Acta Cryst. (1974). A30, 672

Determining Heavy-Atom Positions using Non-Crystallographic Symmetry

BY PATRICK ARGOS* AND MICHAEL G. ROSSMANN

Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, *U.S.A.*

(*Received* 24 *January* 1974; *accepted* 1 *May* 1974)

If there are a number of chemically identical molecular subunits in the asymmetric unit of the crystal cell, related by previously established non-crystallographic symmetry, it is probable that heavy atoms attached specifically to the subunits of the native molecule will show the same symmetry. The Patterson vectors between non-crystallographically equivalent heavy atoms can then be generated for arbitrary trial positions and compared with the actual Patterson synthesis. A search of all positions within the molecular subunit can thus establish the sites of larger heavy-atom substitutions. Once these have been determined, vectors between molecules can be computed and compared with the actual Patterson synthesis in order to establish the position of the molecular center if it is unknown. These methods have been demonstrated in the determination of the major heavy-atom sites, in the presence of non-crystallographic 222 symmetry, for the glyceraldehyde-3-phosphate dehydrogenase molecule.

Introduction

Evidence has accumulated that larger proteins are made up of identical or similar subunits which themselves may contain one or more dissimilar polypeptide chains (Matthews & Bernhard, 1973). A discussion of limitations to the size of protein subunits of viruses was first given by Crick & Watson (1956), while advantages for the aggregation of protein subunits into larger functional particles have been suggested by many authors (*cf.* Monod, Changeux & Jacob, 1963). Greater difficulty must be expected in the determination of heavy-atom sites when these have been multiplied by non-crystallographic symmetry. With the facility of the rotation function (Rossmann & Blow, 1962) in the determination of non-crystallographic symmetry axes, the present paper shows the possibility of solving these more complex difference Patterson syntheses systematically. Lobster glyceraldehyde-3-phosphate dehydrogenase (GAPDH; Buehner, Ford, Moras, Olsen & Rossmann, 1974) has been used here to exemplify methods in the solution of intricate difference Patterson syntheses.

There are a variety of methods used for the deconvolution of the Patterson function, such as Nordman's (1972) search procedure with known rigid structures, superposition techniques as Buerger's (1950, 1951) minimum function and vector verification analyses (Mighell & Jacobsen, 1963). The present discussion is an extension of these vector search methods to heavy-atom difference Patterson functions used in the determination of large protein molecules displaying non-crystallographic symmetry.

The structure of lobster D-glyceraldehyde-3-phosphate dehydrogenase has recently been solved at 3·0 Å resolution (Buehner, Ford, Moras, Olsen & Rossmann, 1974) by utilizing single isomorphous-replacement phases and electron density averaging over the four equivalent subunits. The holo-enzyme of lobster GAPDH crystallizes in the orthorhombic space group $P2_12_12_1$ with $a=149{\cdot}0$, $b=139{\cdot}1$, $c=80{\cdot}7$ Å, containing one tetramer (MW = 143000 Daltons) per asymmetric unit (Watson & Banaszak, 1964). The directions of the three mutually perpendicular non-crystallographic molecular twofold axes with respect to the crystal axes were determined by the rotation function (Rossmann *et al.*, 1972). The position of the molecular center could, however, not be determined with any degree of certainty from Patterson search methods.

The key to the GAPDH structure lay with the solution of the difference Patterson synthesis between the K_2HgI_4 derivative and the native compound at 7·0 Å resolution in terms of four major heavy-atom sites per molecule. This was done essentially by hand, in association with the knowledge of the non-crystallographic symmetry and a tentative molecular center. Single isomorphous phases, calculated from the initial K_2HgI_4 heavy-atom sites, and typical refinement and cross Fourier techniques were applied to the *p*-chloromercuriphenylsulfonate derivatives (PCMS), which eventually led to a complete determination of all major and minor heavy-atom sites.

This paper shows that the above solution could have been obtained readily by computerized systematic search procedures.

The search procedures

The methods discussed here depend on assuming a heavy-atom position within the molecular subunit.

* Present address: Department of Physics, Southern Illinois University, Edwardsville, Illinois 62025, U.S.A.

Specified self- (within the molecule) and cross- (between the molecules) vectors are calculated and compared with the actual difference Patterson synthesis. All reasonable heavy-atom positions within the non-crystallographic asymmetric unit are generated on a grid and tested in turn. A function of the test criterion is then plotted and the position of possible heavy atoms may be discerned.

Three distinct cases can be recognized:

(*a*) When there is no knowledge of the position of the molecular center (the point of intersection of the non-crystallographic symmetry axes representing the molecular point group) with respect to the crystallographic axes, the vectors between non-crystallographically related heavy-atom sites within each molecule can be calculated. Vectors between chemically different sites within the same molecule, or between sites in different molecules related by crystallographic symmetry, must then be neglected. This exploration should be made with respect to the molecular axes.

(*b*) When the molecular center position is known, then the self-vectors as well as the cross-vectors can be considered. However, vectors between chemically different sites will be neglected. This exploration should also be with respect to the molecular axes.

(*c*) When the molecular center is not known but the major heavy-atom sites have been previously determined, then a search for the molecular center can be made with respect to the crystal axes. In this case the cross-vectors between different molecules will be considered in conjunction with the unchanging set of self-vectors. This last procedure is related to the techniques of Nordman (1972) and Mighell & Jacobsen (1963).

In all three of the above procedures the Patterson point vectors are referred to the actual Patterson map. The criterion for fit of the generated vector constellations was taken to be the sum of the Patterson densities at the nearest grid point associated with each vector. While many other criteria might have been considered (Nordman, 1972), the sum function gave satisfactory results. The asymmetric unit of the Patterson synthesis was stored. In order to avoid artifacts due to vectors lying on the origin peak of the Patterson map, a special procedure was devised for any vector within a stated radius, r_p, around the origin. If less than a stated fraction, f_p, of the vectors were within this critical radius, then their value was taken as a mean of all the remaining Patterson peak values; if more than the desired number of vectors were near the origin, then this test for heavy-atom positions was ignored. The latter case would, in general, relate to sites close to the molecular center, an unlikely chemical event. The former case may arise more frequently when heavy atoms are close to molecular symmetry axes.

Vector multiplicities

Multiplicity of vectors at a single Patterson position must also be taken into account. The Patterson density corresponding to a specified vector must be divided by the number of vectors coincident (or nearly coincident) at that position. No correction was, however, made for the absence of an $F^2(0,0,0)$ term in the Patterson map, which would put all peaks at least on a true relative scale. The actual multiplicity factor, m, was computed according to the method of Rossmann (1972) and is given by $m=1+\sum \exp(-a\Delta^2)$ where Δ is the distance from the selected vector to a neighboring vector, a is a suitable constant dependent on the width of a single peak, and the sum is taken over all sufficiently close vectors. Two types of multiplicities occur; (i) when vectors fall close to each other in general positions and (ii) when vectors fall close to symmetry elements in the Patterson map.

Considerable time may be saved by generating self-vectors only within one molecule. Self-vectors in the other molecules are related by crystallographic symmetry and hence are expressed by the Patterson symmetry. Care must be taken, however, in determining the correct multiplicity of all vectors when such time-saving procedures are used.

Tests were made omitting the computation of multiplicity factors between general coincidences for procedures (*b*) and (*c*). No significant change in features of the GAPDH search maps was evident.

Limits of asymmetric unit of search functions

In procedures (*a*) and (*b*) above, exploration of possible heavy-atom sites must be performed with respect to the molecular axes, P, Q, R (*cf.* Rossmann *et al.*, 1973). In most cases these can be defined as an orthogonal system along specified molecular symmetry axes. Exploration may then proceed in terms of either a Cartesian system in equal steps along such axes or in terms of a polar coordinate system referred to the P, Q, R system. In the case of GAPDH, where the molecular symmetry is 222, exploration in planes of constant R was found most suitable. For an icosahedral virus limits of the asymmetric unit are set most easily in terms of polar angles. In either case an arbitrary limiting molecular radius $D=\sqrt{P^2+Q^2+R^2}$ must be set, outside which no heavy atom can possibly be attached to the protein.

As the Patterson synthesis cannot distinguish between enantiomorphic solutions, the asymmetric unit of exploration is determined by the point group of the molecule plus a center of symmetry. Thus in GAPDH, with a molecular point group of 222, the asymmetric unit to be explored is determined by point group *mmm*. Hence the volume of the asymmetric unit was arbitrarily chosen as $P\geq 0$, $Q\leq 0$, $R\geq 0$, $D\leq 42$ Å.

In procedure (*c*) above the position of the molecular center must be placed with respect to the crystallographic axes. Once again it is not possible to distinguish between enantiomorphic solutions. Thus the asymmetric unit of the search function is the same as the

asymmetric unit of the Patterson synthesis, not that of the real cell. In the case of GAPDH, with a space group of $P2_12_12_1$, the Patterson space group is *Pmmm* and hence the molecular center search can be limited to within $0\leq x\leq\frac{1}{2}$, $0\leq y\leq\frac{1}{2}$, $0\leq z\leq\frac{1}{2}$. Time may be saved by omitting the calculation of the search function at those points where the molecules approach unrealistically close to each other. A cut-off, t, can be chosen giving the closest permitted approach between molecular centers.

Computation of Patterson vectors

Consider a position vector $\mathbf{P}$ in the molecular P, Q, R space. Its n non-crystallographically related positions can then be generated in terms of the n molecular rotation matrices M_n such that

$$\mathbf{P}_n=[M_n]\mathbf{P}\,. \tag{1}$$

For instance, a molecule with 222 symmetry, as is the case for GAPDH, would have

$$M_1=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},\quad M_2=\begin{pmatrix}-1&0&0\\0&-1&0\\0&0&1\end{pmatrix},$$

$$M_3=\begin{pmatrix}1&0&0\\0&-1&0\\0&0&-1\end{pmatrix},\quad M_4=\begin{pmatrix}-1&0&0\\0&1&0\\0&0&-1\end{pmatrix}.$$

When a non-crystallographic axis is not coincident with a chosen molecular axis a more general expression for $[M]$ can be used (*cf.* Rossmann & Blow, 1962).

The vectors $\mathbf{P}_n$ must now be referred to the crystal axes by means of the transformation U. Thus

$$\mathbf{X}_n=[U]\mathbf{P}_n\,. \tag{2}$$

It can be shown that

$$U=\begin{pmatrix}u_1&u_2&u_3\\v_1&v_2&v_3\\w_1&w_2&w_3\end{pmatrix} \tag{3}$$

where the direction cosines of the orthogonal molecular P, Q, R axes are $(u_1v_1w_1)$, $(u_2v_2w_2)$, $(u_3v_3w_3)$ referred to the orthogonal crystal axes (X,Y,Z). Values for these direction cosines can be obtained from a determination of the orientation of non-crystallographic symmetry axes with the rotation function (Rossmann & Blow, 1962). Values for such direction cosines were given for GAPDH by Rossmann, Ford, Watson & Banaszak (1972) and examined carefully for orthogonality by Buehner *et al.* (1974).

The vectors $\mathbf{X}_n$ are, however, more conveniently expressed in terms of fractional unit-cell dimensions along the principal crystal axes by the transformation

$$\mathbf{x}_n=[\alpha]\mathbf{X}_n+\mathbf{S} \tag{4}$$

where $\mathbf{x}_n$ and $\mathbf{S}$ are the heavy-atom and molecular-center positions respectively in crystal space. The matrix elements for $[\alpha]$ have been given by Rossmann & Blow (1962). Combining (1), (2) and (4), it is clear that

$$\mathbf{x}_n=[\alpha]\,[U]\,[M_n]\mathbf{P}+\mathbf{S}\,. \tag{5}$$

Self-vectors within a specific molecule, such as between the nth and mth heavy atom, are independent of $\mathbf{S}$ and are given by

$$\mathbf{x}_n-\mathbf{x}_m=[\alpha]\,[U]\{[M_n]-[M_m]\}\mathbf{P}\,. \tag{6}$$

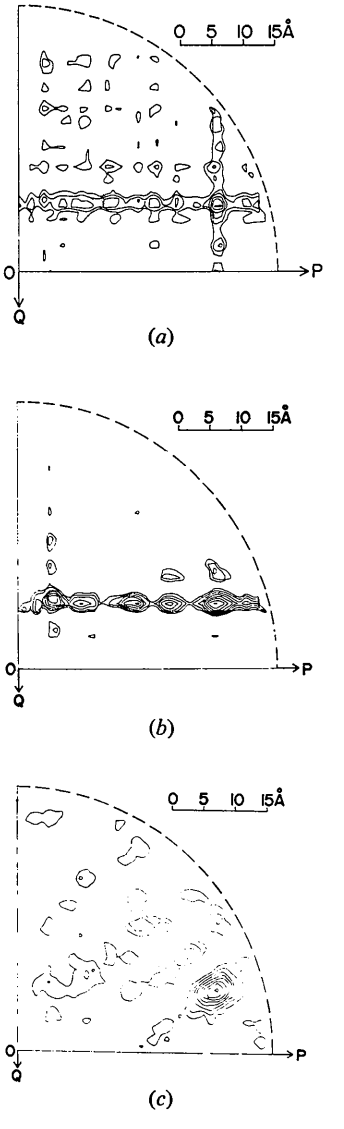


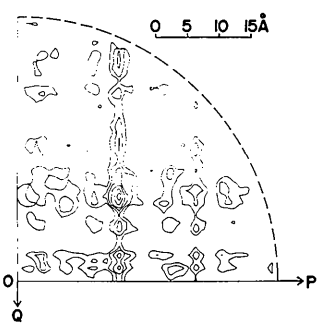
Fig. 1. Sections through the *highest* peak value, nearest the major heavy-atom site B in the PCMS derivative (Table 2). Contours at arbitrary equal intervals. (*a*) Section $R=7{\cdot}0$ Å, resolution 5 Å, for procedure (*a*). (*b*) Section $R=8{\cdot}0$ Å, resolution 8 Å, for procedure (*a*). (*c*) Section $R=8{\cdot}0$ Å, resolution 8 Å, for procedure (*b*).

Cross-vectors between molecules, related to each other by crystallographic symmetry, may be calculated by multiplying the vectors $\mathbf{x}_n$ given by (5) with the crystallographic symmetry operators.

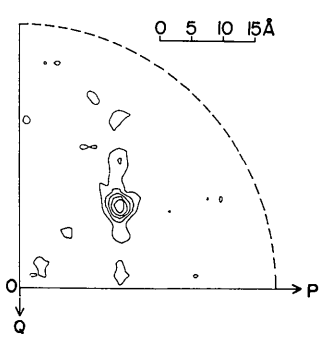
The differences $\{[M_n]-[M_m]\}$ in (6) can take on rather special values corresponding to Harker-type relations within the molecule causing density streaks parallel to non-crystallographic axes in the search function. For instance in GAPDH

$$M_1-M_2=\begin{pmatrix}2&0&0\\0&2&0\\0&0&0\end{pmatrix}.$$

Thus the vector between the heavy atoms in the first and second subunit is independent of R. If then the correct P and Q values are chosen for a given heavy-atom site, significant Patterson densities will be added to the test criterion for all values of R. Hence a streak in the test function, parallel to the R axis, passing through the heavy-atom site, will be found. There will be similar streaks parallel to each of the three non-crystallographic axes, all of which will intersect at the correct heavy-atom site. Similar streaks will occur in procedure (*c*) parallel to the crystallographic axes as a result of Harker relationships between the molecules.



(*a*)



(*b*)

Fig. 2. Section through the highest peak value and close to the major heavy-atom site A in the K_2HgI_4 derivative. Contours at arbitrary equal intervals. (*a*) Section $R=8{\cdot}0$ Å, resolution 5 Å, for procedure (*a*). (*b*) Section $R=8{\cdot}0$ Å, resolution 5 Å, for procedure (*b*).

Results and discussion

The difference Patterson functions for the K_2HgI_4 and PCMS derivatives of GAPDH were computed on an approximately 2 Å grid using 8 Å and also 5 Å resolution data. The U-matrix (4) was based on the rotation function results (Rossmann *et al.*, 1972) rather than the probably slightly more accurate determination based on refined heavy-atom positions (Buehner *et al.*, 1974). Thus no prior knowledge of the heavy-atom positions was permitted to prejudice the calculations. The radius r_p around the Patterson origin was taken as 9 Å, and the fraction f_p of vectors allowed within this radius was chosen as $\frac{1}{2}$. The constant, a, used in the multiplicity calculation was chosen as $[-1n\,(0{\cdot}5)/c^2]$ where c, the half width of a peak, was taken as 5 Å (Rossmann, 1972). The maximum molecular radius D was set as 42 Å. The search grid in P, Q, R space was at 1 Å intervals for procedures (*a*) and (*b*), while the molecular center search of procedure (*c*) used an approximately 3 Å grid. A cut-off $t=50$ Å was also used in procedure (*c*).

The K_2HgI_4 derivative has one major site, A, and one minor site, B, per molecular subunit. The PCMS derivative has one major site at B and one minor site at C per subunit. In both cases the minor site has about one quarter of the occupancy of the major site. Refined heavy-atom parameters as well as the molecular center are given in Table 1 with respect to a single molecule (Buehner *et al.*, 1974).

Sections close to the major heavy-atom sites for procedures (*a*) and (*b*) on the PCMS and K_2HgI_4 derivatives are shown in Figs. 1 and 2 respectively. Searches using 5 and 8 Å PCMS resolution data are compared for procedure (*a*) in Fig. 1(*a*) and (*b*). Although the 8 Å search has its largest peak at the major B site, it is no longer as clear as the 5 Å search.

In procedure (*b*), where a knowledge of the molecular center is assumed, results even at 8 Å resolution are overpowering in the determination of the major sites

Table 1. *Heavy-atom sites after least-squares refinement (Buehner et al., 1974)*

		Averaged coordinates in (Å)			Averaged occupancy (electrons)	Averaged shape factor (Å²)
Derivative	Site	P	Q	R	Z	B
PCMS	B	32·5	−11·1	−7·2	60	11
PCMS	C	20·4	−25·3	13·3	31	39
K_2HgI_4	A	16·7	−13·9	8·1	101	57
K_2HgI_4	B	34·6	−10·3	−6·8	23	40

Molecular center $x=0{\cdot}4196$ $y=0{\cdot}1616$ $z=0{\cdot}1209$

Table 2. *Relative peak height distribution*

Derivative	Resolution (Å)	Procedure	Full multiplicities used	Highest peak	Next-highest peak within streaks	Next-highest peak not in streaks	Peak value at minor site	Figure showing section of this function	Major heavy-atom site (Å) P	Q	R	Molecular-center peak in fractional cell lengths x	y	z
PCMS	5	(*a*)	Yes	927	579	726	432	1(*a*)	32·5	−11·2	±7·1			
PCMS	8	(*a*)	Yes	761	678	524	241	1(*b*)	32·3	−10·9	±8·1			
PCMS	8	(*b*)	No	3603	1291	1600	937	1(*c*)	32·4	−11·4	±7·6	0·419	0·162	0·120
PCMS	8	(*c*)	No	4028	3206	2093	—	3(*a*)						
K_2HgI_4	5	(*a*)	Yes	1041	685	672	416	2(*a*)	16·9	−14·0	±7·8			
K_2HgI_4	5	(*b*)	Yes	6338	2128	1523	1204	2(*b*)	16·9	−13·9	±8·2			
K_2HgI_4	5	(*c*)	No	5994	3399	3300	—	3(*b*)				0·422	0·160	0·121

of both PCMS [Fig. 1(*c*)] and K_2HgI_4 [Fig. 2(*b*)]. Similarly, Fig. 3(*a*) and (*b*) shows good results for the determination of the molecular center both at 8 Å and 5 Å resolution for PCMS and K_2HgI_4 respectively.

In all these results the peak centers fell on the average to within 0·25 Å of the refined heavy-atom positions (Table 2). The 5 Å resolution determinations were

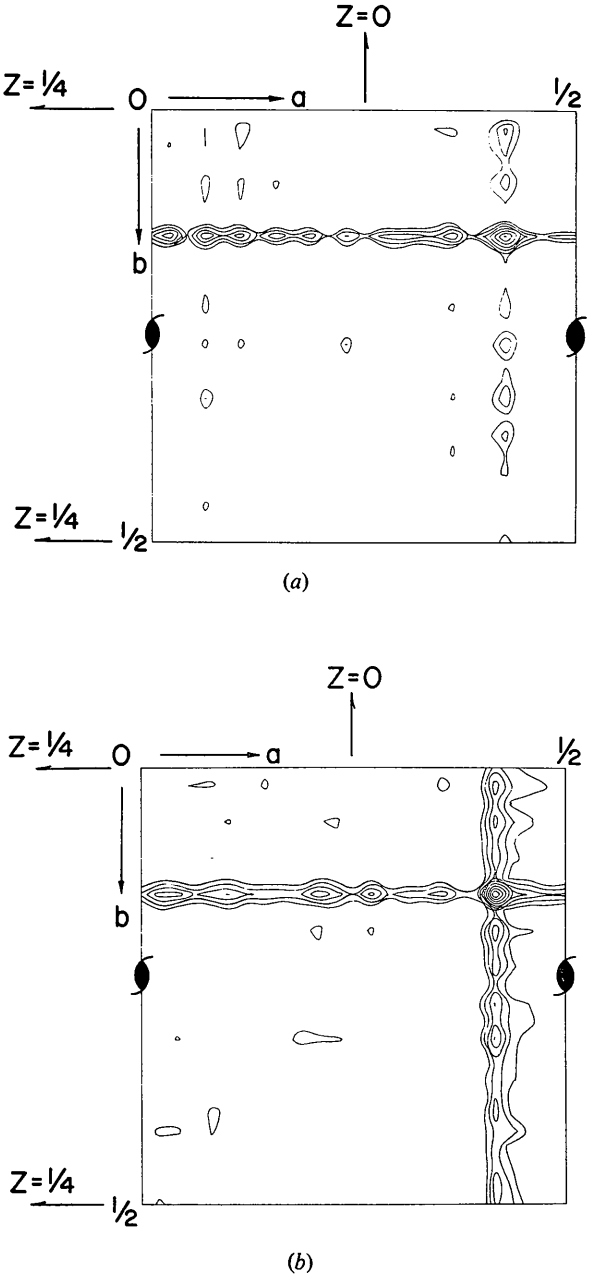


Fig. 3. Section at $z=0{\cdot}120$ through the highest peak value and close to the molecular center in procedure (*c*) search functions for: (*a*) PCMS at 8 Å resolution, (*b*) K_2HgI_4 at 5 Å resolution. Contours at arbitrary equal intervals.

more accurate than the 8 Å determinations, and procedure (*b*) more accurate than procedure (*a*). It should, however, be remarked that it is important [particularly in procedure (*a*)] to choose a sufficiently fine search grid to establish correctly the peak position. In any event the accuracy of these results is more than sufficient for initiation of conventional least-squares refinements.

The functions shown in Figs. 1, 2 and 3 are summarized in Table 2. The peak heights at the major sites, and along the streaks parallel to non-crystallographic axes in procedures (*a*) and (*b*) or parallel to crystallographic symmetry axes in procedure (*c*) are given. They are compared with the highest peak not on a streak and with the value of the functions at the minor sites. It is apparent that the search procedures cannot determine the position of minor sites as small as those found in these GAPDH derivatives.

The highest peak value not appearing in the heavy-atom streaks in the PCMS procedure (*a*) map at 5 Å resolution was tried for a molecular-center search. The procedure (*c*) map indicated six peaks above 2000, although none were more than 2500, which must be compared to a correct peak of 5621. Clearly then, an incorrect heavy-atom position will not lead to an obvious molecular center. Thus, several heavy-atom positions determined from a procedure (*a*) map can be taken as correct if they yield a consistent, obvious molecular center peak.

We are grateful for many helpful discussions with Drs Dino Moras, Geoffrey Ford, Marvin Hackert and John Johnson, to Mrs Kathy Jewell for the help in preparation of the diagrams and to Mrs Sharon Wilder for assistance in the preparation of the manuscript. The work was supported by the National Science Foundation (grant #GB 29596X) and the National Institute of Health (grant #AI 11219).

References

BUEHNER, M., FORD, G. C., MORAS, D., OLSEN, K. W. & ROSSMANN, M. G. (1974). *J. Mol. Biol.* In the press.
BUERGER, M. J. (1950). *Proc. Natl. Acad. Sci.* **36**, 738–742.
BUERGER, M. J. (1951). *Acta Cryst.* **4**, 531–544.
CRICK, F. H. C. & WATSON, J. R. (1956). *Nature, Lond.* **177**, 473–475.
MATTHEWS, B. W. & BERNHARD, S. A. (1973). *Ann. Rev. Biophys. Bioeng.* **2**, 257–317.
MIGHELL, A. D. & JACOBSON, R. A. (1963). *Acta Cryst.* **16**, 443–445.
MONOD, J., CHANGEUX, J. P. & JACOB, F. (1963). *J. Mol. Biol.* **6**, 306–329.
NORDMAN, C. E. (1972). *Acta Cryst.* A**28**, 134–143.
ROSSMANN, M. G. (1972). *J. Mol. Biol.* **64**, 246–249.
ROSSMANN, M. G., ADAMS, M. J., BUEHNER, M., FORD, G. C., HACKERT, M. L., LILJAS, A., RAO. S. T., BANASZAK, L. J., HILL, E., TSERNOGLOU, D. & WEBB, L. (1973). *J. Mol. Biol.* **76**, 533–537.
ROSSMANN, M. G. & BLOW, D. M. (1962). *Acta Cryst.* **15**, 24–31.
ROSSMANN, M. G., FORD, G. C., WATSON, H. C. & BANASZAK, L. J. (1972). *J. Mol. Biol.* **64**, 237–249.
WATSON, H. C. & BANASZAK, L. J. (1964). *Nature, Lond.* **204**, 918–920.