Acta Cryst. (1965). **18**, 576

Two-dimensional least-squares refinement of heavy atom parameters in the determination of protein structures.

By B. LUNDBERG, *Institute of Chemistry, University of Uppsala, Uppsala, Sweden*

(*Received* 26 *June* 1964)

In the method of isomorphous replacement in protein structure determination theoretically two, but in practice four to six, heavy atom derivatives give satisfactory accuracy in the phase angle calculations. Before the phases are determined it is necessary to refine those parameters of the heavy atoms which are independent of phase. In the method to be described all parameters except relative y's are refined. For the centrosymmetric $h0l$ projection of space group $P2_1$ the function chosen for minimization by the method of least squares is:

$$E = \sum_n w_n(|F_o| - |F_c|)^2$$
$$= \sum_n w_n\{(\sum_p (Z_p^s m_{1,p} \cos m_{2,p}) + s(F)|F| + s(F_H)k|F_H|\}^2$$

(Hart, 1961). In this expression $\sum_n$ is taken for all reflexions and $\sum_p$ for p heavy atoms in the asymmetric unit.

$s(F)$ = sign of observed structure factor, $|F|$, of the protein,
$s(F_H)$ = sign of observed structure factor, $|F_H|$, of the derivative,

$$m_{1,p} = \hat{f}_{on} \exp(-B_p \sin^2\theta/\lambda^2),$$

$$m_{2,p} = 2\pi(hx_p + lz_p),$$

$f_H = \sum_p Z_p^s m_{1,p} \cos m_{2,p}$ = calculated structure factor for the heavy atoms,

$$F_o = s(F)|F|,$$

$$|F_c| = |s(F_H)k|F_H| - f_H|,$$

$$s(F_c) = \text{sign of } F_c,$$

k is the scale factor for heavy atom data relative to that of parent compound,
Z^s is the effective number of electrons in the unit cell for heavy atom p and describes the extent of substitution,

$\hat{f}_{on}$ is the unitary scattering factor,
x and z are the fractional coordinates of the heavy atom,
w_n is a weighting factor.

The four possible combinations of signs for $|F|$ and $|F_H|$ are given in Fig. 1.

A minimum is given by $dE/dr_j = 0$; $j = 1, 2, \ldots, q$, where q is the number of parameters, r.

This yields the normal equations:

$$\left[\sum_n w_n \frac{\partial|F_c|}{\partial r_i}\frac{\partial|F_c|}{\partial r_j}\right](\Delta r_i)$$
$$= \sum_n w_n(|F_o| - |F_c|)\frac{\partial|F_c|}{\partial r_i};\ i, j = 1, 2, \ldots, q,$$

which can be expressed in matrix form as

$$(b_{ij})(c_i) = (a_i).$$

The shifts (b_i) in the parameters are given by $(c_i) = (b_{ij})^{-1}(a_i)$.

The derivatives are

$$\frac{\partial|F_c|}{\partial Z_p^s} = -1s(F_c)m_{1,p}\cos m_{2,p}$$

$$\frac{\partial|F_c|}{\partial B_p} = -1s(F_c)(-\sin^2\theta/\lambda^2)Z_p^s m_{1,p}\cos m_{2,p}$$

$$\frac{\partial|F_c|}{\partial x_p} = -1s(F_c)(-2\pi h)Z_p^s m_{1,p}\sin m_{2,p}$$

$$\frac{\partial|F_c|}{\partial z_p} = -1s(F_c)(-2\pi l)Z_p^s m_{1,p}\sin m_{2,p}$$

$$\frac{\partial|F_c|}{\partial k} = +1s(F_c)s(F_H)|F_H|.$$

In the calculations on data from the X-ray investigation of carbonic anhydrase form C (Tilander, Strandberg & Fridborg, 1965), this method was used in a computer program with the designation ROHAP.

During the refinement those reflexions are excluded which do not fulfil the following conditions:

$$C_1 \le [|s(F_H)k|F_H| - s(F)|F||/|f_H|] \le C_2$$
$$||s(F_H)k|F_H| - s(F)|F|| - |f_H|| \le C_3,$$

where the constants C_1, C_2 and C_3 are chosen in an appropriate manner. The program calculates the reliability indices for (*a*) reflexions excluded from the refinement, (*b*) reflexions included in the refinement and (*c*) all reflexions, using the expression

$$R_{a,b,c} = \frac{||s(F_H)k|F_H| - s(F)|F|| - |f_H||}{|s(F_H)k|F_H| - s(F)|F||}.$$

The empirical weighting factor used for carbonic anhydrase form C is

$$w = (|s(F_H)k|F_H| - s(F)|F||)^2[(k|F_H| + |F|)/2]^{-1}.$$

Correlation coefficients (Geller, 1961) are calculated according to

$$\varrho_{ij} = (b_{ij})^{-1}/[(b_{ii})^{-1}(b_{jj})^{-1}]^{\frac{1}{2}}$$

where $(b_{ij})^{-1}$ is the inverse matrix of (b_{ij}) obtained by the elimination method of Gauss.

A full matrix is used in the solution of the normal equations. The approximate position of a heavy atom is taken from difference Patterson and difference Fourier syntheses. The approximate degree of substitution is estimated from the height of the difference Fourier peaks combined with chemical analysis. It is possible to refine a structure in which the asymmetric unit contains eight atoms, involving thirty-three parameters (the number is limited by the storage of the machine), but there will not in general be more than, say, four heavy atoms in a protein derivative. This feature is valuable, however, when it becomes necessary to distinguish between small real peaks and false peaks caused by series termination errors and random errors in the measurements. As an example there were eight possible difference Fourier

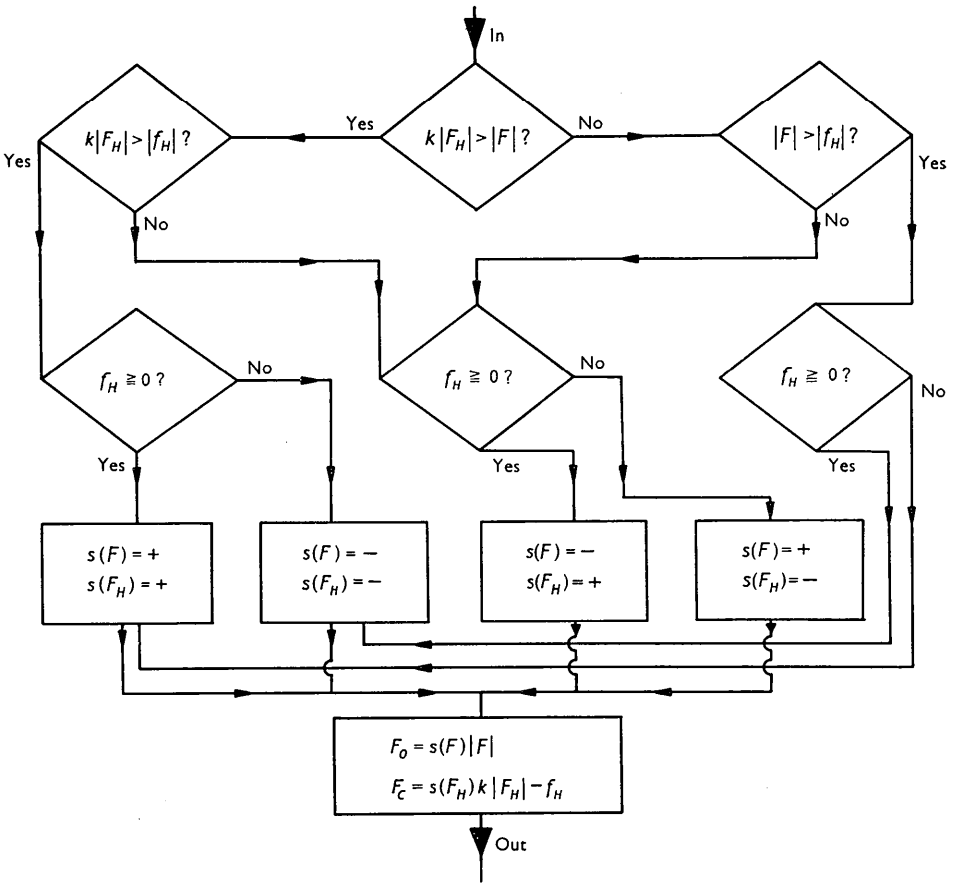

Fig. 1. Schema for sign determination of the observed structure factors $|F|$ and $|F_H|$.

peaks in one of the mercury derivatives of carbonic anhydrase. The approximate substitutions before refinement were respectively 100, 100, 35, 35% and four of 15%. After one cycle of refinement with ROHAP the last four possibilities had a negligible substitution. It is possible to keep any desired type of parameter constant during the refinement, and this has proved useful. In the investigation of carbonic anhydrase, for example, the coupling between Z^s and B is observed to be very strong ($\varrho_{ij} \sim 0{\cdot}9$) and since, at a resolution of 5·5 Å, $\exp(-B \sin^2\theta/\lambda^2)$ is very insensitive to changes in B this factor has so far been set constant at a value approximatively determined from a Wilson plot. Although the coupling between Z^s and k is relatively strong, ($\varrho_{ij} \sim 0{\cdot}5$), refinement of both these parameters at the same time has been quite successful. A further observation made during the use of this program is that relatively large random errors in a few reflexions can dominate the whole refinement. A criterion as to which reflexions should not be used in the refinement, but included in R_c, is given by the deviation of $(w)^{\frac{1}{2}}(|F_o| - |F_c|)$ from the mean value. Further results are given in the paper by Tilander *et al.* (1965).

The electronic computer used with this program was the Swedish built FACIT, which is a binary machine with a ferrite memory for 4096 or 2048 words, each word consisting of 20 or 40 binary digits respectively. Two magnetic tapes are used, one to accommodate four subroutines in the program and one to store the data. Using fixed point calculations the machine has an addition time of 50 μsec. With 150 *h0l* reflexions and nine parameters (two heavy atoms) each cycle required about three minutes computing time and the refinement was usually completed after four to six cycles.

I wish to express my sincere gratitude to Dr B. Strandberg, who introduced me to the field of protein structure investigations, for his continuous interest and valuable discussions. I also thank Prof. G. Hägg for his interest and the facilities placed at my disposal.

This work was supported by grants from the National Institutes of Health, U.S. Public Health Service (GM 11307) and from the Swedish Natural Science Research Council (259–22,24).

References

Geller, S. (1961). *Acta Cryst.* **14**, 1026.
Hart, R. G. (1961). *Acta Cryst.* **14**, 1194.
Tilander, B., Strandberg, B. E. & Fridborg, K. (1965). *J. Mol. Biol.* To be published.