RESOLVE can be applied to other aspects of structure determination as well, suggesting that full automation of the entire structure determination process from scaling diffraction data to a refined model will be possible in the near future.

[3] Automatic Solution of Heavy-Atom Substructures

By Charles M. Weeks, Paul D. Adams, Joel Berendzen,
Axel T. Brunger, Eleanor J. Dodson, Ralf W. Grosse-Kunstleve,
Thomas R. Schneider, George M. Sheldrick, Thomas C. Terwilliger,
Maria G. W. Turkenburg, and Isabel Usón

Introduction

With the exception of small proteins that can be solved by *ab initio* direct methods[1] or proteins for which an effective molecular replacement model exists, protein structure determination is a two-step process. If two or more measurements are available for each reflection with differences arising only from some property of a small substructure, then the positions of the substructure atoms can be found first and used as a bootstrap to initiate the phasing of the complete structure. Historically, substructures were first created by isomorphous replacement in which heavy atoms (usually metals) are soaked into crystals without displacing the protein structure, and measurements were made from both the unsubstituted (native) and substituted (derivative) crystals. When possible, measurements were made also of the anomalous diffraction generated by the metals at appropriate wavelengths. Now, it is common to incorporate anomalous scatterers such as selenium into proteins before crystallization and to make measurements of the anomalous dispersion at multiple wavelengths.

The computational procedures that can be used to solve heavy-atom substructures include both Patterson-based and direct methods. In either case, the positions of the substructure atoms are determined from difference coefficients based on the measurements available from the diffraction experiments as summarized in Table I. The isomorphous difference magnitude, $|\Delta F|$ iso $(=||F_{\mathrm{PH}}|-|F_{\mathrm{P}}||)$, approximates the structure amplitude, $|F_{\mathrm{H}} \cos(\delta)|$, and the anomalous-dispersion difference magnitude, $|\Delta F|$ ano

[1] G. M. Sheldrick, H. A. Hauptman, C. M. Weeks, R. Miller, and I. Usón, *In* "International Tables for Crystallography" (M. G. Rossmann and E. Arnold, eds.), Vol. F, p. 333. Kluwer Academic, Dordrecht, The Netherlands, 2001.

0076-6879/03 $35.00

TABLE I
MEASUREMENTS USED FOR SUBSTRUCTURE DETERMINATION[a]

Acronym	Type of experiment	Measurements
SIR	Single isomorphous replacement	F_P, F_{PH}
SIRAS	Single isomorphous replacement with anomalous scattering	F_P, $F_{PH}+$, $F_{PH}-$
MIR	Multiple isomorphous replacement	F_P, F_{PH1}, F_{PH2}, ...
MIRAS	Multiple isomorphous replacement with anomalous scattering	F_P, $F_{PH1}+$, $F_{PH1}-$, $F_{PH2}+$, $F_{PH2}-$, ...
SAD or SAS	Single anomalous dispersion or single anomalous scattering	$F_{PH}+$, $F_{PH}-$ at one wavelength
MAD	Multiple anomalous dispersion	$F_{PH}+$, $F_{PH}-$ at several wavelengths

[a] The notation used for the structure factors is F_P (native protein), F_{PH} (derivative), F_H or F_A (substructure), $F+$ and $F-$ (for F_{hkl} and $F_{\overline{hkl}}$, respectively, in the presence of anomalous dispersion).

$(=\|\,F+|-|F-\|)$, approximates $2|F''_H \sin(\delta)|$. (The angle δ is the difference between the phase of the whole protein and that of the substructure.) When SIRAS or MAD data are available, the differences can be combined to give an estimate of the complete F_A structure factor.[2,3]

Both Patterson and direct methods require extremely accurate data for the successful determination of substructures. Care should be taken to eliminate outliers and observations with small signal-to-noise ratios, especially in the case of single anomalous differences. Fortunately, it is usually possible to be stringent in the application of appropriate cutoffs because the problem is overdetermined in the sense that the number of available observations is much larger than the number of heavy-atom positional parameters. In particular, it is important that the largest isomorphous and anomalous differences be reliable. The coefficients that are used consider small differences between two or more much larger measurements, so errors in the measurements can easily disguise the true signal. If there are even a few outliers in a data set, or some of the large coefficients are serious overestimates, substructure determination is likely to fail.

Patterson and direct-methods procedures have been implemented in a number of computer programs that permit even large substructures to be determined with little, if any, user intervention. (The current record is 160 selenium sites.) The methodology, capabilities, and use of several such

[2] J. Karle, *Acta Crystallogr. A* **45,** 303 (1989).
[3] W. Hendrickson, *Science* **254,** 51 (1991).

popular programs and program packages are described in this chapter. The SOLVE[4] program, which uses direct-space Patterson search methods to locate the heavy-atom sites, provides a fully automated pathway for phasing protein structures, using the information obtained from MIR or MAD experiments. The two major software packages currently in use in macromolecular crystallography [i.e., the Crystallography and NMR System (CNS[5]) and the Collaborative Computational Project Number 4 (CCP4[6])] provide internally consistent formats that make it easy to proceed from heavy-atom sites to density map, but user intervention is required. CNS employs both direct-space and reciprocal-space Patterson searches. The CCP4 suite includes programs for computing Pattersons as well as the direct-method programs RANTAN[7] and ACORN.[8] The dual-space direct-method programs SnB[9,10] and SHELXD[11,11a] provide only the heavy-atom sites, but they are efficient and capable of solving large substructures currently beyond the capabilities of programs that use only Patterson-based methods. SnB uses a random number generator to assign initial positions to the starting atoms in its trial structures, but SHELXD strives to obtain better-than-random initial coordinates by deriving information from the Patterson superposition minimum function. In some cases, this has significantly decreased the computing time needed to find a heavy-atom solution. Other direct-method programs (e.g., SIR2000[12]), not described in this chapter, also can be used to solve substructures.

Pertinent aspects of data preparation are described in detail in the following sections devoted to the individual programs. Automated or semi-automated procedures for locating heavy-atom sites operate by generating many trial structures. Thus, a key step in any such procedure is the scoring or ranking of trial structures by some measure of quality in such a way that

[4] T. C. Terwilliger and J. Berendzen, *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 849 (1999).

[5] A. T. Brunger, P. D. Adams, G. M. Clore, W. L. DeLano, P. Gross, R. W. Grosse-Kunstleve, J.-S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren, *Acta Crystallogr. D. Biol. Crystallogr.* **54,** 905 (1998).

[6] Collaborative Computational Project Number 4, *Acta Crystallogr. D. Biol. Crystallogr.* **50,** 760 (1994).

[7] J.-X. Yao, *Acta Crystallogr. A* **39,** 35 (1983).

[8] J. Foadi, M. M. Woolfson, E. J. Dodson, K. S. Wilson, J.-X. Yao, and C.-D. Zheng, *Acta Crystallogr. D. Biol. Crystallogr.* **56,** 1137 (2000).

[9] R. Miller, S. M. Gallo, H. G. Khalak, and C. M. Weeks, *J. Appl. Crystallogr.* **27,** 613 (1994).

[10] C. M. Weeks and R. Miller, *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 492 (1999).

[11] G. M. Sheldrick, *in* "Direct Methods for Solving Macromolecular Structures" (S. Fortier, ed.), p. 401. Kluwer Academic, Dordrecht, The Netherlands, 1998.

[11a] T. R. Schneider and G. M. Sheldrick, *Acta Crystallogr. D. Biol. Crystallogr.* **58**, 1772 (2002).

[12] M. C. Burla, M. Camalli, B. Carrozzini, G. L. Cascarano, C. Giacovazzo, G. Polidori, and R. Spagna, *Acta Crystallogr. A* **56,** 451 (2000).

any probable solution can be identified. Therefore, the methods used to accomplish this are described for each program, along with methods for validating the correctness of individual sites. Where applicable, methods used to determine the correct hand (enantiomorph) and refine the substructure also are described. Finally, interesting applications to large selenomethionine derivatives, substructures phased by weak anomalous signals, and substructures created by short halide cryosoaks are discussed.

SOLVE

In favorable cases, the determination of heavy-atom substructures using MAD or MIR data is a straightforward, although often lengthy, process. SOLVE[4] is designed to automate fully the analysis of such data. The overall approach is to link together into one seamless procedure all the steps that a crystallographer would normally do manually and, in the process, to convert each decision-making step into an optimization problem. A somewhat more generalized description of SOLVE, together with a description of RESOLVE, a maximum-likelihood solvent-flattening routine, appear in the chapter by T. Terwilliger (see [2] in this volume[12a]).

The MAD and MIR approaches to structure solution are conceptually similar and share several important steps. In each method, trial partial structures for the heavy or anomalously scattering atoms often are obtained by inspection of difference-Patterson functions or by semiautomated analysis.[13–15] These initial structures are refined against the observed data and used to generate initial phases. Then, additional sites and sites in other derivatives can be found from weighted difference or gradient maps using these phases. The analysis of the quality of potential heavy-atom solutions is also similar for the two methods. In both cases, a partial structure is used to calculate native phases for the entire structure, and the electron density that results is then examined to see whether the expected features of the macromolecule can be found. In addition, the figure of merit of phasing and the agreement of the heavy atom model with the difference Patterson function 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 also an important criterion of quality.[16]

[12a] T. C. Terwilliger, *Methods Enzymol.* **374,** [2], 2003 (this volume).

[13] T. C. Terwilliger, S.-H. Kim, and D. Eisenberg, *Acta Crystallogr. A* **43,** 1 (1987).

[14] G. Chang and M. Lewis, *Acta Crystallogr. D. Biol. Crystallogr.* **50,** 667 (1994).

[15] A. Vagin and A. Teplyakov, *Acta Crystallogr. D. Biol. Crystallogr.* **54,** 400 (1998).

[16] R. E. Dickerson, J. C. Kendrew, and B. E. Strandberg, *Acta Crystallogr.* **14,** 1188 (1961).

Data Preparation

SOLVE prepares data for heavy-atom substructure solution in two steps. First, the data are scaled using the local scaling procedure of Matthews and Czerwinski.[17] Second, MAD data are converted to a pseudo-SIRAS form that permits more rapid analysis.[18]

Systematic errors are minimized by scaling all types of data (e.g., $F+$ and $F-$, native and derivative, and the different wavelengths of MAD data) in similar ways and by keeping different data sets separate until the end of scaling. The scaling procedure is optimized for cases in which 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 from either 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 and Berendzen[4] for additional details).

For MAD data, Bayesian calculations of phase probabilities are slow.[19,20] Consequently, SOLVE uses an alternative procedure for all MAD phase calculations except those done at the final stage. This alternative is to convert the multiwavelength MAD data set into a form that is similar to that used for SIRAS data. The information in a MAD experiment is largely contained in just three quantities: a structure factor $F\text{o}$ corresponding to the scattering from nonanomalously scattering atoms, a dispersive or isomorphous difference at a standard wavelength λo ($\Delta_{\lambda\text{o}}^{\text{ISO}}$), and an anomalous difference ($\Delta_{\lambda\text{o}}^{\text{ANO}}$) at the same standard wavelength.[18] It is easy to see that these three quantities could be treated just like an SIRAS data set with the "native" structure factor F_{P} replaced by $F\text{o}$, the derivative structure factor F_{PH} replaced by $F\text{o} + (\Delta_{\lambda\text{o}}^{\text{ISO}})$, and the anomalous difference replaced by $\Delta_{\lambda\text{o}}^{\text{ANO}}$. 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.[21] 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

[17] B. W. Matthews and E. W. Czerwinski, *Acta Crystallogr. A* **31,** 480 (1975).

[18] T. C. Terwilliger, *Acta Crystallogr. D. Biol. Crystallogr.* **50,** 17 (1994).

[19] T. C. Terwilliger and J. Berendzen, *Acta Crystallogr. D. Biol. Crystallogr.* **53,** 571 (1997).

[20] E. de la Fortelle and G. Bricogne, *Methods Enzymol.* **277,** 472 (1997).

[21] T. C. Terwilliger and D. Eisenberg, *Acta Crystallogr. A* **43,** 6 (1987).

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 assumption is that the structure factor corresponding to anomalously scattering atoms is small compared with the structure factor from all other atoms.

The conversion of MAD to pseudo-SIRAS data is implemented in the program segment MADMRG.[18] In most cases, there is more than one pair of X-ray wavelengths corresponding to a particular reflection. The estimates from each pair of wavelengths are all 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 different weights in their contributions to the total. This can be understood by noting that pairs of wavelengths that differ considerably in dispersive contributions would yield relatively accurate estimates of $\Delta_{\lambda o}^{\mathrm{ISO}}$. In the same way, Bijvoet differences measured at the wavelength with the largest value of f'' will contribute by far the most to estimates of $\Delta_{\lambda o}^{\mathrm{ANO}}$. 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, although it is convenient to choose one of them.

Heavy-Atom Searching and Phasing

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.

To make automated structure solution practical, it is necessary to evaluate trial heavy-atom solutions (typically 300–1000) rapidly. For each potential solution, the heavy-atom sites must be refined and the 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 intermediate refinements and phase calculations. The refinement method chosen on this basis was origin-removed Patterson refinement,[22] which treats each derivative in an MIR data set independently, and which is fast because it does not require phase calculation. The phasing approach used for MIR data throughout SOLVE is Bayesian-correlated phasing,[21,23] a method that takes into account the correlation of nonisomorphism among derivatives without slowing down phase calculations substantially.

Once MIR data have been scaled, or MAD data have been scaled and converted to a pseudo-SIRAS form, automated searches of difference Patterson functions are then used to find a large number (typically 30) of potential one-site and two-site solutions. In the case of 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.[24] 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 initial sites in this way and to find all others by difference Fourier analysis. This initial set of one-site and two-site trial solutions becomes a list of ‘‘seeds’’ for further searching. Once each of the potential seeds is scored and ranked, the top seeds (typically five) are selected as independent starting points in the search for heavy-atom solutions.

For each 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 to rank all existing solutions generated from this seed so far, on the basis of the four criteria discussed below. The second is to take the highest-ranking partial solution that has not yet been analyzed exhaustively 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 reversal of hand. A partial solution is considered to have been analyzed exhaustively when all single-site deletions have been considered, when no more peaks that result in improvement can be found in a difference Fourier, when inversion does not cause improvement, or when the maximum number of sites specified by the user has been reached. In each case, new solutions generated in these ways are refined, scored, and ranked, and the cycle is continued until all the top trial solutions have been analyzed fully and no new possibilities are found. Throughout this process, a tally of the

[22] T. C. Terwilliger and D. Eisenberg, *Acta Crystallogr. A* **39,** 813 (1983).

[23] T. C. Terwilliger and J. Berendzen, *Acta Crystallogr. D. Biol. Crystallogr.* **52,** 749 (1996).

[24] T. C. Terwilliger, *Acta Crystallogr. D. Biol. Crystallogr.* **50,** 11 (1994).

solutions that have already been considered is kept, and any duplicates are eliminated.

In some cases, one clear solution appears early in this process. In other cases, there are several solutions that have similar scores at early (and sometimes even late) stages of the analysis. When no one possibility is much better than the others, all the seeds are analyzed exhaustively. On the other hand, if a promising partial 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 simultaneously so as to build up the solution as quickly as possible. Once the expected number of heavy-atom sites is found, then each site is deleted in turn to see whether the solution can be further improved. If this occurs, then the process is repeated in the same way by addition and deletion of sites and by inversion until no further 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 CCP4[6] suite. In addition, 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.

Scoring, Site Validation, Enantiomorph Determination, and Substructure Refinement

Scoring of potential heavy-atom solutions is an essential part of the SOLVE algorithm because it allows ranking of solutions and appropriate decision-making. Scoring, validation, and enantiomorph determination are all part of the same process, and they are carried out continuously during the solution process. For each trial solution, SOLVE first refines the heavy-atom substructure against the origin-removed Patterson function. Then, it scores the trial solutions using four criteria that are described in detail below: agreement with the Patterson function, cross-validation of heavy-atom sites, the figure of merit, and nonrandomness 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 a so-called *Z* score. The total score for a solution is the sum of its *Z* scores after correction for anomalously high scores in any category. SOLVE identifies the enantiomorph, using the score for the nonrandomness criterion. All the other scores are independent of the hand of the heavy-atom substructure, but the final electron-density map will be just noise if anomalous differences are measured and the hand of the heavy atoms is incorrect.

Consequently, this score can be used effectively in later stages of structure solution to identify the correct enantiomorph.

Patterson Agreement. 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.[25] The score for Patterson function agreement is the average value of the Patterson function at predicted peak locations after multiplication by a weighting factor based on the number of heavy-atom sites in the trial solution. The weighting factor[4] is adjusted such that, if two solutions have the same mean value at predicted Patterson peaks, the one with the larger number of sites receives the higher score. In some cases, predicted Patterson vectors fall on high peaks that are not related to the heavy-atom solution. To exclude these contributions, the 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 larger than their predicted values are truncated. The average values are corrected further for instances in which more than one predicted Patterson vector falls at the same location by scaling that peak height by the fraction of predicted vectors that are unique.

Cross-Validation of Sites. A cross-validation difference Fourier analysis is the basis of the second scoring criterion. 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 the 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, in which a derivative is omitted from phasing and all other derivatives are used to phase a difference Fourier, has been used for many years.[16] The score for cross-validation difference Fouriers is the average peak height after weighting by the same factor used in the difference Patterson analysis.

Figure of Merit. The mean figure of merit of phasing, m,[25] can be a remarkably useful measure of the quality of phasing despite its susceptibility to systematic error.[4] The overall figure of merit is essentially a measure of the internal consistency of the heavy-atom solution with the data. Because heavy-atom refinement in SOLVE is carried out using origin-removed Patterson refinement,[22] occupancies of heavy-atom sites are relatively unbiased. This minimizes the problem of high occupancies leading to inflated figures of merit. In addition, using a single procedure for phasing allows

[25] T. L. Blundell and L. N. Johnson, ‘‘Protein Crystallography.’’ Academic Press, New York, 1976.

comparison among solutions. The score based on figure of merit is simply the unweighted mean for all reflections included in phasing.

Nonrandomness of Electron Density. 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 this criterion is difficult, it is quite straightforward to evaluate instead whether the electron-density map has general 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,[26] the presence of clearly defined regions of protein and solvent,[27–33] and histogram matching of electron densities.[31,34] The identification of solvent and protein regions has been used as the measure of map quality in SOLVE. This requires that there be both solvent and protein regions in the electron-density map. Fortunately, 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 criteria used in scoring by SOLVE are based on the solvent and protein regions each being fairly large, contiguous regions.[33] The unit cell is divided into boxes having each dimension approximately twice the resolution of the map, and the root–mean–square (rms) electron density is calculated within each box without including the F_{000} term in the Fourier synthesis. Boxes within the protein region will typically have high values of this rms electron density (because there will be some points where atoms are located and other points that lie between atoms) whereas boxes in the solvent region will have low values because the electron density will be fairly uniform. The score, based on the connectivity of the protein and solvent regions, is simply the correlation coefficient of the 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. If the electron density is random, there will be little or no correlation. On the other hand, the correlation may be as high as 0.5 or 0.6 for a good map.

[26] D. Baker, A. E. Krukowski, and D. A. Agard, *Acta Crystallogr. D. Biol. Crystallogr.* **49,** 186 (1993).

[27] B.-C. Wang, *Methods Enzymol.* **115,** 90 (1985).

[28] S. Xiang, C. W. Carter, Jr., G. Bricogne, and C. J. Gilmore, *Acta Crystallogr. D. Biol. Crystallogr.* **49,** 193 (1993).

[29] A. D. Podjarny, T. N. Bhat, and M. Zwick, *Annu. Rev. Biophys. Biophys. Chem.* **16,** 351 (1987).

[30] J. P. Abrahams, A. G. W. Leslie, R. Lutter, and J. E. Walker, *Nature* **370,** 621 (1994).

[31] K. Y. J. Zhang and P. Main, *Acta Crystallogr. A* **46,** 377 (1990).

[32] T. C. Terwilliger and J. Berendzen, *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 501 (1998).

[33] T. C. Terwilliger and J. Berendzen, *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 1872 (1999).

[34] A. Goldstein and K. Y. J. Zhang, *Acta Crystallogr. D. Biol. Crystallogr.* **54,** 1230 (1998).

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 have been critical. 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 the calculations required for heavy-atom searching, refinement, and phasing as well as routines to keep track of the lists of current solutions being examined and past solutions that have already been tested.

SOLVE is an easy program to use. Only a few input parameters are needed in most cases, and the SOLVE algorithm carries out the entire process automatically. In principle, the procedure also can be thorough: many starting solutions can be examined, and difficult heavy-atom structures can be determined. In addition, for the most difficult cases, the failure to find a solution can be useful in confirming that additional information is needed.

Crystallography and NMR System

The Crystallography and NMR System (CNS)[5] implements a novel Patterson-based method for the location of heavy atoms or anomalous scatterers.[35] The procedure is implemented using a combination of direct-space and reciprocal-space searches, and it can be applied to both isomorphous replacement and anomalous scattering data. The goal of the algorithm is to make it practical to locate automatically a subset of the heavy atoms without manual interpretation or intervention. Once the sites have been located, CNS provides tools for heavy-atom refinement, phase estimation, density modification, and heavy-atom model completion. These tools, known as task files, are scripts written in the CNS language and are supplied with reasonable default parameters. Using these task files, the process of phasing is greatly simplified and initial electron-density maps, even for large complex structures, can be calculated in a relatively short time. CNS has been used successfully to solve problems with up to 40[36] and 66 selenium sites (see Applications, below).

Data Preparation

Sigma Cutoffs and Outlier Elimination. The peaks in a Patterson map correspond to interatomic vectors of the crystal structure.[37] However, the

[35] R. W. Grosse-Kunstleve and A. T. Brunger, *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 1568 (1999).

[36] M. A. Walsh, Z. Otwinowski, A. Perrakis, P. M. Anderson, and A. Joachimiak, *Struct. Fold. Des.* **8,** 505 (2000).

atoms are not point scatterers, and there are errors associated with experimental data, making the interpretation of the Patterson map difficult. Therefore, steps are taken to minimize the amount of error that is introduced. In practice, the suppression of outliers can be essential to the success of a heavy atom search.[38] In CNS, reflections are first rejected on the basis of their signal-to-noise ratio (‘‘sigma cutoff’’). This is performed on both the observed amplitudes and the computed difference between pairs of amplitudes. For the computation of differences, the observed amplitudes are scaled relative to each other, using overall k-scaling and B-scaling in order to compensate for systematic errors caused by differences between crystals and data collection conditions. Additional reflections are rejected if their amplitudes or difference amplitudes deviate too much from the corresponding root–mean–square (rms) value for all of the data in their resolution shell (‘‘rms outlier removal’’). Empirical observation has led to the values of the rejection criteria shown in Table II. Except for the

TABLE II
DEFAULT PARAMETERS FOR CNS AUTOMATED HEAVY-ATOM SEARCH PROCEDURE

Parameter	Default value[a]	Comment[b]
Number of sites	2/3 of total expected	Typically not all sites are well ordered, and it is easy to add additional sites using gradient map methods once phasing has started with the 2/3 partial solution
Minimum Bragg spacing	4.0 Å	If there are a large number of heavy-atom sites per macromolecule, a higher resolution limit may be required (3.5 Å)
Averaging of Patterson maps	No	If solutions are not found with a single map, then multiple maps can be tried
Special positions	No	Can be set to true if the heavy atoms have been soaked into the crystal
Sigma cutoff on F	1	Decrease to 0 for F_A structure factors
RMS outlier cutoff on F for native or on ΔF for difference Patterson maps	4	Increase to 10 for F_A structure factors
Expected increase in correlation coefficient for dead-end test	0.01	When there are a large number of heavy-atom sites, it may be necessary to decrease this value (to 0.005)

[a] Values present in the heavy_search.inp task file supplied with CNS.
[b] Situations in which the default parameter may require modification.

[37] M. J. Buerger, ‘‘Vector Space.’’ John Wiley & Sons, New York, 1959.

[38] G. M. Sheldrick, *Methods Enzymol.* **276,** 628 (1997).

instances noted in Table II, these values can generally be used without modification.

Combining Patterson Maps. CNS provides the option to average Patterson maps based on different data sets. For example, several MAD wavelengths or a combination of isomorphous and anomalous difference maps can be combined. This is useful if the signal in any individual data set is too weak to locate the heavy atoms unambiguously. A small signal-to-noise ratio in the observed data leads to noise in the Patterson maps. The combination of data increases the signal-to-noise ratio in the resulting Patterson map by averaging out the noise and, therefore, improves the chances of locating the heavy-atom positions (Fig. 1d).

Using F_A Structure Factors. If MAD data are available, it is possible to define structure factors F_A that are approximations to the component of the observed structure factors resulting from the anomalous scatterers.[2,3,18] F_A structure factors can be calculated using programs such as XPREP,[39] MADSYS,[3] or the MADBST module of SOLVE.[4] Although CNS does not perform F_A estimation, the heavy-atom search procedure can make use of this information and that has been found to increase the chances for locating the correct sites (Fig. 1e). Ideally, an algorithm for the estimation of F_A structure factors includes a careful treatment of outliers similar to the sigma cutoff and rms outlier removal outlined above. If this is the case, the parameters for the sigma cutoff and rms outlier removal in CNS should be adjusted to include all data in the heavy-atom search procedure (see Table II).

Heavy-Atom Searching

The CNS heavy-atom search procedure (Fig. 2) consists of four stages that are described in more detail by Grosse-Kunstleve and Brunger.[35] In the first stage, the observed diffraction intensities are filtered by the criteria described above, and two or more Patterson maps (calculated from MIR, MAD, or MIRAS data) can be averaged. The second stage consists of a Patterson search by either a reciprocal-space single-atom fast translation function, by a direct-space symmetry minimum function, or by a combination of both. Combination searches have been shown to be the most accurate.[35] A given number (typically 100) of the 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 as well as Patterson-correlation

[39] Written by G. Sheldrick. Available from Bruker Advanced X-Ray Solutions (Madison, WI).

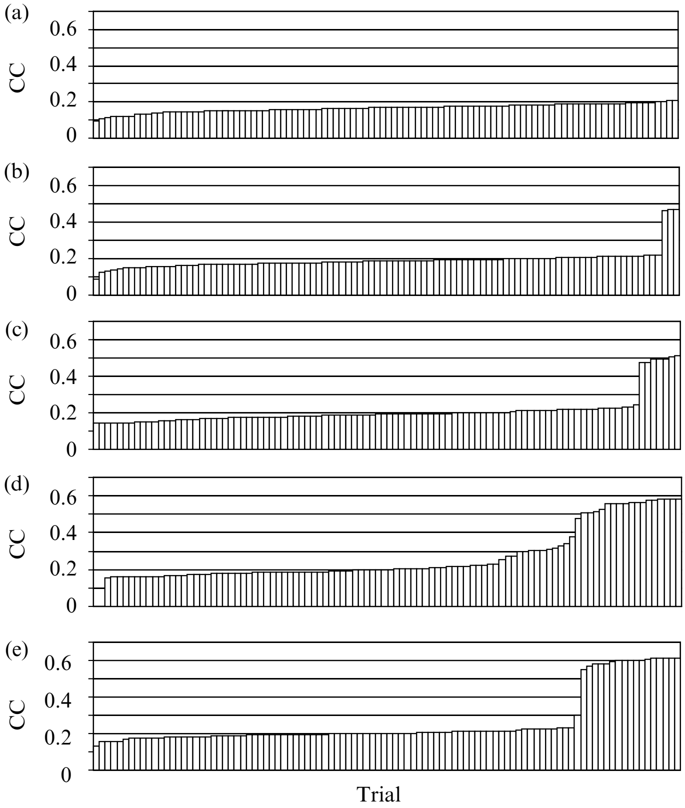


FIG. 1. Results of automated CNS heavy-atom search with the MAD data from 2-aminoethylphosphonate transaminase. Sixty-six selenium sites are present in the asymmetric unit. Automated searches for 44 sites (two-thirds of the expected total) were performed. In all cases, 100 trial solutions were generated and sorted by the correlation coefficient (F2F2). (a) No solutions were found using the anomalous ΔF structure factors at the high-energy remote wavelength as indicated by no separation between the trials. (b) A few solutions were found using the anomalous ΔF structure factors at the peak wavelength. (c) The anomalous ΔF structure factors at the inflection-point wavelength found more solutions, indicating a larger anomalous signal than the peak wavelength. (d) Using combined anomalous ΔF structure factors at the inflection-point wavelength and the dispersive differences between the inflection point and high-energy remote gave an even higher success rate. (e) Finally, the greatest success rate was with F_A structure factors calculated from all three wavelengths, using XPREP.[39]

(PC) refinements[40] starting with each of the initial trial sites. The highest peak is selected that has distances to its symmetrically equivalent points and all preexisting sites larger than the given cutoff distance. If two or more sites already have been placed, a dead-end elimination test is performed.

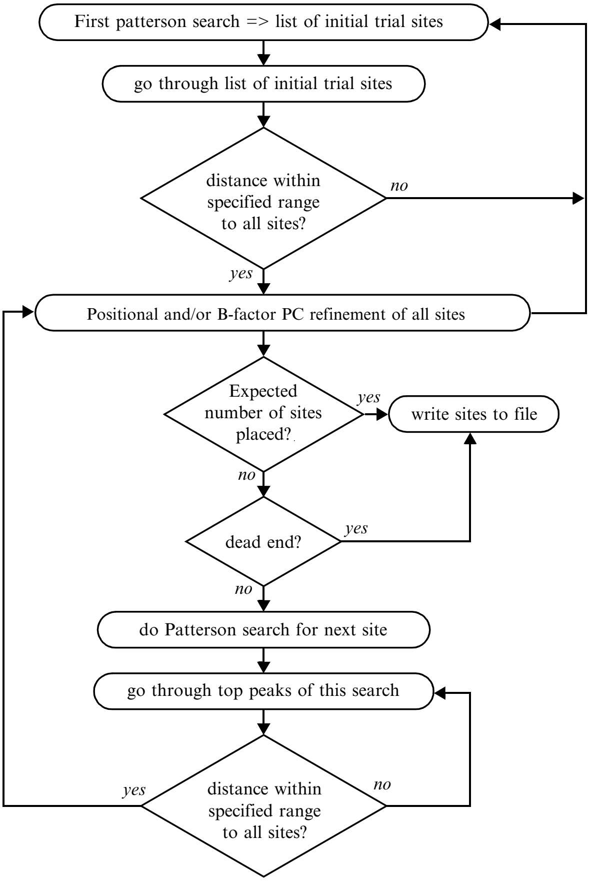

First patterson search => list of initial trial sites
go through list of initial trial sites
distance within specified range to all sites?
no
yes
Positional and/or B-factor PC refinement of all sites
Expected number of sites placed?
yes
write sites to file
no
dead end?
yes
no
do Patterson search for next site
go through top peaks of this search
distance within specified range to all sites?
yes
no

FIG. 2. CNS automated heavy-atom location protocol.

The correlation coefficient computed before placing and refining the last new site is compared with the correlation coefficient computed after the addition of the new site. If the target value does not increase by a specified amount, typically 0.01 (see Table II), then 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 is found. The final stage consists of sorting the solutions ranked by the value of the target function (a correlation coefficient)

[40] A. T. Brunger, *Acta Crystallogr. A* **47,** 195 (1991).

of the PC refinement. If the correct solution has been found, it is normally characterized by the best value of the target function and a significant separation from incorrect solutions (compare, e.g., Fig. 1a and b).

Reciprocal-Space Method: Single-Atom Fast Translation Function. A single heavy-atom site is translated throughout an asymmetric unit, and the standard linear correlation coefficient of F^2_{patt} and $F^2_{\text{calc}}(t)$ (referred to as F2F2) is computed for each position t:

$$F2F2(t) = \frac{\sum\limits_H (F^2_{H,\text{patt}} - \langle F^2_{\text{patt}}\rangle)(F^2_{H,\text{calc}} - \langle F^2_{\text{calc}}\rangle)}{\sqrt{\sum\limits_H (F^2_{H,\text{patt}} - \langle F^2_{\text{patt}}\rangle)^2}\sqrt{\sum\limits_H (F^2_{H,\text{calc}} - \langle F^2_{\text{calc}}\rangle)^2}} \qquad (1)$$

The summations are computed for all Miller indices H, and $\langle F^2\rangle$ denotes the mean of F^2 over all Miller indices. Other target expressions can be used including the correlation coefficient between F_{patt} and $F_{\text{calc}}(t)$, E^2_{patt} and $E^2_{\text{calc}}(t)$, and E_{patt} and E_{patt} and $E_{\text{calc}}(t)$, where the E values are normalized structure factors (see Dual-Space Direct Methods, below). The F2F2 target function is preferred because it permits the use of a fast translation function (FTF),[41] which is 300–500 times faster[35] than the conventional translation function.[42] Thus, the FTF makes the automated reciprocal-space heavy-atom search procedure practical even for large numbers of sites. The reciprocal-space search for an additional site is similar to the search for the initial trial sites, except that the previously placed sites are kept fixed and are included in the structure-factor (F_{calc}) calculation.[41]

Direct-Space Method: Symmetry and Image-Seeking Minimum Functions. The symmetry minimum function (SMF)[43–45] makes maximal use of the information contained in the Harker regions. The computation of an SMF requires a Patterson map as well as a table of the unique Harker vectors and their weights.[43] These Harker vectors and weights are supplied automatically by CNS. The image-seeking minimum function (IMF)[43,45] can be used to locate additional sites once one or more are placed. Computing an IMF map is equivalent to a deconvolution of the Patterson map using knowledge of the already placed heavy-atom sites. Because of coincidental overlap of peaks in the Patterson map, thermal motion of the sites, and noise in the data, the IMF maps typically provide only limited information for macromolecular crystal structures.

[41] J. Navaza and E. Vernoslova, *Acta Crystallogr. A* **51,** 445 (1995).

[42] M. Fujinaga and R. J. Read, *J. Appl. Crystallogr.* **20,** 517 (1987).

[43] P. G. Simpson, R. D. Dobrott, and W. N. Lipscomb, *Acta Crystallogr.* **18,** 169 (1965).

[44] F. Pavelcik, *J. Appl. Crystallogr.* **19,** 488 (1986).

[45] M. A. Estermann, *Nucl. Instr. Methods Phys. Res. A* **354,** 126 (1995).

Peak Search and Special Position Check. The list of initial trial sites is determined by a peak search in the single-atom FTF, the SMF, or their combination. 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. Redundancies due to space-group symmetry and allowed origin shifts are automatically removed. Similarly, additional sites are determined by a peak search in the FTF, the IMF, or their combination. The treatment of redundancies due to symmetry is fully integrated into the search procedure.

Sites at or close to a special position can be accepted or rejected. In the latter case, the shortest distance to all its symmetry equivalent sites is computed for each of the trial sites. If this distance is less than a given cutoff distance (typically 3.5 Å), the site is rejected. Because selenomethionine substitution is the predominant technique for introducing anomalous scatterers into a macromolecule, the rejection of peaks on special positions is set to be the default. However, if heavy atoms have been soaked, cocrystallized, or chemically reacted with the macromolecule, a site could be located on a special position. In such cases, it is appropriate to search for heavy atoms first with special positions rejected and then with them accepted in order to determine whether further sites are found.

Scoring Trial Structures

The result of the CNS heavy-atom search is a number of trial solutions, each containing up to the specified maximum number of sites. There are typically as many of these trial solutions as were requested by the user before running the heavy_search.inp task file. However, when the input Patterson map has only a small number of peaks, it is possible that there will be fewer trial solutions found. The trial solutions can be ranked by the scoring function (which is typically F2F2, the correlation between the squared amplitudes), but other score functions can be used. Although the absolute value of the correlation coefficient could be used as a guide to the correctness of each trial solution, empirical observation has shown that a more informative guide is the presence of solutions with correlation coefficients that are outstanding compared with the rest (Fig. 1). Similar observations have also been made by the authors of other automatic programs for locating heavy atoms.[9]

The heavy_search.inp task file creates a list file (heavy_search.list) that contains an unsorted list of the score function for each trial solution. Each solution with a correlation score that is 1.5σ above the mean of all the solutions is marked with a plus sign (+). To interpret the results easily, the list of configurations can be sorted by correlation coefficient and then plotted graphically (Fig. 1). In the majority of cases encountered to date, if the solution with the highest correlation is also more than 1.5σ above the mean, then all or most of the heavy-atom positions in that solution are correct.

Substructure Refinement, Site Validation, and Enantiomorph Determination

The trial solutions produced by the automated heavy-atom search are used to determine initial phases to generate an electron-density map. Several different tasks must be performed in order to refine the heavy-atom substructure, calculate phases, complete the heavy-atom model, resolve the enantiomorph, and possibly resolve phase ambiguities. A similar approach is followed for MAD, SAD, and (M/S)IR(AS) experiments. In all cases, the following methods are employed.

Substructure Refinement. The heavy-atom sites located automatically with CNS are refined and phase probability distributions generated using the ir_phase.inp or mad_phase.inp task files that deal with isomorphous replacement and anomalous diffraction, respectively. A generalized phase refinement formulation is used when lack-of-closure expressions are calculated between a user-selected reference data set and all other data sets.[46,47] A maximum-likelihood target function[47] is employed that makes use of an error model similar to that of Terwilliger and Eisenberg.[21] Coordinates, B-factors and, when appropriate, occupancies are refined using the Powell conjugate gradient minimization algorithm.[48]

Site Validation. The heavy-atom positions are not extensively validated during the search procedure; instead, the refinement of B-factors during each cycle decreases the contribution from incorrect sites. After phase calculation, the gradient map technique is used to validate the existing sites further, and also to detect sites missing from the current model.[49] The gradient map is a Fourier synthesis calculated from the first derivative of the phasing target function, which can be interpreted as a difference map. A positive peak, clearly separated from any existing atom, corresponds to an atom missing from the heavy-atom model whereas a negative peak, located at the position of an existing atom, indicates that this atom is either incorrectly placed or has been assigned an incorrect chemical type or occupancy. Anisotropic motion of atoms in the substructure also can lead to peaks in the gradient map close to existing sites.

Enantiomorph Determination. The use of the gradient map method in combination with substructure refinement allows the heavy-atom model

[46] J. C. Phillips and K. O. Hodgson, *Acta Crystallogr. A* **36,** 856 (1980).
[47] F. T. Burling, W. I. Weis, K. M. Flaherty, and A. T. Brunger, *Science* **271,** 72 (1996).
[48] M. J. D. Powell, *Math. Program.* **12,** 241 (1977).
[49] G. Bricogne, *Acta Crystallogr. A* **40,** 410 (1984).

to be completed even though the correct hand of the heavy-atom configuration is often still unknown. In CNS, the correct hand is determined by repeating the phase determination with the alternate hand followed by inspection of the two electron-density maps (see below). In the majority of cases, obtaining the alternative hand is achieved simply by inverting the coordinates about the origin. However, in the case of enantiomorphic space groups, the space group must be changed at the same time as the coordinates are inverted (e.g., $P6_1$ is mapped to $P6_5$). In addition, in a small number of space groups, the inversion of the coordinates is not about the origin, but rather some other point in the unit cell. The CNS task file flip_sites.inp automatically takes account of both of these situations.

Once phasing has been performed with the two possible choices of heavy-atom coordinates, the electron-density maps can be compared to determine which hand is correct. Making this decision from the raw experimental phases is feasible only with high-quality MIR(AS) or MAD data sets. In such cases, the solvent boundary, secondary structure elements, or atomic detail in the electron-density map can show clearly which heavy-atom configuration is correct. However, in the general case the raw experimental phases are not sufficient to reveal such features. In particular, in the case of a single anomalous diffraction (SAD) or a single isomorphous replacement (SIR) experiment, it is not possible to distinguish the two hands in this way because of the bimodal phase distributions that are produced. Therefore, it is usually better to perform phase improvement by density modification in the form of solvent flattening or solvent flipping[50] to resolve the phase ambiguity present in the SAD and SIR cases. The CNS task file density_modify.inp should be used to improve the phases irrespective of the type of phasing experiment. After density modification of phases from both heavy-atom hands, the electron-density maps usually identify the correct hand unambiguously and generate maps good enough to begin model building.

Dual-Space Direct Methods: SnB and SHELXD

Direct methods are techniques that use probabilistic relationships among the phases to derive values of the individual phases from the measured amplitudes. The purpose of this section is to give a concise summary of these techniques as they apply to substructure determination. The basic theory underlying direct methods,[51] as well as macromolecular applications

[50] J. P. Abrahams and A. G. W. Leslie, *Acta Crystallogr. D. Biol. Crystallogr.* **52,** 30 (1996).

[51] C. Giacovazzo, *in* "International Tables for Crystallography" (U. Shmueli, ed.), Vol. B, p. 201. Kluwer Academic, Dordrecht, The Netherlands, 1996.

of direct methods,[1] have been reviewed; the reader is referred to these sources for additional details. Historically, direct methods have targeted the determination of complete structures, especially small molecules containing fewer than 100 nonhydrogen atoms. In the early 1990s, the size range of routine direct-methods applications was extended by almost an order of magnitude through a procedure that has come to be known as Shake- and-Bake.[52,53] The distinctive feature of this procedure is the repeated and unconditional alternation of reciprocal-space phase refinement (Shaking) with a complementary real-space process that seeks to improve phases by applying constraints (Baking). This algorithm has been implemented independently in two computer programs, SnB[9,10] and SHELXD[11,11a] (alias Halfbaked or SHELXM). These programs provide default parameters and protocols for the phasing process, but they allow easy user intervention in difficult cases.

It has been recognized for some time that the formalism of direct methods carries over to substructures when applied to single isomorphous[54] (SIR) or single anomalous[55] (SAD or SAS) difference data. MIR data can be accommodated simply by treating the data separately for each derivative, and MAD data can be handled by examining the anomalous differences for each wavelength individually or by combining them together in the form of F_A structure factors.[2,3] The dispersive differences between two wavelengths of MAD data also can be treated as pseudo-SIR differences. If substructure determination were the only concern, it is unclear whether it would be best to measure anomalous scattering data a few times for each of three wavelengths or many times for one wavelength. What is clear is that high redundancy leads to a highly beneficial reduction in measurement errors. SnB and SHELXD can both use either $|\Delta F_{\mathrm{ANO}}|$ or $|F_{\mathrm{A}}|$ values, and so far both approaches have worked well. SnB is normally applied to peak-wavelength anomalous differences computed using the DREAR[56] program suite, and SHELXD is normally applied to $|\Delta F_{\mathrm{ANO}}|$ or $|F_{\mathrm{A}}|$ values that have been calculated using XPREP.[39] It is reassuring to know that one wavelength is generally sufficient for substructure determination when not all wavelengths were measured or when one or more wavelengths were in error. In addition, treating the wavelengths separately allows for useful cross-correlation of sites (see below, Site Validation).

[52] C. M. Weeks, G. T. DeTitta, R. Miller, and H. A. Hauptman, *Acta Crystallogr. D. Biol. Crystallogr.* **49,** 179 (1993).

[53] C. M. Weeks, G. T. DeTitta, H. A. Hauptman, P. Thuman, and R. Miller, *Acta Crystallogr. A* **50,** 210 (1994).

[54] K. S. Wilson, *Acta Crystallogr. B* **34,** 1599 (1978).

[55] A. K. Mukherjee, J. R. Helliwell, and P. Main, *Acta Crystallogr. A* **45,** 715 (1989).

[56] R. H. Blessing and G. D. Smith, *J. Appl. Crystallogr.* **32,** 664 (1999).

The largest substructure solved so far by direct methods contained 160 independent selenium sites.[57] The upper limit of size is unknown, but, by analogy to the complete structure case, it is reasonable to think that it is at least a few hundred sites. In all likelihood, the inherently noisier nature of difference data and the fact that $|\Delta F_{\mathrm{ANO}}|$ and $|F_{\mathrm{A}}|$ values provide imperfect approximations to the substructure amplitudes mean that the maximal substructure size that can be accommodated is probably less than that of complete structures. Although, at present, full structure direct-methods applications require atomic-resolution data of 1.2 Å or better, the resolution of the data typically collected for isomorphous replacement or MAD experiments is sufficient for direct-methods determinations of substructures. Because it is rare for heavy atoms or anomalous scatterers to be closer than 3–4 Å, data having a maximum resolution in this range are adequate.

Data Preparation

Normalization. To take advantage of the probabilistic relationships that form the foundation of direct methods, the usual structure factors, F, must be replaced by the normalized structure factors,[58] E. The condition $\langle |E|^2 \rangle = 1$ is always imposed for every data set. Unlike $\langle |F| \rangle$ which decreases as $\sin(\theta)/\lambda$ increases, the values of $\langle |E| \rangle$ are constant for concentric resolution shells. Similarly, correction factors (ε) are applied that take into account the average intensities of particular classes of reflections as a result of space-group symmetry.[59] The distribution of $|E|$ values is, in principle, and often in practice, independent of the unit cell size and contents, but it does depend on whether a center of symmetry is present. Normalization is a necessary first step in data processing for direct-methods computations. It can be accomplished simply by dividing the data into resolution shells and applying the condition $\langle |E|^2 \rangle = 1$ to each shell. Alternatively, a least-squares-fitted scaling function can be used to impose the normalization condition. The procedures are similar regardless of whether the starting information consists of $|F|$, $|\Delta F|$ (iso or ano), or $|F_{\mathrm{A}}|$ values and leads to $|E|$, $|E_{\Delta}|$, or $|E_{\mathrm{A}}|$ values. Mathematically precise definitions of the SIR and SAD difference magnitudes, $|E_{\Delta}|$, that take into account the atomic scattering factors $|f_{\mathrm{j}}| = |f_{\mathrm{j}}^{\mathrm{o}} + f_{\mathrm{j}}' + if_{\mathrm{j}}''|$ have been presented by Blessing and Smith[56] and implemented in the program DIFFE that is distributed as part

[57] F. von Delft, T. Inoue, S. A. Saldanha, H. H. Ottenhof, F. Schmitzberger, L. M. Birch, V. Dhanaraj, M. Witty, A. G. Smith, T. L. Blundell, and C. Abell, *Struct.* **11,** 985 (2003).

[58] H. A. Hauptman and J. Karle, "Solution of the Phase Problem. I. The Centrosymmetric Crystal." ACA Monograph No. 3. Polycrystal Book Service, Dayton, OH, 1953.

[59] U. Shmueli and A. J. C. Wilson, *in* "International Tables for Crystallography" (U. Shmueli, ed.), Vol. B, p. 190. Kluwer Academic, Dordrecht, The Netherlands, 1996.

of the SnB package. The $|F_A|$ values that are used in SHELXD to form $|E_A|$ values are computed in XPREP,[39] using algorithms similar to those employed in the MADBST component of SOLVE.[4]

Sigma Cutoffs and Outlier Elimination. Direct methods are notoriously sensitive to the presence of even a small number of erroneous measurements. This is especially problematical in the case of difference data, which can be quite noisy. The best antidote is to eliminate any questionable measurement before initiating the phasing process. Fortunately, it is possible to be stringent in the application of cutoffs because the number of difference reflections that must be phased is typically a small fraction of the total available observations. In small-molecule cases in which all reflections accessible to copper radiation have been measured, it is normal to phase about 10 reflections for every atom to be found, and this means that about 15% of the total data are used. In substructure cases, the unit cell for an N-site problem will be much larger than it would be for a small molecule with the same number of atoms to be positioned. Thus, the number of possible reflections will also be much larger, and many more can be rejected if necessary. In fact, only 2–3% of the total possible reflections at 3 Å need be phased in order to solve substructures using direct methods, but these reflections must be chosen from those with the largest $|E_\Delta|$ values.

The DIFFE[56] program rejects data pairs ($|E_1|$, $|E_2|$) [i.e., SIR pairs ($|E_P|$, $|E_{PH}|$), SAD pairs ($|E+|$, $|E-|$), and pseudo-SIR dispersive pairs ($|E_{\lambda 1}|$, $|E_{\lambda 2}|$)] or difference E magnitudes ($|E_\Delta|$) that are not significantly different from zero or deviate markedly from the expected distribution. The following tests are applied when the default values, supplied by the SnB interface for the cutoff parameters (T_{MAX}, X_{MIN}, Y_{MIN}, Z_{MIN}, and Z_{MAX}), are shown in parentheses and are based on empirical tests with known data sets.[60,61]

1. Pairs of data are excluded if $|(|E_1|-|E_2|)-\text{median}(|E_1|-|E_2|)|/\{1.25 \times \text{median}[|(|E_1|-|E_2|)-\text{median}(|E_1|-|E_2|)|]\} > T_{MAX}$ (6.0).
2. Pairs of data are excluded for which either $|E_1|/\sigma(|E_1|)$ or $|E_2|/\sigma(|E_2|) < X_{MIN}$ (3.0).
3. Pairs of data are excluded if $||E_1|-|E_2||/[\sigma^2(|E_1|) + \sigma^2(|E_2|)]^{1/2} < Y_{MIN}$ (1.0).
4. Normalized $|E_\Delta|$ are excluded if $|E_\Delta|/\sigma(|E_\Delta|) < Z_{MIN}$ (3.0).
5. Normalized $|E_\Delta|$ are excluded if $[|E_\Delta|-|E_\Delta|_{MAX}]/\sigma(|E_\Delta|) > Z_{MAX}$ (0.0).

[60] G. D. Smith, B. Nagar, J. M. Rini, H. A. Hauptman, and R. H. Blessing, *Acta Crystallogr. D. Biol. Crystallogr.* **54,** 799 (1998).

[61] P. L. Howell, R. H. Blessing, G. D. Smith, and C. M. Weeks, *Acta Crystallogr. D. Biol. Crystallogr.* **56,** 604 (2000).

The parameter T_{MAX} is used to reject data with unreliably large values of $\|E_1|-|E_2\|$ in the tails of the $(|E_1|-|E_2|)$ distribution. This test assumes that the distribution of $(|E_1|-|E_2|)/\sigma(|E_1|-|E_2|)$ should approximate a zero-mean unit-variance normal distribution for which values less than $-T_{\mathrm{MAX}}$ or greater than $+T_{\mathrm{MAX}}$ are extremely improbable. The quantity $|E_\Delta|_{\mathrm{MAX}}$ is a physical least upper bound such that $|E_\Delta|_{\mathrm{MAX}} = \sum|f|/[\varepsilon \sum|f|^2]^{1/2}$ for SIR data and $|E_\Delta|_{\mathrm{MAX}} = \sum f''/[\varepsilon\sum(f'')^2]^{1/2}$ for SAD data.

Resolution Cutoffs. Before attempting to use MAD or SAD data to locate the anomalous scatterers, a critical decision is to choose the resolution to which the data should be truncated. If data are used to a higher resolution than is supported by significant dispersive and anomalous information, the effect will be to add noise. Because direct methods are based on normalized structure factors, which emphasize the high-resolution data, they are particularly sensitive to this. Because there is some anomalous signal at all the wavelengths in the MAD experiment, a good test is to calculate the correlation coefficient between the signed anomalous differences ΔF at different wavelengths as a function of the resolution. A good general rule is to truncate the data where this correlation coefficient falls below 25–30%. Table III (calculated using XPREP[39]) illustrates three different cases. In case A, the high values involving the peak (PK) and inflection-point (IP) data show that it is not necessary to truncate the data because there is significant MAD information at the highest resolution collected. A poorer correlation would be expected with the low-energy remote data (LR), which has a much smaller anomalous signal. In case B, it is advisable to truncate the data to about 3.9 Å (which indeed led to a successful solution using SHELXD). Case C is clearly hopeless and, in fact, could not be solved. For SAD data collected at a single wavelength, it is still possible to use the correlation coefficient between the anomalous differences collected from two crystals, or from one crystal in two orientations, before merging the two data sets. Such information is also available from the CCP4 programs SCALA and REVISE (see Collaborative Computational Project Number 4, below).

Heavy-Atom Searching and Phasing

The phase problem of X-ray crystallography may be defined as the problem of determining the phases ϕ of the normalized structure factors E when only the magnitudes $|E|$ are given. Owing to the atomicity of crystal structures and the redundancy of the known magnitudes, the phase problem is overdetermined. This overdetermination implies the existence of relationships among the phases that are dependent on the known magnitudes alone, and the techniques of probability theory have identified the linear

TABLE III
Correlation Coefficients (%) Between High-Energy Remote Data and Other Wavelengths as a Function of Resolution Range

A. Apical domain,[a] 1 × (3 SeMet in 144 residues), *C*222$_1$

	Inf – 8.0	– 6.0	– 5.0	– 4.0	– 3.6	– 3.4	– 3.2	– 3.0	– 2.8	– 2.6	– 2.4	– 2.2
PK	91.2	93.9	93.9	89.6	88.6	89.4	89.4	83.9	76.9	65.7	57.0	44.8
IP	89.7	90.0	87.0	84.4	79.8	78.9	79.4	74.7	71.1	54.3	47.2	39.2
LR	48.5	52.8	52.9	38.0	28.4	34.6	14.2	21.1	24.7	9.1	5.4	–3.7

B. Ribosome recycling factor,[b] 1 × (4 SeMet in 185 residues), *P*4$_3$2$_1$2

	Inf – 8.0	– 6.0	– 5.0	– 4.6	– 4.4	– 4.2	– 4.0	– 3.8	– 3.6	– 3.4	– 3.2	– 3.0
PK	69.3	73.1	62.2	56.9	49.6	45.6	48.6	29.6	20.6	24.6	20.1	14.2
IP	59.4	58.3	41.9	43.3	40.7	50.4	34.6	24.7	17.5	16.6	8.1	3.9

C. Unknown protein, 4 × (4 SeMet in 350 residues), *P*2$_1$

	Inf – 8.0	– 6.0	– 5.0	– 4.6	– 4.4	– 4.2	– 4.0	– 3.8	– 3.6	– 3.4	– 3.2	– 3.0
PK	33.2	29.5	19.9	10.6	7.7	17.4	7.6	9.8	9.3	13.4	6.0	2.8
IP	37.6	38.9	37.8	26.5	13.5	24.0	14.2	27.3	25.9	23.1	24.3	22.8

Abbreviations: PK, peak; IP, inflection point; LR, low-energy remote.

[a] M. A. Walsh, I. Dementieva, G. Evans, R. Sanishvili, and A. Joachimiak, *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 1168 (1999).

[b] M. Selmer, S. Al-Karadaghi, G. Hirokawa, A. Kaji, and A. Liljas, *Science* **286,** 2349 (1999).

combinations of three phases whose Miller indices sum to zero (i.e., $\Phi_{\mathbf{HK}} = \phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}}$) as relationships useful for determining unknown structures. (The quantities $\Phi_{\mathbf{HK}}$ are known as structure invariants because their values are independent of the choice of origin of the unit cell.) The conditional probability distribution of the three-phase or triplet invariants depends on the parameter $A_{\mathbf{HK}}$, where $A_{\mathbf{HK}} = (2/N^{1/2})|E_{\mathbf{H}}E_{\mathbf{K}}E_{-\mathbf{H}-\mathbf{K}}|$ and N is the number of atoms, here presumed to be identical, in the asymmetric unit of the corresponding primitive unit cell.[62] Probabilistic estimates of the invariant values are most reliable when the associated normalized magnitudes ($|E_{\mathbf{H}}|$, $|E_{\mathbf{K}}|$, and $|E_{-\mathbf{H}-\mathbf{K}}|$) are large and the number of atoms in the unit cell is small. Thus, it is the largest $|E_{\Delta}|$ or $|E_{\mathrm{A}}|$, remaining after the application of all appropriate cutoffs, that are phased in direct-methods substructure determinations. The triplet invariants involving these reflections are generated, and a sufficient number of those invariants with the highest $A_{\mathbf{HK}}$ values are retained to achieve the desired invariant-to-reflection ratio (e.g., SnB uses a default ratio of 10:1). The inability to obtain a sufficient

[62] W. Cochran, *Acta Crystallogr.* **8,** 473 (1955).

number of accurate invariant estimates is the reason why full-structure phasing by direct methods is possible only for the smallest proteins.

"Multisolution" Methods and Trial Structures. Once the values for some pairs of phases ($\phi_{\mathbf{K}}$ and $\phi_{-\mathbf{H}-\mathbf{K}}$) are known, the triplet structure invariants can be used to generate further phases ($\phi_{\mathbf{H}}$) which, in turn, can be used iteratively to evaluate still more phases. The number of cycles of phase expansion or refinement that must be performed depends on the size of the structure to be determined. Older, conventional, direct-methods programs operate in reciprocal space alone, but the SnB and SHELXD programs alternate phase improvement in both reciprocal and real spaces within each cycle. To obtain starting phases, a so-called multisolution or multitrial approach[63] is taken in which the reflections are each assigned many different starting values in the hope that one or more of the resultant phase combinations will lead to a solution. Solutions, if they occur, must be identified on the basis of some suitable figure of merit. Typically, a random-number generator is used to assign initial values to all phases from the outset.[64] A variant of this procedure employed in SnB is to use the random-number generator to assign initial coordinates to the atoms in the trial structures and then to obtain initial phases from a structure-factor calculation.

The efficiency of direct methods, however, often can be improved considerably by using better-than-random starting trial structures that are, in some way, consistent with the Patterson function. In SHELXD, this is accomplished by computing a Patterson minimum function (PMF)[65] to screen for likely candidates. First, one presumes that the strongest general Patterson peaks may well correspond to a vector between two heavy atoms. For a selected number (e.g., 100) of these vectors, the pair of atoms related by the vector are subjected to a number of random translations (e.g., 99,999). For each of these potential two-atom trial structures, all the symmetry-equivalent atoms are found, the Patterson-function values corresponding to the unique vectors between all of these atoms are calculated and sorted in ascending order, and then the PMF scoring criterion is computed as the mean value of the lowest (e.g., 30%) values in this list. For each two-atom vector, the random translation with the highest PMF is retained. Next, the two-atom trial structures are extended to *N* atoms by using a technique that involves the computation of a full-symmetry Patterson superposition minimum function (PSMF).[37] A list containing all symmetry equivalents of the two starting atoms is generated. Then, each pixel of the PSMF map is

[63] G. Germain and M. M. Woolfson, *Acta Crystallogr. B* **24,** 91 (1968).

[64] R. Baggio, M. M. Woolfson, J.-P. Declercq, and G. Germain, *Acta Crystallogr. A* **34,** 883 (1978).

[65] C. E. Nordman, *Trans. Am. Crystallogr. Assoc.* **2,** 29 (1966).

assigned a value equal to the PMF for all vectors in the list and a dummy atom placed at that pixel. Finally, the $N - 2$ highest peaks in the PSMF map are obtained by interpolation and sorting, and then they are added to the trial structure. Tests using SHELXD have shown that this combination of direct and Patterson methods produces more complete and precise solutions than just using the Patterson methods alone. To make this method applicable in space group *P1*, SHELXD places an extra atom at the origin and performs random translations of the two-atom fragment.

Reciprocal-Space Phase Refinement or Expansion: Shaking. Once a set of initial phases has been chosen, it must be refined against the set of structure invariants whose values are presumed known. So far, two optimization methods (tangent refinement and parameter-shift reduction of the minimal function) have proved useful for extracting phase information in this way. Both of these optimization methods are available in both SnB and SHELXD, but SnB uses the minimal function by default whereas SHELXD uses the tangent formula.

The *tangent formula*[66]

$$\tan(\phi_{\mathbf{H}}) = \frac{-\sum_{\mathbf{K}} |E_{\mathbf{K}} E_{-\mathbf{H}-\mathbf{K}}| \sin(\phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}})}{\sum_{\mathbf{K}} |E_{\mathbf{K}} E_{-\mathbf{H}-\mathbf{K}}| \cos(\phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}})} \tag{2}$$

is the relationship used in conventional direct-methods programs to compute $\phi_{\mathbf{H}}$ given a sufficient number of pairs ($\phi_{\mathbf{K}}$, $\phi_{-\mathbf{H}-\mathbf{K}}$) of known phases. It is also an option within the phase-refinement portion of the dual-space Shake-and-Bake procedure.[67,68] In each cycle, SnB uses the tangent formula to redetermine all the phases, a process referred to as tangent-formula refinement. On the other hand, SHELXD performs a process of tangent expansion in which, during each cycle, the phases of (typically) the 40% highest calculated E magnitudes are held fixed while the phases of the remaining 60% are determined by the tangent formula. The tangent formula suffers from the disadvantage that, in space groups without translational symmetry, it is perfectly fulfilled by a false solution with all phases equal to zero, thereby giving rise to the so-called ''uranium-atom'' solution with one dominant peak in the corresponding Fourier synthesis. In conventional direct-methods programs, the tangent formula is often modified in various ways to include (explicitly or implicitly) information from the so-called negative quartet or four-phase structure invariants[69,70] that are

[66] J. Karle and H. A. Hauptman, *Acta Crystallogr.* **9,** 635 (1956).

[67] C. M. Weeks, H. A. Hauptman, C.-S. Chang, and R. Miller, *Trans. Am. Crystallogr. Assoc.* **30,** 153 (1994).

[68] G. M. Sheldrick and R. O. Gould, *Acta Crystallogr. B* **51,** 423 (1995).

dependent on the smallest as well as the largest E magnitudes. Such modified tangent formulas do indeed largely overcome the problem of false minima for small structures, but because of the dependence of quartet term probabilities on $1/N$, they are little more effective than the normal tangent formula for large structures.

Constrained minimization of an objective function like the minimal function[71,72]

$$R(\Phi) = \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}}[\cos \Phi_{\mathbf{HK}} - I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}})]^2 / \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} \tag{3}$$

provides an alternative approach to phase refinement or phase expansion. $R(\Phi)$ is a measure of the mean-square difference between the values of the triplets calculated using a particular set of phases and the expected probabilistic values of the same triplets as given by the ratio of modified Bessel functions [i.e., $I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}})$]. The minimal function is expected to have a constrained global minimum when the phases are equal to their correct values for some choice of origin and enantiomorph. The minimal function also can be written to include contributions from quartet invariants, although their use is not as imperative as with the tangent formula because the minimal function does not have a minimum when all phases are zero. An algorithm known as parameter shift[73] has proved to be quite powerful and efficient as an optimization method when used within the Shake-and-Bake context to reduce the value of the minimal function. For example, a typical phase-refinement stage consists of three iterations or scans through the reflection list, with each phase being shifted a maximum of two times by 90° in either the positive or negative direction during each iteration. The refined value for each phase is selected, in turn, through a process that involves evaluating the minimal function using the original phase and each of its shifted values.[53] The phase value that results in the lowest minimal-function value is chosen at each step. Refined phases are used immediately in the subsequent refinement of other phases.

Real-Space Constraints: Baking. Peak picking is a simple but powerful way of imposing an atomicity constraint. Karle[74] found that even a relatively small, chemically sensible, fragment extracted by manual interpretation of a small-molecule electron-density map could be expanded

[69] H. Schenk, *Acta Crystallogr. A* **30,** 477 (1974).

[70] H. Hauptman, *Acta Crystallogr. A* **30,** 822 (1974).

[71] T. Debaerdemaeker and M. M. Woolfson, *Acta Crystallogr. A* **39,** 193 (1983).

[72] G. T. DeTitta, C. M. Weeks, P. Thuman, R. Miller, and H. A. Hauptman, *Acta Crystallogr. A* **50,** 203 (1994).

[73] A. K. Bhuiya and E. Stanley, *Acta Crystallogr.* **16,** 981 (1963).

[74] J. Karle, *Acta Crystallogr. B* **24,** 182 (1968).

into a complete solution by transformation back to reciprocal space and then performing additional iterations of phase refinement with the tangent formula. Automatic real-space electron-density map interpretation in the Shake-and-Bake procedure consists of selecting an appropriate number of the largest peaks in each cycle to be used as an updated trial structure without regard to chemical constraints other than a minimum allowed distance between atoms (e.g., 1.0 Å for full structures and 3–3.5 Å for substructures). If markedly unequal atoms are present, appropriate numbers of peaks (atoms) can be weighted by the proper atomic numbers during transformation back to reciprocal space in a subsequent structure-factor calculation. Thus, *a priori* knowledge concerning the chemical composition of the crystal is used, but no knowledge of constitution is required or used during peak selection. It is useful to think of peak picking in this context as simply an extreme form of density modification appropriate when the resolution of the data is small compared with the distance separating the atoms. In theory, under appropriate conditions it should be possible to substitute alternative density-modification procedures such as low-density elimination[75,76] or solvent flattening,[27] but no practical applications of such procedures have yet been made. The imposition of physical constraints counteracts the tendency of phase refinement to propagate errors or produce overly consistent phase sets. For example, the ability to eliminate chemically impossible peaks at special positions using a symmetry-equivalent cutoff distance (similar to the procedure described in the Crystallography and NMR System section) prevents the occurrence of most cases of false minima.[10]

In its simplest form as implemented in the SnB program, peak picking consists of simply selecting the top N E-map peaks, where N is the number of unique nonhydrogen atoms in the asymmetric unit. This is adequate for small-molecule structures. It has also been shown to work well for heavy-atom or anomalously scattering substructures where N is taken to be the number of expected substructure sites.[60,77] For larger structures or substructures (e.g., $N > 100$), the number of peaks selected is reduced to $0.8N$ peaks, thereby taking into account the probable presence of some atoms that, owing to high thermal motion or disorder, will not be visible. An alternative approach to peak picking used in SHELXD is to begin by selecting approximately N top peaks, but then to eliminate some of them (typically one-third) at random. By analogy to the common practice in macromolecular crystallography of omitting part of a structure from a

[75] M. Shiono and M. M. Woolfson, *Acta Crystallogr. A* **48,** 451 (1992).

[76] L. S. Refaat and M. M. Woolfson, *Acta Crystallogr. D. Biol. Crystallogr.* **49,** 367 (1993).

[77] M. A. Turner, C.-S. Yuan, R. T. Borchardt, M. S. Hershfield, G. D. Smith, and P. L. Howell, *Nat. Struct. Biol.* **5,** 369 (1998).

Fourier calculation in hopes of finding an improved position for the deleted fragment, this version of peak picking is described as making a random omit map. It has the potential for being a more efficient search algorithm.

Scoring Trial Structures

SnB and SHELXD compute figures of merit that allow the user to judge the quality of a trial structure and decide whether or not it is a solution. It is worth repeating the caution given above (see Crystallography and NMR System). Although it is sometimes possible to give absolute values that strongly indicate a solution, it is safer to consider relative values. A true solution should have one or more figure-of-merit values that are outstanding relative to the nonsolutions, which generally are in the majority.

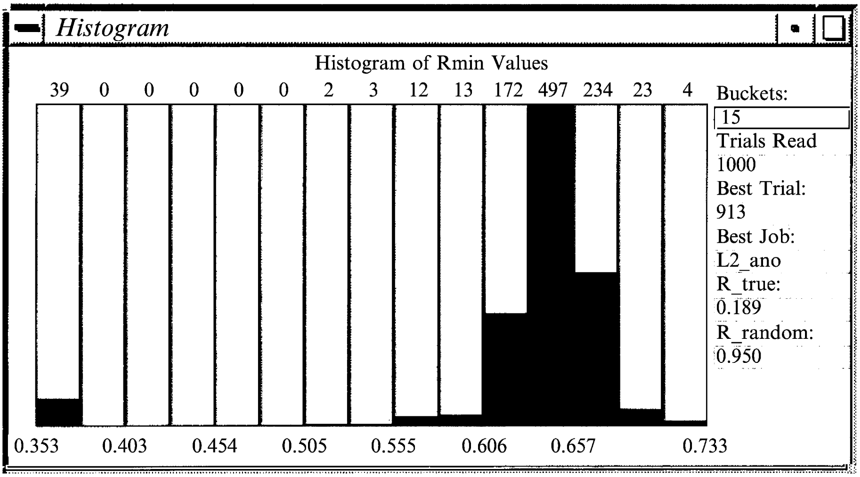
Minimal Function. The minimal function itself, $R(\Phi)$ [Eq. (3)], is a highly reliable figure of merit, provided that it has been calculated directly from the constrained phases corresponding to the final peak positions.[53] This figure of merit is computed by both programs, and solutions typically have the smallest values. The SnB graphical user interface provides an option for checking the status of a running job by displaying a histogram of the minimal-function values for all trials that have been processed so far, as illustrated in Fig. 3 for the peak-anomalous difference data for a 30-site selenomethionyl (SeMet) substructure.[77] A clear bimodal distribution of figure-of-merit values is a strong indication that a solution has, in fact, been found. Confirmation that this is true for trial 913 in the example in Fig. 3 can be obtained by inspecting a trace of the minimal-function value as a function of refinement cycle (Fig. 4). Solutions usually show an abrupt decrease in value over a few cycles, followed by stability at the lower value.

Crystallographic R. SnB and SHELXD compute $R_{\mathrm{CRYST}} = (\sum \| E_{\mathrm{O}}| - |E_{\mathrm{C}}\|)/\sum |E_{\mathrm{O}}|$. This figure of merit, which is also highly reliable, has small values for solutions.

PATFOM. The Patterson figure of merit, PATFOM, is the mean Patterson minimum function value for a specified number of atoms. It is computed by SHELXD. Although the absolute value depends on the structure in question, solutions almost always have the largest PATFOM values.

Correlation Coefficient. The correlation coefficient[42] computed in SHELXD is defined by

$$\mathrm{CC} = \left[\sum wE_{\mathrm{o}}E_{\mathrm{c}} \cdot \sum w - \sum wE_{\mathrm{o}} \cdot \sum wE_{\mathrm{c}}\right] \Big/ \left\{\left[\sum wE_{\mathrm{o}}^2 \cdot \sum w - \left(\sum wE_{\mathrm{o}}\right)^2\right]\left[\sum wE_{\mathrm{c}}^2 \cdot \sum w - \left(\sum wE_{\mathrm{c}}\right)^2\right]\right\}^{1/2} \tag{4}$$



Buckets:
15
Trials Read
1000
Best Trial:
913
Best Job:
L2_ano
R_true:
0.189
R_random:
0.950

FIG. 3. This bimodal histogram of minimal function (R_{MIN}) values for 1000 trials suggests that there are 39 solutions. R_{TRUE} and R_{RANDOM} are theoretical values for true and random phase sets, respectively.[53]

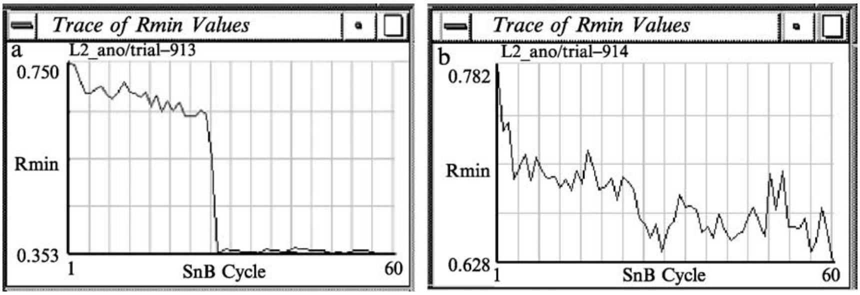


FIG. 4. Plots of the minimal-function value over 60 cycles (a) for a solution (trial 913) and (b) for a nonsolution (trial 914).

with default weights $w = 1/[0.1 + \sigma^2\ (E)]$. Solutions typically have the largest values for this figure of merit. Values of 0.7 or greater when based on all, or almost all, of the $|E|$ data for full structures strongly indicate that a solution has been found. Also, when computed in SHELXD for substructures using $|E_{\mathrm{A}}|$ data, values greater than 0.4 typically indicate a solution. SnB also computes a correlation coefficient, but this criterion has not been found to be reliable for substructures when based on the limited number of $|E_{\Delta}|$ difference data normally used.

Site Validation

Direct-methods programs provide as output a file of peak positions, for one or more of the best trials, sorted in descending order according to the electron density at those positions on the Fourier map. For an N-site substructure, SnB provides $1.5N$ peaks for each trial. The user must then decide which, and how many, of these peaks correspond to actual atoms. The first N peaks have the highest probability of being correct, and in many cases this simple guideline is adequate. Sometimes, there will be a significant break in the density values between true and false peaks, and, when this occurs in the expected place, it is additional confirmation. In other cases, a conservative approach is to accept the $0.8N$ to $0.9N$ top peaks, compute a difference Fourier map, and compare the peaks on this map to the original direct-methods map.

Crossword Tables. The Patterson superposition function is the basis of the crossword table,[78,79] introduced in SHELXS-86[80] and available also in SHELXD, that provides another way to assess which of the heavy-atom sites are correct and, in some cases, to recognize the presence of noncrystallographic symmetry. Each entry in the table links the potential atom forming the row with the potential atom forming the column. For each pair of atoms, the top number is the minimum distance between them, taking the space-group symmetry into account. The bottom number is the Patterson minimum function (PMF) value calculated from all vectors between the two atoms, also taking symmetry into account. The first vertical column is based on the self-vectors (i.e., the vectors between one atom and its symmetry equivalents). In general, wrong sites can be recognized by the presence in the table of several zero PMF values (negative values are replaced by zero). Table IV shows the crossword table for the CuKα anomalous ΔF data for a HiPIP with two Fe_4S_4 clusters in the asymmetric unit.[81] It is easy to find the two clusters (atoms 1–4 and 5–8) by looking for Fe$\cdots$Fe distances of approximately 2.8 Å, and the PMF values for the eight correct atoms are, in general, higher than those involving spurious atoms despite the weakness of the anomalous signal.

Comparison of Trials. When trying to decide which peaks are correct, it is also helpful to compare the peak positions from two or more solutions.

[78] G. M. Sheldrick, Z. Dauter, K. S. Wilson, and L. C. Sieker, *Acta Crystallogr. D. Biol. Crystallogr.* **49,** 18 (1993).

[79] G. M. Sheldrick, *in* "Direct Methods for Solving Macromolecular Structures" (S. Fortier, ed.), p. 131. Kluwer Academic, Dordrecht, The Netherlands, 1998.

[80] G. M. Sheldrick, *J. Mol. Struct.* **130,** 9 (1985).

[81] I. Rayment, G. Wesenberg, T. E. Meyer, M. A. Cusanovich, and H. M. Holden, *J. Mol. Biol.* **228,** 672 (1992).

TABLE IV
Crossword Table for Location of Eight Iron Atoms

Peak	*x*	*y*	*z*	Self	Cross-vectors						
99.9	0.9201	0.0784	0.1133	27.7							
				26.6							
88.4	0.9719	0.1047	0.1356	27.4	2.4						
				39.7	25.1						
85.5	0.9043	0.1258	0.0884	27.7	2.6	3.0					
				27.3	23.3	5.5					
82.7	0.9546	0.0950	0.0503	26.7	2.3	2.5	2.7				
				15.2	28.4	43.5	26.4				
81.1	0.3542	0.5285	0.2615	31.2	14.6	16.6	14.4	14.6			
				20.9	41.4	14.8	9.5	21.5			
80.5	0.4316	0.5144	0.2451	30.0	16.5	18.7	16.4	16.8	3.0		
				25.5	24.6	20.0	21.2	8.9	0.0		
80.4	0.3942	0.5575	0.1995	29.6	14.4	16.4	13.9	14.6	2.7	2.9	
				0.0	31.4	7.7	22.6	33.8	26.6	19.4	
73.9	0.3920	0.5023	0.1694	29.1	14.3	16.6	14.5	14.8	3.2	2.6	3.0
				26.1	22.3	16.0	24.5	18.3	10.9	0.0	17.5
63.8	0.4025	0.4641	0.2218	29.9	16.1	18.4	16.4	16.5	4.0	2.9	5.0
				18.4	17.0	13.1	0.0	4.5	0.0	5.4	0.0
58.9	0.9655	0.0517	0.0945	26.9	2.2	3.0	4.5	2.6	15.2	17.3	15.4
				45.9	7.3	15.8	7.8	5.3	0.0	0.0	6.1

Peaks recurring in several solutions are more likely to be real. However, in order to do this comparison, one must take into account the fact that different solutions may have different origins and/or enantiomorphs. A stand-alone program for doing this is available,[82] and the capability of making such comparisons automatically for all space groups will be available in future versions of SnB and SHELXD. The usefulness of peak correlation is illustrated by an example for a 30-site SeMet substructure.[61,77] Table V presents the relative rankings of peaks, from nine other trials, that correspond to peaks 29–45 of trial 149, which had the lowest minimal-function value for the peak-wavelength difference data for crystal 1. The top 29 peaks for trial 149 were correct selenium positions, but peak 30 (the *N*th peak) was spurious. Peak 33 of trial 149 was found to have a match on every other map, and indeed, it did correspond to the final selenium site. It appears that, in general, the same noise is not reproduced on different maps, especially maps originating from different data sets. Thus, peak correlation can be used to identify correct peaks ranking below the *N*th peak.

[82] G. D. Smith, *J. Appl. Crystallogr.* **35,** 368 (2002).

TABLE V
TRIAL COMPARISON FOR 30-SITE SUBSTRUCTURE

Crystal:	1	1	1	1	1	1	1	2	2	2
Wavelength[a]:	PK	PK	PK	PK	PK	IP	HR	IP	PK	HR
Trial no.:	149	31	158	165	176	104	23	476	93	86
Peak rank:	29	22	29	29	29		21	38	29	28
	31			34						
	33	42	30	30	35	24	22	34	30	30
	34		33		42					
	37			43						
	39		40		38					
	40		42			42				
	45	40								

[a] The wavelengths are peak (PK), inflection point (IP), and high-energy remote (HR).

Enantiomorph Determination

Because all publicly distributed direct-methods programs, including SnB and SHELXD, work with only $|E|$, $|E_\Delta|$, or $|E_A|$ values, they have no way to determine the proper hand. Both enantiomorphs are found with equal frequency among the solutions. If a structure crystallizes in an enantiomorphic space group, either of the space groups may be used during the direct-methods step, but chances are 50% that, at a later stage, the coordinates will have to be inverted and the space group changed to its enantiomorph in order to produce an interpretable protein map. A direct-methods formalism has been proposed[83] that uses both $|E+|$ and $|E-|$ and, in theory, should make it possible to produce only solutions with the proper hand. However, this theory has never been successfully applied to actual experimental data. Similarly, it should be noted that solutions occur at all permitted origin positions with equal frequency. This means that, in the MIR case, cross-phasing is necessary to ensure that all derivatives are referred to the same origin. A direct-methods formalism[84] exists that should automatically do this, but it has never been implemented in a distributed program.

Substructure Refinement

Fourier refinement, often called *E*-Fourier recycling, has been used for many years in direct-methods programs to improve the quality and completeness of solutions.[85] Additional refinement cycles are performed in real

[83] H. Hauptman, *Acta Crystallogr. A* **38,** 632 (1982).
[84] S. Fortier, C. M. Weeks, and H. Hauptman, *Acta Crystallogr. A* **40,** 646 (1984).

space alone, using many more reflections than is possible in the direct-methods steps that are dependent on the accuracy of triplet-invariant relationships. In SHELXD, the final model can be improved further by occupancy or isotropic displacement parameter (B_{iso}) refinement for the individual atoms,[86] followed by calculation of the Sim[87]- or sigma-A[88]-weighted map. The development of a common interface[89] for SnB and the PHASES package[90] permits coordinates determined by direct methods to be passed easily for conventional substructure phase refinement and protein phasing, and for SHELXD this facility is provided by a program SHELXE.[90a]

Collaborative Computational Project Number 4

Unlike many other packages, the Collaborative Computational Project Number 4 (CCP4) suite is a set of separate programs that communicate via standard data files rather than having all operations integrated into one huge program. This has some disadvantages in that it is less easy for programs to make decisions about what operation to do next even though communication is now being coordinated through a graphical user interface (CCP4i). The advantage of loose organization is that it is easy to add new programs or to modify existing ones without upsetting other parts of the suite.

Data Preparation

The CCP4 suite provides a number of programs (i.e., SCALA,[91] TRUNCATE,[92] and SCALEIT) that are useful in preparing data for experimental phasing. SCALA treats scaling and merging as different operations, thereby allowing an analysis of data quality before merging. For isomorphous replacement studies, the native data can be used as the reference set, and all of the derivatives scaled to it. This provides

[85] G. M. Sheldrick, *in* "Crystallographic Computing" (D. Sayre, ed.), p. 506. Clarendon Press, Oxford, 1982.

[86] I. Usón, G. M. Sheldrick, E. de la Fortelle, G. Bricogne, S. di Marco, J. P. Priestle, M. G. Grütter, and P. R. E. Mittl, *Struct. Fold. Des.* **7,** 55 (1999).

[87] G. A. Sim, *Acta Crystallogr.* **12,** 813 (1959).

[88] R. J. Read, *Acta Crystallogr. A* **42,** 140 (1986).

[89] C. M. Weeks, R. H. Blessing, R. Miller, R. Mungee, S. A. Potter, J. Rappleye, G. D. Smith, H. Xu, and W. Furey, *Z. Kristallogr.* **217,** 686 (2002).

[90] W. Furey and S. Swaminathan, *Methods Enzymol.* **277,** 590.

[90a] G. M. Sheldrick, *Z. Kristallogr.* **217,** 644 (2002).

[91] P. R. Evans, *in* "Recent Advances in Phasing." Proceedings of CCP4 Study Weekend (1997).

[92] G. S. French and K. S. Wilson, *Acta Crystallogr. A* **34,** 517 (1978).

well-parameterized "local" scales. For MAD data, all sets are scaled in one pass, gross outliers are rejected (e.g., any measurement four to five times greater than the mean), and then each data set is merged separately to give a weighted mean for each reflection. A detailed analysis of the data is provided in a graphical form. Useful information is given on the scale factors themselves (which can often pinpoint rogue images), on the R_{merge} values, and on the correlation coefficients between wavelengths for MAD data (coefficients <0.4 suggest a resolution cutoff; see discussion of Table III).

Various scaling models related to the experiment can be used. The scale factor is a function of the primary beam direction, treated either as a smooth function of the rotation angle or as an image-by-image correction. In addition, the scale may be a function of the secondary beam direction, acting principally as an absorption correction, expanded either as spherical harmonics or as an interpolated three-dimensional function of the rotation angle and the spatial coordinates of the measured spot on the detector. The secondary beam correction is related to the absorption anisotropy correction described by Blessing,[93] and the interpolated three-dimensional correction is similar to that described by Kabsch.[94] Optimum scaling depends a great deal on exactly how the data were collected, and it is not possible to lay down rules for all cases.

TRUNCATE can convert merged intensities to amplitudes in two ways. The simplest way is just to take the square root of the intensities, setting any negative values to zero. Alternatively, a best estimate of F can be calculated from I, $\sigma(I)$, and the distribution of intensities in resolution shells. This has the effect of forcing all negative observations to be positive and of inflating the weakest reflections ($<\sim 3\sigma$) because an observation significantly smaller than the average intensity is likely to be underestimated. TRUNCATE also analyzes the data to verify that the expected distributions are satisfied. It generates a Wilson plot that should be linear for the resolution shells greater than 4 Å, moments for the intensities (which are excellent indicators of twinning), the cumulative intensity distribution (another clue to both twinning and sometimes noncrystallographic symmetry), and an analysis of anisotrophy. All these criteria need to be examined carefully before using the data.

SCALEIT puts all data sets on the same relative scale and uses normal probability plots[95] to test whether the differences between them are significant. First, the reflections in each resolution bin are sorted according to the value of $\delta(\text{real}) = (F_{\text{P}} - F_{\text{PH}})/[\sigma^2(F_{\text{P}}) + \sigma^2(F_{\text{PH}})]^{1/2}$, where F_{PH} and $\sigma(F_{\text{PH}})$

[93] R. H. Blessing, *Acta Crystallogr. A* **51,** 33 (1995).

[94] W. Kabsch, *J. Appl. Crystallogr.* **21,** 916 (1988).

[95] P. L. Howell and G. D. Smith, *J. Appl. Crystallogr.* **25,** 81 (1992).

are the scaled values for the derivative. For each reflection, the corresponding δ(expected) is then calculated assuming a normal distribution, and δ(real) is plotted against δ(expected). If the native and scaled derivative data sets are essentially identical (in statistical parlance, they represent two samplings of the same population), the normal probability plot will be linear with a slope of unity and an intercept of zero. The size of the substructure contribution can be gauged by the deviation of the slope and intercept from these values, and the variation with resolution indicates to what resolution the heavy-atom contribution extends. A similar analysis can be applied to MAD data to estimate the significance of the dispersive and anomalous differences.

Heavy-Atom Searching and Phasing

The CCP4 suite includes two direct-methods programs that can be used to locate heavy-atom sites using a variety of difference structure-factor coefficients. The simplest approach is to use the best SAD or SIR difference. Program REVISE[96] can be used to estimate F_A or F_M, the full contribution from the substructure. Normalized difference magnitudes, $|E_\Delta|$, are computed using program ECALC.

RANTAN. RANTAN[7] is a classic direct-methods program that performs reciprocal-space phase refinement. The program determines reflections for fixing the origin and enantiomorph, and then assigns a set of random phases with default weights of 0.25 to a starting set of large $|E|$ values. The phases are refined by the tangent formula and expanded to include the whole set of large E magnitudes. Up to five sets of refined phases and weights with the best combined figures of merit are output.

ACORN. ACORN[8] is a fast *ab initio* procedure for solving structures when the data are sufficient to separate atomic sites in the E maps. In the case of substructures, 4-Å data (sometimes even lower) will usually suffice. The initial phase sets are generated from the atomic coordinates of a putative structural fragment. The fragment can be made up in various ways. In simple cases, such as metalloproteins or heavy-atom substructures, it is sufficient to generate many trial structures starting from a single randomly placed atom. The reflections are divided into three groups (strong, medium, and weak) according to their $|E|$ values. Correlation coefficients (CC; see Dual-Space Direct Methods, above), between the observed and calculated E values for each class are used in different ways throughout the procedure. All reflections are used to select likely trial sets. The strong and weak reflections are used in the phase refinement, and the CC for the

[96] H.-F. Fan, M. M. Woolfson, and J.-X. Yao, *Proc. R. Soc. Lond. A* **442,** 13 (1993).

medium reflections provides a simple criterion of correctness for a phase set. The starting phase sets are refined primarily using dynamic density modification, supplemented by Patterson superposition and real-space Sayre equation refinement.

Dynamic density modification (DDM) eliminates the negative densities and truncates the highest density. For the first cycle, this truncation will occur at the sites of the starting coordinate(s). During later cycles, the density is modified according to a formula based on the standard deviation of the map and the cycle number.

Patterson superposition generates a semisharpened Patterson sum-function map from the starting fragment.

Sayre equation refinement is carried out in real space, using fast Fourier transforms instead of working directly with the phase relationships. The equations are identical, but the real-space formulation is much faster.

ACORN first uses DDM for many cycles. Then, if no solution can be found, a few cycles of Sayre equation refinement are performed. This may modify the phase set sufficiently to allow the DDM algorithm to function more effectively.

Scoring Trial Structures

ACORN will stop automatically if the value of CC for the medium E values becomes greater than a preset value during DDM, thereby indicating that a probable solution has been found. This CC value needs to be adjusted according to the data quality, particularly when searching for anomalous scatterers using SAD or MAD data. Another criterion for success, similar to that used in SnB, is that the same solution is found more than once. In CCP4, this is checked using the phased translation function, a function that detects similar solutions after taking both hands (enantiomorphs) and alternative origins into account. The third, and most significant, criterion is whether the trial solution gives the appropriate number of sites with more-or-less appropriate peak heights.

Site Validation, Enantiomorph Determination, and Substructure Refinement

Within CCP4, the program MLPHARE is used to refine the substructure sites and to generate protein phases. Initially, it is usually sufficient to refine putative sites against the centric data or some other subset. Typically, the refinement is enormously overdetermined (i.e., there are many more observations than parameters), and the refined phases are sufficiently good to allow the cross-checking of sites and the choice of hand. Numerical criteria are the figure of merit and phasing power, both of which are useful criteria for assessing whether a new site is improving the solution or not. However, it is difficult to define an absolute required value for either of these quantities. Another useful criterion is the extended Cullis R factor, defined as the $\langle$Lack of closure$\rangle$/$\langle$Isomorphous difference$\rangle$. (The isomorphous difference is $|F_{\mathrm{PH}}-F_{\mathrm{P}}|$; lack of closure is $|F_{\mathrm{PH}}-|F_{\mathrm{P}}+F_{\mathrm{H}}||$, where $|F_{\mathrm{P}}+F_{\mathrm{H}}|$ is a vector sum of the calculated F_{H} and F_{P} using the current best protein phases.) This is the most reliable signal for a usable derivative. For centric data, values less than 0.6 are excellent, and values greater than 0.9 indicate that something is not right. If a new site does not reduce the existing Cullis R value, it is probably not correct.

Applications

This section contains a discussion of applications of the programs described above to substructures that can be regarded in some way as being at the cutting edge. These applications include large selenomethionine derivatives, substructures phased by weak anomalous signals, and substructures created by soaking protein crystals in cryobuffers containing concentrated halide salts. The tabulations presented below should not be regarded as a complete survey of the literature. The intention here is to focus on how to use the programs effectively in these challenging situations.

Large Substructures

Improvements in data collection instruments and methods have permitted macromolecular diffraction data, especially small anomalous-scattering differences, to be measured much more accurately. At the same time, the use of genetic engineering to replace methionine by selenomethionine[97] (SeMet) has provided a convenient means for inserting many anomalous scatterers into large proteins. In the last 3 or 4 years, this has resulted in a dramatic increase in the size of the substructures that have challenged phasing methods and the programs that implement them. So far, as shown in Table VI, the programs described above (especially those that employ direct methods) have met this challenge well, and the upper size of substructures manageable with current software has clearly not been reached yet.

[97] W. A. Hendrickson, J. R. Horton, and D. M. LeMaster, *EMBO J.* **9,** 1665 (1990).

In general, recognizing when a solution has occurred has not been a problem, but selecting all the correct sites has been difficult in a few cases (e.g., carrier protein reductase and lactonizing enzyme) and was aided by a careful consideration of the noncrystallographic symmetry. On the other hand, some of the largest studies have proceeded smoothly once a solution was identified. For example, in the case of the bifunctional enzyme DmpFG, the 86 top selenium sites found by SnB were put into the ADDSOLVE component of the SOLVE package for refinement and to search for additional sites. ADDSOLVE found 14 more sites (total of 100 of 108 Se), and it was followed by solvent flattening using RESOLVE. Then, two rounds of ARP/wARP[98] tracing, extending the resolution to 1.7 Å for the native data, found 2330 residues (88%) automatically.

KPHMT. The 2.8-Å, peak-wavelength, anomalous data set (136,609 unique reflections) for the largest SeMet substructure, ketopantoate hydroxymethyltransferase (KPHMT) from *Escherichia coli*, was highly redundant (average multiplicity per Friedel mate, 10.6), complete (all data, 99.9%; anomalous completeness, 99.8%), accurate ($R_{\mathrm{sym}} = 0.120$; $R_{\mathrm{anom}} = 0.073$), and had a good signal-to-noise ratio [$I/\sigma(I) = 25.6$ overall; $I/\sigma(I) = 6.0$ in the highest resolution shell]. It is assumed that the high quality of the data was important for a successful outcome. Although the KPHMT substructure was originally solved by SnB, it can also be solved by SHELXD using the peak data alone. In fact, the use of Patterson-based trial structures in SHELXD improves the success rate (percentage of trials going to solution) by perhaps an order of magnitude, resulting in one solution every 14 h on an 800-MHz Athlon PC when the full 2.8-Å data set is used. On the other hand, an experimental version of SnB that uses the sine-enhanced minimal function[99] also gives a significantly improved success rate relative to the distributed program (SnB version 2.1).

The best SHELXD solution (2.8-Å data, 0.75-Å Fourier map grid) for KPHMT had 145 of the top 160 peaks, and 149 of the top 200 peaks, within 2 Å of the methionine sulfurs in the native structure. These matches could be improved to 152 and 157 peaks, respectively, by combining the phases for the best 16 solutions. In this case, 97 of the peaks were actually within 0.5 Å of the sulfur positions. In comparison, the original SnB solution (3.5-Å data, 2-Å grid) gave corresponding matches of 122 and 127 peaks. The KPHMT structure consists of two independent decamers with N-terminal methionines that have never been found. The strategy followed by von Delft[57] in solving the structure was to take the top 120 SnB peaks

[98] A. Perrakis, R. Morris, and V. S. Lamzin, *Nat. Struct. Biol.* **6,** 458 (1999).

[99] H. Xu, H. A. Hauptman, and C. M. Weeks, *Acta Crystallogr. D. Biol. Crystallogr.* **58,** 90 (2002).

TABLE VI
Twenty Selenomethionine Substructures with 40 or More Sites

Protein	Space group	d (Å)	kDa/ asymmetric unit	Program used[a]	Actual sites	Sites found
Cyanase[b]	*P1*	3.0	170	SHELXD	40	40
Pyruvate dehydrogenase: E1[c]	$P2_1$	3.5	200	SnB	40	40
EphB2 receptor SAM domain[d]	$P4_1$	1.95	78	SnB	48	?
Arylamine *N*-acetyltransferase[e]	$P2_12_12$	4.0	240	SnB	48	48
MutS repair protein[f]	$P2_12_12_1$	3.0	230	SnB	48	32
Target protein MP883[g]	$P2_12_12_1$	3.0	180	SHELXD	50	50
Confidential	$P2_12_12_1$	2.4	183	SOLVE	52	52
D-Hydantoinase[h]	$C222_1$	3.0	300	SOLVE	54	54
Confidential	$P2_1$	4.0	61	SOLVE	56	56
Cap-binding complex[i]	$P2_12_12_1$	3.0	300	SHELXD	57	57
Nicotinamide nucleotide transhydrogenase[j]	$P2_1$	3.0	160	SHELXD	59	58
Tryparedoxin peroxidase[k]	$P2_1$	3.2	230	SOLVE	60	46
Gastroenteritis viral protease[l]	$P2_1$	2.9	198	SnB	60	37
2-Aminoethylphosphonate transaminase[m]	$P2_1$	2.55	270	SHELXD	66	66
Human HMG-CoA reductase[n]	$P2_1$	2.6	200	SnB	68	45
Acyl carrier protein reductase[o]	$P2_1$	3.0	204	SnB	69	31
D-Mannoheptose 6-epimerase[p]	$P2_1$	3.0	370	SnB	70	65
Muconate lactonizing enzyme[q]	$P2_12_12_1$	4.0	112	SnB	80	57
Pseudomonas sp. DmpFG[r]	$P2_12_12_1$	2.2	280	SnB	108	86
Ketopantoate hydroxymethyltransferase[s]	$P2_1$	3.5	567	SnB	160	120

[a] Program used for the original solution. SnB applications used peak-wavelength anomalous $|E_\Delta|$ data. SOLVE and SHELXD applications used MAD $|E_A|$ data.

[b] M. A. Walsh, Z. Otwinowski, A. Perrakis, P. M. Anderson, and A. Joachimiak, *Struct. Fold. Des.* **8,** 505 (2000).

[c] P. Arjunan, N. Nemeria, A. Brunskill, K. Chandrasekhar, M. Sax, Y. Yan, F. Jordan, J. R. Guest, and W. Furey, *Biochemistry* **41,** 5213 (2002).

[d] C. D. Thanos, K. E. Goodwill, and J. U. Bowie, *Science* **283,** 833 (1999).

(*continued*)

(two-thirds of the originally expected 180 sites), refine them with SHARP,[20] and locate the other 40 sites using difference Fouriers. This strategy resulted in a map that could be interpreted easily.

AEP Transaminase. Table VII compares the application of several programs to the data for the 66-site SeMet substructure of 2-aminoethylphosphonate (AEP) transaminase. The data are of high quality, but the selenium absorption edge was missed because of problems with the beamline at the time of data collection. As a result, what was thought to be the inflection-point data actually had the strongest anomalous signal. Despite this complication, all the programs tested could solve the structure although there is variation with respect to the data set that gives the highest success rate. The superiority of the combined (direct methods and Patterson) approach that uses Patterson-based seeds to generate the starting structures is apparent. Because CNS uses a dead-end criterion to terminate the Patterson search and, typically, the search is abandoned early when the anomalous signal is poor, the average time per trial will usually be less for the less successful runs. The CCP4 program ACORN runs trials in an order dependent on the scoring for a single randomly positioned

[e] J. C. Sinclair, J. Sandy, R. Delgoda, E. Sim, and M. E. Noble, *Nat. Struct. Biol.* **7,** 560 (2000).

[f] M. H. Lamers, A. Perrakis, J. H. Enzlin, H. H. K. Winterwerp, N. deWind, and T. K. Sixma, *Nature* **407,** 711 (2000).

[g] Berkeley Structural Genomics Center, personal communication.

[h] J. Abendroth, K. Niefind and D. Schomburg, *J. Mol. Biol.* **320,** 143 (2002).

[i] C. Mazza, M. Ohno, A. Segref, I. W. Mattaj, and S. Cusack, *Mol. Cell* **8,** 383 (2001).

[j] P. A. Buckley, J. B. Jackson, T. R. Schneider, S. A. White, D. W. Rice, and P. J. Baker, *Struct. Fold. Des.* **8,** 809 (2000).

[k] M. S. Alphey, C. S. Bond, E. Tetaud, A. H. Fairlamb, and W. N. Hunter, *J. Mol. Biol.* **300,** 903 (2000).

[l] K. Anand, G. J. Palm, J. R. Mesters, S. G. Siddell, J. Ziebuhn, and R. Hilgenfeld, *Embo. J.* **21,** 3213 (2002).

[m] C. C. H. Chen, A. Kim, H. Zhang, A. J. Howard, G. Sheldrick, D. Dunaway-Mariano, and O. Herzberg, *Biochemistry* **41,** 13162 (2002).

[n] E. S. Istvan, M. Palnitkar, S. K. Buchanan, and J. Deisenhofer, *EMBO J.* **19,** 819 (2000).

[o] A. C. Price, Y.-M. Zhang, C. O. Rock, and S. W. White, *Biochemistry* **40,** 12772 (2001).

[p] A. M. Deacon, Y. S. Ni, W. G. Coleman, Jr., and S. E. Ealick, *Struct. Fold. Des.* **8,** 453 (2000).

[q] M. Merckel, T. Kajander, A. M. Deacon, A. Thompson, J. G. Grossman, N. Kalkkinen, and A. Goldman, *Acta Crystallogr. D. Biol. Crystallogr.* **58,** 727 (2002).

[r] B. A. Manjasetty, J. Powlowski, and A. Vrielink, *Proc. Natl. Acad. Sci. USA* **100,** 6992 (2003)

[s] F. von Delft, T. Inoue, S. A. Saldanha, H. H. Ottenhof, F. Schmitzbergera, L. M. Birch, V. Dhanaraj, M. Witty, A. G. Smith, T. L. Blundell, and C. Abell, *Struct.* **11,** 985 (2003).

TABLE VII

Success Rates for 2-Aminoethylphosphonate Transaminase Data Sets[a]

Program:	CNS	SnB	SHELXD[b]	SHELXD[c]	ACORN
Trials run:	100	1000	1000	1000	Variable
Time per trial[d]:	600[e]	250	90	40	—
Success rate					
IP	7%	12.1%	15.0%	42.4%	1 of 17
PK	3	4.1	9.3	38.0	1 of 81
HR	0	0.2	2.5	14.8	0
IP/HR	—	16.0	6.8	12.7	—
PK/HR	—	0.1	0.0	0.0	—
IP + IP/HR	13	—	13.7	65.3	—
F_A	17	3.8[f]	6.1	56.4	1 of 26

[a] The data sets are as follows: inflection point (IP), peak (PK), high-energy remote (HR), IP and HR dispersive differences (IP/HR), PK and HR dispersive differences (PK/HR), combined IP and IP/HR, and F_A structure factors computed using XPREP.[39]

[b] Random-atom trial structures.

[c] Patterson-seeded trial structures.

[d] Seconds on a 300-MHz SGI R12000.

[e] Average time per trial for the F_A data set (estimated from a run on a 833-MHz Compaq Alpha).

[f] Optimum parameters differ from the default values used for single-wavelength differences.

starting atom. ACORN terminates as soon as it finds what it regards as a solution.

SOLVE builds trial structures in ways that make an exact comparison with the other programs difficult. The inflection-point data for AEP transaminase were input to SOLVE, and the automatic protocol for SAD data (specifying a maximum number of 66 sites) was used. SOLVE found 66 sites in 7 h on a 500-MHz Compaq Alpha (~10 h on a 300-MHz SGI R12000), and 65 of these matched the 66 known Se sites with distances in the range of 0.06–0.75 Å. RESOLVE then took the 66 sites and found all six NCS operators automatically, carried out NCS averaging and solvent flattening, and autobuilt a model including side chains for 78% of the 2232 residues.

Weak Anomalous Signals

It has long been the dream of crystallographers to use the resonant scattering from naturally occurring elements, in particular sulfur, to phase protein structures. However, the K absorption edges of sulfur and other, smaller atoms such as phosphorus and chlorine correspond to wavelengths longer than 4 Å, well beyond the tunable range (0.8–2.0 Å) of most synchrotrons. Furthermore, the severe absorption and radiation-damage problems encountered at such long wavelengths are likely to be insurmountable in most cases. It is fortunate, then, that elements such as sulfur retain some anomalous scattering effect even at wavelengths far removed from their absorption edges. It has been 20 years since Hendrickson and Teeter pioneered the use of sulfur anomalous diffraction to solve the structure of a small protein, crambin.[100] Similar applications have been slow to follow, principally because of the difficulty in measuring the small anomalous signal with sufficient accuracy. However, as the applications summarized in Table VIII attest, the ways and means are now being found to conduct the necessary experiments successfully. Tetragonal hen egg-white lysozyme[101] and the metalloprotease thermolysin[102] are previously known test structures used to demonstrate feasibility. Obelin was the first *de novo* structure determined by sulfur anomalous-scattering data and solvent flattening with the latter step carried out at 3.0 Å using the iterative single-wavelength anomalous scattering method first proposed by Wang.[27] In the second *de novo* determination, that of the C_1 subunit of α-crustacyanin, the top six peaks corresponded to a single member of each of the six disulfide moieties present in the asymmetric unit. In some cases, it was necessary to deviate from the default parameters used for the determination of substructures with stronger signals (e.g., use larger phase-to-atom ratios or decrease sigma cutoffs). (See also [5] in this volume[103]).

Two facts stand out regarding the examples in Table VIII. First, X-rays in the wavelength range of 1.5 to ~2.0 Å are chosen to reach a workable compromise that minimizes absorption and radiation-damage effects while maintaining some anomalous signal. (At $\lambda = 1.54$ Å, the $\delta f''$ values are 0.56 electrons for sulfur and 0.70 for chlorine.) Second, highly redundant data are measured in an attempt to maximize accuracy. For example, in the lysozyme study by Weiss,[104] no solutions were obtained when the data were truncated such that the redundancy factor was ~13 or less. One solution out of 5000 trials was obtained with a redundancy of ~16, but this increased to 40 per 5000 trials (0.8% success) when the redundancy was ~25.

[100] W. A. Hendrickson and M. M. Teeter, *Nature* **290,** 107 (1981).

[101] C. C. F. Blake, G. A. Mair, A. C. T. North, D. C. Phillips, and V. R. Sarma, *Proc. R. Soc. B* **167,** 365 (1967).

[102] B. W. Matthews, J. N. Jansonius, P. M. Colman, B. P. Schoenborn, and D. Duporque, *Nat. New Biol.* **238,** 37 (1972).

[103] R. A. P. Nagem, I. Polikarpov, and Z. Dauter, *Methods Enzymol.* **374,** [5], 2003 (this volume).

[104] M. S. Weiss, *J. Appl. Crystallogr.* **34,** 130 (2001).

TABLE VIII
Substructure Determinations Using Weak Anomalous Signals

Protein	Wavelength used (Å)	Redundancy	Space group	d (Å)	kDa/asymmetric unit	Program used	Actual sites	Sites found
Lysozyme[a]	1.54	~23	$P4_32_12$	1.8	14	SHELXD	$S_{10}Cl_8$	17
Lysozyme[b]	1.54	>16	$P4_32_12$	1.63	14	SnB	$S_{10}Cl_8$	?
Thermolysin[c]	1.5–2.1	35–40	$P6_122$	1.83	35	SnB	$ZnCa_5S_3$	15[d]
Obelin[e]	1.74	6	$P6_2$	3.5	22	SOLVE	S_8Cl	9
α-Crustacyanin[f]	1.77	11	$P2_12_12_1$	2.6	40	SnB	S_{12}	6(S–S)

[a] Z. Dauter, M. Dauter, E. de la Fortelle, G. Bricogne, and G. M. Sheldrick, *J. Mol. Biol.* **289,** 83 (1999).

[b] M. S. Weiss, *J. Appl. Crystallogr.* **34,** 130 (2001).

[c] M. S. Weiss, T. Sicker, and R. Hilgenfeld, *Structure* **9,** 771 (2001).

[d] Selected for semiautomated refinement using MLPHARE,[6] DM,[6] and ARP/wARP.[98]

[e] Z.-J. Liu, E. S. Vysotski, C.-J. Chen, J. P. Rose, J. Lee, and B.-C. Wang, *Protein Sci.* **9,** 2085 (2000).

[f] E. J. Gordon, G. A. Leonard, S. McSweeney, and P. F. Zagalsky, *Acta Crystallogr. D Biol. Crystallogr.* **57,** 1230 (2001).

Short Halide Soaks

Pioneering work has shown that the phasing power of the chloride anions present in tetragonal lysozyme can be exploited further by substituting their higher homologs bromine and iodine, either by replacing the NaCl in the crystallization buffer by NaBr[105] or by a quick soak (less than 1 min) of crystals in a cryobuffer containing concentrated (e.g., 0.25–1.0 *M*) halide salt.[106] The latter method appears to be generally applicable, and it leads to incorporation of anomalous scatterers into the ordered solvent regions around protein molecules.[106,107] The bromine K absorption edge at 0.92 Å can be employed for MAD experiments, and either bromine or iodine can be used in the SAD or SIRAS approach. In practice, the use of a single, near-remote wavelength has been used effectively to solve structures of bromine-soaked crystals. Prolonging the soak time beyond about 20 s does not seem to lead to greater incorporation of halide ions, but a higher concentration of salt leads to more sites with higher occupancies.

Table IX contains a listing of some previously unknown protein structures determined with the aid of halide cryosoaks. Direct methods were used to locate the halide substructures. The primary difference between these applications and those described in the previous sections is that the total number of sites to be found was uncertain. (Fortunately, the formula $\Delta F_{\mathrm{ANOM}}/F = 2^{1/2} \times [(f'' \times N_{\mathrm{A}}^{1/2})/(6.7 \times N_{\mathrm{P}}^{1/2})]$, where N_{A} and N_{P} are the numbers of anomalously scattering and protein atoms, respectively, gives an indication of the equivalent number of fully occupied anomalous sites when applied to the low-resolution data.[107] It appears, however, that this uncertainty has not been a significant problem, and the number of sites selected from the direct-methods map can be arbitrary. There is no sharp boundary between the strong, highly occupied sites and noise. In general, it is a good idea to underestimate the number of sites initially so that figures of merit do not become "diluted" by the inclusion of incorrect sites. Additional sites can be located easily using appropriate residual maps.

Obtaining the Programs

Detailed information about each of the programs described in this chapter, including instructions for downloading, can be obtained from their respective Web sites (Table X).

[105] Z. Dauter and M. Dauter, *J. Mol. Biol.* **289,** 93 (1999).

[106] Z. Dauter, M. Dauter, and K. R. Rajashankar, *Acta Crystallogr. D. Biol. Crystallogr.* **56,** 232 (2000).

[107] Z. Dauter and M. Dauter, *Structure* **9,** R21 (2001).

TABLE IX

SUBSTRUCTURE DETERMINATIONS USING HALIDE SOAKS

Protein	Salt concentration	Soak time	Space group	d (Å)	kDa/asymmetric unit	Program used	Sites used	Total sites
β-Defensin-2[a]	0.25 *M* KBr[b]	60	$P2_12_12$	?	16	SHELXS	9	?
Yeast YKG9[c]	0.5 *M* NaBr	45	$P4_32_12$	2.8	36	SnB[d]	7	?
PCP[e]	1.0 *M* NaBr	30	$P6_2$	1.8	37	SHELXD	9	22
Thioesterase 1[f]	1.0 *M* NaBr	20	$P2_1$	1.8	56	SnB	7	40

[a] D. M. Hoover, K. R. Rajashankar, R. Blumenthal, A. Puri, J. J. Oppenheim, O. Chertov, and J. Lubkowski, *J. Biol. Chem*. **275,** 32911 (2000).

[b] 0.25 *M* KI also used.

[c] Y.-S. J. Ho, L. M. Burden, and J. H. Hurley, *EMBO J*. **19,** 5288 (2000).

[d] SHELXS also used.

[e] Z. Dauter, M. Li, and A. Wlodawer, *Acta Crystallogr. D Biol. Crystallogr*. **57,** 239 (2001).

[f] Y. Devedjev, Z. Dauter, S. R. Kuznetsov, T. L. Z. Jones, and Z. S. Derewenda, *Struct. Fold. Des*. **8,** 1137 (2000).

TABLE X
Software Web Sites for Substructure Determination

Program	Web site
SOLVE	http://www.solve.lanl.gov
CNS	http://cns.csb.yale.edu
SnB	http://www.hwi.buffalo.edu/SnB/
SHELXD	http://shelx.uni-ac.gwdg.de/ SHELX/
CCP4	http://www.ccp4.ac.uk

The various sites feature a variety of instructional material. For example, the CNS distribution contains Web-based tutorials describing the steps required for MIR, MAD, and SAD phasing. The SnB site features a short tutorial on direct methods. Most of these programs are available at no cost to nonprofit organizations.

Acknowledgments

The authors thank Stephen Potter, who created and maintained a Web site for reporting substructure results; the many users who responded with information about their applications; and Osnat Herzberg and Celia Chen, who allowed us to use their data for 2-aminoethylphosphonate transaminase as test data. We are also grateful for financial support: NIH GM-46733 (C.M.W.), NIH 1P50GM-62412 (P.D.A.), and U.S. Department of Energy Contract No. DE-AC03-76SF00098 (P.D.A). The CCP4 programs have been collected and developed under the auspices of Collaborative Computing Project Number 4, in Protein Crystallography, supported by the U.K. Science and Engineering Research Councils and coordinated at Daresbury Laboratory.

[4] Use of Noble Gases Xenon and Krypton as Heavy Atoms in Protein Structure Determination

By Marc Schiltz, Roger Fourme, and Thierry Prangé

Introduction

Xenon and krypton derivatives of proteins can be obtained by subjecting a native protein crystal to a xenon or krypton gas atmosphere pressurized in the range of 1–100 bar.[1] The noble gas atoms are able to diffuse rapidly toward potential interaction sites in proteins via the solvent channels that are always present in crystals of macromolecules. The number and occupancies of xenon/krypton-binding sites vary with the

0076-6879/03 $35.00