

In the case of both MAD and MIR data, a key element of the procedure is the scoring and ranking of possible solutions. This scoring procedure makes it possible to treat structure solution as an optimization procedure, rather than a decision-making one. In the case of MAD data, a second key element of the procedure is the conversion of MAD data to a pseudo-SIRAS form (Terwilliger, 1994*b*) that allows much more rapid analysis than one involving the full MAD data set.

14.2.2.2. *MAD and MIR structure solution*

The MAD and MIR approaches to structure solution are conceptually very similar and share several important steps. Two of these are the identification of possible locations of heavy or anomalously scattering atoms and an analysis of the quality of each of these potential heavy-atom solutions. In each method, trial partial structures for these heavy or anomalously scattering atoms are often obtained by inspection of difference Patterson functions or by semiautomated analysis (*e.g.* Terwilliger *et al.*, 1987; Chang & Lewis, 1994; Vagin & Teplyakov, 1998). In other cases, direct-methods approaches have been used to find heavy-atom sites (Sheldrick, 1990; Miller *et al.*, 1994). Potential heavy-atom solutions found in any of these approaches are often just a starting point for structure solution, with additional sites found by difference Fourier or other approaches.

The analysis of the quality of potential heavy-atom solutions is also very similar in the MIR and MAD methods. In both cases a partial structure is used to calculate native phases for the entire structure, and the electron density that results is examined to see if the expected features of the macromolecule are found. Additionally, the agreement of the heavy-atom model with the difference Patterson function and the figure of merit of phasing are commonly used to evaluate the quality of a solution. In many cases, an analysis of heavy-atom sites by sequential deletion of individual sites or derivatives is often an important criterion of quality as well (Dickerson *et al.*, 1961).

14.2.2.3. *Decision making and structure solution*

The process of structure solution can be thought of largely as a decision-making process. In the early stages of solution, a crystallographer must choose which of several potential trial solutions may be worth pursuing. At a later stage, the crystallographer must choose which peaks in a heavy-atom difference Fourier are to be included in the heavy-atom model, and which hand of the solution is correct. At a final stage, the crystallographer must decide whether the solution process is complete and which of the possible heavy-atom models is the best. The most important feature of the *Solve* software is the use of a consistent scoring algorithm as the basis for making all these decisions.

14.2.2.4. *The need for rapid refinement and phasing during automated structure solution*

In order to make automated structure solution practical, it was necessary to be able to evaluate heavy-atom solutions very rapidly. This is because the automated approach used by *Solve* requires analysis of many heavy-atom solutions (typically 300–1000). For each heavy-atom solution examined, the heavy-atom sites have to be refined and phases calculated. In implementing automated structure solution, it was important to recognize the need for a trade-off between the most accurate heavy-atom refinement and phasing at all stages of structure solution and the time required to carry it out. The balance chosen for *Solve* was to use the most accurate available methods for final phase calculations, and to use approximate but much faster methods for all refinements and phase calculations. The refinement method chosen on this basis was origin-removed Patterson refinement (Terwilliger & Eisenberg, 1983), which treats each derivative in an MIR data set independently and which is very fast because it does not require phase calculation. The phasing approach used for MIR data thoughout *Solve* is Bayesian correlated phasing (Terwilliger & Berendzen, 1996; Terwilliger & Eisenberg, 1987), which takes into account the correlation of non-isomorphism among derivatives without substantially slowing down phase calculations.

For MAD data, Bayesian calculations of phase probabilities are very slow (*e.g.* Terwilliger & Berendzen, 1997; de La Fortelle & Bricogne, 1997). Consequently, we have used an alternative procedure for all MAD phase calculations except those done at the very final stage. This alternative is to convert the MAD data set into a form that is similar to one obtained in the single isomorphous replacement with anomalous scattering (SIRAS) method. In this way, a single data set with isomorphous and anomalous differences is obtained that can be used in heavy-atom refinement by the origin-removed Patterson refinement method and in phasing by conventional SIRAS phasing (Terwilliger & Eisenberg, 1987).

14.2.2.5. *Conversion of MAD data to a pseudo-SIRAS form*

The conversion of MAD data to a pseudo-SIRAS form that has almost the same information content requires two important assumptions. The first assumption is that the structure factor corresponding to anomalously scattering atoms in a structure varies in magnitude but not in phase at various X-ray wavelengths. This assumption will hold when there is one dominant type of anomalously scattering atom. The second is that the structure factor corresponding to anomalously scattering atoms is small compared to the structure factor from all other atoms. As long as these two assumptions hold, the information in a MAD experiment is largely contained in just three quantities: a structure factor (F_o) corresponding to the scattering from non-anomalously scattering atoms, a dispersive or isomorphous difference at a standard wavelength λ_o ($\Delta^{\rm ISO}_{\lambda_o}$), and an anomalous difference ($\Delta^{\rm ANO}_{\lambda_o}$) at the same standard wavelength (Terwilliger, 1994*b*). It is easy to see that these three quantities could be treated just like a SIRAS data set with the 'native' structure factor F_P replaced by F_o, the derivative structure factor F_{PH} replaced by $F_o + \Delta^{\rm ISO}_{\lambda_o}$, and the anomalous difference replaced by $\Delta^{\rm ANO}_{\lambda_o}$ (Terwilliger, 1994*b*). This is the approach taken by *Solve*. In this section, it is briefly shown how these three quantities can be estimated from MAD data.

For a particular reflection and a particular wavelength λ_j, we can write the total normal (*i.e.*, non-anomalous) scattering from a structure ($\mathbf{F}_{{\rm tot},\,\lambda_j}$) as the sum of two components. One is the scattering from all non-anomalously scattering atoms ($\mathbf{F}_o$). This scattering is wavelength-independent. The second is the normal scattering from anomalously scattering atoms ($\mathbf{F}_{H_{\lambda_j}}$) at wavelength λ_j. This term includes wavelength-dependent dispersive shifts in atomic scattering due to the f' term in the scattering factor, but not the anomalous part due to the f'' term. The magnitude of the total scattering factor can then be written in the form

$$F_{{\rm tot},\,\lambda_j} = |\mathbf{F}_o + \mathbf{F}_{H_{\lambda_j}}|. \qquad (14.2.2.1)$$

Here $\mathbf{F}_o$ and $\mathbf{F}_{{\rm tot},\,\lambda_j}$ can be thought of corresponding, respectively, to the native structure factor, F_P, and the derivative structure factor, F_{PH}, as used in the method of isomorphous replacement (Blundell & Johnson, 1976). If the scattering from anomalously scattering atoms is small compared to that from all other atoms, equation (14.2.2.1) can be rewritten in the approximate form

$$F_{{\rm tot},\,\lambda_j} \simeq F_o + F_{H_{\lambda_j}} \cos(\alpha), \qquad (14.2.2.2)$$

where α is the phase difference between the structure factors corresponding to non-anomalously and anomalously scattering atoms in the unit cell, $\mathbf{F}_o$ and $\mathbf{F}_{H_{\lambda_j}}$, respectively, at this X-ray wavelength.

The data in a MAD experiment consist of observations of structure-factor amplitudes for Bijvoet pairs, $F^+_{\lambda_j}$ and $F^-_{\lambda_j}$, for several X-ray wavelengths λ_j. These can be rewritten in terms of an average structure-factor amplitude $\overline{F}_{\lambda_j}$ and an anomalous difference $\Delta^{\rm ANO}_{\lambda_j}$ (*cf.* Blundell & Johnson, 1976). We would like to convert these into estimates of the amplitude of the structure factor corresponding to the non-anomalously scattering atoms alone, the amplitude of the structure factor corresponding to the entire structure at a standard wavelength, and the anomalous difference at the standard wavelength.

The normal scattering due to anomalously scattering atoms ($\mathbf{F}_{H_{\lambda_j}}$) changes in magnitude but not direction as a function of X-ray wavelength. We can therefore write (Terwilliger, 1994*b*)

$$\mathbf{F}_{H_{\lambda_j}} = \mathbf{F}_{H_{\lambda_o}} \frac{f_o + f'(\lambda_j)}{f_o + f'(\lambda_o)}, \tag{14.2.2.3}$$

where λ_o is an X-ray wavelength arbitrarily defined as a standard, and the real part of the scattering factor for the anomalously scattering atoms at wavelength λ_o is $f_o + f'(\lambda_j)$. A corresponding approximation for the anomalous differences at various wavelengths can also be written (Terwilliger & Eisenberg, 1987)

$$\Delta^{\rm ANO}_{\lambda_j} = \Delta^{\rm ANO}_{\lambda_o} \frac{f''(\lambda_j)}{f''(\lambda_o)}, \tag{14.2.2.4}$$

where $f''(\lambda_j)$ is the imaginary part of the scattering factor for the anomalously scattering atoms at wavelength λ_j. Based on equation (14.2.2.4), anomalous differences at any wavelength can be estimated using measurements at the standard wavelength.

An estimate of the structure-factor amplitude (F_o) corresponding to the scattering from non-anomalously scattering atoms and of the dispersive difference at standard wavelength λ_o ($\Delta^{\rm ISO}_{\lambda_o}$) can be obtained from average structure-factor amplitudes ($\overline{F}_{\lambda_j}$) at any pair of wavelengths λ_i and λ_j by proceeding in two steps. Using equations (14.2.2.2) and (14.2.2.3), the component of $\mathbf{F}_{H_{\lambda_o}}$ along $\mathbf{F}_o$, which we term $\Delta^{\rm ISO}_{\lambda_o}$, can be estimated as

$$\Delta^{\rm ISO}_{\lambda_o} \simeq F_{H_{\lambda_o}} \cos(\alpha) \tag{14.2.2.5}$$

or

$$\Delta^{\rm ISO}_{\lambda_o} \simeq (\overline{F}_{\lambda_i} - \overline{F}_{\lambda_j}) \frac{f_o + f'(\lambda_o)}{f'(\lambda_i) - f'(\lambda_j)}. \tag{14.2.2.6}$$

Then, in turn, this estimate of $\Delta^{\rm ISO}_{\lambda_o}$ can be used to obtain F_o:

$$F_o \simeq \overline{F}_{\lambda_j} - \Delta^{\rm ISO}_{\lambda_o} \frac{f_o + f'(\lambda_j)}{f_o + f'(\lambda_o)}. \tag{14.2.2.7}$$

This set of F_o, $F_o + \Delta^{\rm ISO}_{\lambda_o}$ and $\Delta^{\rm ANO}_{\lambda_j}$ can then be used just as F_P, F_{PH} and $\Delta^{\rm ANO}$ are used in the SIRAS (single isomorphous replacement with anomalous scattering) method.

The algorithm described above is implemented in the program segment *MADMRG* as part of *Solve* (Terwilliger, 1994*b*). In most cases, there are more than one pair of X-ray wavelengths corresponding to a particular reflection. The estimates from each pair of wavelengths are averaged, using weighting factors based on the uncertainties in each estimate. Data from various pairs of X-ray wavelengths and from various Bijvoet pairs can have very different weights in their contributions to the total. This can be understood by noting that pairs of wavelengths that yield a large value of the denominator in equation (14.2.2.6) (*i.e.*, those that differ considerably in dispersive contributions) would yield relatively accurate estimates of $\Delta^{\rm ISO}_{\lambda_o}$. In the same way, Bijvoet differences measured at the wavelength with the largest value of f'' will contribute the most to estimates of $\Delta^{\rm ANO}_{\lambda_j}$.

The standard wavelength choice in this analysis is arbitrary, because values at any wavelength can be converted to values at any other wavelength. The standard wavelength does not even have to be one of the wavelengths in the experiment, though it is convenient to choose one of them.

14.2.2.6. *Scoring of trial heavy-atom solutions*

Scoring of potential heavy-atom solutions is an essential part of the *Solve* algorithm because it allows ranking of solutions and appropriate decision making. *Solve* scores trial heavy-atom solutions (or anomalously scattering atom solutions) using four criteria: agreeement with the Patterson function, cross-validation of heavy-atom sites, figure of merit, and non-randomness of the electron-density map. The scores for each criterion are normalized to those for a group of starting solutions (most of which are incorrect) to obtain *Z* scores. The total score for a solution is the sum of its *Z* scores after correction for anomalously high scores in any category.

The first criterion used by *Solve* for evaluating a trial heavy-atom solution is the agreement between calculated and observed Patterson functions. Comparisons of this type have always been important in the MIR and MAD methods (Blundell & Johnson, 1976). The score for Patterson-function agreement is the average value of the Patterson function at predicted locations of peaks, after multiplication by a weighting factor based on the number of heavy-atom sites in the trial solution. The weighting factor (Terwilliger & Berendzen, 1999*b*) is adjusted so that if two solutions have the same mean value at predicted Patterson peaks, the one with the larger numbers of sites receives the higher score. Typically the weighting factor is approximately given by $(N)^{1/2}$, where there are *N* sites in the solution.

In some cases, predicted Patterson vectors fall on high peaks that are not related to the heavy-atom solution. To exclude these contributions, occupancies of each heavy-atom site are refined so that the predicted peak heights approximately match the observed peak heights at the predicted interatomic positions. Then all peaks with heights more than 1σ higher than their predicted values are truncated at this height. The average values are further corrected for instances where more than one predicted Patterson vector falls on the same location by scaling that peak height by the fraction of predicted vectors that are unique.

A 'cross-validation' difference Fourier analysis is the basis of the second criterion used to evaluate heavy-atom solutions. One at a time, each site in a solution (and any equivalent sites in other derivatives for MIR solutions) is omitted from the heavy-atom model and phases are recalculated. These phases are used in a difference Fourier analysis and the peak height at the location of the omitted site is noted. A similar analysis where a derivative is omitted from phasing and all other derivatives are used to phase a difference Fourier has been used for many years (Dickerson *et al.*, 1961). The score for cross-validation difference Fouriers is the average peak height, after weighting by the same factor used in the difference Patterson analysis.

The mean figure of merit of phasing (*m*) (Blundell & Johnson, 1976) can be a remarkably useful measure of the quality of phasing despite its susceptibility to systematic error (Terwilliger & Berendzen, 1999*b*). The overall figure of merit is essentially a measure of the internal consistency of the heavy-atom solution and the data, and is used as the third criterion for solution quality in *Solve*. As heavy-atom refinement in *Solve* is carried out using origin-removed Patterson refinement (Terwilliger & Eisenberg, 1983), occupancies of heavy-atom sites are relatively unbiased. This minimizes the problem of high occupancies leading to inflated figures of merit. Additionally, using a single procedure for phasing allows comparison between solutions. The score based on figure of merit is simply the unweighted mean for all reflections included in phasing.

The most important criterion used by a crystallographer in evaluating the quality of a heavy-atom solution is the interpretability of the resulting electron-density map. Although a full implementation of such a criterion is difficult, it is quite straightforward to evaluate instead whether the electron-density map has features that are expected for a crystal of a macromolecule. A number of features of electron-density maps could be used for this purpose, including the connectivity of electron density in the maps (Baker *et al.*, 1993), the presence of clearly defined regions of protein and solvent (Wang, 1985; Podjarny *et al.*, 1987; Zhang & Main, 1990; Xiang *et al.*, 1993; Abrahams *et al.*, 1994; Terwilliger & Berendzen, 1999*a*,*c*), and histogram matching of electron densities (Zhang & Main, 1990; Goldstein & Zhang, 1998). We have used the identification of solvent and protein regions as the measure of map quality in *Solve*. This requires that there be both solvent and protein regions in the electron-density map, but for most macromolecular structures the fraction of the unit cell that is occupied by the macromolecule is in the suitable range of 30–70%. The criterion used in scoring by *Solve* is based on the connectivity of the solvent and protein regions (Terwilliger & Berendzen, 1999*c*). The unit cell is divided into boxes approximately twice the resolution of the map on a side, and within each box the r.m.s. electron density is calculated, without including the F_{000} term in the Fourier synthesis. For boxes within the protein region, this r.m.s. electron density will typically be high (as there are some points where atoms are located and other points between atoms), while for those in the solvent region it will be low (as the electron density is fairly uniform). The score based on the connectivity of the protein and solvent regions is simply the correlation coefficient of this r.m.s. electron density for adjacent boxes. If there is a large contiguous protein region and a large contiguous solvent region, then adjacent boxes will have highly correlated values of their r.m.s. electron densities. If the electron density is random, there will be little or no correlation. In practice, for a very good electron-density map, this correlation of local r.m.s. electron density may be as high as 0.5 or 0.6.

14.2.2.7. *Automated MIR and MAD structure determination*

The four-point scoring scheme described above provides the foundation for automated structure solution. To make it practical, the conversion of MAD data to a pseudo-SIRAS form and the use of rapid origin-removed Patterson-based heavy-atom refinement has been nearly essential. The remainder of the *Solve* algorithm for automated structure solution is largely a standardized form of local scaling, an integrated set of routines to carry out all of the calculations required for heavy-atom searching, refinement and phasing, and routines to keep track of the lists of current solutions being examined and past solutions that have already been tested.

Scaling of data in the *Solve* algorithm is done by a local scaling procedure (Matthews & Czerwinski, 1975). Systematic errors are minimized by scaling F^+ and F^-, native and derivative, and wavelengths of MAD data in very similar ways and by keeping different data sets separate until the end of scaling. The scaling procedure is optimized for cases where the data are collected in a systematic fashion. For both MIR and MAD data, the overall procedure is to construct a reference data set that is as complete as possible and that contains information either from a native data set (for MIR) or for all wavelengths (for MAD data). This reference data set is constructed for just the asymmetric unit of data and is essentially the average of all measurements obtained for each reflection. The reference data set is then expanded to the entire reciprocal lattice and used as the basis for local scaling of each individual data set [see Terwilliger & Berendzen (1999*b*) for additional details].

Once MIR data have been scaled, or MAD data have been scaled and converted to a pseudo-SIRAS form, difference Patterson functions are used to identify plausible one-site or two-site heavy-atom solutions. For MIR data, difference Patterson functions are calculated for each derivative. For MAD data, anomalous and dispersive differences are combined to yield a Bayesian estimate of the Patterson function for the anomalously scattering atoms (Terwilliger, 1994*a*). An automated search of the Patterson function is then used to find a large number (typically 30) of potential single-site and two-site solutions. In principle, Patterson methods could be used to solve the complete heavy-atom substructure, but the approach used in *Solve* is to find just the first one or two heavy-atom sites in this way and to find all others by difference Fourier analysis. This initial set of one-site and two-site solutions becomes the initial list of potential solutions ('seeds') for automated structure solution. Once each of the potential seeds is scored and ranked, the top seeds (typically five) are selected as independent starting points for the search for heavy-atom solutions.

For each starting solution (seed), the main cycle in the automated structure-solution algorithm used by *Solve* consists of two basic steps. The first is to refine heavy-atom parameters and rank all existing solutions generated so far from this seed based on the four criteria discussed above. The second is to take the highest-ranking solution that has not yet been exhaustively analysed and use it in an attempt to generate a more complete solution. Generation of new solutions is carried out in three ways: by deletion of sites, by addition of sites from difference Fouriers, and by inversion. A partial solution is considered to have been exhaustively analysed when all single-site deletions have been considered, when no more peaks in a difference Fourier can be found that improve upon it, and when inversion does not improve it, or when the maximum number of sites input by the user has been reached. In each case, new solutions generated in these three ways are refined, scored and ranked, and the cycle is continued until all the top solutions have been fully analysed and no new solutions are found. Throughout this process, a tally of the solutions that have already been considered is kept, and any time a solution is a duplicate of a previously examined solution it is dropped.

In some cases, one very clear solution appears early in the structure-solution process, while in others, there are several solutions that have similar scores at early (and sometimes even late) stages of structure solution. In cases where no one solution is much better than the others, all the seeds are exhaustively analysed. On the other hand, if a very promising solution emerges from one seed, then the search is narrowed to focus on that seed, deletions are not carried out until the end of the analysis, and many peaks from the difference Fourier analysis are added at a time so as to build up the solution as quickly as possible. Once the expected number of heavy-atom sites are found, then each site is deleted in turn to see if the solution can be further improved. If this occurs, then the new solutions are analysed in the same way by addition and deletion of sites and by inversion until no improvement is obtained.

At the conclusion of the *Solve* algorithm, an electron-density map and phases for the top solution are reported in a form that is compatible with the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). Additionally, command files that can be modified to look for additional heavy-atom sites or to construct other electron-density maps are produced. If more than one possible solution is found, the heavy-atom sites and phasing statistics for all of them are reported.

14.2.2.8. *Generation of model X-ray data sets*

An important feature of *Solve* is the inclusion of modules for the generation of model data. *Solve* can construct model raw X-ray data for either MIR or MAD cases. The macromolecular structure can be defined by a file in PDB format (Bernstein *et al.*, 1977) with heavy-atom parameters defined by the user. Any degree of 'experimental' uncertainty in measurement of intensities can be included, and limited non-isomorphism for MIR data in which cell dimensions differ for native and any of the derivative data sets (but in which the macromolecular structure is identical) can be included. This automatic generation of model data is very useful in evaluating what can and what cannot be solved. Once a data set has been generated, the *Solve* algorithm can be used to attempt to solve it. *Solve* generates a model electron-density map based on the input coordinates, and during the structure-solution process all maps calculated with trial solutions can be compared to the model map. In many cases, heavy-atom solutions can be related to different origins (and to different handedness as well). The origin shift is identified by *Solve* by finding the shift that best maps the trial solution onto the (known) correct solution.

14.2.2.9. *Conclusions*

The *Solve* algorithm is very useful for solving macromolecular structures by the MIR and MAD methods. It has been used to solve MAD structures with as many as 56 selenium atoms in the asymmetric unit (W. Smith & C. Janson, personal communication). From the user's point of view, the algorithm is very simple. Only a few input parameters are needed in most cases, and the *Solve* algorithm carries out the entire process automatically. In principle, the procedure can be very thorough as well, so that many trial starting solutions can be examined and difficult heavy-atom structures can be found. Additionally, for the most difficult structure-solution cases, the failure to find a solution can be useful in confirming that additional information is needed.

14.2.2.10. *Software availability*

The *Solve* software and complete documentation can be obtained from the web site http://solve.lanl.gov.

Acknowledgements

TCT and JB gratefully acknowledge support from the National Institutes of Health and the US Department of Energy.

References

14.1

Bernal, J. D. (1939). *Structure of proteins. Nature (London)*, **143**, 663–667.

Bijvoet, J. M. (1954). *Structure of optically active compounds in the solid state. Nature (London)*, **173**, 888–891.

Blow, D. M. (1957). *X-ray analysis of haemoglobin: determination of phase angles by isomorphous substitution*. PhD thesis, University of Cambridge.

Blow, D. M. (1958). *The structure of haemoglobin. VII. Determination of phase angles in the non-centrosymmetric [100] zone. Proc. R. Soc. London Ser. A*, **247**, 302–336.

Blow, D. M. & Crick, F. H. C. (1959). *The treatment of errors in the isomorphous replacement method. Acta Cryst*. **12**, 794–802.

Blow, D. M. & Rossmann, M. G. (1961). *The single isomorphous replacement method. Acta Cryst.* **14**, 1195–1202.

Bokhoven, C., Schoone, J. C. & Bijvoet, J. M. (1951). *The Fourier synthesis of the crystal structure of strychnine sulphate pentahydrate. Acta Cryst.* **4**, 275–280.

Cork, J. M. (1927). *The crystal structure of some of the alums. Philos. Mag.* **4**, 688–698.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *The structure of haemoglobin. IV. Sign determination by the isomorphous replacement method. Proc. R. Soc. London Ser. A*, **225**, 287–307.

Harker, D. (1956). *The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement. Acta Cryst.* **9**, 1–9.

Hendrickson, W. A. (1979). *Phase information from anomalous-scattering measurements. Acta Cryst.* A**35**, 245–247.

Hendrickson, W. A. (1991). *Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science*, **254**, 51–58.

Kartha, G. & Parthasarathy, R. (1965*a*). *Combination of multiple isomorphous replacement and anomalous dispersion data for protein structure determination. I. Determination of heavy-atom positions in protein derivatives. Acta Cryst.* **18**, 745–749.

Kartha, G. & Parthasarathy, R. (1965*b*). *Combination of multiple isomorphous replacement and anomalous dispersion data for protein structure determination. II. Correlation of the heavy-atom positions in different isomorphous protein crystals. Acta Cryst.* **18**, 749–753.

Matthews, B. W. (1966*a*). *The determination of the position of the anomalously scattering heavy atom groups in protein crystals. Acta Cryst.* **20**, 230–239.

Matthews, B. W. (1966*b*). *The extension of the isomorphous replacement method to include anomalous scattering measurements. Acta Cryst.* **20**, 82–86.

Matthews, B. W. (1970). *Determination and refinement of phases for proteins*. In *Crystallographic computing*, edited by F. R. Ahmed, S. R. Hall & C. P. Huber, pp. 146–159. Copenhagen: Munksgaard.

North, A. C. T. (1965). *The combination of isomorphous replacement and anomalous scattering data in phase determination of non-centrosymmetric reflexions. Acta Cryst.* **18**, 212–216.

Okaya, Y. & Pepinsky, R. (1960). *New developments in the anomalous dispersion method for structure analysis.* In *Computing methods and the phase problem in X-ray crystal analysis*, pp. 273–299. London: Pergamon Press.

Perutz, M. F. (1956). *Isomorphous replacement and phase determination in non-centrosymmetric space groups. Acta Cryst.* **9**, 867–873.

Ramachandran, G. N. & Raman, S. (1956). *A new method for the structure analysis of non-centrosymmetric crystals. Curr. Sci.* **25**, 348–351.

Ramaseshan, S. (1964). *The use of anomalous scattering in crystal structure analysis*. In *Advanced methods of crystallography*, edited by G. N. Ramachandran, pp. 67–95. London: Academic Press.

Rossmann, M. G. (1960). *The accurate determination of the position and shape of heavy-atom replacement groups in proteins. Acta Cryst.* **13**, 221–226.

Rossmann, M. G. (1961). *The position of anomalous scatterers in protein crystals. Acta Cryst.* **14**, 383–388.

Singh, A. K. & Ramaseshan, S. (1966). *The determination of heavy atom positions in protein derivatives. Acta Cryst.* **21**, 279–280.

14.2

Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. (1994). *Structure at 2.8-angstrom resolution of f1-ATPase from bovine heart-mitochondria. Nature (London)*, **370**, 621–628.

Als-Nielsen, J. & McMorrow, D. F. (2001). *Elements of modern X-ray physics*. New York: John Wiley & Sons.

Baker, D., Krukowski, A. E. & Agard, D. A. (1993). *Uniqueness and the ab initio phase problem in macromolecular crystallography. Acta Cryst.* D**49**, 186–192.

Bellizzi, J. J. III, Widom, J., Kemp, C. W. & Clardy, J. (1999). *Producing selenomethionine-labeled proteins with a baculovirus expression vector system. Structure*, **7**, R263–R267.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *Protein data bank: computer-based archival file for macromolecular structures. J. Mol. Biol.* **112**, 535–542.

Bijvoet, J. M. (1949). *Phase determination in direct Fourier-synthesis of crystal structures. Proc. Acad. Sci. Amst.* B**52**, 313–314.

Blow, D. M. (1958). *The structure of haemoglobin. VII. Determination of phase angles in the non-centrosymmetric [100] zone. Proc. R. Soc. London Ser. A*, **247**, 302–335.

Blundell, T. L. & Johnson, L. N. (1976). *Protein crystallography.* p. 368. New York: Academic Press.

Burling, F. T., Weis, W. I., Flaherty, K. M. & Brünger, A. T. (1996). *Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. Science*, **271**, 72–77.

Chang, G. & Lewis, M. (1994). *Using genetic algorithms for solving heavy-atom sites. Acta Cryst.* D**50**, 667–674.