

## 15. Location of Heavy Atoms from Protein $\Delta F$ Data

In principle both the Patterson interpretation and direct methods are suitable for the location of heavy atoms from protein or oligonucleotide isomorphous or anomalous  $\Delta F$  data-sets.

### 15.1 Data preparation

For both the anomalous and isomorphous cases the user must prepare a file name.hkl containing  $h$ ,  $k$ ,  $l$ ,  $\Delta F$  and  $\sigma(\Delta F)$  [or  $(\Delta F)^2$  and  $\sigma((\Delta F)^2)$ ] in the usual format (3I4,2F8.2), terminated by the dummy reflection with  $h = k = l = 0$ . The sign of  $\Delta F$  is ignored. The auxiliary program SHELXPRO provides some facilities for the generation of this file, as does for example the CCP4 system.

Careful scaling of the derivative and native data, pruning of statistically unreasonable  $\Delta F$ -values, and good estimated standard deviations are essential to the success of this approach. It should be emphasised that treating  $\Delta F$  as if it were  $F$  involves an approximation which, at best, will add appreciable 'noise'.

SHELXS-96 will usually recognize that it has been given macromolecular  $\Delta F$  data (from the cell volume and contents) and will then set appropriate defaults, so as with small molecules the *.ins* file will often simply consist of TITL..UNIT, then TREF (for direct methods) or PATT (Patterson interpretation) and finally HKLF 3 (because the *.hkl* file contains  $\Delta F$  (HKLF 3) or  $(\Delta F)^2$  (HKLF 4). The UNIT instruction should contain the correct number of heavy atoms and the **square root** of the number of light atoms in the cell; they may conveniently be assumed to be nitrogen. The mean atomic volume and density printed by the program should of course be ignored. It is strongly recommended that these standard TREF and PATT jobs are tried first before any parameters are varied.

### 15.2 Limitations of $\Delta F$ -data

Unfortunately there are two fundamental difficulties with the application of direct methods to  $\Delta F$  data. The first is that the negative quartets are meaningless, because the  $\Delta F$ -values represent lower bounds on their true values, and so are unsuitable for identifying the very small  $E$ -values which are required for the cross-terms of the negative quartets. On the other hand the  $\Delta F$  values do correctly identify the **largest**  $E$ -values, and so the old triplet formula works well. The second problem is that the estimation of probabilities for the triplet formula for the use in figures of merit: what should replace the  $1/N$  term (where  $N$  is the number of atoms per cell) when  $\Delta F$ -data are used?

### 15.3 Direct methods

Most of the recent advances in direct methods exploit either the weak reflections or more sophisticated formulas for probability distributions, so are wasted on  $\Delta F$  data. Nevertheless, direct methods will tend to perform better in space groups with (a) translation symmetry (not counting lattice centering), (b) a fixed rather than a floating origin and (c) no special

positions; thus  $P2_12_12_1$  (the only space group to fulfill all three criteria) is good but  $P1$ ,  $C2$ ,  $R3$  and  $I4$  are unsuitable.

If the standard direct methods run fails to find convincing heavy-atom sites, it should first be checked that the program has put out a comment that it has set the defaults for macromolecular data. The number of phase permutations may have to be increased (the first TREF parameter) or the number of large  $E$ -values for phase refinement may have to be changed (one should aim for at least 20 triplets per refined phase), but if too many phases are refined the performance is degraded because the  $\Delta F$ -values only identify the strongest  $E$ -values reliably. The probability estimates may be changed by modifying the UNIT instruction, or more simply by changing the third TREF parameter, which multiplies the products of the three  $E$ -values in the triplet probability formula; for small molecules a value in the range 0.75 to 0.95 gives the best probability estimates, but it may be necessary to go outside this range for  $\Delta F$ -data.

#### 15.4 Patterson interpretation

For location of the heavy-atom site by Patterson interpretation of  $\Delta F$ -data it may well be necessary to increase the number of superposition vectors to be tried (the first parameter on the PATT instruction), since the heavy-atom to heavy-atom vectors may be well down the Patterson peak-list. This number can be made negative to increase the 'depth of search' at the cost of a significant increase in computer time. The second number (the minimum vector length for the superposition vector) should be set to at least 8 Å (and to a larger value if the cell is large), and it can usually be made negative to indicate that special positions are not to be considered as possible heavy atom sites. An advantage of Patterson as opposed to direct methods is that such false solutions can be eliminated at a much earlier stage.

The third PATT parameter is also fairly critical for macromolecular  $\Delta F$ -data; it is the apparent resolution, and is used to set the tolerances for deconvoluting the superposition map. If - as can easily happen with area detector data - a few  $\Delta F$ -values are at appreciably higher resolution than the rest of the data, this may fool the program into setting too high an effective resolution. In such cases it is worth experimenting with several different values, e.g. 3.5 Å instead of 3.0 etc. The only other parameter which may need to be altered is maxat, if more than 8 sites are expected.

A typical  $\Delta F$  PATT run (e.g. PATT 10 -12 2.5) will produce a relatively large number of possible solutions, some of which may be equivalent. The 'correlation coefficient' (which is defined in the same way as in most molecular replacement programs) is the only useful figure of merit for comparison purposes. Hand interpretation of the 'crossword table' is not as easy as for small molecules, because the minimum interatomic distances are not so useful; it is however still necessary to find a set of atoms for which the Patterson minimum function values are consistently high for at least most of the pairs of sites involved. This information tends to be more decisive for the higher symmetry space groups, because when there are more vectors between symmetry equivalents, it is unlikely that all will be associated with large Patterson values simultaneously by accident.