Acta Cryst. (1976). A**32**, 320

Crystallographic Relationship between Human and Hen-Egg Lysozymes. I. Methods for the Establishment of Molecular Orientational and Positional Parameters

BY P. E. NIXON* AND A. C. T. NORTH†

Laboratory of Molecular Biophysics, South Parks Road, Oxford OX1 3PS, England

(*Received* 10 *September* 1975; *accepted* 22 *September* 1975)

In order to determine what values of rotational and translational parameters will best allow a large fragment of a protein molecule to explain the diffraction from another protein, a correlation function between F_o and F_c must be defined. Various such correlation functions are discussed; the scalar product $\sum F_o^2(\mathbf{h})F_c^2(\mathbf{h})$ is recommended for establishing the rotation, but the residual is better for the translation. It is shown that the calculation of the latter is not computationally impracticable. A large fragment of the hen-egg-white lysozyme molecule was used as a model for human lysozyme; rotational and translational searches were successful, and the unrefined residual was 49 %. Wilson's distribution functions are shown to apply to lysozyme surprisingly well. One consequence of this is that the results of Parthasarathy & Parthasarathi [*Acta Cryst.* (1972). A**28**, 426–432] may be used to derive a value of the average error in the coordinates.

1. Introduction

The isomorphous replacement method has now resulted in the elucidation of the structures of a number of proteins, but at the cost of a great deal of labour. However, it is becoming increasingly clear that families of proteins exist, members of which show great similarity in all or part of their structures. If such a similarity is expected it may be possible to use the structure of the known protein to derive phase information for the unknown protein by the equation

$$\mathbf{F}_c(\mathbf{h}) = \sum_j f_j \exp(2\pi i \mathbf{h} \cdot \mathbf{x}_j), \qquad (1)$$

where the $\mathbf{x}_j$ are the coordinates and $\mathbf{f}_j$ the scattering factors of the known atoms. When the known and unknown crystal structures are not isomorphous, it is not possible to use (1) until $\mathbf{h}$ and the $\mathbf{x}_j$ refer to the same coordinate system: in general one has no prior knowledge of how the structure proposed as a model will be packed into the unit cell of the unknown structure.

The only general way to determine which orientation and position of the model best fits the X-ray diffraction intensities from the unknown is to try all the different orientations and positions, and choose the one where the F_c best fit the F_o. If this is to be an automatic procedure a correlation function between the sets of observed and calculated structure factor amplitudes must be defined: one hopes that only one set of values of the rotational and translational parameters will give a good value of this function (*i.e.* a maximum or minimum, as appropriate).

In general, unless there is more than one molecule per asymmetric unit, we have to find six molecular parameters, three rotational and three translational (except that there are only two translational parameters in monoclinic space groups and none in triclinic). The rotational parameters may be found initially from Patterson-type functions in which intramolecular vectors predominate. Once the orientations of the molecules are known, their positions relative to the crystallographic axes may be found from a search in which the three translational parameters are varied.

2. Correlation functions

Determination of the required parameters may be carried out directly by comparison of observed and calculated structure factors or alternatively by attempting to match the corresponding Patterson functions. Some correlation functions are more conveniently described in reciprocal space, and others in real or vector space. Some successful examples of real-space correlation functions are described by Nordman (1972) and Huber (1970).

2(*a*) *Residuals*

Vand & Pepinsky (1956) suggested the use of various reciprocal-space residuals:

$$R_1 \equiv \sum |F_o - F_c| / \sum F_o \qquad (2)$$

$$R_2 \equiv \sum |F_o^2 - F_c^2| / \sum F_o^2 \qquad (3)$$

$$R_3 \equiv \sum (F_o - F_c)^2 / \sum F_o^2 \qquad (4)$$

$$R_4 \equiv \sum (F_o^2 - F_c^2)^2 / \sum F_o^4 . \qquad (5)$$

We may use some results of Parthasarathy & Parthasarathi (1972) to investigate how these residuals might perform as correlation functions.

Our model is likely to be in error, firstly because chemical differences suggest that some atoms of the

* Present address: Department of Chemistry, University of Auckland, Auckland, New Zealand.

† Present address: Astbury Department of Biophysics, University of Leeds, Leeds LS2 9JT, England.

known molecule should be omitted, secondly because of positional errors in coordinates. If the unknown structure contains N atoms, and our model contains C atoms (usually $C \leq N$) each having an error $\Delta\mathbf{x}_i$, the incompleteness of the model may be defined by

$$\sigma_1^2 \equiv \sum_i^C \mathbf{f}_i^2 / \sum_i^N \mathbf{f}_i^2 . \quad (6)$$

If the distribution of atom types is the same in the model and in the complete structure, σ_1^2 is the ratio of assumed known to total scattering matter. The error in the model may be defined (Luzzati, 1952) by

$$D \equiv \frac{1}{C} \sum_i^C \cos 2\pi\mathbf{h} \,.\, \Delta\mathbf{x}_i .$$

$D=1$ for a perfectly accurate model, $D=0$ for a randomly wrong one, and D decreases with increasing $\sin\theta$ (unless $D=1$). Parthasarathy & Parthasarathi (1972) derived expressions for R_3 and R_4 as functions of σ_1^2 and D. They assume that Wilson's (1949) distribution functions apply (see §4), and find that:

centric case

$$R_3 = 1+\sigma_1^2-(4\sigma_1/\pi)\ [(1-\sigma_1^2D^2)^{1/2}\sigma_1 D \sin^{-1}(\sigma_1 D)] , \quad (7)$$

$$R_4 = 1+\sigma_1^4-\tfrac{2}{3}\sigma_1^2-\tfrac{4}{3}\sigma_1^4D^2 , \quad (8)$$

acentric case

$$R_3 = 1+\sigma_1^2-2\sigma_1\left[\int_0^{\pi/2}\xi^{1/2}\mathrm{d}\varphi - \frac{1-\sigma_1^2D^2}{2}\int_0^{\pi/2}\xi^{-1/2}\mathrm{d}\varphi\right], \quad (9)$$

where

$$\xi = 1-\sigma_1^2D^2\sin^2\varphi ,$$

$$R_4 = 1-\sigma_1^2(1-\sigma_1^2)-\sigma_1^4D^2 . \quad (10)$$

These functions are not readily visualized and are therefore plotted in Fig. 1. They may be used to assess the sensitivity of the residuals to errors in the trial structure.

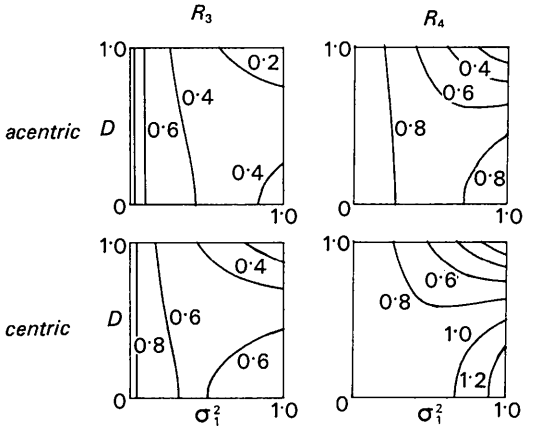

Fig. 1. Theoretical values of R_3 and R_4 as functions of σ_1^2 and D, plotted at contour intervals of 0·2.

The difference between the values of a residual for a trial structure at the correct orientation and position, and at wrong orientations or positions is the difference between, say, $R(0{\cdot}8,\sigma_1^2)$ and $R(0,\sigma_1^2)$ if the model has random errors corresponding to $D=0{\cdot}8$ (see §5). Fig. 1 shows that this difference is large (*i.e.* the contours are nearly horizontal) only when $\sigma_1^2 < ca.\ 0{\cdot}75$; R_3 or R_4 should only be used, therefore, when well over $\frac{3}{4}$ of the structure is thought to be known. Although Luzzati (1952) has derived values for $R_1(D,1)$ (*i.e.* the variation of the residual with the error in a trial structure containing all the atoms of the unknown structure) general functions $R_1(D,\sigma_1^2)$ and $R_2(D,\sigma_1^2)$ are not available: we expect however that it is it necessary to know a similar proportion of the structure before R_1 or R_2 are useful.

In general, residuals of this kind are only useful in translational searches once the molecular orientations have been established; Hackert, Ford & Rossmann (1973) have successfully used R_1 as a translation function in one dimension. However, when there is only one molecule in the unit cell this approach may be applied to rotational parameters and Joynson, North, Sarma, Dickerson & Steinrauf (1970) have successfully used R_1 as a rotation function for triclinic lysozyme.

2(*b*) *Scalar product*

Probably the most widely-used correlation function [see for example Vand & Pepinsky (1956), Rossmann & Blow (1962), Tollin (1966)] is the scalar product:

$$R = \sum_{\mathbf{h}} F_o^2(\mathbf{h})F_c^2(\mathbf{h}) . \quad (11)$$

As Vand & Pepinsky pointed out, the numerator of (5) may be written as

$$\sum_{\mathbf{h}} [F_o^4 - 2F_o^2F_c^2 + F_c^4] . \quad (12)$$

The first term of (12) does not depend on rotational parameters, and the third term should not vary strongly and so the scalar product R is closely related to R_4.

When the model is best aligned R will have a large value, proportional to $\langle F_o^2F_c^2\rangle$: when the model is quite wrongly aligned (*i.e.* it is a randomly wrong structure), R will be proportional to $\langle F_o^2\rangle\langle F_c^2\rangle$, and will have a smaller value. The peak-to-background ratio will be

$$\frac{\langle F_o^2F_c^2\rangle}{\langle F_o^2\rangle\langle F_c^2\rangle} , \quad (13)$$

which, according to Parthasarathy & Parthasarathi (1972) is equal to $1+D^2$ if the F's are distributed according to Wilson's (1949) acentric distribution and is equal to $1+2D^2$ if the centric distribution applies (see §4). Thus for protein work we would expect the peak to background ratio never to be greater than 2 and if $D \approx 0{\cdot}8$ to be equal to about 1·6. This value does not depend on σ_1^2, the completeness of the model, and so R, [equation (11)], rather than a residual, should be

used when only a small part of the structure is used as a model.

3. Human and hen-egg-white lysozymes

The structure of hen-egg-white lysozyme in space group $P4_32_12$ is known from the work of Blake, Koenig, Mair, North, Phillips & Sarma (1965), and we have used the atomic positions published by Imoto, Johnson, North, Phillips & Rupley (1972). X-ray diffraction intensities from human lysozyme in space group $P2_12_12_1$ have been measured by Blake & Swan (1971) and by Banyard, Blake & Swan (1974), and a copy of these data was kindly made available to us by these workers.

3(*a*) *The model used for human lysozyme*

The amino acid residue sequences of human and hen lysozymes are compared in Table 1 of the review by Imoto *et al.* (1972). For our model we used the non-hydrogen atoms of the hen lysozyme structure, with certain omissions. Residues 47 and 48 were omitted entirely because of the extra glycine residue inserted between them and the consequent local rearrangement. Residues which were different in the two proteins had all the side chain atoms beyond C_β omitted (but glutamine and glutamic acid were considered indistinguishable in this context), except that for some conservative changes one or two more atoms were included in the side chain; an example of this is residue 129 where Leu(hen) → Val(human), and the C_γ atom was also used in the model for human lysozyme. This choice produced a set of 858 atoms, 83% of the non-hydrogen covalently bound atoms in human lysozyme. No solvent molecules were included.

3(*b*) *Rotation searches*

When calculating the F_c^2 for use in (11) from a large molecule it is computationally more efficient to calcu-

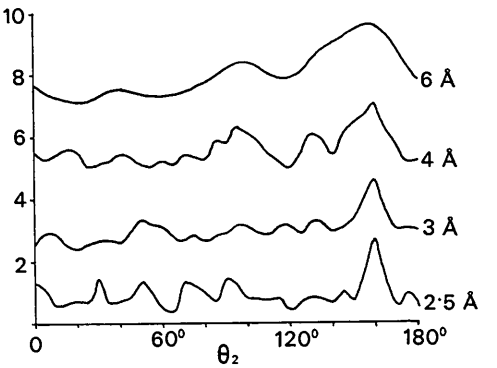

Fig. 2. Sections through the human/hen lysozyme rotation peak at constant θ_1 and θ_3, for shells of structure factors of various resolutions. The vertical scale is arbitrary.

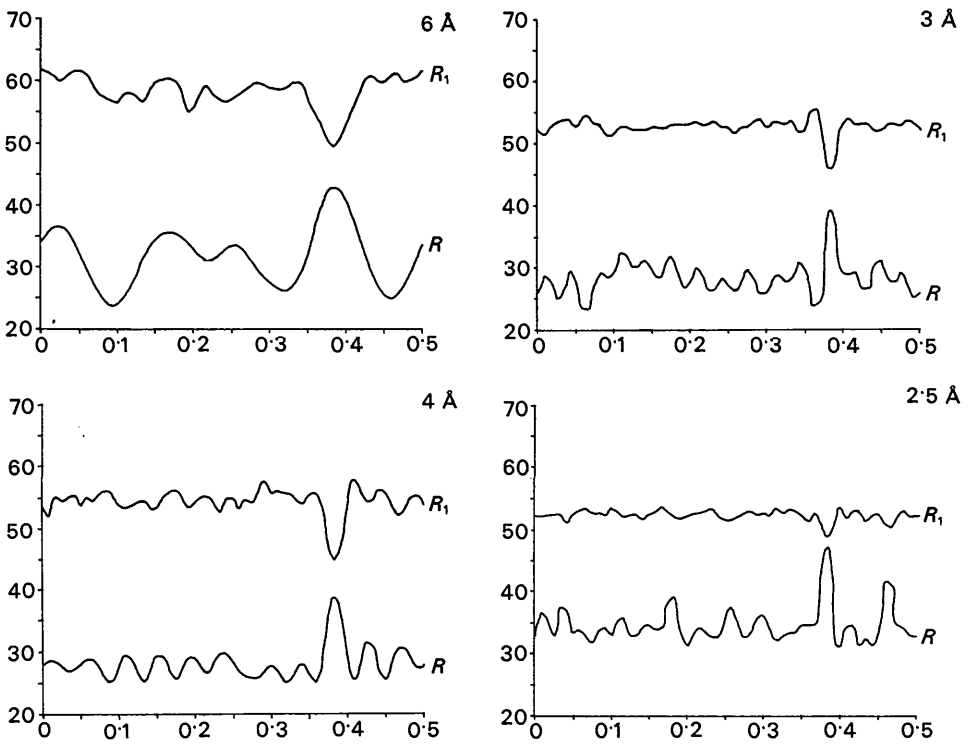

Fig. 3. Sections through the main feature of the human/hen lysozyme translation functions R_1 and R for different resolutions. The ordinate is % for R_1, but arbitrary for R. The abscissa represents a distance of 28·6 Å.

late the Fourier transform of one isolated molecule of the model structure, and then to use an interpolation procedure at each orientation, rather than to calculate a Fourier transform at each orientation. Following Joynson *et al.* (1970), we calculated the Fourier transform of the model at points in reciprocal space corresponding to an 80 Å cube. We used a program based on the 'fine mesh program' described by Tollin & Rossmann (1966), and the strongest 35 of the 269 unique measured structure factors of human lysozyme in the $\lambda/(2 \sin \theta)$ range 10 to 6 Å.

There was one peak which stood 4·5 standard deviations above the mean value, with a peak/background ratio of 1·36; the next highest peak was only 3·0 standard deviations above the mean.

The position of this highest peak was refined by using shells of structure factors with $\lambda/(2 \sin \theta)$ in the ranges 10 to 6, 6 to 4, 4 to 3 and 3 to 2·5 Å, the 100 strongest reflexions in each range being used. Since only a small range of orientations was now being investigated, and because of the higher resolution of the data, it was more convenient to calculate the F_c^2 from a rotated set of atomic coordinates as they were required (Tollin, 1969). The angular parameters used were $\theta_+, \theta_2, \theta_-$ (Lattman, 1972). The three shells of higher-resolution data gave values for the position of the peak in good agreement: the three orientations indicated for the molecule differed by a rotation corresponding to a r.m.s. atomic displacement of 0·3 Å. The angles indicated by the low-resolution data were about 3° from this, giving a r.m.s. atomic displacement of 1·8 Å. Fig. 2 shows sections through this rotation function calculated with the four shells of data.

3(*c*) *Translation searches*

The set of atomic coordinates used in the rotation search, rotated according to the average of the positions indicated by the rotation functions using the three higher-resolution sets of data, was used as the model for the translation search.

If after rotation an atom is at $\mathbf{x}_j$ with respect to a local arbitrary origin, its position with respect to the crystallographic origin will be $\mathbf{x}_j + \mathbf{X}$; the purpose of the translation function is to find $\mathbf{X}$. If there are n molecules in the unit cell there are also similar atoms at $\mathbf{A}_i(\mathbf{x}_j + \mathbf{X}) + \mathbf{d}_i$, $1 \le i \le n$, where the $\mathbf{A}$ are rotation matrices and the $\mathbf{d}$ are translation vectors; if only one molecule is present per asymmetric unit (as in this case) all the elements of all the $\mathbf{A}$ and $\mathbf{d}$ are known from the space-group symmetry. ($\mathbf{A}_1$ is a unitarity matrix and $\mathbf{d}_1$ is a null vector.) (1) may then be written

$$\mathbf{F}_c(\mathbf{h}, \mathbf{X}) = \sum_i^n \sum_j^m \mathbf{f}_j \exp \{2\pi i \mathbf{h}[\mathbf{A}_i(\mathbf{x}_j + \mathbf{X}) + \mathbf{d}_i]\}$$

$$= \sum_i^n \mathbf{F}_i(\mathbf{h}) \exp (2\pi i \mathbf{h} \mathbf{A}_i \mathbf{X}) , \qquad (14)$$

where the molecular transforms F_i are given by

$$\mathbf{F}_i(\mathbf{h}) = \sum_j^n \mathbf{f}_j \exp [2\pi i \mathbf{h}(\mathbf{A}_i \mathbf{x}_j + \mathbf{d}_i)] . \qquad (15)$$

The $\mathbf{F}_i$ are calculated first and stored, and residual type translation functions may then be calculated from (41) *without* needing to loop over all the atoms for each value of $\mathbf{h}$ for each value of $\mathbf{X}$.

All the human lysozyme structure factors with $\lambda/(2 \sin \theta) \ge 6$ Å were used in a search of all unique values of $\mathbf{X}$; both R_1 and R were calculated; steps of about 1·5 Å were used. R_1 showed a minimum of 0·54 and a mean of 0·69 with standard deviation 0·03. According to Nixon (1976) a randomly wrong structure with 36% of the data centric should give a residual of 0·674, in good agreement with the figure obtained here. Finer sampling in $\mathbf{X}$ defined the minimum more precisely with a value of 0·49. Although R showed a maximum at this point, it was only the sixth highest. (None of the 5 higher maxima in R corresponded to a R_1 of less than 0·67.)

The region around this hole in R_1 was then explored using all the structure factors, divided into shells as for the rotation searches. All four shells gave minima in R_1 within 0·2 Å. Fig. 3 compares R_1 and R for different resolutions, while Fig. 4 compares R_1, R_3, R_4 and R for 10 to 6 Å resolution data; R_1 is seen to be the sharpest and least noisy function.

Finally an automatic minimization of R_3 for 6 Å data was performed with six variable parameters (three rotational and three translational); the minimum was only 0·15 Å (r.m.s. on each atom) away from that found using the higher resolution data from the two-stage search just described.

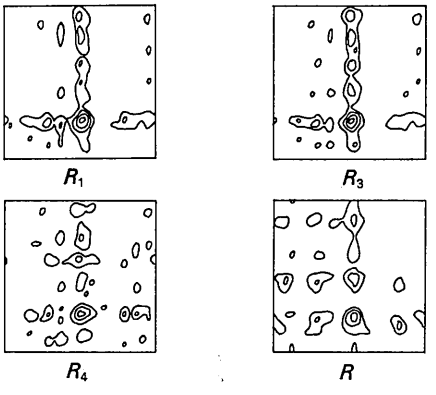

Fig. 4. Contours of sections through the main feature of several translation functions at 6 Å resolution. The contours are in units of one standard deviation from the mean in each case. Only contours below the mean are shown, except for the function $R = \sum F_o^2 F_c^2$ where only contours above the mean are shown.

4. Wilson statistics

In crystallographic theory the distribution function of the structure factors is often required to be known, or

at least assumed. Equations (7) to (10) and the values given for (13) depend on the distribution functions given by Wilson (1949):

$$\textit{acentric} \quad P(F)=\frac{2F}{\Sigma}\exp\left(-F^2/\Sigma\right), \qquad (16)$$

$$\textit{centric} \quad P(F)=2(2\pi\Sigma)^{-1/2}\exp\left(-F^2/2\Sigma\right). \qquad (17)$$

The mean value of F^2 from both these distributions is $\Sigma(=\sum f_i^2)$.

Fig. 5 shows histograms of the distributions of observed and calculated structure factors; the solid line is derived from (16) or (17), Σ being taken as the mean of F^2 for each shell. The agreement of the distributions of F_o and F_c with (16) and (17) is good – much better than we had expected in the light of the commonly expressed view that Wilson statistics cannot readily be applied to protein data. Thus provided that in (16) or (17) we use a value of Σ given by $\langle F^2\rangle$, the average being taken over shells of data, Wilson's distributions may be applied to human lysozyme data.

5. Discussion

We believe that the good agreement between the rotational and translational parameters found from different shells of data is evidence of a real correspondence between the model based upon hen-lysozyme and the human-lysozyme diffraction data. The value of the residual, 0·49, is encouraging considering the omission of many of the side-chain atoms and all the solvent molecules from the calculations. [Compare, for example, with the value of 0·48 at the start of refinement of the trypsin/trypsin inhibitor complex (Huber *et al.*, 1974)]. The use of the residual as a translation function has been shown to be both practicable from a computational aspect and also less noisy and more precise than the use of the scalar product type of function.

Table 1 shows the values of R_1, R_3 and R_4 for the various shells of data. The corresponding values of D were obtained (by a numerical method where necessary) from (7) to (10) and from Table 2 of Luzzati (1952). The values obtained for D were not very sensitive to the value chosen for σ_1^2; in the case of R_4 only, the scaling factor between F_o and F_c was critical: the last column of Table 1 is not therefore as trustworthy as the other determinations of D. The values of D obtained from different residuals agree satisfactorily and lead by use of equations (50) and (51) of Luzzati (1952) to values of an average error in the coordinates ranging from 0·4 to 0·9 Å. The variation of D with $\lambda/2\sin\theta$ or between two- and three-dimensional structure factors is not as great as Luzzati's theory predicts.

The results of this paper have demonstrated a close structural relationship between human and hen lysozymes and have established the coordinate transformations between the two unit cells; in the following paper (Nixon & North, 1976) we investigate the electron density map derived for human lysozyme and compare it with that determined by independent isomorphous replacement.

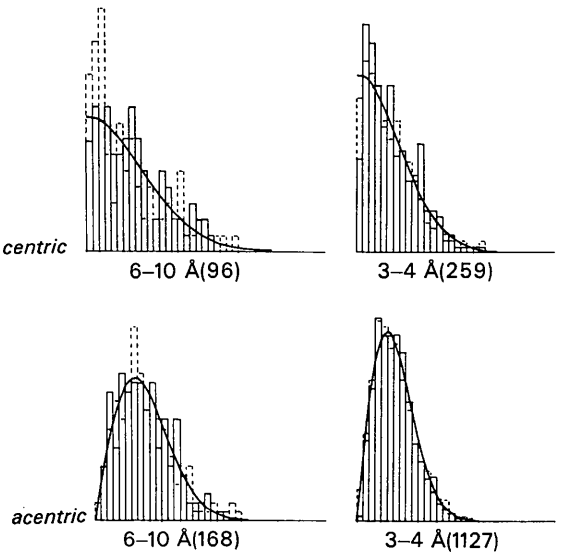

Fig. 5. Distribution functions for structure factors of human lysozyme. The units of F are arbitrary; the ordinate $P(F)$ is the number of structure factors with value F. The solid histogram is F_o; the dotted histogram is F_c; the solid line is calculated from (16) or (17). The number in brackets shows the number of reflexions in each set. The shell of structure factors with $\lambda/(2\sin\theta)$ between 6 and 4 Å gave similar histograms.

Table 1. *The values of R_1, R_3 and R_4 for the* 858 *atom model for human lysozyme, and the values of D derived from them*

Resolution Å	Symmetry	R_1	D	R_3	D	R_4	D
10 to 6	acentric	0·370	0·89	0·174	0·85	0·296	0·90
	centric	0·527	0·80	0·294	0·85	0·566	0·79
6 to 4	acentric	0·406	0·74	0·206	0·79	0·431	0·79
	centric	0·509	0·82	0·282	0·87	0·525	0·81
4 to 3	acentric	0·405	0·75	0·211	0·78	0·464	0·76
	centric	0·581	0·75	0·385	0·75	0·669	0·71
3 to 2·5	acentric	0·451	0·66	0·266	0·66	0·609	0·60
	centric	0·557	0·77	0·390	0·74	0·826	0·58

We thank Drs S. H. Banyard, C. C. F. Blake and I. D. A. Swan for allowing us access to their data. P. E. N. is grateful to the Medical Research Council, London, and to Wolfson College, Oxford, for financial support.

References

BANYARD, S. H., BLAKE, C. C. F. & SWAN, I. D. A. (1974). In *Lysozyme* edited by E. OSSERMAN, R. E. CANFIELD & S. BEYCHOK. New York: Academic Press.

BLAKE, C. C. F., KOENIG, D. F., MAIR, G. A., NORTH, A. C. T., PHILLIPS, D. C. & SARMA, V. R. (1965). *Nature, Lond.* **206**, 757–761.

BLAKE, C. C. F. & SWAN, I. D. A. (1971). *Nature, New Biol.* **232**, 12–15.

HACKERT, M. L., FORD, G. C. & ROSSMANN, M. G. (1973). *J. Mol. Biol.* **78**, 665–673.

HUBER, R. (1970). *Crystallographic Computing*, edited by F. R. AHMED, pp. 96–102. Copenhagen: Munksgaard.

HUBER, R., KUKLA, D., BODE, W., SCHWAGER, P., BARTELS, K., DEISENHOFER, J. & STEIGEMANN, W. (1974). *J. Mol. Biol.* **89**, 73–101.

IMOTO, I., JOHNSON, L. N., NORTH, A. C. T., PHILLIPS, D. C. & RUPLEY, J. A. (1972). *The Enzymes.* Vol. VII, 3rd ed. pp. 665–868. Edited by P. BOYER. New York: Academic Press.

JOYNSON, M. A., NORTH, A. C. T., SARMA, V. R., DICKERSON, R. E. & STEINRAUF, L. K. (1970). *J. Mol. Biol.* **50**, 137–142.

LATTMAN, E. (1972). *Acta Cryst.* B**28**, 1065–1068.

LUZZATI, V. (1952). *Acta Cryst.* **5**, 802–810.

NIXON, P. E. (1976). *Acta Cryst.* A**32**, 344.

NIXON, P. E. & NORTH, A. C. T. (1976). *Acta Cryst.* A**32**, 325–333.

NORDMAN, C. E. (1972). *Acta Cryst.* A**28**, 134–143.

PARTHASARATHY, S. & PARTHASARATHI, V. (1972). *Acta Cryst.* A**28**, 426–432.

ROSSMANN, M. G. & BLOW, D. M. (1962). *Acta Cryst.* **15**, 24–31.

TOLLIN, P. (1966). *Acta Cryst.* **21**, 613–614.

TOLLIN, P. (1969). Private communication.

TOLLIN, P. & ROSSMANN, M. G. (1966). *Acta Cryst.* **21**, 872–876.

VAND, V. & PEPINSKY, R. (1956). *Z. Kristallogr.* **108**, 1–14.

WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.

Acta Cryst. (1976). A**32**, 325

Crystallographic Relationship between Human and Hen-Egg Lysozymes. II. Weighting of Electron-Density Maps Phased from an Incomplete Model Structure and Comparison with Map Obtained by Isomorphous Replacement

BY P. E. NIXON* AND A. C. T. NORTH†

Laboratory of Molecular Biophysics, South Parks Road, Oxford OX1 3PS, England

(*Received* 10 *September* 1975; *accepted* 22 *September* 1975)

Methods suitable for completing and refining a protein structure are investigated both theoretically and with human and hen lysozymes as an example. Sim (or Woolfson) weighting and the α-synthesis are compared with unweighted maps, and while the former is an improvement over unweighted maps, the α-synthesis is less clearly an improvement. Both $WF_o \exp(i\alpha_c)$ and difference maps were found to be useful, and a comparison between the isomorphous replacement map of human lysozyme and our maps based on hen lysozyme was encouraging.

1. Introduction

In the previous paper we proposed a model for the structure of human lysozyme, based on the known molecular structure of hen-egg-white lysozyme; we showed that it was possible to determine the orientation and translation of this model to give a reasonable fit to the measured X-ray diffraction intensities of human lysozyme crystals, both at low (6 Å) and medium (2·5 Å) resolutions. It is more interesting, however, to investigate whether or not we can describe the difference between the lysozymes of two species without further phase information. The human/hen lysozyme situation is particularly useful as a test of methods because the human lysozyme structure was being determined by isomorphous replacement methods simultaneously with the work described in this paper (Banyard, Blake & Swan, 1974).

2. Weighting schemes

We have a set of observed structure factor amplitudes, F_o, and a corresponding set of structure factors $\mathbf{F}_c$ (complex quantities) calculated from only some of the atoms in the crystal. This situation is equivalent to that encountered in the 'heavy-atom' method, for which Woolfson (1956) and Sim (1959, 1960) have proposed weighting schemes for centrosymmetrical

* Present address: Department of Chemistry, University of Auckland, Auckland, New Zealand.

† Present address: Astbury Department of Biophysics, University of Leeds, Leeds LS2 9JT, England.