

- EPELBOIN, Y. (1974). *J. Appl. Cryst.* **7**, 372–377.
 EPELBOIN, Y. (1977). *Acta Cryst.* **A33**, 758–767.
 KATO, N. (1968). *Acta Cryst.* **A24**, 157–160.
 KATO, N. & LANG, A. R. (1959). *Acta Cryst.* **12**, 787–794.
 LANG, A. R. (1959). *Acta Cryst.* **12**, 249–250.
 LAUE, M. VON (1935). *Ann. Phys. (Leipzig)*, **23**, 705.
 LAUE, M. VON (1960). *Röntgenstrahlinterferenzen*. Frankfurt am Main: Akademische Verlagsges.
 PETRASHEN', P. V. (1976). *Fiz. Tverd. Tela*, **18**, 3729–3731.
 PETRASHEN', P. V. & CHUKHOVSKII, F. N. (1978). *Fiz. Tverd. Tela*, **20**, 1104–1108.
 PETRASHEN', P. V., CHUKHOVSKII, F. N. & SHULPINA, I. L. (1978). *Dokl. Akad. Nauk SSSR*, **240**, 84–87.
 PINSKER, Z. G. (1974). *The X-ray Dynamical Scattering by Perfect Crystals*. Moscow: Nauka. (In Russian.)
 TAKAGI, S. (1962). *Acta Cryst.* **15**, 1311–1312.
 TAKAGI, S. (1969). *J. Phys. Soc. Jpn.* **26**, 1239–1253.
 TAUPIN, D. (1964). *Bull. Soc. Fr. Minéral. Cristallogr.* **87**, 469–511.
 WONCIEWICZ, B. C. & PATEL, J. R. (1975). *J. Appl. Cryst.* **8**, 67–68.
 WONCIEWICZ, B. C. & PATEL, J. R. (1976). *J. Appl. Phys.* **47**, 1837–1845.

Acta Cryst. (1980). **A36**, 295–298

On the Fourier Refinement of Protein Structures

BY M. VIJAYAN

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India and Laboratory of Molecular Biophysics, South Parks Road, Oxford, England*

(Received 20 April 1979; accepted 12 November 1979)

Abstract

The situation normally encountered in the high-resolution refinement of protein structures is one in which the inaccurate positions of P out of a total of N atoms are known whereas those of the remaining atoms are unknown. Fourier maps with coefficients $(F_N - F'_p) \times \exp(ia'_p)$ and $(mF_N - nF'_p) \exp(ia'_p)$, where F_N is the observed structure factor and F'_p and α'_p are the magnitude and the phase angle of the calculated structure factor corresponding to the inaccurate atomic positions, are often used to correct the positions of the P atoms and to determine those of the Q unknown atoms. A general theoretical approach is presented to elucidate the effect of errors in the positions of the known atoms on the corrected positions of the known atoms and the positions of the unknown atoms derived from such maps. The theory also leads to the optimal choice of parameters used in the different syntheses. When the errors in the positions of the input atoms are systematic, their effects are not taken care of automatically by the syntheses.

Introduction

Fourier methods are extensively used at different stages of the high-resolution refinement of proteins to calcu-

late the shifts in structural parameters as well as to check the results obtained during the course of the refinement (Freer, Alden, Carter & Kraut, 1975; Watenpaugh, Sieker, Herriot & Jensen, 1973; Bode & Schwager, 1975; Dodson, Dodson, Hodgkin, Isaacs & Vijayan, 1978). The problems associated with these methods become more pronounced when they are applied to protein crystallography for several reasons. They often give rise to phenomena which cannot be anticipated or easily defined, and one can arrive at a crystallographically acceptable, but erroneous, refined structure (Dodson *et al.*, 1978). Hence the need for a fresh theoretical look at Fourier methods.

Modulus synthesis, phase synthesis and their convolution

Fourier methods have been analysed by many workers (*e.g.* Luzzati, 1953; Ramachandran & Srinivasan, 1970; Dodson & Vijayan, 1971). The formulation of Ramachandran & Srinivasan is used in the present analysis and the relevant results from their work are outlined in this section.

If the structure consists of N atoms with positions \mathbf{r}_j and form factors f_j ($j = 1, \dots, N$), a Fourier synthesis with the structure factors $F \exp(ia)$ as coefficients obviously has peaks at \mathbf{r}_j with strengths proportional to f_j . If the moduli of the structure factors are used as Fourier coefficients, Ramachandran & Srinivasan have shown, to a first approximation, that the resulting

* Contribution No. 144.

synthesis would have the following peak positions and strengths:

$$0 \quad (\sum f_j^2)^{1/2}$$

$$\mathbf{r}_j - \mathbf{r}_k \quad f_j f_k / 2 (\sum f_j^2)^{1/2}$$

$$(j \neq k)$$

Likewise, a synthesis with the phase factors, $\exp(i\alpha)$, as coefficients has, to a first approximation, the following peak positions and strengths:

$$\mathbf{r}_j \quad f_j / (\sum f_j^2)^{1/2}$$

$$\mathbf{r}_j + \mathbf{r}_k - \mathbf{r}_l \quad -f_j f_k f_l / 2 (\sum f_j^2)^{3/2}$$

$$(k \neq l)$$

The convolution of the above two syntheses should then have the following peak positions and strengths:

$$\mathbf{r}_j \quad f_j$$

$$\mathbf{r}_j + \mathbf{r}_k - \mathbf{r}_l \quad f_j f_k f_l / 2 \sum f_j^2$$

$$(k \neq l)$$

$$\mathbf{r}_j + \mathbf{r}_k - \mathbf{r}_l \quad -f_j f_k f_l / 2 \sum f_j^2$$

$$(k \neq l)$$

$$\mathbf{r}_j - \mathbf{r}_k + \mathbf{r}_l + \mathbf{r}_m - \mathbf{r}_n \quad -f_j f_k f_l f_m f_n / 4 (\sum f_j^2)^2$$

$$(j \neq k, m \neq n)$$

The second and third of the above terms cancel out. Thus, one is left with the first term which gives the atomic peaks and the fourth term which gives the general background. The convolution of the Fourier transforms of F and $\exp(i\alpha)$ should, however, be identical to the Fourier transform of $F \exp(i\alpha)$. The latter has peaks only at \mathbf{r}_j with strengths f_j . Therefore, the fourth term in the convolution must have resulted from the approximations involved in deriving the Fourier transforms of F and $\exp(i\alpha)$. Thus, when considering the convolution of a modulus synthesis and a phase synthesis, this term can be neglected as will be done in the following.

The situation in proteins

The situation encountered in the high-resolution refinement of protein structures is one in which the inaccurate positions of P out of a total of N atoms are known whereas those of the remaining Q atoms are unknown. Following Ramachandran & Srinivasan (1970), we define

$$S_N^2 = \sum_{j=1}^N f_{Nj}^2, \quad S_P^2 = \sum_{j=1}^P f_{Pj}^2 \quad \text{and} \quad S_Q^2 = \sum_{j=1}^Q f_{Qj}^2;$$

$$N = P + Q \quad \text{and} \quad S_N^2 = S_P^2 + S_Q^2.$$

The correct positions and the inaccurate (known) positions of the P atoms may be denoted by \mathbf{r}_{Pj} and \mathbf{r}'_{Pj} , and the corresponding structure factors by $F_P \exp(i\alpha_P)$ and $F'_P \exp(i\alpha'_P)$. Now, the peak positions and strengths of syntheses with different types of coefficients can be easily derived and are listed below.

F_N :

$$0 \quad S_N$$

$$\mathbf{r}_{Pj} - \mathbf{r}_{Pk} \quad f_{Pj} f_{Pk} / 2 S_N$$

$$(j \neq k)$$

$$\mathbf{r}_{Pj} - \mathbf{r}_{Qk} \quad f_{Pj} f_{Qk} / 2 S_N$$

$$\mathbf{r}_{Qj} - \mathbf{r}_{Pk} \quad f_{Qj} f_{Pk} / 2 S_N$$

$$\mathbf{r}_{Qj} - \mathbf{r}_{Qk} \quad f_{Qj} f_{Qk} / 2 S_N$$

$$(j \neq k)$$

$\exp(i\alpha'_P)$:

$$\mathbf{r}'_{Pj} \quad f_{Pj} / S_P$$

$$\mathbf{r}'_{Pj} + \mathbf{r}'_{Pk} - \mathbf{r}'_{Pl} \quad -f_{Pj} f_{Pk} f_{Pl} / 2 S_P^3$$

$$(k \neq l)$$

$F'_P \exp(i\alpha'_P)$:

$$\mathbf{r}'_{Pj} \quad f_{Pj}$$

Difference Fourier synthesis

Difference Fourier maps with coefficients

$$(F_N - F'_P) \exp(i\alpha'_P)$$

are often used to correct the positions of P atoms and to determine those of the Q atoms. Using the results mentioned in the previous sections, it can be readily shown that such a map would consist of peaks with the following positions and strengths:

$$\mathbf{r}'_{Pj} \quad [(S_N - S_P) / S_P] f_{Pj} \quad (\text{I})$$

$$\mathbf{r}'_{Pj} + \mathbf{r}'_{Pk} - \mathbf{r}'_{Pl} \quad - (S_N / 2 S_P^3) f_{Pj} f_{Pk} f_{Pl}$$

$$(j \neq l) \quad (\text{II})$$

$$\mathbf{r}'_{Pj} + \mathbf{r}_{Qk} - \mathbf{r}_{Ql} \quad (1/2 S_N S_P) f_{Pj} f_{Qk} f_{Ql}$$

$$(k \neq l) \quad (\text{III})$$

$$\mathbf{r}'_{Pj} + \mathbf{r}_{Pk} - \mathbf{r}_{Ql} \quad (1/2 S_N S_P) f_{Pj} f_{Pk} f_{Ql} \quad (\text{IV})$$

$$\mathbf{r}_{Pj} + \mathbf{r}'_{Pk} - \mathbf{r}_{Pl} \quad (1/2 S_N S_P) f_{Pj} f_{Pk} f_{Pl}$$

$$(j \neq l) \quad (\text{V})$$

$$\mathbf{r}_{Qj} + \mathbf{r}'_{Pk} - \mathbf{r}_{Pl} \quad (1/2 S_N S_P) f_{Qj} f_{Pk} f_{Pl} \quad (\text{VI})$$

These terms can be thought of as sets of vectors centred around different atomic positions.

I: Peaks at the inaccurate positions of the P atoms.

II: 'Patterson maps' of \mathbf{r}'_{pj} placed on each \mathbf{r}'_{pj} . The peaks have negative strengths. When $k = l$, we have negative origin peaks at \mathbf{r}'_{pj} . When $k \neq l$, general background results.

III: Origin-removed 'Patterson maps' of the unknown atoms \mathbf{r}_{Qj} placed at each \mathbf{r}'_{pj} . No contribution at peak positions. General background results.

IV: Vector sets between \mathbf{r}_{pj} and \mathbf{r}_{Qj} centred around \mathbf{r}'_{pj} . Background results.

V: Vector sets between \mathbf{r}'_{pj} and \mathbf{r}_{pj} placed at each \mathbf{r}_{pj} . Significant peaks occur at correct atomic positions \mathbf{r}_{pj} when $k = l$ if the positional errors are small. When $k \neq l$, background results.

VI: Vector sets between \mathbf{r}'_{pj} and \mathbf{r}_{pj} placed at each \mathbf{r}_{Qj} . Significant peaks occur at positions \mathbf{r}_{Qj} of the unknown atoms when $k = l$ if the positional errors are small. When $k \neq l$, background results.

Following the above arguments and assuming N and P to be large, which is true in the case of proteins, we can rewrite the peak strengths at different atomic positions:

$$\mathbf{r}'_{pj} \quad [(S_N - 2S_P)/2S_P] f_{pj}$$

$$\mathbf{r}_{pj} + \langle (\mathbf{r}'_{pk} - \mathbf{r}_{pk}) \rangle_{j \neq k} \quad (S_P/2S_N) f_{pj}$$

$$\mathbf{r}_{Qj} + \langle (\mathbf{r}'_{pk} - \mathbf{r}_{pk}) \rangle \quad (S_P/2S_N) f_{Qj}$$

In addition, there are several terms contributing to the background. These are unimportant in normal circumstances.

It can be readily seen that all the known properties of the difference Fourier synthesis follow from these expressions. For example, when almost all the atoms are included in the calculations, peaks corresponding to the unknown atoms appear with nearly half their normal strengths. The peak strengths steadily decrease as the proportion of the atoms included in phase-angle calculations decreases. In addition, they also tell us how errors in the positions of input atoms affect the positions of the atoms which we seek to determine. This is an aspect of great importance in the refinement of protein structures. When the errors are randomly distributed, the peaks corresponding to the second and the third terms will be centred around the correct positions. Also, when \mathbf{r}_{pj} and \mathbf{r}'_{pj} are close to each other, the combined effect of the first two terms is to produce a density gradient, the magnitude of which depends on the proportion of the scattering matter included in the calculation of structure factors. It can be easily seen that non-random errors lead to shifts in peak positions. The background is also then likely to become important as it consists of components of various shifted Patterson maps.

'Mixed' syntheses

The theory developed above can also be applied to rationalize, and to determine the parameters in, the syntheses employing coefficients of the type

$$(mF_N - nF'_P) \exp(i\alpha'_P)$$

or

$$(kF_N - F'_P) \exp(i\alpha'_P)$$

used by many workers (Freer, Alden, Carter & Kraut, 1975; Bode & Schwager, 1975). These two syntheses are obviously identical except for a scale factor. On the basis of the earlier discussion, it can be readily seen that the best results in terms of peak strengths are obtained when

$$m = 2S_N/S_P \quad \text{and} \quad n = S_N^2/S_P^2$$

or

$$k = m/n = 2S_P/S_N.$$

Errors in the magnitudes of the observed structure factors are not considered in this treatment. These errors are likely to make the effective scattering power of the known atoms less than that calculated theoretically. Therefore, one should perhaps use a lower value of k than that given by the above equation.

The area where the greatest problems are faced in refining protein structures is concerned with poorly defined atoms which usually belong to residues occurring on the surface of the protein molecule or to solvent molecules. They are often associated with high temperature factors arising from large thermal-vibration amplitudes as well as static disorder corresponding to different structural or conformational possibilities. Consequently, such poorly defined atoms are associated with weak and diffuse electron densities. Most often, the positions of well-defined atoms are determined in the early stages of refinement. Attempts are then made to locate the poorly defined atoms. When a difference Fourier synthesis or a normal Fourier synthesis [with $F_N \exp(i\alpha'_P)$ as the coefficients] is used for this purpose, it follows from theory that the peaks corresponding to these atoms appear with less than half or half their normal strengths. The diffuse nature of the peaks coupled with their low strengths often makes it difficult to distinguish them from the general background resulting from various errors. However, the corresponding peaks in a synthesis with $(mF_N - nF'_P) \exp(i\alpha'_P)$ as the coefficients, though diffuse, would appear with their normal strengths, thus making it easier to distinguish them from the background.

It can be readily seen that S_N^2/S_P^2 would not correspond to the ratio between the scattering power of the whole structure (including solvent molecules) and that of the known atoms when the positions of the well

defined atoms have been determined and attempts are being made to locate poorly defined atoms on account of the high temperature factors associated with the latter. Therefore, it would be more realistic to replace S_N^2 and S_P^2 by $\langle F_N^2 \rangle$ and $\langle F_P^2 \rangle$, respectively. The parameter k , for example, can then be redefined as

$$k = 2\langle F_P^2 \rangle^{1/2} / \langle F_N^2 \rangle^{1/2}.$$

The poorly defined atoms scatter significantly at low resolution whereas their contribution to scattering is likely to taper off with increasing Bragg angle. The effect of this variation in scattering power can be accounted for in a simple manner by the following procedure. The parameters m and n or k can be evaluated in different convenient ranges of Bragg angle, and the parameters appropriate to each range can then be used for constructing the Fourier coefficients for reflections belonging to that range.

The formulae usually employed for the calculation of shifts in positional parameters from conventional difference Fourier maps (Stout & Jensen, 1968) are derived with the tacit assumption that the negative peaks at the incorrect atomic positions and the positive peaks at the correct atomic positions (which combine to give a density gradient) have equal strengths. This assumption is true only when all the atoms are included in the calculation of F_P' . In such a situation, the peaks at the incorrect and correct positions are expected to have strengths of $-f_{pj}/2$ and $f_{pj}/2$, respectively, when all the reflections used are acentric. But the two sets of peaks have reduced and unequal strengths when only part of the structure is used for phase-angle calculations, which is often the case in proteins. However, it follows from the expressions derived earlier that a modified map with

$$(k_1 F_N - k_2 F_P') \exp(i\alpha_P'),$$

where $k_1 = S_N/S_P$ and $k_2 = 1/2 + S_N^2/2S_P^2$, as coefficients could be expected to have peak strengths of $-f_{pj}/2$ and $f_{pj}/2$ at the incorrect and the correct positions, respectively, when all the reflections are acentric. Therefore, in principle, one is justified in using the usual formulae, referred to earlier, for calculation of shifts in positional parameters only when the map is computed with the above coefficients. It may, however, be noted that k_1 and k_2 have nearly equal values even when the ratio between S_N and S_P departs substantially from unity. Therefore, in most cases, it is sufficient to multiply the parameter shifts obtained from conventional difference Fourier maps by

a factor of $2S_N/S_P$ instead of two as is normally done when all reflections are acentric.

It must be emphasized that even 'mixed' syntheses cannot completely cope with systematic errors. When errors in the known atomic positions are randomly distributed, they are very effective. When the errors are non-random, these syntheses still tend to give correct peak strengths, but there is no inherent mechanism for correcting positional errors resulting from the positional errors in the input atoms.

Conclusion

There exists a definite, though complicated, relationship between the errors in the positions of the known atoms on the one hand and the errors in the new positions of the known atoms as well as the positions of the unknown atoms derived from Fourier syntheses on the other. When the errors in the input positions are random, these are easily and automatically taken care of by the syntheses themselves if the parameters used in the Fourier coefficients are chosen judiciously. However, when they are systematic, as is often the case in proteins, they are not taken care of automatically by different syntheses.

The author wishes to thank Professor Dorothy Hodgkin for her interest in the work. His thanks are also due to the referee who made some useful suggestions. Financial assistance from the University Grants Commission, India, is acknowledged.

References

- BODE, W. & SCHWAGER, P. (1975). *J. Mol. Biol.* **98**, 693–717.
 DODSON, E. J., DODSON, G. G., HODGKIN, D. C., ISAACS, N. & VIJAYAN, M. (1978). Unpublished results.
 DODSON, E. & VIJAYAN, M. (1971). *Acta Cryst.* **B27**, 2402–2411.
 FREER, S. T., ALDEN, R. A., CARTER, C. W. & KRAUT, J. (1975). *J. Biol. Chem.* **250**, 46–54.
 LUZZATI, V. (1953). *Acta Cryst.* **6**, 142–152.
 RAMACHANDRAN, G. N. & SRINIVASAN, R. (1970). *Fourier Methods in Crystallography*. New York: John Wiley.
 STOUT, G. H. & JENSEN, L. H. (1968). *X-ray Structure Determination*. New York: Macmillan.
 WATENPAUGH, K. D., SIEKER, L. C., HERRIOT, J. R. & JENSEN, L. H. (1973). *Acta Cryst.* **B29**, 943–956.