abrupt increase in correct peaks occurs when Fourier refinement is started.

Since the correlation coefficient is a relatively absolute figure of merit (given atomic resolution, values greater than 65% almost invariably correspond to correct solutions), it is usually clear when *SHELXD* has solved a structure. The current version of *SHELXD* includes an option for calculating it using the full data every 10 or 20 internal loop cycles, and jumping to the external loop if the value is high enough. Recalculating it every cycle would be computationally less efficient overall.

16.1.8. Applying dual-space programs successfully

The solution of the (known) structure of triclinic lysozyme by *SHELXD* and shortly afterwards by *SnB* (Deacon *et al.*, 1998) finally broke the 1000-atom barrier for direct methods (there happen to be 1001 protein atoms in this structure!). Both programs have also solved a large number of previously unsolved structures that had defeated conventional direct methods; some examples are listed in Table 16.1.8.1. The overall quality of solutions is generally very good, especially if appropriate action is taken during the Fourier-

Table 16.1.8.1. *Some large structures solved by the Shake-and-Bake method*

Previously known test data sets are indicated by an asterisk (*). When two numbers are given in the resolution column, the second indicates the lowest resolution at which truncated data have yielded a solution. The program codes are *SnB* (S) and *SHELXD* (D).

(*a*) Full structures (>300 atoms).

Compound	Space group	N_u (molecule)	N_u + solvent	N_u (heavy)	Resolution (Å)	Program	Reference
Vancomycin	$P4_32_12$	202	258	8Cl	0.9–1.4	S	[1]
			312	6Cl	1.09	D	[2]
Actinomycin X2	$P1$	273	305	—	0.90	D	[3]
Actinomycin Z3	$P2_12_12_1$	186	307	2Cl	0.96	D	[4]
Actinomycin D	$P1$	270	314	—	0.94	D	[4]
Gramicidin A*	$P2_12_12_1$	272	317	—	0.86–1.1	S, D	[5]
DMSO d6 peptide	$P1$	320	326	—	1.20	S	[6]
Er-1 pheromone	$C2$	303	328	7S	1.00	S	[7]
Ristocetin A	$P2_1$	294	420	—	1.03	D	[8]
Crambin*	$P2_1$	327	423	6S	0.83–1.2	S, D	[9], [10]
Hirustasin	$P4_32_12$	402	467	10S	1.2–1.55	D	[11]
Cyclodextrin derivative	$P2_1$	448	467	—	0.88	D	[12]
Alpha-1 peptide	$P1$	408	471	Cl	0.92	S	[13]
Rubredoxin*	$P2_1$	395	497	Fe, 6S	1.0–1.1	S, D	[14]
Vancomycin	$P1$	404	547	12Cl	0.97	S	[15]
BPTI*	$P2_12_12_1$	453	561	7S	1.08	D	[16]
Cyclodextrin derivative	$P2_1$	504	562	28S	1.00	D	[17]
Balhimycin*	$P2_1$	408	598	8Cl	0.96	D	[18]
Mg-complex*	$P1$	576	608	8Mg	0.87	D	[19]
Scorpion toxin II*	$P2_12_12_1$	508	624	8S	0.96–1.2	S	[20]
Amylose-CA26	$P1$	624	771	—	1.10	D	[21]
Mersacidin	$P3_2$	750	826	24S	1.04	D	[22]
Cv HiPIP H42Q*	$P2_12_12_1$	631	837	4Fe	0.93	D	[23]
HEW lysozyme*	$P1$	1001	1295	10S	0.85	S, D	[24], [25]
rc-WT Cv HiPIP	$P2_12_12_1$	1264	1599	8Fe	1.20	D	[23]
Cytochrome c3	$P3_1$	2024	2208	8Fe	1.20	D	[26]

(*b*) Se substructures (> 25 Se) solved using peak-wavelength anomalous-difference data.

Protein	Space group	Molecular weight (kDa)	Se located	Se total	Resolution (Å)	Program	Reference
SAM decarboxylase	$P2_1$	77	20	26	2.25	S	[27]
AIR synthetase	$P2_12_12_1$	147	28	28	3.0	S	[28]
FTHFS	$R32$	200	28	28	2.5	D	[29]
AdoHcy hydrolase	$C222$	95	30	30	2.8–5.0	S	[30]
Epimerase	$P2_1$	370	64	70	3.0	S	[31]

References: [1] Loll *et al.* (1997); [2] Schäfer *et al.* (1996); [3] Schäfer (1998); [4] Schäfer, Sheldrick, Bahner & Lackner (1998); [5] Langs (1988); [6] Drouin (1998); [7] Anderson *et al.* (1996); [8] Schäfer & Prange (1998); [9] Stec *et al.* (1995); [10] Weeks *et al.* (1995); [11] Usón *et al.* (1999); [12] Aree *et al.* (1999); [13] Prive *et al.* (1999); [14] Dauter *et al.* (1992); [15] Loll *et al.* (1998); [16] Schneider (1998); [17] Reibenspiess (1998); [18] Schäfer, Sheldrick, Schneider & Vértesy (1998); [19] Teichert (1998); [20] Smith *et al.* (1997); [21] Gessler *et al.* (1999); [22] Schneider *et al.* (2000); [23] Parisini *et al.* (1999); [24] Deacon *et al.* (1998); [25] Walsh *et al.* (1998); [26] Frazão *et al.* (1999); [27] Ekstrom *et al.* (1999); [28] Li *et al.* (1999); [29] Radfar *et al.* (2000); [30] Turner *et al.* (1998); [31] Deacon & Ealick (1999).

Copyright © 2006 International Union of Crystallography

Table 16.1.8.2. *Overall success rates for full structure solution for hirustasin using different two-atom search vectors chosen from the Patterson peak list*

Resolution (Å)	Two-atom search fragments	Solutions per 1000 attempts
1.2	Top 100 general Patterson peaks	86
1.2	Top 300 general Patterson peaks	38
1.2	One vector, error = 0.08 Å	14
1.2	One vector, error = 0.38 Å	41
1.2	One vector, error = 0.40 Å	219
1.2	One vector, error = 1.69 Å	51
1.4	Top 100 general Patterson peaks	10
1.5	Top 100 general Patterson peaks	4
1.5	One vector, error = 0.29 Å	61

refinement stage. Most of the time, the *Shake-and-Bake* method works remarkably well, even for rather large structures. However, in problematic situations, the user needs to be aware of options that can increase the chance of success.

16.1.8.1. *Utilizing Pattersons for better starts*

When slightly heavier atoms such as sulfur are present, it is possible to start the *Shake-and-Bake* recycling procedure from a set of atomic positions that are consistent with the Patterson function. For large structures, the vectors between such atoms will correspond to Patterson densities around or even below the noise level, so classical methods of locating the positions of these atoms unambiguously from the Patterson are unlikely to succeed. Nevertheless, the Patterson function can still be used to filter sets of starting atoms. This filter is currently implemented as follows in *SHELXD*. First, a sharpened Patterson function (Sheldrick *et al.*, 1993) is calculated, and the top 200 (for example) non-Harker peaks further than a given minimum distance from the origin are selected, in turn, as two-atom translation-search fragments, one such fragment being employed per solution attempt. For each of a large number of random translations, all unique Patterson vectors involving the two atoms and their symmetry equivalents are found and sorted in order of increasing Patterson density. The sum of the smallest third of these values is used as a figure of merit (PMF). Tests showed that although the globally highest PMF for a given two-atom search fragment may not correspond to correct atomic positions, nevertheless, by limiting the number of trials, some correct solutions may still be found. After all the vectors have been used as search fragments (*e.g.* after 200 attempts), the procedure is repeated starting again with the first vector. The two atoms may be used to generate further atoms using a full Patterson superposition minimum function or a weighted difference synthesis (in the current version of *SHELXD*, a combination of the two is used).

In the case of the small protein BPTI (Schneider, 1998), 15 300 attempts based on 100 different search vectors led to four final solutions with mean phase error less than 18°, although none of the globally highest PMF values for any of the search vectors corresponded to correct solutions. Table 16.1.8.2 shows the effect of using different two-atom search fragments for hirustasin, a previously unsolved 55-amino-acid protein containing five disulfide bridges first solved using *SHELXD* (Usón *et al.*, 1999). It is not clear why some search fragments perform so much better than others; surprisingly, one of the more effective search vectors deviates considerably (1.69 Å) from the nearest true S–S vector.

16.1.8.2. *Avoiding false minima*

The frequent imposition of real-space constraints appears to keep dual-space methods from producing most of the false minima that plague practitioners of conventional direct methods. Translated molecules have not been observed (so far), and traditionally problematic structures with polycyclic ring systems and long aliphatic chains are readily solved (McCourt *et al.*, 1996, 1997). False minima of the type that occur primarily in space groups lacking translational symmetry and are characterized by a single large 'uranium' peak do occur frequently in *P*1 and occasionally in other space groups. Triclinic hen egg-white lysozyme exhibits this phenomenon regardless of whether parameter-shift or tangent-formula phase refinement is employed. An example from another space group (*C*222) is provided by the Se substructure data for AdoHcy hydrolase. In this case, many trials converge to false minima if the feature in the *SnB* program that eliminates peaks at special positions is not utilized.

The problem with false minima is most serious if they have a 'better' value of the figure of merit being used for diagnostic purposes than do the true solutions. Fortunately, this is not the case with the uranium 'solutions', which can be distinguished on the basis of the minimal function [equation (16.1.4.2)] or the correlation coefficient [equation (16.1.6.1)]. However, it would be inefficient to compute the latter in each dual-space cycle since it requires that essentially all reflections be used. To be an effective discriminator, the figure of merit must be computed using the phases calculated from the point-atom model, not from the phases directly after refinement. Phase refinement can and does produce sets of phases, such as the uranium phases, which do not correspond to physical reality. Hence, it should not be surprising that such phase sets might appear 'better' than the true phases and could lead to an erroneous choice for the best trial. Peak picking, followed by a structure-factor calculation in which the peaks are sensibly weighted, converts the phase set back to physically allowed values. If the value of the minimal function computed from the refined or *unconstrained* phases is denoted by $R_{\rm unc}$ and the value of the minimal function computed using the *constrained* phases resulting from the atomic model is denoted by $R_{\rm con}$, then a function defined by

$$R \text{ ratio} = (R_{\rm con} - R_{\rm unc})/(R_{\rm con} + R_{\rm unc}) \qquad (16.1.8.1)$$

can be used to distinguish false minima from other nonsolutions as well as the true solutions. Once a trial falls into a false minimum, it never escapes. Therefore, the *R* ratio can be used, within *SnB*, as a criterion for early termination of unproductive trials. Based on data for several *P*1 structures, it appears that termination of trials with *R* ratio values exceeding 0.2 will eliminate most false minima without risking rejection of any potential solutions. In the case of triclinic lysozyme, false minima can be recognized, on average, by cycle 25. Since the default recommendation would be for 1000 cycles, a substantial saving in CPU time is realized by using the *R* ratio early-termination test. It should be noted that *SHELXD* optionally allows early termination of trials if the second peak is less than a specified fraction (*e.g.* 40%) of the height of the first. Generally, but not always, the *R*-ratio and peak-ratio tests eliminate the same trials.

Recognizing false minima is, of course, only part of the battle. It is also necessary to find a real solution, and essentially 100% of the triclinic lysozyme trials were found to be false minima when the standard parameter-shift conditions of two 90° shifts were used. In fact, significant numbers of solutions occur only when single-shift angles in the range 140–170° are used (Fig. 16.1.8.1), and there is a surprisingly high *success rate* (percentage of trial structures that go to solutions) over a narrow range of angles centred about 157.5°. It is also not surprising that there is a correlated decrease in the percentage of false minima in the range 140–150°. This suggests that a fruitful strategy for structures that exhibit a large percentage

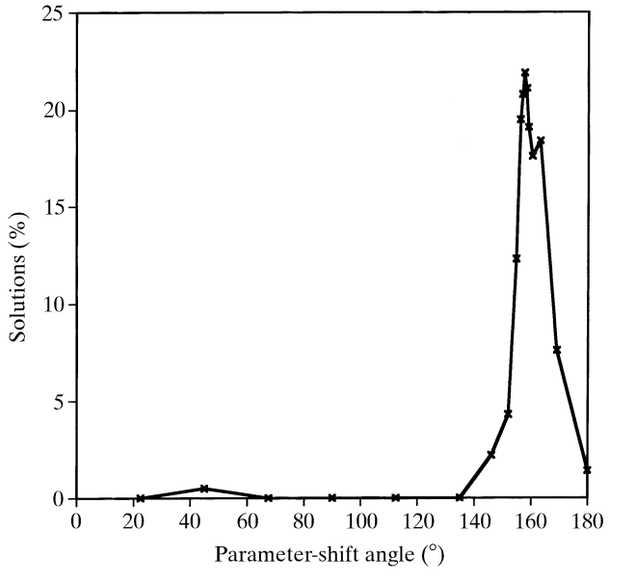

Fig. 16.1.8.1. Success rates for triclinic lysozyme are strongly influenced by the size of the parameter-shift angle. Each point represents a minimum of 256 trials.

of false minima would be the following. Run 100 or so trials at each of several shift angles in the range 90–180°, find the smallest angle which gives nearly zero false minima, and then use this angle as a single shift for many trials. Balhimycin is an example of a large non-*P*1 structure that also requires a parameter shift of around 154° to obtain a solution using the minimal function.

16.1.8.3. *Data resolution and completeness*

The importance of the presence of several atoms heavier than oxygen for increasing the chance of obtaining a solution by *SnB* at resolutions less than 1.2 Å was noticed for truncated data from vancomycin and the 289-atom structure of conotoxin EpI (Weeks & Miller, 1999*b*). The results of *SHELXD* application to hirustasin are consistent with this (Usón *et al.*, 1999). The 55-amino-acid protein hirustasin could be solved by *SHELXD* using either 1.2 Å low-temperature data or 1.4 Å room-temperature data; however, as shown in Fig. 16.1.8.2(*a*), the mean phase error (MPE) is significantly better for the 1.2 Å data over the whole resolution range. The MPE is determined primarily by the data-to-parameter ratio, which is reflected in the smaller number of reliable triplet invariants at lower resolution. Although small-molecule interpretation based on peak positions worked well for the 1.2 Å solution (overall MPE = 18°), standard protein chain tracing was required for the 1.4 Å solution (overall MPE = 26°). As is clear from the corresponding electron-density map (Fig. 16.1.8.2*b*), the *Shake-and-Bake* procedure produces easily interpreted protein density even when bonded atoms are barely resolved from each other. The hirustasin structure was also determined with *SHELXD* using 1.55 Å truncated data, and this endeavour currently holds the record for the lowest-resolution successful application of *Shake-and-Bake*.

The relative effects of accuracy, completeness and resolution on *Shake-and-Bake* success rates using *SnB* for three large *P*1 structures were studied by computing error-free data using the known atomic coordinates. The results of these studies, presented in Table 16.1.8.3, show that experimental error contributed nothing of consequence to the low success rates for vancomycin and lysozyme. However, completing the vancomycin data up to the maximum measured resolution of 0.97 Å resulted in a substantial increase in success rate which was further improved to an astounding success rate of 80% when the data were expanded to 0.85 Å.

On account of overload problems, the experimental vancomycin data did not include any data at 10 Å resolution or lower. A total of 4000 reflections were phased in the dual-space loop in the process of solving this structure with the experimental data. Some of these data were then replaced with the largest error-free magnitudes chosen from the missing reflections at several different resolution limits. The results in Table 16.1.8.4 show a tenfold increase in success rate when only 200 of the largest missing magnitudes were supplied, and it made no difference whether these reflections had a maximum resolution of 2.8 Å or were chosen randomly from the whole 0.97 Å sphere. The moral of this story is that, *when collecting data for Shake-and-Bake, it pays to take a second pass using a shorter exposure to fill-in the low-resolution data*.

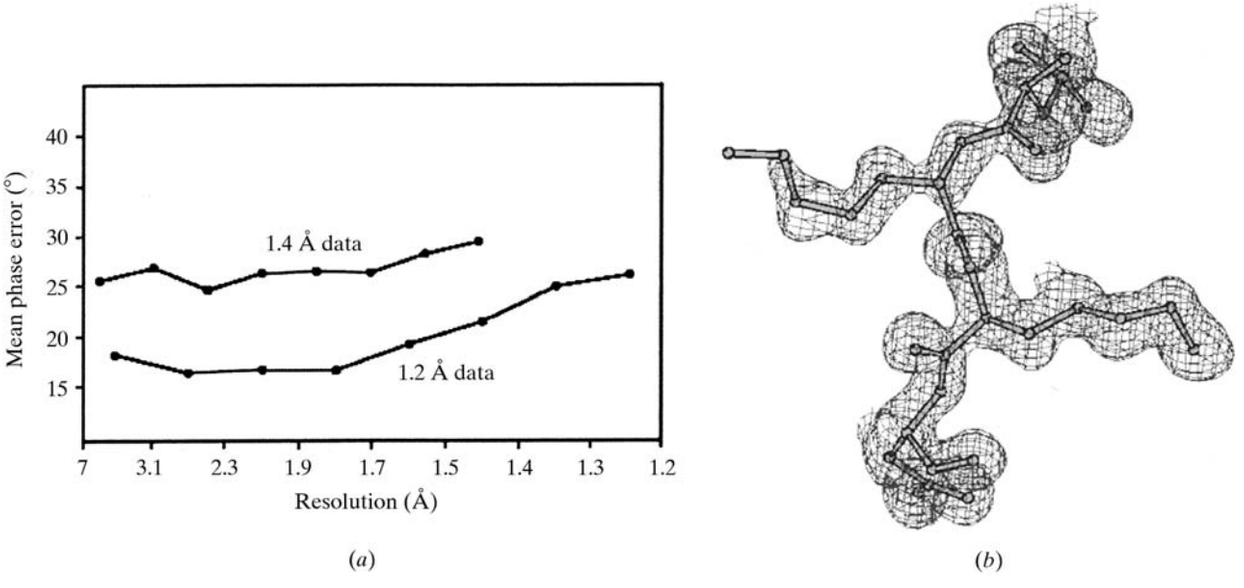

Fig. 16.1.8.2. (*a*) Mean phase error as a function of resolution for the two independent *ab initio SHELXD* solutions of the previously unsolved protein hirustasin. Either the 1.2 Å or the 1.4 Å native data set led to solution of the structure. (*b*) Part of the hirustasin molecule from the 1.4 Å room-temperature data after one round of *B*-value refinement with fixed coordinates.

Table 16.1.8.3. *Success rates for three P1 structures illustrate the importance of using complete data to the highest possible resolution*

	Vancomycin	Alpha-1	Lysozyme
Atoms	547	471	~1200
Completeness (%)	80.2	85.6	68.3
Resolution (Å)	0.97	0.90	0.85
Parameter shift	112.5°, 1	90°, 2	90°, 2
Success rates (%)			
Experimental	0.25	14	0
Error-free	0.2	19	0
Error-free complete	14	29	0.8
Error-free complete extended to 0.85 Å	80	42	—

16.1.8.4. *Choosing a refinement strategy*

Variations in the computational details of the dual-space loop can make major differences in the efficacy of *SnB* and *SHELXD*. Recently, several strategies were combined in *SHELXD* and applied to a 148-atom *P*1 test structure (Karle *et al.*, 1989) with the results shown in Fig. 16.1.8.3. The CPU time requirements of parameter-shift (PS) and tangent-formula expansion (TE) are similar, both being slower than no phase refinement (NR). In real space, the random-omit-map strategy (RO) was slightly faster than simple peak picking (PP) because fewer atoms were used in the structure-factor calculations. Both of these procedures were much faster than iterative peaklist optimization (PO). The original *SHELXD* algorithm (TE + PO) performs quite well in comparison with the *SnB* algorithm (PS + PP) in terms of the percentage of correct solutions, but less well when the efficiency is compared in terms of CPU time per solution. Surprising, the two strategies involving random omit maps (PS + RO and TE + RO), which had been calculated to give reference curves, are much more effective than the other algorithms, especially in terms of CPU efficiency. Indeed these two runs appear to approach a 100% success rate as the number of cycles becomes large. The combination of random omit maps and Karle-type tangent expansion appears to be even more effective (Fig. 16.1.8.4) for gramicidin A, a $P2_12_12_1$ structure (Langs, 1988). It should be noted that conventional direct methods incorporating the tangent formula tend to perform better in $P2_12_12_1$ than in *P*1, perhaps because there is less risk of a uranium-atom pseudosolution.

Subsequent tests using *SHELXD* on several other structures have shown that the use of random omit maps is much more effective than picking the same final number of peaks from the top of the peak list. However, it should be stressed that it is the combination TE + RO that is particularly effective. A possible special case is when a very small number of atoms is sought (*e.g.* Se atoms from MAD data). Preliminary tests indicate that peaklist optimization (PO) is competitive in such cases because the CPU time penalty associated with it is much smaller than when many atoms are involved.

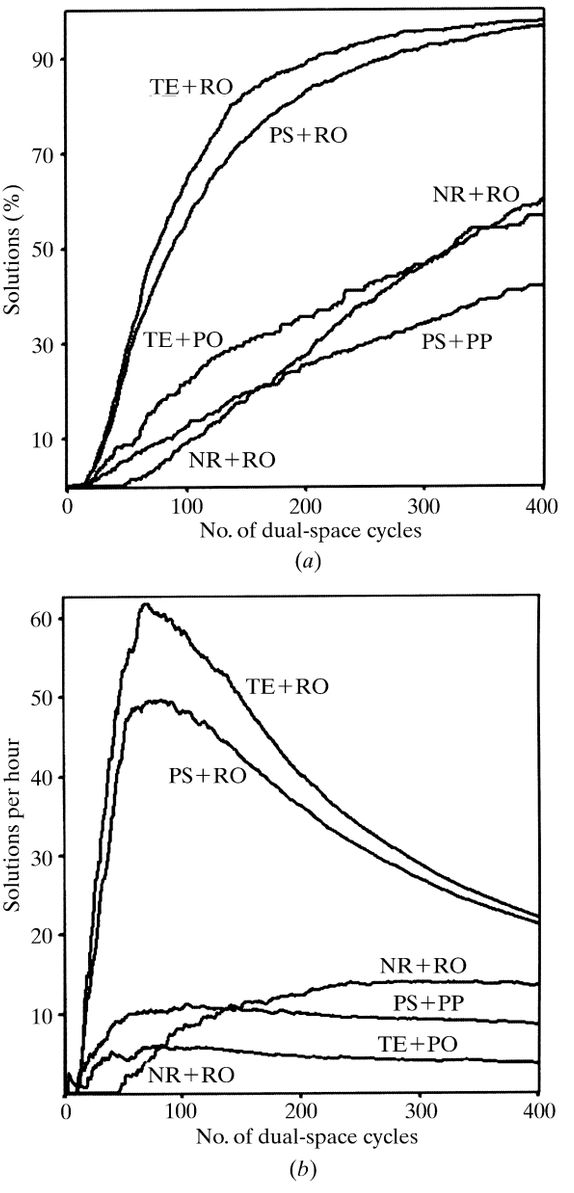

Fig. 16.1.8.3. (*a*) Success rates and (*b*) cost effectiveness for several dual-space strategies as applied to a 148-atom *P*1 structure. The *phase-refinement strategies* are: (PS) parameter-shift reduction of the minimal-function value, (TE) Karle-type tangent expansion (holding the top 40% highest E_c fixed) and (NR) no phase refinement but Sim (1959) weights applied in the *E* map (these depend on E_c and so cannot be employed after phase refinement). The *real-space strategies* are: (PP) simple peak picking using $0.8N_u$ peaks, (PO) peaklist optimization (reducing N_u peaks to $2N_u/3$), and (RO) random omit maps (also reducing N_u peaks to $2N_u/3$). A total of about 10 000 trials of 400 internal loop cycles each were used to construct this diagram.

Table 16.1.8.4. *Improving success rates by 'completing' the vancomycin data*

Error-free reflections added	Success rate (%)
0	0.25
100 (3.5 Å)	0.3
200 (2.8 Å)	2.1
200 (0.97 Å)	2.4
400 (1.3 Å)	8.2
800 (1.1 Å)	11.1

With hindsight, it is possible to understand why the random omit maps provide such an efficient *search algorithm*. In macromolecular structure refinement, it is standard practice to omit parts of the model that do not fit the current electron density well, to perform some refinement or simulated annealing (Hodel *et al.*, 1992) on the rest of the model to reduce memory effects, and then to calculate a new weighted electron-density map (omit map). If the original features reappear in the new density, they were probably correct; in other cases the omit map may enable a new and better

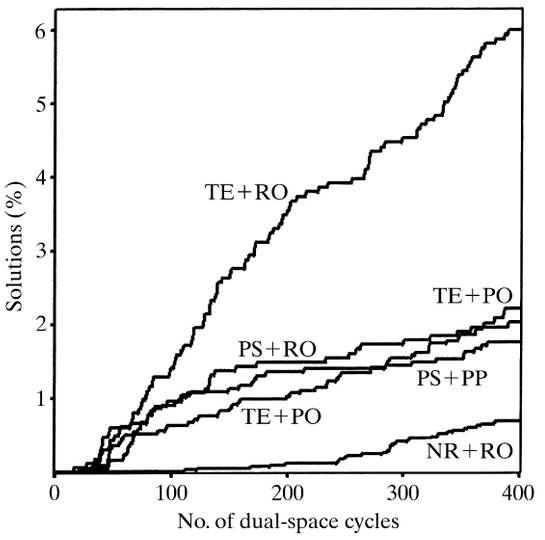

Fig. 16.1.8.4. Success rates for the 317-atom $P2_12_12_1$ structure of gramicidin A.

interpretation. Thus, random omit maps should not lead to the loss of an essentially correct solution, but enable efficient searching in other cases. It is also interesting to note that the results presented in Figs. 16.1.8.3 and 16.1.8.4 show that it is possible, albeit much less efficiently, to solve both structures using random omit maps without the use of any phase relationships based on probability theory (curves NR + RO).

16.1.8.5. *Expansion to P1*

The results shown in Table 16.1.8.4 and Fig. 16.1.8.3 indicate that success rates in space group $P1$ can be anomalously high. This suggests that it might be advantageous to expand all structures to $P1$ and then to locate the symmetry elements afterwards. However, this is more computationally expensive than performing the whole procedure in the true space group, and in practice such a strategy is only competitive in low-symmetry space groups such as $P2_1$, $C2$ or $P\bar{1}$ (Chang *et al.*, 1997). Expansion to $P1$ also offers some opportunities for starting from 'slightly better than random' phases. One possibility, successfully demonstrated by Sheldrick & Gould (1995), is to use a rotation search for a small fragment (*e.g.* a short piece of α-helix) to generate many sets of starting phases; after expansion to $P1$ the translational search usually required for molecular replacement is not needed. Various Patterson superposition minimum functions (Sheldrick & Gould, 1995; Pavelčík, 1994) can also provide an excellent start for phase determination for data expanded to $P1$. Drendel *et al.* (1995) were successful in solving small organic structures *ab initio* by a Fourier recycling method using data expanded to $P1$ without the use of probability theory.

16.1.8.6. *Substructure applications*

It has been known for some time that conventional direct methods can be a valuable tool for locating the positions of heavy-atom substructures using isomorphous (Wilson, 1978) and anomalous (Mukherjee *et al.*, 1989) difference structure factors. Experience has shown that successful substructure applications are highly dependent on the accuracy of the difference magnitudes. As the technology for producing selenomethionine-substituted proteins and collecting accurate multiple-wavelength (MAD) data has improved (Hendrickson & Ogata, 1997; Smith, 1998), there has been an increased need to locate many selenium sites. For larger structures (*e.g.* more than about 30 Se atoms), automated Patterson interpretation methods can be expected to run into difficulties since the number of unique peaks to be analysed increases with the square of the number of atoms. Experimentally measured difference data are an approximation to the data for the hypothetical substructure, and it is reasonable to expect that conventional direct methods might run into difficulties sooner when applied to such data. Dual-space direct methods provide a more robust foundation for handling such data, which are often extremely noisy. Dual-space methods also have the added advantage that the expected number of Se atoms, N_u, which is usually known, can be exploited directly by picking the top N_u peaks. Successful applications require great care in data processing, especially if the F_A values resulting from a MAD experiment are to be used.

All successful applications of *SnB* to previously unknown SeMet data sets, as reported in Table 16.1.8.1, actually involved the use of peak-wavelength anomalous difference data ($|E_\Delta|$). The amount of data available for substructure problems is much larger than for full-structure problems with a comparable number of atoms to be located. Consequently, the user can afford to be stringent in eliminating data with uncertain measurements. Guidelines for rejecting uncertain data have been suggested (Smith *et al.*, 1998). Consideration should be limited to those data pairs ($|E_1|, |E_2|$) [*i.e.*, isomorphous pairs ($|E_{\rm nat}|, |E_{\rm der}|$) and anomalous pairs ($|E_{+\mathbf{H}}|, |E_{-\mathbf{H}}|$)] for which

$$\min[|E_1|/\sigma(|E_1|), |E_2|/\sigma(|E_2|)] \geq x_{\min} \qquad (16.1.8.2)$$

and

$$\frac{\|E_1| - |E_2\|}{[\sigma^2(|E_1|) + \sigma^2(|E_2|)]^{1/2}} \geq y_{\min}, \qquad (16.1.8.3)$$

where typically $x_{\min} = 3$ and $y_{\min} = 1$. The final choice of maximum resolution to be used should be based on inspection of the spherical shell averages $\langle|E_\Delta|^2\rangle_s$ *versus* $\langle s\rangle$. The purpose of this precaution is to avoid spuriously large $|E_\Delta|$ values for high-resolution data pairs measured with large uncertainties due to imperfect isomorphism or general fall-off of scattering intensity with increasing scattering angle. Only those $|E_\Delta|$ for which

$$|E_\Delta|/\sigma(|E_\Delta|) \geq z_{\min} \qquad (16.1.8.4)$$

(typically $z_{\min} = 3$) should be deemed sufficiently reliable for subsequent phasing. The probability of very large difference $|E|$'s (*e.g.* > 5) is remote, and data sets that appear to have many such measurements should be examined critically for measurement errors. If a few such data remain even after the adoption of rigorous rejection criteria, it may be best to eliminate them individually. A later paper (Blessing & Smith, 1999) elaborates further data-selection criteria.

On the other hand, it is also important that the phase:invariant ratio be maintained at 1:10 in order to ensure that the phases are overdetermined. Since the largest $|E|$'s for the substructure cell are more widely separated than they are in a true small-molecule cell, the relative number of possible triplets involving the largest reciprocal-lattice vectors may turn out to be too small. Consequently, a relatively small number of substructure phases (*e.g.* $10N_u$) may not have a sufficient number (*i.e.*, $100N_u$) of invariants. Since the number of triplets increases rapidly with the number of reflections considered, the appropriate action in such cases is to increase the number of reflections as suggested in Table 16.1.7.1. This will typically produce the desired overdetermination.

It is rare for Se atoms to be closer to each other than 5 Å, and the application of *SnB* to AdoHcy data truncated to 4 and 5 Å has been successful. Success rates were less for lower-resolution data, but the CPU time required per trial was also reduced, primarily because much smaller Fourier grids were necessary. Consequently, there was no net increase in the CPU time needed to find a solution.

A special version of *SHELXD* is being developed that makes extensive use of the Patterson function both in generating starting atoms and in providing an independent figure of merit. It has already successfully located the anomalous scatterers in a number of structures using MAD F_A data or simple anomalous differences. A recent example was the unexpected location of 17 anomalous scatterers (sulfur atoms and chloride ions) from the 1.5 Å-wavelength anomalous differences of tetragonal HEW lysozyme (Dauter *et al.*, 1999).

16.1.9. Extending the power of direct methods

The *Shake-and-Bake* approach has increased, by an order of magnitude, the size of structures solvable by direct methods. In addition, a routine application of the *SnB* program to peak-wavelength anomalous difference data has revealed 64 of the 70 Se sites in a selenomethionine-substituted protein (Deacon & Ealick, 1999). Although there is no indication that maximum size limitations have been reached, the fact that the reliability of invariant estimates is known to decrease with increasing structure size suggests that such limitations may exist; based on preliminary tests, it is conjectured that the limit is a few thousand unique atoms for conventional full-structure experiments. Thus, it is natural to wonder what can be done in situations where direct methods are not now routinely applicable. These cases include (1) macromolecules that lack heavy-atom or anomalous-scattering sites with sufficient phasing power for present techniques, (2) macromolecules for which no derivatives are available or for which selenium substitution is impossible, and (3) structures of any size which fail to diffract at sufficiently high resolution. 'Sufficiently high' typically means about 1.2 Å in non-substructure situations.

The requirement for data to very high resolution is, of course, troublesome for macromolecules. One approach to lowering resolution requirements might be to replace the peak search by a search for small common fragments (*e.g.* the five atoms of a peptide unit or an aromatic residue). Furthermore, it should also be possible to integrate the *wARP* procedure (Lamzin & Wilson, 1993; Perrakis *et al.*, 1997) into the real-space part of the *Shake-and-Bake* cycle. The Patterson function (Pavelčík, 1994; Sheldrick & Gould, 1995) and large Karle–Hauptman determinants (Vermin & de Graaff, 1978) might also improve the success rate in borderline cases by providing better-than-random starting coordinates or phases.

However, it is not necessarily true that peak picking is the primary limitation to lower-resolution applications. The lack of enough sufficiently accurate triplet-invariant values appears to be a more fundamental problem. Simulation experiments have shown that the *SnB* program can solve the crambin structure even at 2.0 Å if the invariants used are accurate enough (Weeks *et al.*, 1998). Therefore, the primary breakdown of *Shake-and-Bake* occurs in reciprocal space and could likely be overcome if correct individual invariant values were used instead of the rather crude estimates provided by the Cochran (1955) distribution for the cosines of the triplet invariants. Individual invariant estimates, $\omega_{\mathbf{HK}}$, can be accommodated by a *modified tangent formula*,

$$\tan\varphi_{\mathbf{H}} = \frac{\sum_{\mathbf{K}} W_{\mathbf{HK}} \sin(\omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}})}{\sum_{\mathbf{K}} W_{\mathbf{HK}} \cos(\omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}})}, \qquad (16.1.9.1)$$

or by a *modified minimal function*,

$$R(\Phi) = (1/2\textstyle\sum_{\mathbf{H},\mathbf{K}} W_{\mathbf{HK}}) \sum_{\mathbf{H},\mathbf{K}} W_{\mathbf{HK}} \{[\cos(\Phi_{\mathbf{HK}}) - \cos(\omega_{\mathbf{HK}})]^2 + [\sin(\Phi_{\mathbf{HK}}) - \sin(\omega_{\mathbf{HK}})]^2\}, \qquad (16.1.9.2)$$

where $W_{\mathbf{HK}}$ are appropriately chosen weights. Either of these relationships can serve as the basis for a modified *Shake-and-Bake* procedure.

One approach to providing better invariant values is to estimate them individually from the known structure-factor magnitudes ($|E|$'s). Several methods for doing this have been proposed over the years for the small-molecule case (*e.g.* Hauptman *et al.*, 1969; Langs, 1993), and this approach has met with limited success. In the macromolecular case, however, better options for estimating invariant values are available whenever supplemental information in the form of isomorphous-replacement or anomalous-dispersion data is provided. In addition, the development of multiple-beam diffraction raises the possibility of measuring invariant values experimentally. The modified tangent and minimal-function formulas provide the foundation for a unified treatment of all such supplemental information.

16.1.9.1. *Integration with isomorphous replacement*

The integration of traditional direct methods with isomorphous replacement was initiated by Hauptman (1982*a*), who studied the conditional probability distribution of triplet invariants comprised jointly of native and derivative phases assuming as known the six magnitudes associated with reciprocal-lattice vectors **H**, **K** and $-\mathbf{H}-\mathbf{K}$. It was shown that many triplets, whose true values were near either 0 or π, could be identified and reliably estimated. Later it was shown that cosine estimates could be obtained anywhere in the range -1 to $+1$ (Fortier *et al.*, 1985). In a series of six recent papers, Giacovazzo and collaborators utilized a combined direct-methods/isomorphous-replacement approach, with limited success, to devise procedures for the *ab initio* solution of the phase problem for macromolecules (Giacovazzo, Siliqi & Ralph, 1994; Giacovazzo, Siliqi & Spagna, 1994; Giacovazzo, Siliqi & Zanotti, 1995; Giacovazzo & Platas, 1995; Giacovazzo, Siliqi & Platas, 1995; Giacovazzo *et al.*, 1996). Their methods depend only on diffraction data for a pair of isomorphous structures and do not require any prior structural knowledge. Hu & Liu (1997) have generalized the earlier work to obtain the conditional distribution of the general (n-phase) structure invariant when diffraction data are available for any number (m) of isomorphous structures. Finally, it has been shown that, provided the heavy-atom substructure is known, Hauptman's triplet distribution leads to unique values for the triplets and the individual phases (Langs *et al.*, 1995).

16.1.9.2. *Integration with anomalous dispersion*

In a manner analogous to the SIR case, Hauptman (1982*b*) derived the conditional probability distribution for triplet invariants given six magnitudes ($|E_{\mathbf{H}}|, |E_{-\mathbf{H}}|, |E_{\mathbf{K}}|, |E_{-\mathbf{K}}|, |E_{\mathbf{H}+\mathbf{K}}|, |E_{-\mathbf{H}-\mathbf{K}}|$) in the presence of anomalous dispersion. It was shown that unique estimates, lying anywhere in the whole interval 0–2π, could be obtained for the triplet values. This result was unanticipated since all earlier work had led to the conclusion that a twofold ambiguity in the value of an individual phase was intrinsic to the SAS approach. Later, it was demonstrated how the probabilistic estimates led to individual phases by means of a system of SAS tangent equations (Hauptman, 1996). Although the initial application of this tangent-based approach to the previously known macromomycin structure (750 non-H protein atoms plus 150 solvent molecules) was encouraging, it has not yet been applied to unknown macromolecules.

The conditional probability distributions of the quartet invariants, in both the SIR and SAS cases, have been derived based on corresponding difference structure factors rather than on the individual structure factors themselves (Kyriakidis *et al.*, 1996). Fan and his collaborators (Fan *et al.*, 1984; Fan & Gu, 1985; Fan *et al.*, 1990; Sha *et al.*, 1995; Zheng *et al.*, 1996) have also extensively studied the use of direct methods in the SAS case. Applications to the known small protein avian pancreatic polypeptide at 2 Å

references