Acta Cryst. (1995). B**51**, 423–431

Structure Solution by Iterative Peaklist Optimization and Tangent Expansion in Space Group *P*1

By George M. Sheldrick and Robert O. Gould*

Institut für Anorganische Chemie der Universität Göttingen, Tammannstrasse 4, *D*-37077 *Göttingen, Germany*

(*Received* 20 *December* 1994; *accepted* 13 *March* 1995)

Abstract

An extension to the peaklist optimization procedure is proposed, in which one overall phase refinement cycle consists of tangent expansion, E-map, peaksearch and elimination of peaks to achieve a maximum correlation coefficient between E_o and E_c. This procedure appears to be able to solve large structures from random phases given data to atomic resolution. The power of the method can be substantially increased by starting with slightly better than random phases, obtained for example from threefold Patterson vector superposition minimum functions or rotation searches using a fragment of known geometry. These two sources of phase information require expansion of the data to the space group $P1$, which also appears to be a useful strategy when starting from random phases. This real/reciprocal space recycling procedure was successful in solving two small known proteins and three unknown 200+-atom small-molecule structures. An investigation of the influence of the resolution on the peaklist optimization algorithm shows that there is a marked deterioration in the effectiveness as the resolution of the data is truncated, the deterioration being particularly marked between 1.2 and 1.3 Å

Introduction

It has become common practice to improve and extend the trial structures obtained by direct methods by some sort of automatic Fourier recycling before attempting a chemical interpretation. A scheme described by Sheldrick (1982) and subsequently incorporated in the program *SHELXS*86 (Sheldrick, 1985) is shown in Fig. 1. The starting phases were taken from the direct methods solution with the best figures of merit, or from tangent expansion (Karle, 1968) of a partial structure. These phases were used to calculate an E-map (Fourier synthesis using observed E-magnitudes and calculated phases), which was then searched for the highest M independent peaks, where M was usually *ca* 30% greater than the expected number of unique atoms. Starting with the lowest peak, peaks were eliminated if so doing caused the index $R_E = \Sigma(E_o - kE_o)^2/\Sigma E_o^2$ (where k was a scale factor chosen to minimize R_E) to fall, and otherwise retained. Three scans of the peaklist were performed, after which this stage, which will be referred to henceforth as 'peaklist optimization', had usually converged. The resulting peaks were then used to calculate new phases for the strongest E-values (say $E_o > 1.2$), assuming that all atoms were point atoms with the same atomic number, and the cycle of peaklist optimization, E-map and peaksearch repeated several times. The final peaklist was used to construct a picture of the structure, which often was rather complete. It was probably the completeness of this structure solution which, more than any other factor, led to the wide acceptance of the program. Indeed, in some cases there were reasons to suspect that the peaklist optimization procedure had somehow succeeded in extracting the correct structure from a very dubious or even totally incorrect direct-methods solution. Lamzin & Wilson (1993) have proposed a scheme for automated protein refinement (ARP), which also involves the rejection and addition of atoms, but the selection of these atoms is based on analysis of the $3F_o - 2F_c$ and $F_o - F_c$ maps, respectively.

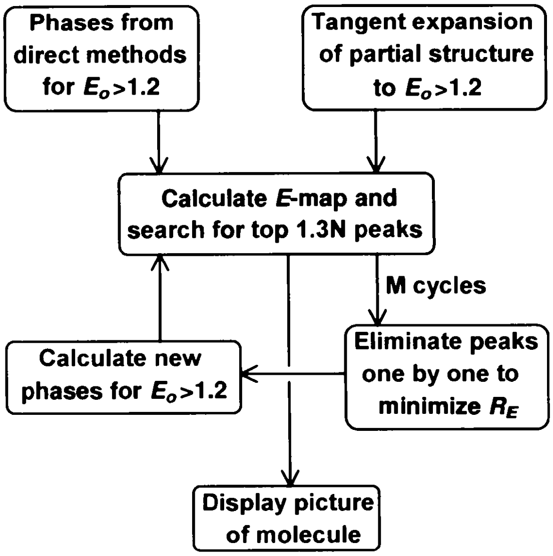

Phases from direct methods for E_o>1.2
Tangent expansion of partial structure to E_o>1.2
Calculate E-map and search for top 1.3N peaks
M cycles
Eliminate peaks one by one to minimize R_E
Calculate new phases for E_o>1.2
Display picture of molecule

Fig. 1. The peaklist optimization procedure as implemented in *SHELXS*86 to improve the quality of the E-map obtained from direct methods or from tangent expansion of a partial structure. Typically 1–5 iterations were performed using only E-values greater than 1.2.

* Permanent address: Department of Chemistry, University of Edinburgh, West Mains Road, Edinburgh EH9 3JJ, Scotland.

Two recent developments indicated that it might be practical to extend this real/reciprocal space recycling so that the first stage – the original 'direct methods' – could be dispensed with entirely. The first of these is the series of spectacular successes that Weeks, DeTitta, Miller & Hauptman (1993), DeTitta, Weeks, Thuman, Miller & Hauptman (1994) and Weeks, DeTitta, Hauptman, Thuman & Miller (1994) have achieved with a related real/reciprocal space recycling scheme, in which phases are refined to minimize a function of the estimated cosines of structure invariants, followed by E-map calculation, peaksearch and calculating new phases based on the top N peaks (where N is the number of unique atoms expected). Alternation between real and reciprocal space was also always a fundamental feature of the *DIRDIF* system; see, for example, Beurskens, Gould, Bruins Slot & Bosman (1987). The second important development is the enormous increase in computer number-crunching power over the last few years, which makes it possible to employ 'brute force' algorithms which would have been out of the question a few years ago. To put this in a crystallographic context, we shall define the unit of computer power as the 'VAX', since the Digital Equipment VAX/780 and the approximately equally powerful MicroVAXII were widely used in crystallographic laboratories in the second half of the 1980's. Current inexpensive RISC workstations benchmark in the range 30–200 VAX, and a 90 MHz Pentium PC achieves *ca* 60 VAX (benchmarks based on crystallographic least-squares refinement). One 'VAX-year' is then the amount of number crunching that a VAX could (theoretically) have achieved in 1 year of continuous operation. The VAX-year proves to be a convenient unit to measure the computer resources required by the methods presented in this paper.

Methods

In this paper we restrict ourselves to the space group $P1$; except where explicitly stated to the contrary, data for structures in other space groups were expanded to triclinic first (all the test structures are noncentrosymmetric). At first sight this procedure has the serious disadvantage that the time taken is substantially increased (by a factor corresponding approximately to the number of symmetry operations). On the other hand, it has been observed frequently that the success rate of conventional direct methods is much higher (per starting random phase set) in lower-symmetry space groups; our tests have shown that the success rate is often increased by an order of magnitude in going from $P\bar{1}$ to $P1$. For many years it has been standard practice in Göttingen to solve $P\bar{1}$ structures in $P1$ and then to find by inspection the translation necessary to place the ($P\bar{1}$) inversion center at the origin. Possibly the presence of spatially fixed symmetry elements enables a phase set to become trapped in a false minimum, whereas in the absence of symmetry a continuous phase-improvement path is available which leads to the correct solution. Although the real/reciprocal space recycling scheme proposed here is often successful starting from random phases, we were primarily interested in using it to refine slightly better than random starting phase sets obtained by either Patterson vector superposition or rotation search using a fragment of known geometry. Both these approaches are conventionally followed by a translation search to locate the position of the origin of the true space group, but they can also be regarded as providing approximate phases directly for data expanded to the space group $P1$.

The method proposed and tested in this work is summarized in Fig. 2. If 'heavier' atoms such as phosphorus, sulfur or chlorine are present (in addition to carbon *etc.*), starting maps can be obtained in the form of threefold Patterson vector superposition minimum functions. If the structure contains a relatively rigid fragment, for which the geometry can be predicted by semiempirical methods or taken from a related crystal structure, the best rotation function solutions can provide starting atoms lists. If neither of these approaches is practical, we enter the procedure with random phases; it would have been just as convenient to start from random atoms, and Weeks, Hauptman, Chang & Miller (1994) have shown that

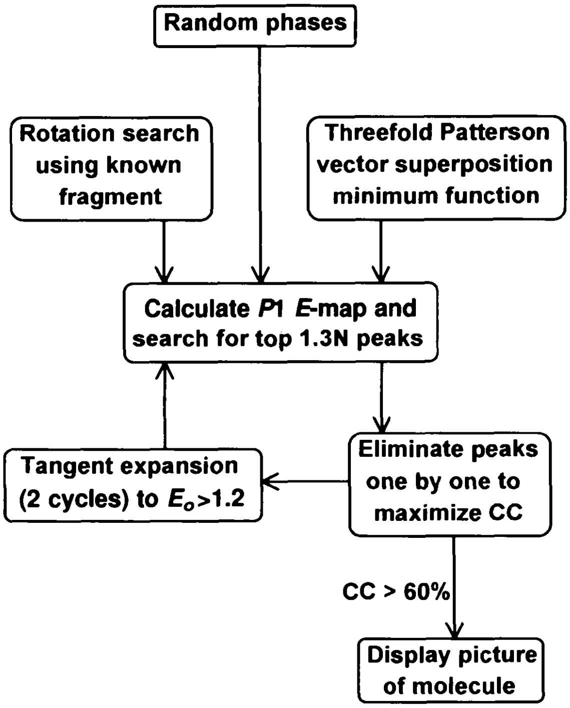

Random phases
Rotation search using known fragment
Threefold Patterson vector superposition minimum function
Calculate $P1$ E-map and search for top 1.3N peaks
Tangent expansion (2 cycles) to E_o>1.2
Eliminate peaks one by one to maximize CC
CC > 60%
Display picture of molecule

Fig. 2. The peaklist optimization procedure in space group $P1$ starting from (almost) random phases. All E-values are employed in calculating the correlation coefficient CC; 50% of the reflections with E_o > 1.4 which have the highest E_c values are typically used as input to the tangent formula. The recycling is usually repeated until at least one of the parallel trials has a correlation coefficient that clearly identifies it as being a correct solution.

random atoms give a slightly higher success rate than random phases in their procedure.

One full cycle of the procedure consists of tangent expansion, E-map, peaksearch and peaklist optimization. The current atom list is used to calculate E-values for all E_o greater than (say) 1.4. For example, 50% of these reflections with the highest calculated E-magnitudes are then used on the right-hand side of a tangent formula summation. In the tests described here we performed two tangent iterations per overall cycle, developing phases for all E_o values greater than (say) 1.4 and 1.2, respectively. The phases were applied to the observed E-values at the end of each iteration. In the first overall cycle starting from the superposition minimum function or rotation search, four tangent iterations were performed rather than two. We have not yet made systematic attempts to improve the algorithm for tangent expansion, which is based closely on that proposed by Karle (1968), but a number of tests showed that it is relatively optimal and not very sensitive to small variations in the parameters. The phases from the tangent expansion and the observed E-magnitudes are used to calculate an E-map, which is then searched for M peaks (where M is *ca* 30% greater than the number of atoms in the cell). Peaklist optimization (two scans from the lowest to the highest peak were found to be the most cost effective) is then followed by applying the calculated phases to the observed E-values and the cycle is repeated as often as required. Usually *ca* ten phase sets are processed in parallel; this has the advantage in the tangent expansion stage that it is not necessary to store the (possibly a very large number of) triple phase invariants; instead they are found on the fly and applied to all phase sets in parallel.

Vector superposition

The Patterson vector superposition minimum function is discussed in detail in Buerger's (1959) book, where it is referred to as the 'vector shift' method, but has been almost forgotten as a method of solving structures because of the almost omnipotent direct methods. Sheldrick (1992) and Sheldrick, Dauter, Wilson, Hope & Sieker (1993) showed that it can be a very effective way of locating heavier atoms such as sulfur even for small proteins, provided that data have been measured to very high resolution. Since we are interested here in obtaining initial phases rather than a few heavier atoms, we have chosen to make a threefold vector superposition rather than the twofold superposition used in our previous work. The threefold superposition is obtained by overlaying three copies of the three-dimensional sharpened Patterson, shifted from one another by the three sides of a vector triangle, all three vectors corresponding to strong Patterson peaks. The minimum function of the three Pattersons is calculated numerically, and is then searched for the highest peaks. The threefold superposition map is noisier than a twofold map, but for a general triangle consisting of single-weight Patterson vectors it should theoretically correspond (in the absence of Patterson overlap) to a single image of the structure; a twofold superposition would consist of two images related by a center of symmetry. It will be seen later that it can be difficult to extract the correct structure from a centrosymmetric double image. In our tests we calculated 'super-sharp' Pattersons with coefficients $(E^3F)^{1/2}$. The Patterson peak list is searched for unique vector triangles, taking the symmetry of Patterson space into account: the triangles with the highest values of

$$PT = P_1P_2P_3/[d^2 + (gr)^2],$$

where P_1, P_2 and P_3 are the values of the Patterson function for each of the three vectors, d is the lack of closure of the vector triangle (Å), r is the maximum resolution of the reflection data (Å), and g is an empirical constant which was set to 0.6 in the tests reported here. Only Patterson peaks further than a specified minimum distance (typically 6 Å) from the nearest lattice point were employed; Patterson functions of larger molecules typically contain high density at smaller distances from the origin arising from repeated secondary structure *etc.*

Rotation search

For testing purposes we chose simply to formulate the rotation search as a search for maxima of the function $\Sigma[(E_o{}^2 - 1)E_c{}^2]$, calculated for the highest observed E-values (typically $E_o > 1.8$). Powell's (1965) method was used to find the maximum starting from random orientations since it does not require analytical derivatives. The E_c values were calculated by a conventional structure-factor summation assuming point atoms. Comparisons showed that the results were at least as good as those obtained with the more sophisticated Patterson space rotation search employed in the program *PATSEE* (Egert & Sheldrick, 1985). The unique orientations (taking the rotation and reflection but not translation components of the symmetry operators into account to eliminate equivalents) with the highest values of this function were used to generate several (typically ten) sets of initial atoms for the recycling procedure.

Peaklist optimization

Three small changes were made to the original peaklist optimization scheme. Firstly, instead of using a reduction in R_E as a criterion for deleting a peak, an increase in the correlation coefficient proposed by Fujinaga & Read (1987) was used instead; this appears to be more sensitive in the critical early stages. This correlation coefficient is defined as follows

$$CC = [\Sigma wE_o^2E_c^2\Sigma w - \Sigma wE_o^2\Sigma wE_c^2]/ \{[\Sigma wE_o^4\Sigma w - (\Sigma wE_o^2)^2][\Sigma wE_c^4\Sigma w - (\Sigma wE_c^2)^2]\}^{1/2}.$$

After some experimentation we used weights $w = 1/[0.04 + \sigma^2(E_o)]$ for all the tests reported here. Secondly, instead of assuming that all atoms have the same point-atom scattering factors, the peaklist is compared with the anticipated chemical contents of the unit cell, assigning the N_1 highest peaks to the N_1 atoms expected for the element with the highest atomic number *etc*. We also tried allowing the peaklist optimization algorithm to vary the scattering factor types so as to maximize the correlation coefficient, but this proved unreliable and was not used in the work reported here. The point-atom scattering factors were normalized as a function of diffraction angle by dividing by the square root of the total scattering power of the estimated unit-cell contents, so that, for example, the values for heavier elements tended to rise with increasing 2θ. The third change is the use of all data in the peaklist optimization rather than (typically) $E_o > 1.2$, as in *SHELXS*86 (Sheldrick, 1985); a decade ago it was felt necessary to restrict the number of reflections in order to minimize the computer time required.

An additional option is the elimination of fragments consisting of a very small number of atoms; this is a simple way of incorporating the idea of connectivity. In the peaklist optimization tests reported here, fragments containing less than four atoms were eliminated (except where stated to the contrary). This proved to be slightly beneficial for the JFA and balhimycin tests, but slightly disadvantageous for rubredoxin.

Mean phase errors (MPE)

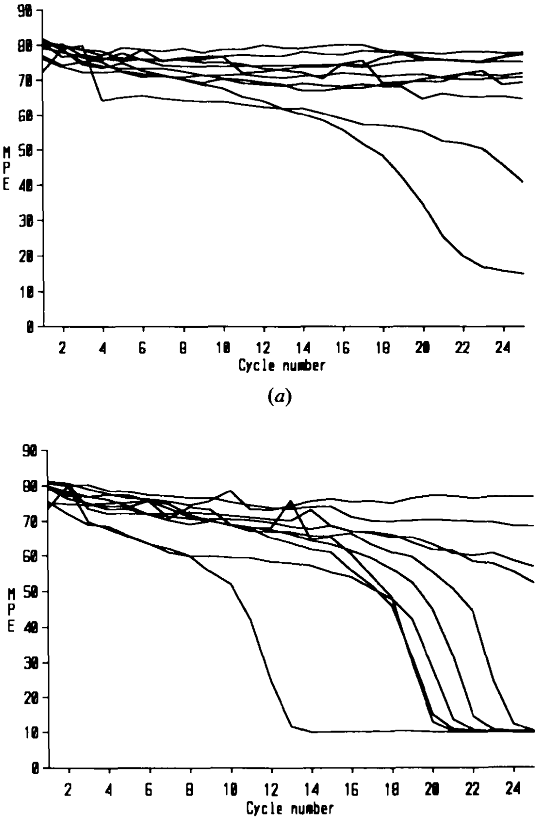
To monitor the progress for known test structures, an E-weighted mean phase error (MPE) was calculated (for E_o greater than 1.4) by finding the translation necessary to give the best agreement between the calculated phases and either the 'true' phases or the 'true' phases subtracted from 360° (for the enantiomorph). The 'true' phases were based on a full anisotropic least-squares refinement, including H atoms. Since an exhaustive search was made for the translation with the lowest weighted mean phase error, random phases must give values less than 90° and typically resulted in mean phase errors of around 80°. Progress in solving unknown structures is monitored by writing the atom list to file for the solution with the best correlation coefficient thus far; this can be inspected by interactive computer graphics whilst the structure solution job is still in progress. Weeks, Hauptman, Chang & Miller (1994) described a similar interactive monitoring scheme for their real/reciprocal space recycling.

Tests on known structures

To test the method we chose two genuine $P1$ 'small-molecule' structures and two small proteins. In all cases excellent experimental data had been measured to 1 Å resolution or better. The 148-atom $P1$ structure (JFA) was determined by Karle, Flippen-Anderson, Uma, Balaram & Balaram (1989) by direct methods (Sheldrick, 1990). The 263-atom $P1$ structure (balhimycin) was solved (Sheldrick, Paulus, Vertesey & Hahn, 1995) *via* location of the four Cl atoms by Patterson vector superposition methods, followed by partial structure expansion (as summarized in Fig. 1); we have not succeeded in solving the structure by conventional direct methods. For the two small proteins crambin and rubredoxin, the same data sets were used as for the Patterson and direct-methods tests reported by Sheldrick *et al.* (1993). Relevant crystallographic details of these four test structures are summarized in Table 1. These structures, even without expanding the two proteins to $P1$, are large enough to provide a realistic test of any *ab initio* structure solution method.

Table 1. *Known test structures*

Codename	Space group	N (unique atoms)	N (atoms in $P1$ cell)
JFA	$P1$	148	148
Balhimycin	$P1$	263	263
Crambin	$P2_1$	*ca* 420	*ca* 840
Rubredoxin	$P2_1$	*ca* 500	*ca* 1000



(*a*)

(*b*)

Fig. 3. The E-weighted MPE as a function of cycle number for the 148-atom $P1$ structure JFA starting from random phases. In (*a*) no peaklist optimization was employed; the top 148 unique peaks were reinput as atoms. In (*b*) peaklist optimization was employed as summarized in Fig. 2.

Random starting phases

Fig. 3 shows the E-weighted MPE (for $E_o > 1.4$) as a function of the overall cycle number for the structure JFA and the procedure defined in Fig. 2 starting from random phases. In (*a*) no peaklist optimization was applied; the top $N = 148$ unique peaks were reinput as atoms. In (*b*) peaklist optimization was employed as described above. It will be seen that after 24 cycles, (*a*) has yielded one correct solution out of ten trials, whereas (*b*) has produced seven correct solutions. In fact all the phase sets converged to correct solutions after ~ 70 cycles in (*b*). The peaklist optimization runs converged to significantly lower mean phase errors than the runs in which simply the top N peaks were recycled.

Starting phases from rotation search

The unit cell of JFA contains two linear peptide molecules with slightly different but predominantly α-helical structures. Fig. 4 shows the dramatic effect of starting with phases from a rotation search using a standard 12-atom α-helical triglycine fragment. Five of the top ten unique rotation function maxima have converged to correct solutions within three cycles, and after 16 cycles all the phase sets have led to solutions. There are many good ways of fitting this small fragment to the structure, but it should be noted that one of the solutions (indicated with an asterisk in Fig. 4) corresponds to the opposite enantiomorph, *i.e.* must be regarded as a fortuitous accident (it also has the largest initial MPE).

Starting phases from Patterson superposition

The balhimycin structure can be solved by the procedure described in Fig. 2 starting from random phases, but the first correct solution (out of ten parallel trials) does not emerge until about cycle 70. It is much more effective to take advantage of the presence of four Cl atoms in the unit cell, which should generate four vector triangles. The top ten vector triangles generated as described above do indeed include the four correct triangles (ranked 2, 4, 8 and 9, according to the PT figure of merit, and indicated by asterisks in Fig. 5*a*), three of which leading to a solution within five cycles, and two triangles consisting of two Cl atoms and one light atom converge by cycles 8 and 14, respectively. One correct vector triangle has a lower initial MPE than the other solutions, but does not refine; it turns out that it is locked into a centrosymmetric false minimum.

A similar pattern is observed for crambin (Fig. 5*b*, expanded to $P1$), which has 12 S atoms in the unit cell which should generate 220 possible vector triangles, some of which are, however, related by symmetry. The ten triangles with the best figures of merit PT include four correct S3 triangles (indicated by asterisks), which all converge within five cycles. Three other triangles which each involve one correct S—S vector also converge by cycles 10, 13 and 17. To put this

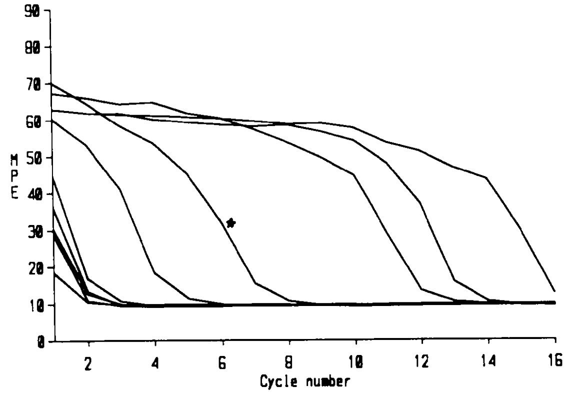
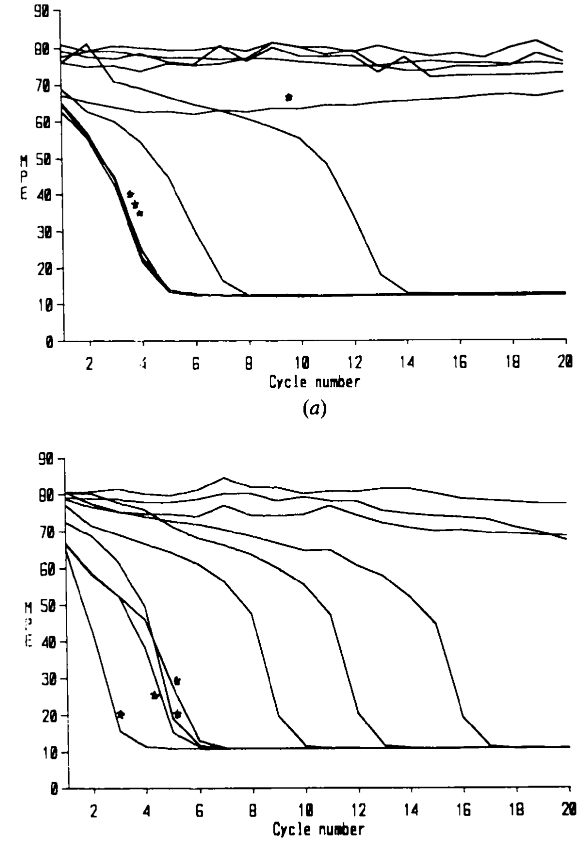


Fig. 4. The same structure and procedure as in Fig. 3(*b*), but starting from 'almost random' phases obtained from the best independent rotation search solutions for a 12-atom α-helical triglycine fragment. The improvement is dramatic, and all trials have converged to the correct solution after 16 cycles. However, one solution (marked with an asterisk) corresponds to the enantiomorph of the true structure (and α-helix fragment).



(*a*)

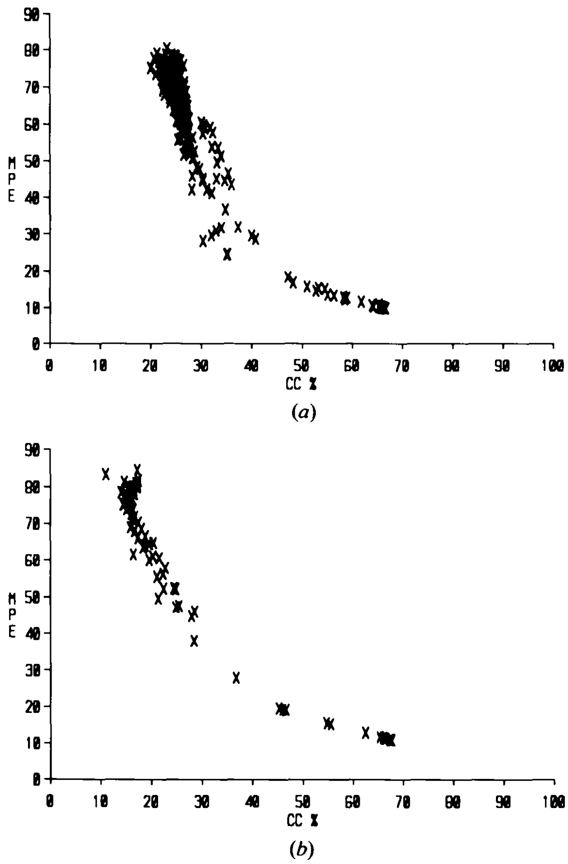
(*b*)

Fig. 5. Peaklist optimization in $P1$ starting from 'almost random' phases obtained from threefold vector superposition minimum functions for (*a*) balhimycin and (*b*) crambin. The vector triangles were derived automatically by an analysis of the supersharp Patterson; those which correspond to three 'heavy' atoms (Cl or S, respectively) are marked with asterisks.

70% success rate in the *ab initio* solution of a small protein into perspective, it must be pointed out that it is only possible because high-quality experimental data are available to extremely high resolution, and the test reported here required 8.4 VAX-years of computer time! In this example the selection of vector triangles could have been improved by downweighting triangles involving vectors with $y = 0$, since Patterson density will tend to accumulate in this plane in space group $P2_1$. On the other hand, it is desirable to retain the Harker vectors with $y = \frac{1}{2}$.

Correlation coefficient (CC) and MPE

In all the tests reported here, the correlation coefficient was strongly correlated with the MPE and provided a very reliable indication as to whether a correct solution was emerging. In general, a CC of greater than 50% corresponds to a correct solution. Fig. 6 shows scatterplots of MPE against CC for JFA and crambin; in both cases the points lie close to a smooth curve, especially as CC increases. When, rather than using all data, only E_o values greater than 1.4 were used to calculate the correlation coefficient, the discrimination was much poorer; incipient correct solutions only stood out from the rest when the MPE had fallen below *ca* 30°. Beurskens *et al.* (1987) have also observed that the correlation coefficient is more effective when all data are used.



(*a*)

(*b*)

Fig. 6. Scatterplots of MPE against correlation coefficient for (*a*) JFA and (*b*) crambin. It will be seen that the correlation coefficient provides an excellent indication of the quality of the phase set and it is clearly appropriate for the identification of correct solutions of unknown structures.

Enantiomorph resolution

In the tests reported so far, when the MPE dropped to below *ca* 60°, the recycling procedure was able to determine the full structure within a few additional cycles in which the MPE dropped monotonically. In the case of rubredoxin (expanded to $P1$), there are two Fe atoms in the cell, so it is necessary to start from an Fe_2S vector triangle; in fact, the triangle with the best figure of merit PT and several other high-ranking triangles were of this type. When the full data and peaklist optimization were used with a superposition minimum function generated from the 'best' vector triangle, the recycling procedure reduced the MPE to 45.1° after two cycles, but the MPE of the enantiomorph was also reduced to 51.0°. In the third cycle, the MPE was reduced further to 27.0°, but the MPE from the inverted structure rose to 68.0°, and after five cycles the values had essentially converged to 13.7 and 78.3°, respectively. Clearly the presence of only two heavy atoms leads to a pseudocentrosymmetric structure (with an inversion center midway between the two Fe atoms), but the recycling procedure is able to break this pseudosymmetry rather effectively. Because peaks are eliminated one-by-one, the peaklist optimization would be expected to lead to one enantiomorph or the other; each peak removed would be more likely to tip the scales further in the same direction. When, instead of the peaklist optimization, the top N peaks were accepted in each cycle, the MPE values were still 47.6 and 44.6° after 20 cycles, and there was little sign of progress. There is no reason why simply recycling the top N peaks should break the pseudosymmetry, especially since the tangent formula tends to drift towards a centrosymmetric (pseudo)solution anyway.

The effect of resolution on peaklist optimization

The peaklist optimization/tangent formula recycling from the 'best' rubredoxin vector triangle was also tested at different resolutions by truncating the data; experience with the direct methods solution of this structure suggested that the enantiomorph resolution would depend very sensitively on the resolution (Sheldrick *et al.*, 1993). Truncated data are of course more favorable than data collected out to a diffraction limit at the same resolution, since the signal-to-noise ratio of the latter will be worse. With data truncated to 1.1 Å, the enantiomorph resolution proceeded more slowly, but was essentially complete after ten cycles, with MPE's of 16.2 and 76.0°. Truncating the data to 1.2 Å led to no convincing resolution of the pseudosymmetry; after 20 cycles the MPE values were 52.0 and 61.8°.

We decided to investigate the effect of resolution on peaklist optimization alone (without the intervening tangent iterations) by expanding rubredoxin from the positions of the Fe and four S atoms (which can be obtained by automated Patterson interpretation, even with data truncated to 1.5 Å; Sheldrick *et al.*, 1993), but in the true space group $P2_1$ rather than with data expanded to $P1$. This is a more favorable situation because the extra atoms help to break the pseudosymmetry. We varied the number of peaks M input to the peaklist optimization, but otherwise the procedure was as described above. It appears that it is of advantage to reduce M a little as the resolution becomes worse, otherwise the procedure can become 'bogged down' in a false solution with a large number of spurious atoms. Table 2 shows the percentage of correct peaks (within 0.3 Å of correct atomic positions) as the peaklist is descended in blocks of 100 peaks. It will be seen that peaklist optimization has produced a substantial improvement over the initial peaklist at 1.2 Å, but not at 1.3 Å. In this example, elimination of fragments containing less than four atoms was not particularly effective. The results are presented in Fig. 7 in the form of contoured electron-density maps for the same thin slab of space around the planar residue Trp37, calculated as Sim-weighted $(2wE_o - E_c)$ maps (Sim, 1959) before and after peaklist optimization. The program '*O*' (Jones, Zou, Cowan & Kjeldgaard, 1991) was used to prepare these pictures. These maps resemble normal small-molecule maps with isolated correct atoms at high resolution, but as the resolution deteriorates the peaks remain fairly sharp but the percentage of correct sites falls. Connectivity, the most important feature in the interpretation of low-resolution protein F-maps obtained by, for example, MIR methods, is conspicuous by its absence. The peaklist optimization is particularly good at removing the many spurious peaks (a feature of E-maps) at high resolution, but becomes less effective as the data are truncated further. There is a marked deterioration in the interpretability of the maps on going from 1.2 to 1.3 Å, a point which should be borne in mind when collecting high-resolution data of unknown metalloproteins with a view to *ab initio* structure solution from the native data alone!

Table 2. *Numbers of correct peaks (within* 0.3 Å *of the true positions) before and after peaklist optimization starting from* 1 Fe *and* 4 S *for rubredoxin in* $P2_1$

The peaklist was sorted in order of descending peak height.

	Truncated to	Peaks			
	(Å)	1–100	101–200	201–300	301–400
$2wE_o - E_c$ Sim Fourier	1.2	68	41	28	13
Peaklist optimization	1.2	97	94	89	38
PO – small fragment elimination	1.2	99	94	87	29
$2wE_o - E_c$ Sim Fourier	1.3	60	21	18	12
Peaklist optimization	1.3	40	31	18	6
PO – small fragment elimination	1.3	52	19	15	8

Table 3. *Previously unsolved structures*

True space group	N (atoms in $P1$)	Initial phases	N (cycles)	VAX-years
$P\bar{1}$	2 × 220 = 440	Random	4 × 12	0.3
$I4_1$	8 × 52 = 416	Random	5 × 335	1.3
$P3_221$	6 × 200 = 1200	Rotational search	4 × 45	1.3
$P2_12_12_1$	4 × 270 = 1080	Vector superposition	15 × 80	2.9

The first number under N(cycles) is the number of parallel permutations, the second is the number of cycles after which the first correct solution appeared.

Tests on unknown structures

Table 3 illustrates some of the unknown structures solved by this procedure; all had defeated exhaustive attempts using conventional direct methods. Except for two unique Cl atoms in the fourth structure, the heaviest atoms were O. All were expanded to triclinic and the translation necessary to place the origin correctly relative to the symmetry elements of the correct space group was determined by inspection. In principle it would be possible to automate this step; the program *MISSYM* (LePage, 1987) could also be used. In the case of the $I4_1$ structure, this inspection also led to the deduction of the correct space group; the direct methods attempts had all been performed in the space group $I4$, because one of the five reflections with $h = k = 0$ and $l = 4n + 2$ had significant intensity [*ca* $8\sigma(I)$]. Probably this structure would have been solved eventually by conventional direct methods if the correct space group had been used, although the resolution and quality of the data are scarcely adequate. The other three structures have 200 or more unique atoms, and the success rate of conventional direct methods is still rather low for structures of this size. In addition, the $P3_221$ structure showed pronounced pseudosymmetry (the odd l reflections were systematically weak), and the other two data sets were of mediocre quality. The method presented in this paper has also already proved useful in helping us to solve a variety of more trivial structures, especially in cases where the space group assignment was uncertain. An interesting possibility is that the peaklist optimization should, at least in theory, be capable of solving twinned structures, given the twin law but not the space group; it is often easier to guess the former than the latter.

Concluding remarks

The results reported here make it clear that alternation between real and reciprocal space is capable of solving very large structures, given data to atomic resolution (in practice *ca* 1.2 Å or better). The method is particularly effective when slightly better than random

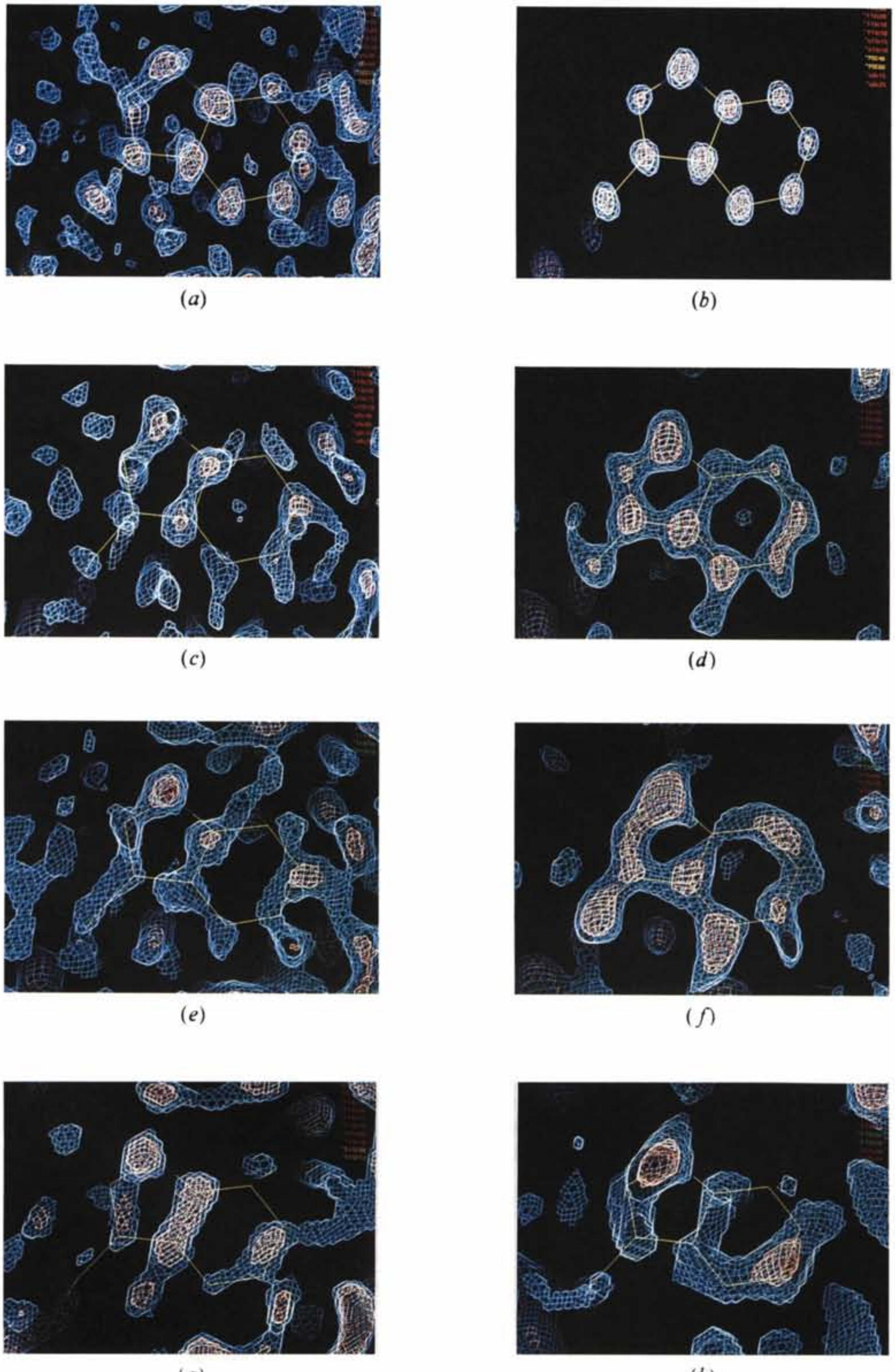


Fig. 7. Sim-weighted E-maps before (a, c, e and g) and after (b, d, f and h) peaklist optimization starting from 1 Fe + 4 S for rubredoxin in space group $P2_1$ as a function of the resolution. The same section through the residue Trp37 was used for all maps. The full data (0.92 Å) were used for (a) and (b), and the data were truncated to 1.2 Å for (c) and (d), 1.3 Å for (e) and (f), and 1.5 Å in (g) and (h). With high-resolution data the procedure is very efficient at removing the noise peaks, but the performance deteriorates markedly between 1.2 and 1.3 Å. At 1.5 Å there is no discernible improvement in the map.

starting phases are available, for example, from threefold Patterson vector superposition or a rotation search to find the orientation of a known fragment. These two sources of phase information required us to apply the recycling procedure with the data expanded to the space group $P1$; it is not clear whether this is also most effective when starting from random phases (or random atoms), although it appears that the higher success rate per starting phase set may compensate for the extra computational time required. However, in higher symmetry space groups, the presence of centrosymmetric projection reflections would be an asset in reciprocal space, and – except where atoms are expected to occupy special positions – one could also eliminate peaks close to special positions before entering the peaklist optimization. It would require tests on many structures and space groups to establish whether it is cost-effective to expand to $P1$; at least if the space group is uncertain, it will also be established by the structure solution in $P1$.

In this work we have concentrated on the use of peaklist optimization in the real space part of the procedure, whilst employing the well established tangent formula in reciprocal space. Various modified tangent formulae have been proposed which would undoubtedly improve the performance of the reciprocal space counterpart; see, for example, the Sayre tangent formula (Debaerdemaeker, Tate & Woolfson, 1985) or a very similar formula proposed later by Giacovazzo (1993), a tangent formula derived from an $E^2 - 1$ Patterson function by Rius (1993), or a tangent formula incorporating negative quartets (Sheldrick, 1990). All these modified tangent formulae utilize the weakest as well as the strongest E-values; however, it is conceivable that such reflections are exploited more effectively by the peaklist optimization stage (but not by recycling the top N peaks, because the weak reflections do not contribute significantly to the E-map). At the meeting at which this work was first presented, Weeks, Hauptman, Chang & Miller (1994) reported that the tangent formula was highly competitive with the minimal principle in their own real/reciprocal space recycling procedure. They referred to their method as 'shake and bake', so the work presented in this paper could perhaps be described as 'half-baked'!

Although peaklist optimization is much more effective than reinserting the top N peaks in terms of success rate per cycle, much of this advantage is compensated by that fact that it requires appreciably more computation time, especially when all data are used rather than just the largest E-values. The correlation coefficient based on all the data is much more reliable as a figure of merit, and neither the tangent formula nor reinputting the strongest peaks are able to escape from a false solution with a centrosymmetric distribution of atoms. Perhaps it is essential to include either the peaklist optimization in real space or a method such as the minimal principle in reciprocal space in order to be able to resolve enantiomorph ambiguities. The most cost-effective approach may well be to use just the largest and smallest E-values in the initial cycles, and then the full data for the final cycles in order to obtain as complete a structure as possible.

We are grateful to the Deutsche Akademische Austauschdienst for study leave support (to ROG). We should like to thank Drs Judith Flippen-Anderson, Zbigniew Dauter, Håkon Hope, Erich Paulus and Keith Wilson for providing us with excellent test data, and Drs Karin Schmidt-Bäse, Robin English, Axel Göhrt and Ehmke Pohl for supplying the data for the unsolved structures (and for not solving them first!).

References

Beurger, M. J. (1959). *Vector Space*, Ch. 11. New York: Wiley.

Beurskens, P. T., Gould, R. O. G., Bruins Slot, H. J. J. & Bosman, W. P. (1987). *Z. Krist.* **179**, 127–159.

Debaerdemaeker, T., Tate, C. & Woolfson, M. M. (1985). *Acta Cryst.* A**41**, 286–290.

DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.

Egert, E. & Sheldrick, G. M. (1985). *Acta Cryst.* A**41**, 262–268.

Fujinaga, M. & Read, R. J. (1987). *J. Appl. Cryst.* **20**, 517–521.

Giacovazzo, C. (1993). *Z. Krist.* **206**, 161–171.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Karle, J. (1968). *Acta Cryst.* B**24**, 182–186.

Karle, I. L., Flippen-Anderson, J. L., Uma, K., Balaram, H. & Balaram, P. (1989). *Proc. Natl. Acad. Sci. USA*, **86**, 765–769.

Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.

LePage, Y. (1987). *J. Appl. Cryst.* **20**, 264–269.

Powell, M. J. D. (1965). *Comput. J.* **7**, 303–307.

Rius, J. (1993). *Acta Cryst.* A**49**, 406–409.

Sheldrick, G. M. (1982). In *Crystallographic Computing*, edited by D. Sayre, pp. 506–514. Oxford: Clarendon Press.

Sheldrick, G. M. (1985). *SHELXS86. Program for the Solution of Crystal Structures*. Univ. of Göttingen, Germany.

Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.

Sheldrick, G. M. (1992). In *Crystallographic Computing* 5, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 145–157. Oxford Univ. Press.

Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.

Sheldrick, G. M., Paulus, E., Vertesey, L. & Hahn, F. (1995). *Acta Cryst.* B**51**, 89–98.

Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.

Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.

Weeks, C. M., DeTitta, G. T., Miller, R. & Hauptman, H. A. (1993). *Acta Cryst.* D**49**, 179–181.

Weeks, C. M., Hauptman, H. A., Chang, C.-S. & Miller, R. (1994). Proc. of the Am. Crystallogr. Assoc. Meeting. Atlanta, GA, USA. Abstract TRN13.