Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Substructure solution with *SHELXD*

Thomas R. Schneider and George M. Sheldrick*

Department of Structural Chemistry, Göttingen University, Tammannstrasse 4, D-37077 Göttingen, Germany

Correspondence e-mail: gsheldr@shelx.uni-ac.gwdg.de

Received 30 May 2002
Accepted 2 July 2002

Iterative dual-space direct methods based on phase refinement in reciprocal space and peak picking in real space are able to locate relatively large numbers of anomalous scatterers efficiently from MAD or SAD data. Truncation of the data at a particular resolution, typically in the range 3.0–3.5 Å, can be critical to success. The efficiency can be improved by roughly an order of magnitude by Patterson-based seeding instead of starting from random phases or sites; Patterson superposition methods also provide useful validation. The program *SHELXD* implementing this approach is available as part of the *SHELX* package.

1. Introduction

Recent advances in synchrotron technology, cryocrystallography and the incorporation of selenomethionine into proteins make the multiwavelength anomalous diffraction (MAD) method (Hendrickson, 1991; Smith, 1998) an effective approach to the solution of protein structures. Soaking with bromides or iodides combined with MAD or single-wavelength anomalous diffraction (SAD; Dauter *et al.*, 1999, 2000; Dauter & Dauter, 1999) or even SAD applied to native anomalous scatterers such as sulfur or phosphorus (Weiss *et al.*, 2001; Dauter & Adamiak, 2001) are potential alternatives. These approaches can result in relatively large numbers (50 or more) of heavy-atom sites, especially when large protein complexes are investigated.

In principle, the MAD approach, in which data are collected at two or more wavelengths for which the f' and f'' anomalous scattering factors are non-zero for at least one of the elements present, determines experimental phases directly. There is however a hidden *phase problem*: it is still necessary to find the positions of the anomalous scatterers in order to calculate the reference phases. Without these heavy-atom reference phases the protein phases cannot be found. Hand interpretation of the Patterson function is hardly a viable option, but conventional small-molecule direct methods (Wilson, 1978; Yao, 1981; Mukherjee *et al.*, 1989; Sheldrick, 1990) and automated computer Patterson interpretation (Sheldrick *et al.*, 1993; Sheldrick, 1998; Terwilliger & Berendzen, 1999; Grosse-Kunstleve & Brunger, 1999) are often successful in locating the heavy-atom sites. An alternative direct-methods approach to substructure solution has been described recently by de Graaff *et al.* (2001). Although small-molecule direct methods require data to atomic resolution (1.2 Å or better), direct methods are still effective for the solution of heavy-atom substructures because the distances between the heavy atoms are usually appreciably greater than the resolution of the data (typically about 3 Å). Nevertheless, these methods do not

always succeed when there are a large number of sites. Conventional direct methods tend to be upset by a few aberrant reflections which are common when working with weak anomalous differences; probability formulas such as negative quartets that depend on the weak E values cannot be used because these E values are unreliable for difference structure factors. Patterson interpretation methods can suffer from the effect of accumulated coordinate errors and false assignments if the atoms are found in a stepwise manner. In contrast, dual-space direct methods (Miller *et al.*, 1993; Weeks & Miller, 1999; Sheldrick *et al.*, 2001) appear to be robust and efficient for large substructures. We describe here the implementation of the dual-space strategy and the integration with Patterson methods in the program *SHELXD* and analyze the factors critical for the success of this approach.

2. Methods

2.1. Data-quality control

The main problem with SAD, SIR or MAD data is that they are noisy because they are based on small differences between observed structure factors; the best antidote is to collect highly redundant data (Weiss *et al.*, 2001; Dauter & Adamiak, 2001). On the other hand, the resolution and completeness of the F_A data are less critical: 3.5 Å is adequate, since the anomalous atoms are more than 3.5 Å apart, and the problem is still highly over-determined. Although higher resolution and completeness are not required to find the anomalous scatterers, they do have a major influence on the quality of the resulting electron-density maps (Brodersen *et al.*, 2000).

Before attempting to use MAD or SAD data to locate anomalous scatterers, a critical decision to be made is to decide which resolution the data should be truncated. If data are used to a higher resolution than there is significant dispersive and anomalous information, the effect will be to add noise. Since direct methods are based on normalized structure factors, which emphasize the high-resolution data, they are particularly sensitive to this. An effective test is to calculate the correlation coefficient between the signed anomalous differences ΔF at different wavelengths as a function of the resolution. To a first-order approximation, assuming that the anomalous differences are small compared with the native structure factors, the anomalous differences at different wavelengths should be related by a positive proportionality constant given by the ratio of their f'' values. The fact that correlation coefficients between the anomalous ΔF values at different wavelengths can be greater than 95% for very high quality data indicates that the approximations involved are acceptably small. The high-energy remote wavelength is usually the best choice as a reference for calculating the correlation coefficients, because it still contains significant anomalous signal and is insensitive to wavelength drift. A good general rule is to truncate the data where this correlation coefficient falls below about 25–30%. This procedure can also indicate if there is a major problem with the data set. 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. If only one set of anomalous data is available, the correlation coefficient cannot be calculated, but it is still possible to calculate the ratio of ΔF to its estimated standard deviation as a function of the resolution. It is recommended that the data are not merged when they are scaled, so that the agreement of the equivalent reflections provides additional statistical information. Provided that it has been possible to propagate good estimates of standard deviations through each stage of the data processing, the data can be truncated at the resolution at which this ratio drops to below about 1.3. If even this information is not reliably available, then a useful rule of thumb is to truncate the data to about 0.5 Å less than the diffraction limit of the crystal employed for data collection. The program *XPREP* (Bruker Nonius, 2001) was used for all the preliminary data processing and statistics in this work; *XPREP* employs local scaling (Matthews & Czerwinski, 1975) and uses essentially the same approximations in deriving MAD F_A values as described by Terwilliger (1994).

2.2. Dual-space recycling

The *dual-space* recycling approach (also known as *Shake-and-Bake*) was implemented in the computer programs *SnB* (Miller *et al.*, 1994) and more recently in *SHELXD*. It has been reviewed recently in detail (Usón & Sheldrick, 1999; Sheldrick *et al.*, 2001) and the optimization of the *SnB* program for substructures has been described by Howell *et al.* (2000). The *Shake-and-Bake* algorithm is of necessity based on the strongest normalized difference structure factors E, typically corresponding to the largest ∼15% of the observed difference structure factors ΔF (SIR or SAD) or F_A (MAD or SIRAS) in each resolution shell, because the probability formulas only provide significant phase information for the strongest E values. For SIR or SAD phasing, the smaller E values are in any case unreliable because they represent lower limits on the normalized heavy-atom structure factors. The use of only a fraction of the total number of reflections is also a main reason for the speed and efficiency of dual-space recycling.

The dual-space approach alternates between real and reciprocal space. In reciprocal space, phases are refined in *SnB* by reducing the *minimal function* (Miller *et al.*, 1993) or expanded in *SHELXD* from the ∼40% most reliable using the *tangent formula* (Karle & Hauptman, 1956; Karle, 1968). Both these techniques appear to be good ways of propagating phase information throughout reciprocal space; however, both would, if used exclusively, lead to phase divergence away from a chemically sensible (*e.g.* equal-atom) arrangement of sites. This is why the alternate real-space cycles are required to impose the strong constraint that we expect to find N sites with approximately equal scattering power. The real/reciprocal-space combination appears to be a particularly effective searching algorithm. Peak-picking can be regarded as an extreme (and computationally efficient) form of density

modification and enables a minimum distance criterion to be applied to eliminate unreasonably close heavy atoms. Similarly, it is usually desirable (and is the default option in *SHELXD*) to ignore peaks on special positions, although it is not unknown for heavy-atom derivatives produced by soaking to have sites on special positions. Ghost peaks on special positions are often characteristic of false solutions in small-molecule direct methods. The number of unique substructure sites N is assumed to be known approximately and the selection of N peaks probably provides a useful constraint on the structure solution. The dual-space recycling is typically performed for several hundred or more sets of N random starting atoms, with typically $2N$ cycles for each. The application of dual-space recycling to substructures is summarized in Fig. 1.

In the course of testing *SHELXD* for *ab initio* solution of structures from native data, it was discovered by accident that a very effective procedure is to leave out about 30% of the peaks at random when calculating phases for the next cycle. In retrospect, it is possible to understand why this is an effective search strategy by analogy with the *omit maps* (Hodel *et al.*, 1992) frequently used in macromolecular crystallography. If the deleted atoms are part of an essentially correct solution, they will probably be regenerated; if not, they will be replaced by different, possibly better, potential atoms.

2.3. Starting atoms consistent with the Patterson function

The efficiency of the dual-space algorithm can be improved appreciably by using starting atoms consistent with the Patterson function rather than random starting atoms. Our algorithm for generating starting atoms that are consistent with the Patterson makes extensive use of a special form of the Patterson minimum function (PMF) proposed by Nordman (1966). Two atoms are placed in a unit cell and all their symmetry equivalents generated. The Patterson function values corresponding to all unique vectors involving these atoms are sorted into ascending order and the PMF is then calculated as the mean value of the lowest (say) 30% of the values in this list. Since it is unlikely that this PMF will have a high value for wrong atom positions, especially when the symmetry is high and there are many vectors, it may be used as a criterion for a translational search for a two-atom fragment. Each strong general Patterson peak is in principle a suitable two-atom 'fragment' for this translational search, because it may well correspond to a vector between two heavy atoms. Since we are only interested in generating many different sets of atom coordinates consistent with the Patterson function, there is no need to determine the global maximum PMF; indeed, often this does not give good starting atoms for the dual-space recycling. A simple and effective approach is to try a fixed number (usually in the range 9999–99 999) of random translations for a vector and retain the one with the highest PMF. A random selection of vectors from the Patterson peak list (excluding Harker peaks), biased so that the high peaks are chosen more often, is an effective way to pick the two-atom search fragment. For substructure solution we use an unsharpened Patterson, though for locating heavy atoms from native data without the use of the anomalous signal it may be better to sharpen the Patterson.

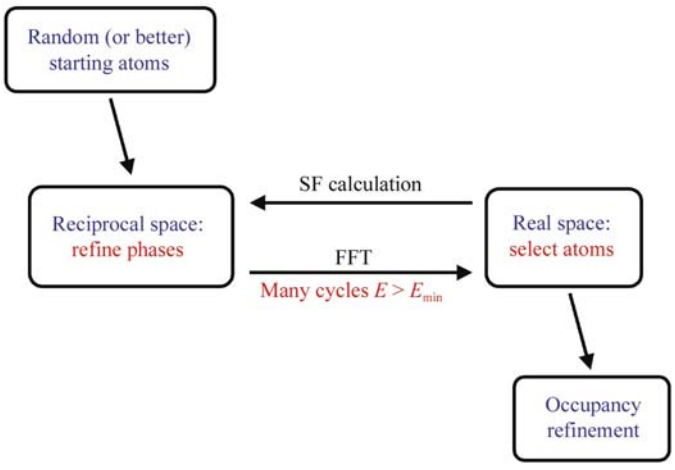

Figure 1
Flow diagram for dual-space recycling substructure solution.

Before the first dual-space cycle, the two starting atoms need to be extended to N atoms. A difference Fourier synthesis would be effective for a small number of heavy atoms, but a better technique for a large number is to calculate a full-symmetry Patterson superposition minimum function (PSMF; Buerger, 1959). Firstly, all symmetry equivalents are generated for the two starting atoms. Each pixel of the PSMF map is assigned a value equal to the PMF for all vectors between these atoms and a dummy atom placed at the pixel. Peaks are then obtained by map interpolation and sorted in the usual way.

By applying this procedure before each run through the dual-space recycling, it is possible to generate an unlimited number of different sets of starting atoms, all more or less consistent with the Patterson function. Our tests have shown that this combination of direct and Patterson methods produces more complete and precise solutions than using Patterson methods alone. It appears that iterative Patterson-only procedures suffer from an accumulation of atomic coordinate errors each time a new atom is added. Because it includes phase refinement, the dual-space approach does not suffer from this degradation as the number of atoms increases.

2.4. Occupancy refinement

SHELXD provides the option of refining the occupancies of the atoms after the peak-search in the final dual-space cycles. Although originally intended to handle the problem of the fractional occupancies encountered in derivatives obtained by soaking (especially halide soaks), as described below it also helps to identify the number of anomalous scatterers in the equal-atom case, *e.g.* when the number of selenomethionine sites is less than expected because of disorder or when unexpected anomalous scatterers are present.

2.5. Validation of the solutions

In *SHELXD*, potential solutions are identified by high values of the correlation coefficient CC between E_o and E_c (Fujinaga & Read, 1987),

$$\mathrm{CC} = 100\,\frac{\left(\sum wE_oE_c\sum w - \sum wE_o\sum wE_c\right)}{\left\{\left[\sum wE_o^2\sum w - \left(\sum wE_o\right)^2\right]\left[\sum wE_c^2\sum w - \left(\sum wE_c\right)^2\right]\right\}^{1/2}}.$$

For *ab initio* applications, these potential solutions can be improved and extended by means of *peak-list optimization* (Sheldrick & Gould, 1995), which finds the set of potential sites that maximizes CC for all reflections. This is not used for substructures because of the unreliable weak E values. Nevertheless, the CC values calculated both with all E values and with only the E values not used directly for substructure solution (analogous to the use of the free R factor; Brünger, 1992) provide good indications as to whether the substructure sites are correct. The weights w can be used to weight down the less reliable ΔF estimates; in *SHELXD* w is set to $[1 + g\sigma^2(E)]^{-1}$, with a default value of 0.1 for g.

The Patterson superposition function is also the basis of the *crossword table* (Sheldrick *et al.*, 1993; Sheldrick, 1998) introduced in *SHELXS*86, which provides a convenient way to assess which of the heavy-atom sites are correct and also in some cases to recognize the presence of non-crystallographic symmetry. In this table, the rows and columns correspond to the potential atoms. For each pair of atoms the top number is the minimum distance between them, taking the space-group symmetry into account, and the bottom number is the PMF 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.* between one atom and its symmetry equivalents. In general, wrong sites can be recognized in this table by the presence of several zero PMF values (negative values are replaced by zero). The mean PMF value for a specified number of atoms provides a figure of merit PATFOM that can be useful for selecting the best solution, although the absolute value depends on the structure in question and tends to be smaller for larger structures.

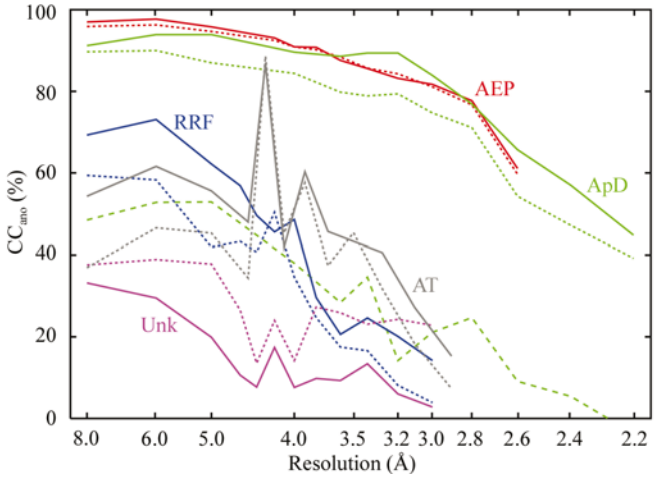

Figure 2
Correlation coefficients (calculated using *XPREP*) expressed as percentages between the high-energy remote data and the two or three other wavelengths collected in MAD experiments on five different proteins. For AEP (red) and ApD (green) the high values involving the peak (solid line) and inflection-point (dotted line) data show that it is not necessary to truncate the data; there is significant MAD information up to the highest resolution collected. A poorer correlation would be expected with the low-energy remote data for ApD (dashed green line) which has a much smaller anomalous signal. For RRF (blue) it would be advisable to truncate the data to about 3.9 Å (which indeed led to a successful solution using *SHELXD*) and for AT (gray) 3.5 Å is appropriate (see Fig. 3). Unk (purple) is clearly hopeless and in fact could not be solved.

3. Results

The approach described above has been tested on MAD (and SAD) data for a set of known substructures, some of which were originally solved with *SHELXD* and some with other programs. The overall results are summarized in Table 1. With two exceptions, the top N peaks corresponded to the N correct known sites and had peak heights that were significantly higher than the highest noise peaks. The smaller structures required a few seconds per solution and, except for the 160-site SAD problem, all gave an appreciable number of solutions per hour on a 1 GHz PC using default settings for all parameters. In all cases, the correlation coefficients decisively identified the correct solutions. This was also in general true of the PATFOM values (not shown), but since the actual values obtained depend on the size of the substructure and on the complexity of the space group, PATFOM does not provide a good indication of whether a substructure has been solved or not. In our experience with substructure solution, when a group of correlation coefficients are well clear of the rest and have values greater than 35% they always correspond to correct solutions; for *ab initio* solution of complete structures at atomic resolution the corresponding threshold is about 65%. In the case of SAD, correct substructure solutions have been found with correlation coefficients of less than 25%. The minimal function (Miller *et al.*, 1993), the tangent-formula residual R_α (Sheldrick, 1990) and the conventional R factor were also successful at identifying the correct solutions for a given structure, but not quite as useful as the correlation coefficient as absolute figures of merit.

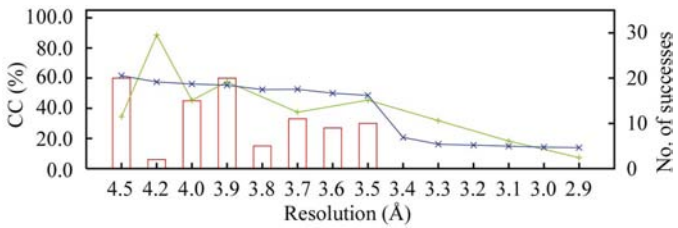

Figure 3
The effect of different resolution cutoffs on the success rate of *SHELXD* in solving the AT substructure (Hensgens *et al.*, 2002). The correlation coefficient between the signed $|F_{hkl}| - |F_{\overline{hkl}}|$ differences of the high-energy remote and peak-wavelength data (CC_{ano}) is shown in green. For each resolution cutoff, *SHELXD* was run for 100 tries. The correlation coefficient between E_o and E_c for the best solution (CC_{max}) is shown in blue and the number of solutions per 100 tries (#succ) as red boxes.

Table 1
Some applications of integrated Patterson/direct methods to the location of the anomalous scatterers from MAD data.

Protein	No. of sites†	Space group	MW‡ (kDa)	d_{min}§ (Å)	*CC* (%)	*P* ratio¶	Solns††	Ref.‡‡
ApD	3/3 Se	$C222_1$	16	2.2	45	16	512	i
RRF	3/4 Se	$P4_32_12$	20	4.0	60	1.4	566	ii
ModE	6/6 Se	$P2_12_12$	57	3.0	66	7.3	326	iii
9hem	18/18 Fe	$P2_1$	64	2.9	73	4.0	480	iv
AT	32/32 Se	*C*2	160	3.5	49	5.1	7	v
Cyanase	40/40 Se	*P*1	170	2.4	57	1.0	132	vi
TH	51/60 Se	$P2_1$	161	2.5	52	2.8	26	vii
AEP	66/66 Se	$P2_1$	270	2.55	61	13	49	viii
KPHMT	145/160 Se	$P2_1$	567	2.8	39	35	0.1	ix

† The first number is the number found and the second is the total number that should be present. ‡ Molecular weight of the asymmetric unit. § The limiting resolution to which the data were processed. ¶ The *P* ratio is the speed-up obtained by using starting atoms consistent with the Patterson function. †† The number of solutions per hour on a 1 GHz Pentium PC. ‡‡ References: (i) Walsh *et al.* (1999), (ii) Selmer *et al.* (1999), (iii) Hall *et al.* (1999), (iv) Mathias *et al.* (1999), (v) Hensgens *et al.* (2002), (vi) Walsh *et al.* (2000), (vii) Buckley *et al.* (2000), (viii) Chen *et al.* (2000), (ix) von Delft (2001).

With the exception of the structure in space group *P*1, Patterson seeding appreciably increased the number of solutions per hour, though this improvement varied widely with the structure and also to some extent with the other parameter settings. In fact, Patterson seeding as explained above cannot work in the space group *P*1 because the PMF depends only on a single vector that does not change if a translation search is performed. To circumvent this problem in *P*1 an extra atom is placed on the origin, so that the PMF depends on three vectors, two of which change when the two-atom fragment is translated. However, the reason for the lack of speed-up in the *P*1 example (cyanase) probably lies in the very high success rate per attempt of the dual-space algorithm in this space group starting even from random atoms (Xu *et al.*, 2000).

Table 2 shows the results obtained with the 66-site AEP problem when the data, originally collected as a three-wavelength MAD experiment, are treated in different ways: two- or three-wavelength MAD, pure anomalous differences, pure dispersive differences or pseudo-SIRAS. With the exception of the pure peak minus high-energy remote dispersive differences, all lead to a satisfactory number of correct solutions per 1000 tries. The wavelength used for the peak data collection was known to be unreliable and our failure to solve the structure with the pure dispersive differences between peak and high-energy remote suggests that these are smaller or less accurate than expected. In general, it requires very high wavelength stability to obtain accurate peak data in a MAD experiment; the white line (which may have been enhanced by oxidation of some of the Se atoms; Sharff *et al.*, 2000) can be very sharp. In the AEP example, the best results are obtained by two-wavelength (inflection-point and high-energy remote) MAD or pseudo-SIRAS treatment; the difference is that in the pseudo-SIRAS analysis only one wavelength (the inflection point) contributes to the anomalous differences, whereas in the MAD analysis both wavelengths contribute. The results show a good correlation between the number of correct solutions per 1000 attempts, the correlation coefficients and the differences in peak height between correct sites and noise peaks. The latter show some memory effects from the use of the random omit procedure, although this is not applied to the final dual-space cycle; the separation between the weakest correct site and the highest noise peak is greater when the random omit is switched off, but then the frequency of correct solutions decreases because the searching is less effective. In this example, the success rate when Patterson seeding is used is already so high that the random omit procedure results in only a minor improvement; however, when the Patterson seeding is not used the random omit procedure improved the success rate by a factor of nearly six.

Table 2
Tests of different ways of treating the experimental data using the 66-site AEP structure (Chen *et al.*, 2000) showing the higher success rate when Patterson seeding is used.

HR, high-energy remote, PK, peak; IP, inflection point. The ΔF or F_A values were calculated using *XPREP*. In each case, 1000 attempts were made. The correlation coefficients were calculated using all data and also using only the weak reflections, *i.e.* those not used directly in the dual-space recycling ($E < 1.5$). The peak heights from the final dual-space cycle have been normalized so that the highest peak has a height of 1.0. CC_{weak} is the correlation coefficient between E_o and E_c calculated using only those reflections not used to locate the heavy atoms.

Source of ΔF	Solutions (with Patterson)	Solutions (without)	Best CC (%)	Best CC_{weak} (%)	Height of peak 66	Height of peak 67
IP/HR (MAD)	697	175	62.9	52.7	0.559	0.282
IP/HR (MAD)†	655	30	62.9	52.7	0.563	0.176
IP/HR (SIRAS)	652	137	62.1	52.2	0.512	0.273
IP/PK/HR (MAD)	564	61	60.9	51.4	0.517	0.302
IP (SAD)	424	150	54.1	31.2	0.342	0.195
PK (SAD)	380	93	51.9	30.0	0.484	0.222
HR (SAD)	148	25	47.2	27.3	0.547	0.291
IP/HR (energy-dispersive only)	127	68	43.6	27.1	0.346	0.332
PK/HR (energy-dispersive only)	0	0	9.9	2.0	0.350	0.347

† In this row only, the random omit option was switched off.

In the above example, the anomalous and dispersive signals are relatively strong and it was possible to use the full resolution range in all the tests. When these signals are weak, the choice of the resolution at which to truncate the data can be critical. Fig. 2 shows the utility of the correlation coefficient between signed $|F_{hkl}| - |F_{\overline{hkl}}|$ differences in deciding where to truncate. The effects of different resolution thresholds on the success of the substructure solution are illustrated in Fig. 3. In the case of AT (Hensgens *et al.*, 2002), *SHELXD* did not find any solutions when any data to higher resolution than 3.5 Å were included. For different cutoffs between 3.5 and 4.5 Å, the quality of the best solution as measured by the highest correlation coefficient achieved (CC_{max}) in a run of 100 tries is very similar. However, the success rate shows a strong variation with the resolution cutoff, which probably arises from a resolution dependence of the data quality (possibly caused by different levels of diffuse X-ray scatter).

Since the correlation coefficient (CC) and PATFOM figures of merit are physically independent, they can be used to produce a two-dimensional scatter plot that sometimes shows clusters of solutions both for the correct solution and for

pseudo-solutions. An example is shown in Fig. 4. Correct solutions and false solutions show a characteristic bimodal distribution, in this case with one outlier; however, a bimodal distribution is not a necessary condition for structure solution because for straightforward problems it sometimes happens that all solutions are correct.

The exact number of ordered anomalously scattering sites in a given crystal is not always known. To test the robustness of *SHELXD* when searching for an incorrect number of sites, the program was run requesting different numbers of sites for the AT substructure (Fig. 5). All scenarios, except when 20 sites were requested without occupancy refinement, gave complete solutions of the substructure (for 20 sites, the program will only output the highest 28 peaks). Without the occupancy refinement against all the substructure structure factors after the final dual-space cycle, a sharp drop in peak height is observed right after the number of peaks requested. In this case, it is not possible to deduce the correct number of sites. However, with peak-height refinement, the sharp drop occurs between site number 32 and site number 33, independent of the actual number of sites requested. In fact, if no drop in peak height is observed (as here for the case of 20 requested sites), this can be taken as an indication that the substructure consists of more sites than expected. In such cases *SHELXD* should be re-run with an increased number of sites requested.

For large and difficult substructure problems, the presence of a potential site in most of the 'correct' solutions can be an indication that this site is correct; sites that only appear in a few solutions are more likely to be noise peaks. This can be exploited by selecting all the solutions that form the cluster judged best in terms of CC, converting the phase sets to the same origin and enantiomorph and combining them by vectorial addition of the transformed normalized structure factors. This might be expected to improve both the percentage of correct sites and also their accuracy. Table 3 illustrates the results of this procedure for the 160-site KPHMT problem. Although the selenium sites would not be expected to agree perfectly with the methionine sulfur positions in the native structure, it is clear that reciprocal-space vector averaging of the best 16 solutions has resulted in a small but significant improvement in the completeness and accuracy of the substructure.

Table 3
Cumulative numbers of selenium sites from the SAD substructure solution within the given distance from a methionine S atom in the native structure for the 160-site KPHMT test (von Delft, 2001).

The numbers are given as N_{160}/N_{200}, where N_{160} is based on the top 160 peaks and N_{200} on the top 200 peaks.

	<0.5 Å	<1.0 Å	<1.5 Å	<2.0 Å
Solution with the best CC	85/87	137/140	143/147	145/149
Best 16 solutions combined	97/97	140/143	149/154	152/157

In addition to its use in verifying the overall solution and the individual sites, the crossword table is also a valuable source of information about non-crystallographic symmetry. This provides a further indication as to which sites are correct and enables rotation matrices and translation vectors to be derived that can be used subsequently for density modification with NCS averaging (Bricogne, 1976; Cowtan & Zhang, 1999). An example is illustrated in Fig. 6.

4. Conclusions

The procedure described in this paper is robust and sufficiently fast so that it is unlikely to become the rate-determining stage of structure determination, except possibly

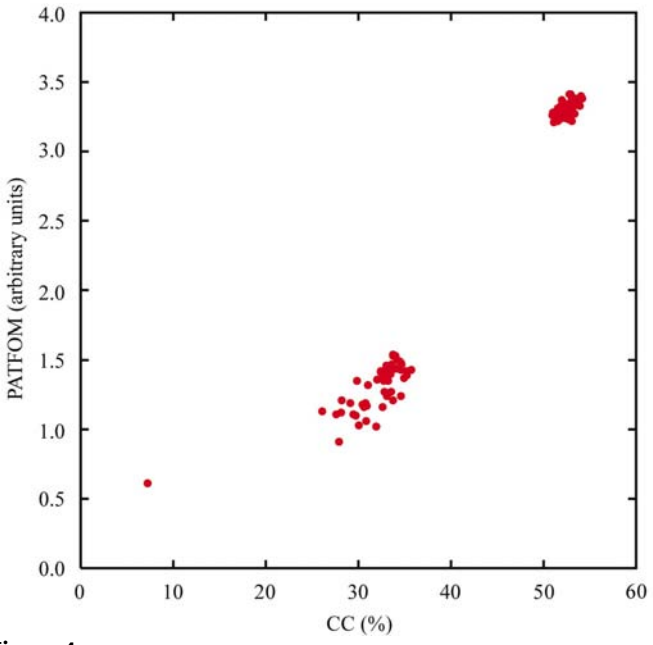

Figure 4
Scatterplot of CC *versus* PATFOM scatter plot for structure TH (Buckley *et al.*, 2000) showing clusters of correct solutions and pseudo-solutions.

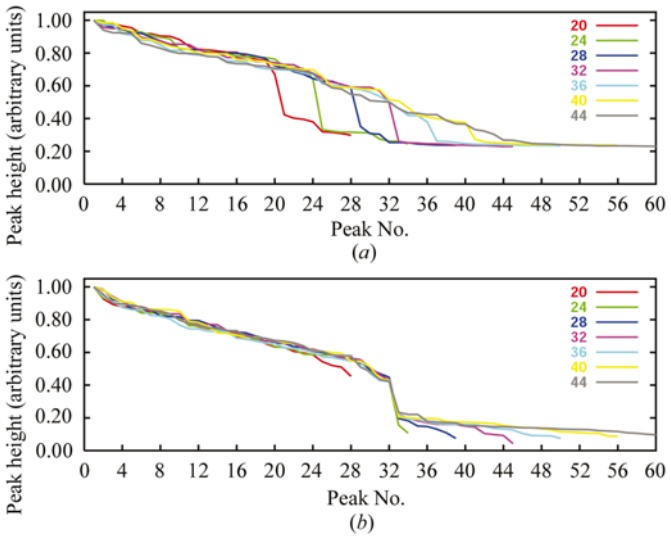

Figure 5
Peak height *versus* peak number of the best solution out of 100 attempts to solve the substructure of AT using MAD F_A values limited to 3.5 Å resolution (*a*) without occupancy refinement and (*b*) with occupancy refinement. For each curve, a different number of sites was requested. The pink lines correspond to the correct number of sites (32).

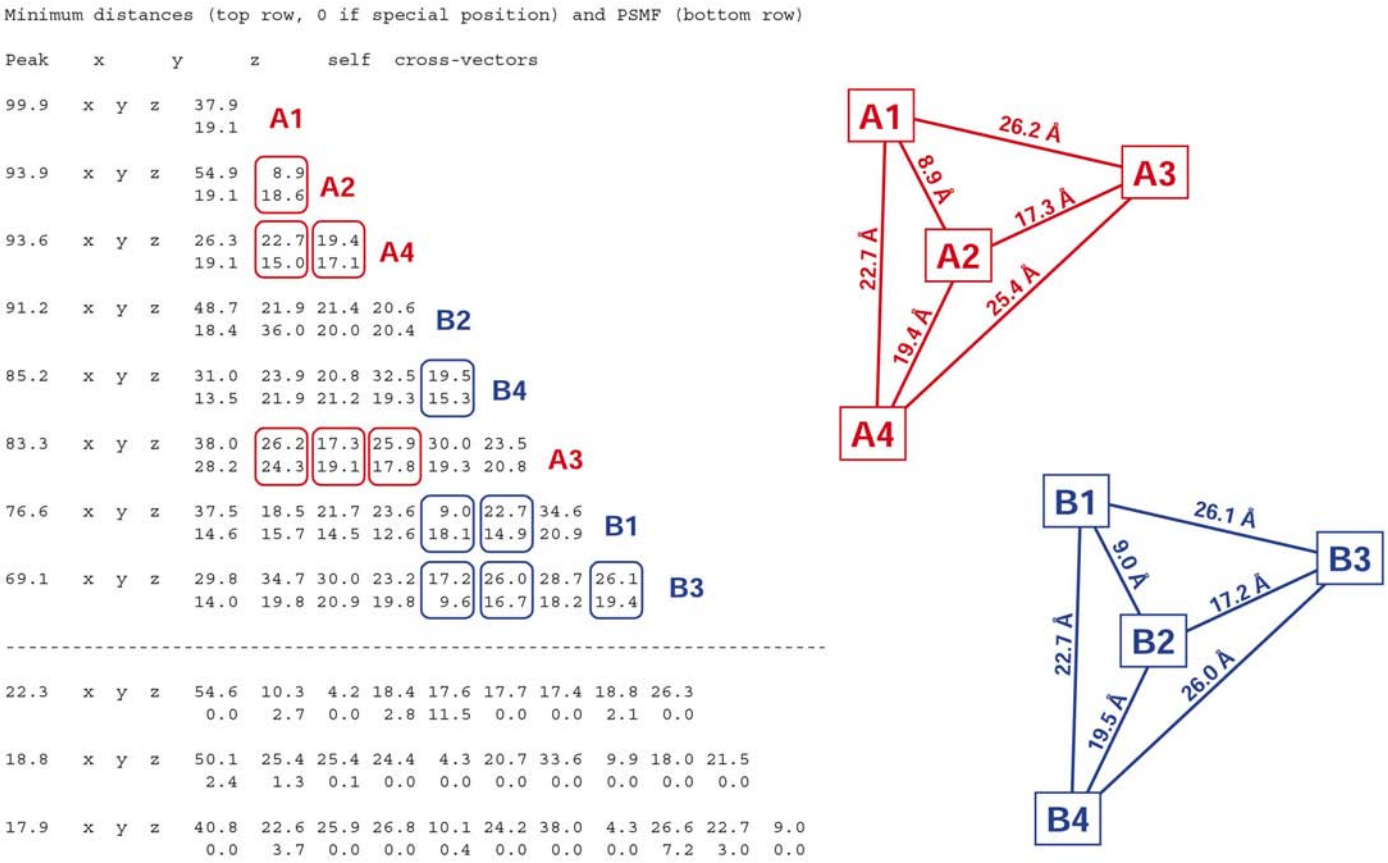

Figure 6
Example of a crossword table, showing the interpretation in terms of two similar molecules each containing four anomalous scatterers. Each entry in the lower triangular matrix consists of two numbers: the minimum distance between two atoms (upper number) and the PMF between them (lower), taking all symmetry equivalents into account. The row corresponds to one atom and the column to the other. The first column corresponds to the vectors between one atom and its equivalents.

for very large substructures with weak anomalous signals. If necessary, it can very simply and efficiently be 'parallelized' by running the program with different random number seeds on a large number of processors at the same time. In the large majority of applications, the only critical starting parameter is the resolution at which to truncate the ΔF data, although if the data-to-parameter ratio becomes too low (because the data are incomplete or when it is necessary to truncate at rather low resolution) it may be necessary to reduce the minimum E value for the reflections used in the dual-space recycling (usually set to 1.5). It is also of advantage if the number of unique sites N is approximately known, as is normally the case with selenomethionine MAD experiments; in unclear cases, it might be a good idea to try different values for N. The substructure determination appears to work almost as well with single-wavelength (SAD) data as with multiple-wavelength (MAD) data, although in some cases a two-wavelength MAD experiment may give the best results. It is possible that if a SAD data set were collected for the same total time (*i.e.* higher redundancy) the results would be as good as MAD and less accident-prone. On the other hand, it should be remembered that a MAD (or SIRAS) experiment will give more precise phases than SAD: the MAD data can be analyzed to obtain the phase shift between the substructure and the protein, whereas the SAD phase shifts are subject to a twofold ambiguity and there are no SAD estimates for the phases of reflections in centrosymmetric projections.

The robustness and efficiency of the dual-space approach with Patterson seeding and its ability to solve larger substructures make it an eminently suitable approach for high-throughput structural genomics. Solving the substructure is of course a necessary but not sufficient requirement for solving the complete structure!

We are grateful to Patrick Baker, Patrick Buckley, Bauke Dijkstra, Zbigniew Dauter, Frank von Delft, Charles Hensgens, Osnat Herzberg, Gordon Leonard, Maria Selmer and Martin Walsh for providing test data and encouragement, to Isabel Usón for discussions and to the Fonds der Chemischen Industrie and to the European Union (QLRI-CT-2000-00398) for support.

References

Bricogne, G. (1976). *Acta Cryst.* A**32**, 832–847.
Brodersen, D. E., de La Fortelle, E., Vonrhein, C., Bricogne, G., Nyborg, J. & Kjeldgaard, M. (2000). *Acta Cryst.* D**56**, 431–441.
Bruker Nonius (2001). *XPREP*, Version 6.12. Bruker Nonius, Madison, Wisconsin, USA.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Buckley, P. A., Jackson, J. B., Schneider, T. R., White, S. A., Rice, D. W & Baker, P. J. (2000). *Structure*, **8**, 809–815.
Buerger, M. J. (1959). *Vector Space.* New York: Wiley.
Chen, C. C. H., Kim, A., Zhang, H., Howard, A. J., Sheldrick, G. M., Dunaway-Mariano, D. & Herzberg, O. (2000). Abstr. Am. Crystallogr. Assoc. Meet. Astract 02.06.03.
Cowtan, K. & Zhang, K. Y. J. (1999). *Prog. Biophys. Mol. Biol.* **72**, 245–270.
Dauter, Z. & Adamiak, D. A. (2001). *Acta Cryst.* D**57**, 990–995.
Dauter, Z. & Dauter, M. (1999). *J. Mol. Biol.* **289**, 93–101.
Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.
Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Acta Cryst.* D**56**, 232–237.
Delft, F. von (2001). Personal communication.
Fujinaga, M. & Read, R. J. (1987). *J. Appl. Cryst.* **20**, 517–521.
Graaff, R. A. G. de, Hilge, M., van der Plas, J. L. & Abrahams, J. P. (2001). *Acta Cryst.* D**57**, 1857–1862.
Grosse-Kunstleve, R. W. & Brunger, A. T. (1999). *Acta Cryst.* D**55**, 1568–1577.
Hall, D. R., Gourley, D. G., Duke, E. M., Leonard, G. A., Anderson, L. A., Pau, R. N., Boxer, D. H. & Hunter, W. N. (1999). *Acta Cryst.* D**55**, 542–543.
Hensgens, C. M. H., Kroezinga, E. A., van Montfort, B. A., van der Laan, J.-M., Sutherland, J. D. & Dijkstra, B. W. (2002). *Acta Cryst.* D**58**, 716–718.
Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Hodel, A. A., Kim, S.-H. & Brünger, A. T. (1992). *Acta Cryst.* A**48**, 851–858.
Howell, P. L., Blessing, R. H., Smith, G. D. & Weeks, C. M. (2000). *Acta Cryst.* D**56**, 604–617.
Karle, J. (1968). *Acta Cryst.* B**24**, 182–186.
Karle, J. & Hauptman, H. (1956). *Acta Cryst.* **9**, 635–651.
Mathias, P. M., Coelho, R., Pereira, I. A. C., Coelho, A., Thompson, A. W., Sieker, L. C., Le Gall, J. & Carrondo, M. A. (1999). *Structure*, **7**, 119–130.
Matthews, B. W. & Czerwinski, E. W. (1975). *Acta Cryst.* A**31**, 480–487.
Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.
Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *J. Appl. Cryst.* **27**, 613–621.
Mukherjee, A. K., Helliwell, J. R. & Main, P. (1989). *Acta Cryst.* A**45**, 715–718.
Nordman, C. E. (1966). *Trans. Am. Cryst. Assoc.* **2**, 29–38.
Selmer, M., Al-Karadaghi, S., Hirokawa, G., Kaji, A. & Liljas, A. (1999). *Science*, **286**, 2349–2352.
Sharff, A. J., Koronakis, E., Luisi, B. & Koronakis, V. (2000). *Acta Cryst.* D**56**, 785–788.
Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Sheldrick, G. M. (1998). *Direct Methods for Solving Macromolecular Structure*, edited by S. Fortier, pp. 131–141. Dordrecht: Kluwer Academic Publishers.
Sheldrick, G. M., Dauter, Z., Wilson, K. S. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.
Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423–431.
Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R. & Usón, I. (2001). *International Tables for Crystallography*, Vol. F, edited by E. Arnold & M. Rossmann, pp. 333–351. Dordrecht: Kluwer Academic Publishers.
Smith, J. L. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 211–225. Dordrecht: Kluwer Academic Publishers.
Terwilliger, T. C. (1994). *Acta Cryst.* D**50**, 17–23.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Usón, I. & Sheldrick, G. M. (1999). *Curr. Opin. Struct. Biol.* **9**, 643–648.
Walsh, M. A., Dementieva, I., Evans, G., Sanishvili, R. & Joachimiak, A. (1999). *Acta Cryst.* D**55**, 1168–1173.
Walsh, M. A., Otwinowski, Z., Perrakis, A., Anderson, P. M. & Joachimiak, A. (2000). *Structure*, **8**, 505–514.
Weeks, C. M. & Miller, R. (1999). *Acta Cryst.* D**55**, 492–500.
Weiss, M. S., Sicker, T. & Hilgenfeld, R. (2001). *Structure*, **9**, 771–777.
Wilson, K. S. (1978). *Acta Cryst.* B**34**, 1599–1608.
Xu, H., Hauptman, H. A., Weeks, C. M. & Miller, R. (2000). *Acta Cryst.* D**56**, 238–240.
Yao, J.-X. (1981). *Acta Cryst.* A**37**, 642–644.