

16. DIRECT METHODS

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 *P1* 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 (*C222*) 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_{unc} and the value of the minimal function computed using the *constrained* phases resulting from the atomic model is denoted by R_{con} , then a function defined by

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

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 *P1* 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

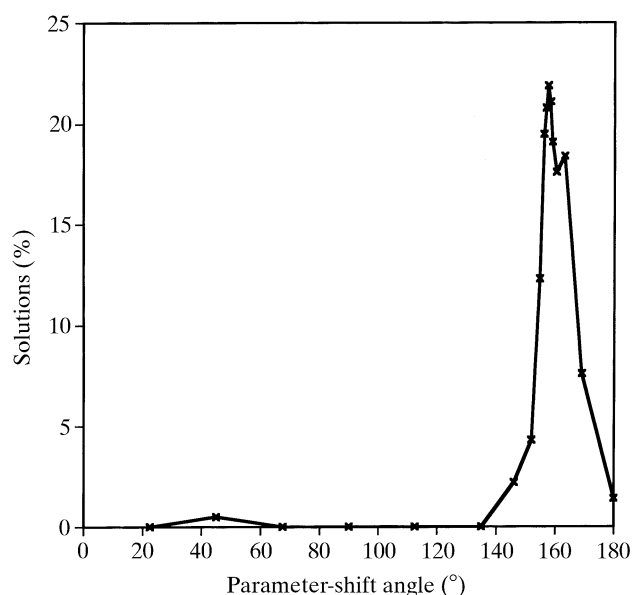


Fig. 16.1.8.1. Success rates for triclinic lysozyme are strongly influenced by the size of the parameter-shift angle. Each point represents a minimum of 256 trials.

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-*P1* 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, 1999b). 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. 16.1.8.2b), 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 *P1* 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.*

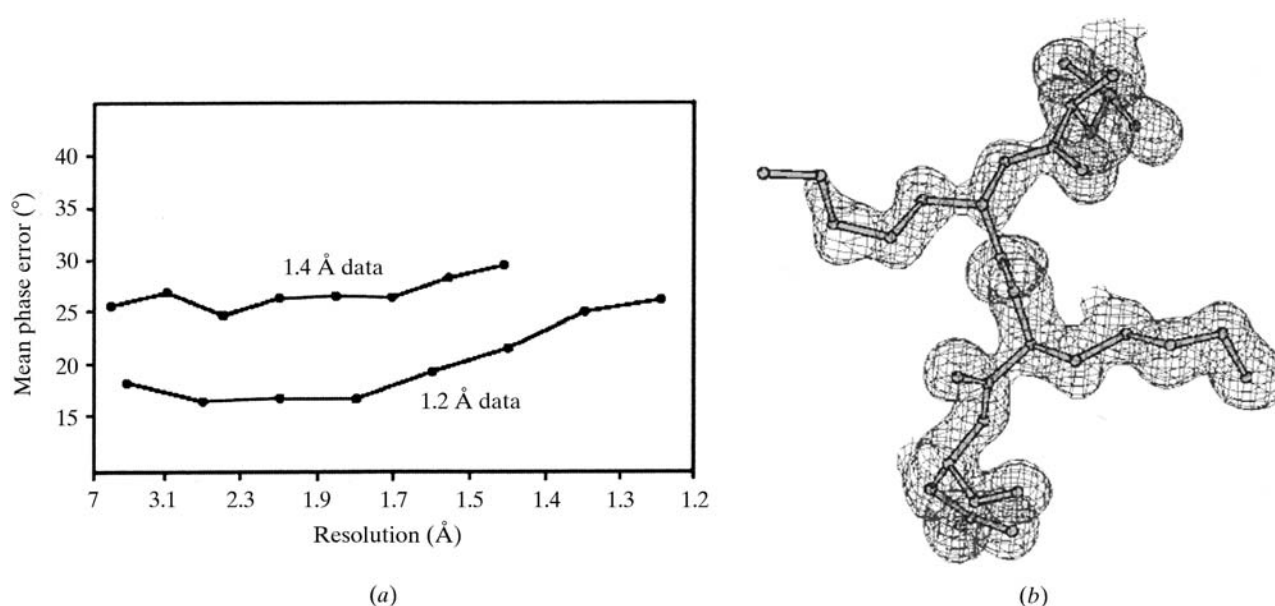


Fig. 16.1.8.2. (a) Mean phase error as a function of resolution for the two independent *ab initio* *SHELXD* solutions of the previously unsolved protein hirustasin. Either the 1.2 Å or the 1.4 Å native data set led to solution of the structure. (b) Part of the hirustasin molecule from the 1.4 Å room-temperature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