Proc. Natl. Acad. Sci. USA
Vol. 74, No. 7, pp. 2835–2839, July 1977
Biochemistry

Method for obtaining a high resolution protein map starting from a low resolution map

(protein crystallography/phase extension technique)

R. C. AGARWAL AND N. W. ISAACS*

International Business Machines Corporation, T. J. Watson Research Center, Yorktown Heights, New York 10598

Communicated by Richard L. Garwin, May 9, 1977

ABSTRACT **A method is described for estimating the phases of high resolution single-crystal diffraction data from proteins, by using as a starting point a set of low resolution phases (about 3 Å) derived by multiple isomorphous replacement (or other) methods. The method consists in refining by least-squares the positions and thermal parameters of a set of dummy atoms placed in the initial low resolution electron density map, so as to minimize the discrepancy between the calculated scattering intensities and the scattering intensities observed in the high resolution data set. Phases calculated from these refined atomic positions are used to extend the resolution and to improve the quality of the electron density map. The success of the method depends on a new least-squares algorithm that has a radius of convergence of about 0.75 Å. This large radius of convergence, together with the severe restrictions placed on the initial positions of the dummy atoms by the requirement that they lie within limited regions of the isomorphous electron density map, and the constraint imposed by the polymeric nature of a polypeptide chain account for the success of the method. The method has been successfully used to phase the structure factors of 2-zinc insulin at a resolution of 2 Å and 1.5 Å, starting from a set of isomorphous phases at 3-Å resolution.**

The primary objective of a protein crystal structure analysis is to obtain an interpretable electron density map at the highest possible degree of resolution. Because x-ray diffraction measurements give only the magnitudes and not the phases of the Fourier spectrum of the electron density, the phases have to be approximated by some method in order to compute the electron density. These approximate phases are usually obtained by using multiple isomorphous replacement (m.i.r.) derivatives of the protein.

Several authors have suggested phase extension techniques whereby high resolution phases can be estimated and refined, starting from low resolution m.i.r. phases (1–5). One such technique due to Sayre (6) has been shown to be effective by Cutfield *et al.* (7). Frequently, these phase extension computations are expensive, and in most cases a significant number of m.i.r. phases are required to start the phase extension process. In this paper, we propose a new technique by which an interpretable protein map can be obtained starting from low resolution m.i.r. phases (about 3 Å). The method has been applied to insulin and the results obtained are presented.

METHODS

Recently a new least-squares atomic parameter refinement technique has been developed (8). This refinement technique makes use of the fast Fourier transform analysis at all stages of the calculation. Compared with previous refinement techniques that require computation proportional to NM^2, the requirement for this new technique is computation proportional to $N\log N$, in which N is the number of reflections and M is the number of atoms ($N \gg M$). Another characteristic of this refinement technique, significant for our present purpose, is its large radius of convergence. Tests have shown that it is possible to successfully refine structures in which the initial root-mean-square displacement of the atoms from their true positions is about 0.75 Å. This large radius of convergence provides the basis of this technique that essentially consists of the following steps: (*i*) placing dummy atoms into the m.i.r.-phased low resolution map, (*ii*) refining positions and thermal parameters of these dummy atoms by the least-squares method, and (*iii*) using the calculated phases from the refined dummy atoms to obtain an electron density map of higher resolution.

Placing the Dummy Atoms. Let us assume that we have m.i.r. phases out to a resolution of 3 Å. The map computed with these phases will not show atomic resolution but, if the isomorphous phases are reasonably accurate, it will definitely show significant electron density in the region of space where there are groups of atoms. Provided that the electron density map is generally correct, we can expect that any point within the more concentrated region of density will be close to a true atomic position. We can then place atoms at positions throughout the density and, although these "dummy" atoms will bear no stereochemical resemblance to a protein structure, we can expect, because of the continuous nature of the electron density in a polypeptide chain, that a large number of them will lie close to a real atomic position and that all but a few will lie within the radius of convergence of the least-squares program mentioned above. In placing the dummy atoms, our only considerations need be the following: (*i*) that distances between pairs of atoms be of the order of interatomic distances, (*ii*) that no single atom have more than three close neighbors, and (*iii*) that, as far as possible, the number of electrons for the dummy atoms placed should roughly match the local electron density in each region of the map. For the purposes of the refinement, all the dummy atoms can be of the same type (with seven electrons), unless there are heavier atoms in the real structure that can be located. The positioning of the dummy atoms has been done by a computer program.

Refining the Dummy Structure. The coordinates and isotropic thermal parameters (Bs) of the dummy atoms are refined by the above least-squares refinement method with higher resolution structure factors as observations. Because of the large radius of convergence of this technique, most of the dummy atoms will move toward some actual protein atom position. This is particularly true for atoms having low B values, because for these atoms the corresponding electron density peaks are very high. The least-squares refinement technique is very effective in reducing the R factor (agreement factor or residual) to the range of 20–25%, even if the starting set of dummy atoms had an unrealistic geometry and accounted for only a part of the

Abbreviation: m.i.r., multiple isomorphous replacement.

* Present address: Department of Chemistry, University of York, Heslington, York, England.

structure. At this stage, the dummy atoms usually account for only those atoms that have low thermal parameters.

With a significant reduction in the R factor, calculated structure factors ($|F_{calc}|$) better match the observed structure factors ($|F_{obs}|$) and, consequently, the calculated phases obtained from the refined dummy atoms are likely to improve. As a result of the improved phasing, a map computed by using $|F_{obs}|$ and the calculated phases will have sharper peaks and the features of the structure will be better resolved. At this stage, the crystallographer can do one of several things as follows: (*i*) attempt to interpret the map as a protein structure, imposing the stereochemical properties of proteins, (*ii*) repeat the process of dummy structure building and refinement till he obtains a more easily interpreted map, or (*iii*) refine the calculated phases by using a phase refinement technique (1–3). The protein model obtained from the map would then be used as a starting point for the refinement of the atomic positions to best represent all observed diffraction intensities.

In interpreting such a map, mistakes will inevitably be made in constructing a protein model. But this is not a serious problem if a large portion (say 80–90%) of the good parts (i.e., regions with low thermal parameters, such as helical segments) of the structure are correctly interpreted. Errors are most likely in the interpretation of long side chains and ill-defined main chain segments. If the least-squares technique is used for the refinement of this model, the mistakes in interpretation can be corrected during the refinement process. Incorrectly placed atoms tend to refine to very large B values (>50 Å^2) and as the refinement progresses and the R factor decreases, missing atoms (these are usually the side chain atoms with larger Bs) become evident in electron density difference maps.

EXPERIENCE WITH INSULIN

Our choice of insulin for the structure on which to test the technique was governed by the fact that we have extensively refined the insulin structure to an R factor of 11.3% (for 11,889 diffraction intensities that satisfy the condition that $0.55 < |F_{obs}|/|F_{calc}| < 1.8$) at 1.5-Å resolution and have compared the map obtained by the present technique with the map of a refined insulin structure.

The 2-zinc insulin data used in this example was in space group R_3 with cell parameters $a = b = 82.5$ Å and $c = 34.0$ Å. The asymmetric unit (1/3, 1/3, 1) had 808 protein atoms including 2 zinc and 12 sulphur atoms, and about 280 water molecules. We took as a starting point a 3-Å m.i.r.-phased map of insulin that was calculated with the same isomorphous phases as were used to calculate the original 2.8-Å map (9). For the purpose of computing the electron density maps, the value of F(0,0,0) was obtained from the content of the unit cell. The map was calculated on a grid spacing of approximately 0.7 Å. The ratio of the peak density (corresponding to zinc atoms) to the average density was only 5.83. In the asymmetric unit, 766 atoms were placed by using a program that positioned atoms on the grid starting from the largest peaks. No atoms were placed at grid positions with a density less than twice the average and the atoms were separated by at least 1.15 Å for larger peaks and 1.4 Å for smaller peaks. The two zinc atoms on the *c* axis were labeled as such, whereas all other atoms were labeled as nitrogens. Thermal parameters were assigned according to the corresponding peak density and ranged from 15 to 25 Å^2.

We refined the positions and the thermal parameters of these dummy atoms by using the least-squares refinement procedure, initially by using 2.25-Å data ($|F_{obs}|$) and then gradually by increasing the resolution to a reciprocal lattice spacing corresponding to 2 Å. At the start of the refinement, low resolution reflections were given high weights and high resolution reflections were given low weights. As the refinement progressed, the weighting scheme was progressively relaxed to give unit weights. The initial R factor for the 2.25-Å data was about 36%. After six cycles of coordinate refinement and two cycles of thermal parameter refinement, the R factor for 2-Å data was reduced to about 22%. The computer time required for each cycle of refinement on an IBM 370/168 was approximately 90 sec.

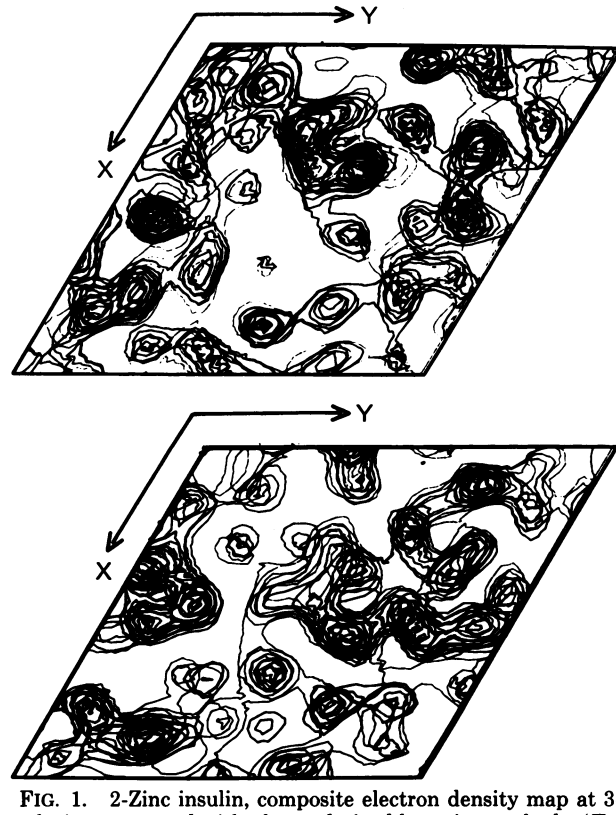

FIG. 1. 2-Zinc insulin, composite electron density map at 3-Å resolution computed with phases derived by m.i.r. methods. (*Top*) Sections $z = -3/48$–$0/48$. (*Bottom*) Sections 1/48–4/48.

We then calculated a 2-Å map by using those 3-Å m.i.r. phases that had figures of merit [the figure of merit of an m.i.r. phase is defined as $\cos(\epsilon)$, in which ϵ is the expected error in the phase] greater than 0.8 and the phases calculated from the refined dummy structure for the remaining data, with the $|F_{obs}|$ weighted according to Sim's procedure (10). For this map, we used a somewhat finer grid spacing of approximately 0.57 Å. The ratio of the peak density to the average density was 11.02, a considerable improvement over the 3-Å map. The dummy model-building procedure was repeated on this map with an added constraint that for each atom placed there be no more than two neighboring atoms within approximately 1.75 Å with electron density peaks greater than its own. This constraint does not guarantee that the number of neighboring atoms will be less than two or three but it helps in reducing them. Of the 780 dummy atoms placed in the asymmetric unit, two atoms on the c axis were labeled zinc as before, 12 other large peaks were labeled sulphur, and the rest were labeled oxygen or carbon depending on their peaks. Similarly, B values of 10–25 were assigned based on their peak heights. (As we shall see, except for the zinc atoms, all other dummy atoms should probably have been labeled as being of the same type. For the protein work they could all be labeled nitrogens.)

These 780 atoms were refined initially with 2-Å data and

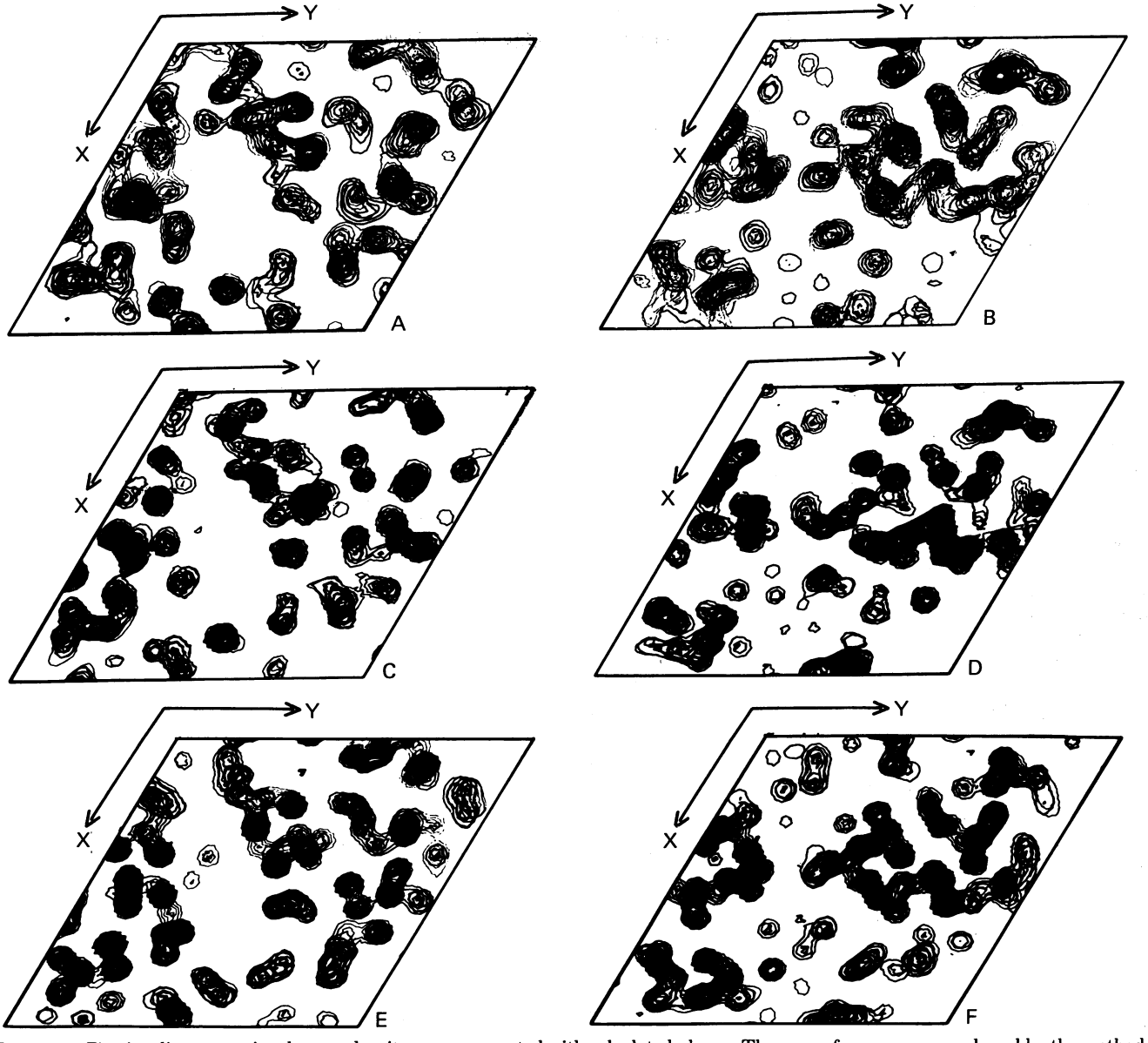

FIG. 2. 2-Zinc insulin, composite electron density maps computed with calculated phases. The upper four maps were phased by the method described in this paper: (*A*) and (*B*) at 2 Å, (*C*) and (*D*) at 1.5 Å. For comparison, the bottom pair (*E*) and (*F*) were computed with phases calculated from the least-squares refined structure at 1.5 Å. The maps in the left column show sections $z = -3/72$–$1/72$, and those in the right column show sections $z = 2/72$–$6/72$. These sections are roughly equivalent to those shown in Fig. 1.

toward the end with 1.7-Å data. A weighting scheme as described above was used. The initial R factor for 2-Å data was about 30% and, after six cycles of coordinate refinement and two cycles of thermal parameter refinement, it dropped to about 21% for 1.7-Å data.

A 1.7-Å map was computed by using these calculated phases and observed diffraction intensities with all reflections given unit weights. For this map, the grid spacing was about 0.5 Å and the ratio of the peak density to the average density had further increased to 29.5. The dummy model-building program used on this map was same as before but the minimum distance between dummy atoms was increased to 1.2 Å. Of the 831 dummy atoms placed in the asymmetric unit, 2 were labeled as zinc; 12 were labeled sulphur; and the rest were labeled oxygen, nitrogen, or carbon depending on their peaks. Similarly, B values ranging from 10 to 25 were assigned. The parameters for these 831 atoms were refined initially, by using only 1.7-Å data and towards the end by using all the data (1.5 Å). The same weighting scheme was used. The initial R factor for 1.7-Å data was about 30%, and seven cycles of coordinate refinement and two cycles of thermal parameter refinement reduced it to about 20% for the 1.5-Å data. The phases calculated from this refined dummy structure were used to compute a 1.5-Å map. In this map, the ratio of the peak density to the average density was 28.93, indicating no appreciable change from the last map. For comparison, an electron density map calculated by using the phases from the refined insulin structure had a ratio of the peak density to the average density of 38.01.

The method does not essentially require high resolution data, though for meaningful refinement the number of observations used should be sufficient to overdetermine the problem. This effectively requires at least 2-Å data. In a separate experiment, we extended the phase set from the original 3-Å m.i.r. set to 2 Å (using only the 2-Å diffraction intensities). The procedure used was essentially that described above. In the 3-Å map, 826 dummy atoms were placed and all were labeled as carbon ex-

cept for the two zinc atoms. Ten cycles of coordinate refinement and four of thermal parameters reduced the R factor from 40 to 17% for the 2-Å data. The map computed with these calculated phases and unit weights gave a ratio of 22.8 for the peak density to the average density.

Comparison of the Maps. The initial 3-Å map, 2-Å and 1.5-Å phase-extended maps, and the 1.5-Å map calculated with the phases derived from the refined structure of insulin (R = 11.3%) are shown in Figs. 1 and 2. These maps show the electron density in the hexagonal cell viewed down the *c* axis. Each set of stacked sections covers a depth of about 2.4 Å in *c* and 33% of the cell in both *a* and *b* directions. It is obvious that the phase extension to both 2-Å (Fig. 2 *A* and *B*) and 1.5 Å (Fig. 2 *C* and *D*) resolution has produced maps that show a considerable increase in clarity, while retaining the essential features of the isomorphous (or even the refined structure) map. The success of this method can be judged by the interpretability of the final electron density map, and we feel that the quality of these maps allows a sufficiently correct interpretation to provide a set of refinable coordinates of protein atoms. In both of the phase-extended maps, regions of structure (such as helical segments), in which the thermal parameters are low, are well defined; whereas regions with high thermal parameters are poorly defined or even absent. In the 2-Å map (Fig. 2 *A* and *B*), the main chain is reasonably well defined and generally continuous. On the other hand, the 1.5-Å extended map (Fig. 2 *C* and *D*) shows breaks in the main chain in unexpected places, such as in the helical region. This feature may well be due to an oversharpening effect caused by an incorrect assignment of the atom type in the dummy structure. For the 2-Å extension, all atoms except the zincs were assigned as carbons, but for the 1.5-Å extension these atoms were assigned as a mixture of C, N, O, and S; and, of the 12 atoms designated as S, only 2 were in fact close to true S positions. As mentioned before, we feel that except for the zincs, all other dummy atoms should have been given the same label (probably nitrogens). In placing dummy atoms at density peaks, our dummy atom placement program did not attempt to match the number of electrons for the dummy atoms placed with the local electron density (constraint *iii*). This may also in part explain the break in the helical region of Fig. 2*D*. A more sophisticated program should make use of the above constraint which will help in correct refinement of the dummy atoms.

Phase Comparison. We compared various sets of phases obtained by different techniques. As stated before, we have extensively refined the insulin structure to an R factor of 11.3% for 1.5-Å data. The calculated phases from this refined structure can be assumed to be a good approximation of the actual phases. One of us (N.W.I.) used Sayre's phase extension and refinement technique (1) to improve and extend 1.9-Å m.i.r. phases to 1.5-Å resolution (7) and this phase set was available. We compared phases from the refined structure (assumed to be the correct phases) ϕ_c with 1.9-Å m.i.r. phases (ϕ_i), with 1.5-Å phases obtained by Sayre's refinement technique (ϕ_s), and with 1.5-Å calculated phases from the refined dummy structure (ϕ_d). The results of this comparison are summarized in Table 1. This table gives the average phase error ($\overline{|\Delta\phi|}$) between ϕ_c and various other phase sets.

One interesting feature of Table 1 is that for ϕ_d, the average phase error decreases very sharply with the magnitude of the reflections, much more sharply than for the other two phase sets. This is expected because ϕ_d is obtained by a least-squares refinement technique that tends to match strong $|F_{obs}|$ and $|F_{calc}|$ closely. For strong reflections, the R factor is small and, consequently, it is expected that the phase error should also be small. On the other hand, for weak reflections the match between $|F_{obs}|$ and $|F_{calc}|$ is not very good and, consequently, the phase error is also large. In calculating a density map, the correct phasing of strong reflections is very important and therefore, although the average error for ϕ_d is large, the map phased with ϕ_d closely resembles the one phased with ϕ_c.

Table 1. Average phase error ($\overline{|\Delta\phi|}$) in degrees between ϕ_c and various other phase sets

Phases compared	1.9-Å m.i.r. phases ϕ_i	1.5-Å phases Sayre's technique (1) ϕ_s	1.5-Å phases dummy structure refinement ϕ_d
All 1.5-Å phases ~13,400 reflections			70
All Sayre phases ~10,000 reflections		55	68
All 1.9-Å phases ~6,300 reflections	60	52	65
~2,000 strongest 1.9-Å reflections	48	38	47
~1,000 strongest 1.9-Å reflections	43	33	39
~500 strongest 1.9-Å reflections	37	32	32
~250 strongest 1.9-Å reflections	34	30	27

From the table, it is clear that ϕ_d values are at least as good as ϕ_i values. Actually, they are considerably better for strong reflections. It is also clear that ϕ_d values are not as good as ϕ_s values, except for the 500 strongest reflections. But this is expected, as the starting point for ϕ_s was the 1.9-Å m.i.r. phase set whereas that for ϕ_d was the 3-Å set.

CONCLUSION

The method presented allows for the extension of phases from a set of low resolution diffraction data to a set of high resolution data at comparatively small cost. We emphasize that the protein model fitted into the phase-extended map should be used only as a starting point for further crystallographic refinement. Our experience and that of others (G. G. Dodson, personal communication) indicate that difference electron density maps are able to show gross errors in interpretation at a very early stage and, as a method of procedure, we would recommend calculating a number of difference Fouriers on the starting model to correct the gross errors, before beginning the structure refinement by least-squares.

In this paper, we have demonstrated an effective new tool for protein crystallography. This technique is by no means in its final form. There is a good deal of scope for improvements in various stages of the method, particularly in placing the dummy atoms.

We are grateful to Prof. D. C. Hodgkin and Dr. G. G. Dodson for allowing us to use the insulin data. The continued support and encouragement of Drs. R. L. Garwin, K. D. Hardman, and D. Sayre has been a stimulus for this work. N.W.I. was a recipient of an International Business Machines Corporation World Trade Research Fellowship.

The costs of publication of this article were defrayed in part by the payment of page charges from funds made available to support the

research which is the subject of the article. This article must therefore be hereby marked "advertisement" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

1. Sayre, D. (1972) *Acta Crystallogr. Sect. A* **28,** 210–212.
2. De Rango, C., Mauguen, Y. & Tsoucaris, G. (1975) *Acta Crystallogr. Sect. A* **31,** 227–233.
3. Hendrickson, W. A. (1973) *Trans. Am. Crystallogr. Assoc.* **9,** 61–83.
4. Hoppe, W., Gassman, J. & Zechmeister, K. (1969) in *Crystallographic Computing,* ed. Ahmed, F. R. (Munksgaard, Copenhagen), pp. 26–36.
5. Collins, D. M., Cotton, F. A., Hazen, E. E., Jr., Meyer, E. F., Jr. & Morimoto, C. N. (1975) *Science* **190,** 1047–1053.
6. Sayre, D. (1974) *Acta Crystallogr. Sect. A* **30,** 180–184.
7. Cutfield, J. F., Dodson, E. J., Dodson, G. G., Hodgkin, D. C., Isaacs, N. W., Sakabe, K. & Sakabe, N. (1975) *Acta Crystallogr. Sect. A* **31,** S21.
8. Agarwal, R. C. & Isaacs, N. W. (1976) *Abstr. Am. Crystallogr. Assoc.* **4,** 48.
9. Adams, M. J., Blundell, T. L., Dodson, E. J., Dodson, G. G., Vijayan, M., Baker, E. N., Harding, M. M., Hodgkin, D. C., Rimmer, B. & Sheat, S. (1969) *Nature* **224,** 491–495.
10. Sim, G. A. (1959) *Acta Crystallogr.* **12,** 813–815.