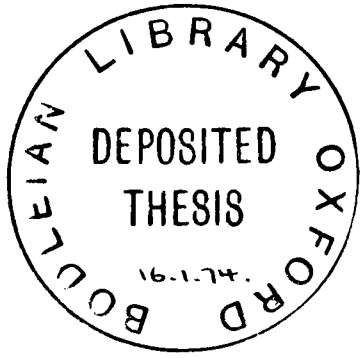
4

D.Phil. Thesis

A Crystallographic Comparison of Protein Molecules with Related Conformations

by

Phillip Nixon



Laboratory of
Molecular Biophysics

Wolfson College
Oxford
October 1973

ABSTRACT

The multiple isomorphous replacement method of protein crystallography has revealed the structures of several proteins, but the method is limited by the need for many isomorphous crystals. This difficulty could be eased to a large extent by using a molecular fragment as a model for the unknown structure, when only one set of X-ray intensities need be measured. This thesis discusses ways to determine the orientation of such a partial structure, and recommends a new method to determine the location of a properly oriented partial structure in the crystal unit cell.

These methods were applied to the proteins human lysozyme, hen lysozyme crystallised at a high temperature, and goat α-lactalbumin, for all of which part of the published structure of hen lysozyme was used as a model. Only the application to human lysozyme was successful, and the failure in the case of the high-temperature hen lysozyme is surprising. An application of these methods to the antibiotic thiostrepton, using an almost complete structure from another crystal form as a model was successful.

Having found a partial structure which explains the X-ray data moderately well, it is necessary to complete and refine the model. Several of the theories discussed in this thesis assume Wilson statistics, which are shown to apply to human lysoz yme surprisingly

well. Three types of completion / refinement technique were applied, a least squares refinement of atomic occupancies, direct methods, and the use of maps with Fourier coefficients with phases derived from the partial structure. The first two methods failed, but the third had partial success. The relationship between "difference maps", Sim weighting, and the α-synthesis is discussed; an extension of Luzzati's treatment of maps with Fourier coefficients $F \exp(i\alpha_c)$ shows that of these three Fourier syntheses, the α-synthesis has the best(unknown peak height)/(known peak height) ratio, whereas Sim weighting has the best signal/noise ratio — which, incidentally, never exceeds unity for acentric data — where "noise" is defined in a more realistic way than that adopted by Sim. A corrected and extended version of a previous semi-quantitative account of difference maps is given.

The application of various Fourier syntheses to human lysozyme structure factor amplitudes proceeded as a test of methods in parallel with the isomorphous replacement study by other workers; it is recommended that both F_{obs} and difference maps be calculated, using Sim weighting. Small differences in the main chain conformations of hen and human lysozymes can be determined, missing side chain atoms can often be placed, but rearrangements of the strucutre as determined by the isomorphous replacement study are not always seen on difference maps.

The value of the residual for a randomly wrong structure when a mixture of acentric and centric data is used is shown to be within 0.007 of the value which corresponds to a linear interpolation between the all centric and all acentric limiting cases.

CLOWN: I cannot do't without computers.

Winters Tale IV ii

CONTENTS

PROLOGUE

(i) Acknowledgements 1

(ii) Protein crystallography 2

(iii) The proteins discussed in this thesis 4

CHAPTER 1 THEORY of ROTATION and TRANSLATION FUNCTIONS

(i) The phase problem 9

(ii) The Patterson function 10

(iii) The use of a model structure 12

(iv) Correlation functions 15

(v) Rotation functions 28

(vi) Translation functions 34

(vii) Refinement of rotational and translational parameters 39

CHAPTER 2 APPLICATIONS of ROTATION & TRANSLATION FUNCTIONS

(i) Human lysozyme 43

(ii) Orthorhombic hen lysozyme 58

(iii) Goat α-lactalbumin 60

(iv) Thiostrepton 67

CHAPTER 3 REFINEMENT of HUMAN LYSOZYME

(i) The first approximation to human lysozyme 71

(ii) The average error in the atomic coordinates 73

(iii) Wrongly placed atoms in the model 77

CHAPTER 3 (continued)

(iv) Tangent formula techniques 83

(v) Application of the tangent formula to human lysozyme 90

(vi) Difference syntheses (introduction) 95

(vii) Sim weighting 98

(viii) Difference maps (theories) 105

(ix) The α-synthesis or sum function 110

(x) Expectation values of signal to noise ratios, &c. 114

(xi) Human lysozyme difference maps 122

(xii) Interpretation of human lysozyme maps 139

CHAPTER 4 CONCLUSIONS

(i) Locating the model 157

(ii) Completing and refining the model 161

APPENDIX A The difference Fourier 163

APPENDIX B Mixed centric and acentric data 169

REFERENCES 173

SYMBOLS AND ABBREVIATIONS

Lower case bold face letters are vectors.

Upper case bold face letters are either matrices or complex numbers (except X).

$a,b,c,$ α,β,γ		unit cell lengths and angles
A		real part of $\mathbf{F}$
$\mathbf{A}_i$		matrix representing rotations (usually proper) between molecules in unit cell.
B		imaginary part of $\mathbf{F}$
C		no. of known atoms; as subscript, refers to known atoms.
$\mathbf{d}_i$		translational shift associated with $\mathbf{A}_i$
$D(\mathbf{h})$	=	$< \cos 2\pi \mathbf{h}.\Delta\mathbf{x}_i) >_i$
E		normalised $\mathbf{F}$ ($E^2 = F^2/\Sigma$)
f_j		an atomic scattering factor
$\mathbf{F}$		a structure factor
F		a structure factor amplitude
$\mathbf{F}_c$		structure factor corresponding to C atoms, i.e. a calculated structure factor.
$\mathbf{F}_i$		Fourier transform of the i^{th} molecule or part thereof.
$\mathbf{F}_Q$		structure factor corresponding to missing atoms, i.e. $\mathbf{F}_Q = \mathbf{F} - \mathbf{F}_c$
G		the Fourier transform of a function restricting the volume in Patterson space over which an integral is taken.

$\mathbf{h} = h,k,\ell$	three indices in reciprocal space
$\mathbf{h}'$	three indices after rotation
I_0, I_1	Bessel functions
M.I.R.	= multiple isomorphous replacement
n	no. of molecules in unit cell
N	no. of atoms in molecule
NAG	Nottingham Algorithms Group library - a library of subroutines available on the 1906A
P	probability function
P(x/y)	means the probability of x having a given value, given that y has a given value.
P	a Patterson function
P_c	the Patterson function corresponding to the set of C atoms, i.e. a "calculated Patterson" (as opposed to observed)
q	refers to one of the unknown atoms
Q	no. of unknown atoms
Q	equation (18) as defined by Tollin as a translation function
R	the scalar product of F and F_c (equation (18))
$R_{n+2(m-1)}$	$= \sum\|F^n - F_c^n\|^m / \sum F^{nm}$ equation (7) to (10)
t	one of the C known atoms
w_i	a set of weights
x	$= F_Q/F_c$
x	$= F\sum_Q^{-\frac{1}{2}}$

$\mathbf{x}$	= x,y,z a vector, usually $0 < x < 1$
$\mathbf{X}$	= X,Y,Z a vector representing the shift necessary to transform from an arbitrary to a crystallographic origin.
X	$= 2FF_c/\Sigma_Q$ the argument of Sim's and of Woolfson's functions
y	$= F_c \Sigma_Q^{-\frac{1}{2}}$
z	$= F_c/F_Q$
α	a phase angle
β	$= \alpha - \alpha_c$ (see figure (13))
Δx_i	error in coordinates of an atom
θ	Bragg angle
θ	$= \alpha_Q - \alpha_c$ (see figure (13))
$\boldsymbol{\theta}$	$= \theta_1, \theta_2, \theta_3$ three Eulerian angles
$\theta_{\pm}$	$= \theta_1 \pm \theta_3$
λ	wavelength of X-rays
μ	signal to noise ratio in difference map
ρ	electron density
σ_1^2	$= \Sigma_N^c/\Sigma$
Σ	$= \sum_i^N f_i^2$
Σ_c	$= \sum_t^C f_t^2$
τ	(expected height of known atom)$/f_t$
χ	(expected height of unknown atom)$/f_q$
ψ	$= \Sigma_Q/\Sigma_c$
*	complex conjugate
< >	mean value

PROLOGUE

(1) Acknowledgements

The very first thing is to acknowledge the help and friendship shown by my supervisor, Professor Tony North. He has sometimes hinted when he thought I was wasting my efforts, but has never forbidden me from doing so. I am also grateful to Professor Phillips for his encouragement, and for providing such a stimulating bunch of colleagues for me. Nearly all of these have provided me with useful suggestions. I am particularly grateful to Stephen Banyard for allowing me such free access to his work, upon which the best part of this thesis depends, and to Simon French for explaining probabilities to me. I have had many useful discussions with Eleanor Dodson and Sabesan, and thank them for letting me mention their unpublished results in this thesis. I also thank Drs Michael Joynson, Brian Anderson and Jean Berthou for giving me copies of their data. The latter looked after me very hospitably while I was in Paris. Alain Lifschitz of the Laboratoire de Minéralogie-Cristallographie, Université de Paris, kindly gave me a copy of his fast Fourier transform routine. The staff of the computing centre, Oxford, have also been very helpful.

I am indebted to my wife for her encouragement and help in the production of this thesis, and to my parents who gave me a chemistry

set when I was 11. The Medical Research Council and Wolfson College, Oxford, have helped financially.

(ii) Protein Crystallography

To date crystallography is the only method which can independently show the three-dimensional structure of protein molecules. As is shown for example in the reviews of the subject by Phillips, and by North and Phillips, the method of heavy atom isomorphous replacement has been the basis of all the structure determinations so far. It is, however, a very laborious method, and often fails because of lack of suitable crystals or isomorphous derivatives. Stryer, Kendrew and Watson pointed out in 1964 that the differences between closely related structures could be determined by measuring a new set of structure factor amplitudes and using the already determined phases. When the related structures are not isomorphous, the problem of determining the phases corresponding to the known structure and to the coordinate system of the unknown structure is not trivial.

Many proteins are likely to be related; proteins from different species which correspond in physico-chemical and, if appropriate, enzymic properties are likely to have similar tertiary structures, particularly if their primary sequences are similar. In fact, a similarity of primary sequence in otherwise unrelated proteins has led to suggestions of similarities in tertiary structure - see α - lactalbumin for example. A method which needed one protein to be

Amino Acid Sequence of Different Lysozymes

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Hen	Lys	Val	Phe	Gly	Arg	Cys	Glu	Leu	Ala	Ala	Ala	Met	Lys	Arg	His	Gly	Leu	Asp	Asn	Tyr
Human				Glu						Arg	Thr	Leu			Leu		Met		Gly	
B α-Lac	Glu	Gln	Leu	Thr	Lys			Val	Phe	Arg	Glu	Leu				Asp		Lys	Gly	
	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	.	36	37	38	39
Hen	Arg	Gly	Tyr	Ser	Leu	Gly	Asn	Trp	Val	Cys	Ala	Ala	Lys	Phe	Glu	.	Ser	Asn	Phe	Asn
Human			Ile			Ala			Met		Leu			Trp		.		Gly	Tyr	
B α-Lac	Gly		Val			Pro	Glu				Thr	Thr	.		His	Thr		Gly	Tyr	Asp
	40	41	42	43	44	45	46	47	.	48	49	50	51	52	53	54	55	56	57	58
Hen	Thr	Gln	Ala	Thr	Asn	Arg	Asn	Thr	.	Asp	Gly	Ser	Thr	Asp	Tyr	Gly	Ile	Leu	Gln	Ile
Human		Arg				Tyr		Ala	Gly		Arg							Phe		
B α-Lac		Glu		Ile	Val	Glu				Asn	Gln						Leu	Phe		
	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78
Hen	Asn	Ser	Arg	Trp	Trp	Cys	Asn	Asp	Gly	Arg	Thr	Pro	Gly	Ser	Arg	Asn	Leu	Cys	Asn	Ile
Human	Ser						Gln	Asp	Asn		Ala	Asp	Ala		Ala				Arg	Val
B α-Lac	Ser		Asp	Lys	Phe		Asn	Asn		Leu		Asn	Asn	Ile	Met	Val				
	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118
Hen	Val	Ser	Asp	Gly	Asn	Gly	Met	Asn	Ala	Trp	Val	Ala	Trp	Arg	Asn	Arg	Cys	Lys	Gly	Thr
Human				Tyr						Lys				Ala	Val		Ala		His	Leu
B α-Lac		Asn	Lys	Ile			Lys	Asn	Asp	Gln	Asp		His		Ser		Ile			
	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98
Hen	Pro	Cys	Ser	Ala	Leu	Leu	Ser	Ser	Asp	Ile	Thr	Ala	Ser	Val	Asn	Cys	Ala	Lys	Lys	Ile
Human	Val	Arg		Pro	Gln		Ile	Arg										Gln	Asn	Arg
B α-Lac	Leu	.		Lys	Val		Ile		Tyr		Leu		His	Lys	Ala	Leu		Ser	Glu	Lys
	119	120	121	122	123	124	125	126	127	128	.	129								
Hen	Asp	Val	Gln	Ala	Trp	Ile	Arg	Gly	Cys	Arg	:	Leu								
Human			Arg	Gln	Tyr	Val	Gln			Gly	.	Val								
B α-Lac	Leu	Asp		.		Leu	.	.		Lys	Glu									

Table (1)

known in atomic detail, and only one set of structure factor amplitudes for the unknown protein, would save many problems of shortage of crystals and lack of isomorphous derivatives.

(iii) The proteins discussed in this thesis

The structure of hen-egg white lysozyme - often called simply "lysozyme" hereafter - was determined in atomic detail by Blake, Koenig, Mair, North, Phillips and Sarma in 1965. The space group was $P4_32_12$ ($a = b = 79.1$Å, $c = 37.9$Å) with one molecule per asymmetric unit. The structure factor amplitudes had been measured on a three-counter linear diffractometer. The sequence of amino acid residues is given in table (1), and the atomic coordinates used in this thesis are those of table IV of the review "Vertebrate Lysozymes" by Imoto, Johnson, North, Phillips and Rupley. Figure (16) on page 133, which is adapted from this review, shows the hydrogen bonds between main-chain atoms; it can be seen that there are four short helical regions and one three-strand sheet region, but quite a lot of the structure cannot readily be described in a general way.

Lysozyme crystallises in many crystal forms, some of them tabulated in the review by Imoto et al.. Jollès and Berthou have found another form which crystallises apparently only at temperatures above 25°C (pH 4.7 to 7.3); if the structure is found to be different from the room-temperature tetragonal form, this may be significant. The space group is $P2_12_12_1$ ($a = 56.3$Å, $b = 73.8$Å, $c = 30.4$Å) with

one molecule per asymmetric unit. Structure factor amplitudes were measured from precession photographs using a partially automatic scanning microdensitometer.

Lysozyme can be extracted from a number of human tissues; large quantities have been obtained from the urine of a single leukaemic patient. The sequence of this leukaemic human lysozyme (henceforth "human lysozyme") is given in table (1); it is the same as the amino acid sequence of human milk lysozyme. Low resolution X-ray crystallographic studies by Blake and Swan of human lysozyme have shown that it is generally similar to hen-egg white lysozyme. The extension of this work to 2.5Å resolution by Banyard and Blake proceeded simultaneously with the work in this thesis. The space group is $P2_12_12_1$ ($a = 57.10 \pm 0.02$Å, $b = 60.82 \pm 0.02$Å, $c = 32.89 \pm 0.01$Å) with one molecule per asymmetric unit. Two equivalents had been measured for most of the structure factors and the ratio $\langle |F(hk\ell) - F(hk\bar{\ell})| \rangle / 2 \langle \bar{F} \rangle = 2.3\%$ indicated that the quality of the measurements was comparable to some other recent protein studies (see Banyard). Structure factors within the cylinder $(k^2/b^2 + \ell^2/c^2)^{-\frac{1}{2}} > 8$Å had been measured on a four-circle diffractometer, and those outside this cylinder had been measured on a five-counter linear diffractometer. This high resolution isomorphous replacement study has at the time of writing reached the stage of a wire model, from which preliminary atomic coordinates have been measured, and

some qualitative comparisons with the hen lysozyme have been made (see section 3 (xii))

α-lactalbumin is a milk whey protein which plays an important part in the lactose synthetase system, which has been reviewed by Brew. The amino acid sequence of the cow protein is known and is given in table (1). It is largely similar to the lysozyme sequence, except at the carboxy terminus end. This similarity, especially at the cysteine bridges, led Browne, North, Phillips, Brew, Vanaman and Hill to show that to incorporate the amino acid residue changes into a model of lysozyme required remarkably little disturbance of the main features of the model. The X-ray crystallography of α-lactalbumin has not been easy. Various workers, as discussed by Fenna, have found that the cow protein crystallises with between 4 and 24 molecules per asymmetric unit. Aschaffenburg, Fenna, Handford and Phillips have found that various forms of goat α-lactalbumin can be found with but two molecules per asymmetric unit. "α-lactalbumin" hereafter refers to goat α-lactalbumin in space group $P2_1$ ($a = 45.0 \pm 0.1$Å, $b = 89.0 \pm 0.1$Å, $c = 32.1 \pm 0.1$Å, $\beta = 92.6^{\circ}$) with two molecules per asymmetric unit. X-ray diffraction structure factor amplitudes had been measured by Handford and Joynson in the resolution range 6 - 3Å using a five-counter linear diffractometer; those with resolution out to 6Å (and the linear blind region) had been measured using a four-circle diffractometer. Only one suitable crystal

could be found, and only one equivalent of the four ($hk\ell$, $\bar{h}k\bar{\ell}$, $\bar{h}\bar{k}\bar{\ell}$, $h\bar{k}\ell$) had been measured, as described by Joynson.

The antibiotic thiostrepton is briefly mentioned in section 2(iv).

I. THEORY OF ROTATION AND TRANSLATION FUNCTIONS

(i) The Phase Problem

X-rays exhibit interference effects (see International Tables, III, p.43) characteristic of waves and so may be assigned an amplitude and a phase. If crystals are assumed to be regularly repeating structures, the scattering of X-rays by crystals is explained by diffraction theory; that is, the wave $\mathbf{F}(\mathbf{h})$ at a point in reciprocal space $\mathbf{h}$ scattered from a specimen with electron density $\rho(x)$ is given by the Fourier transform:-

$$F(\mathbf{h}) = V\int_0^1 \rho(x) \exp(2\pi i \mathbf{h}.x)\, dx \qquad (1)$$

where V is the volume of the unit cell. The inverse relationship is

$$\rho(x) = \frac{1}{V} \sum_{h=-\infty}^{\infty} F(\mathbf{h}) \exp(-2\pi i \mathbf{h}.x) \qquad (2)$$

and if the electron density has an exact periodicity $F(\mathbf{h})$ is zero except when all three elements of $\mathbf{h}$ are integers. The concept of atoms (Democritus, c 400 B.C.) is often useful, when equation (1) becomes for N atoms in the unit cell:-

$$F(\mathbf{h}) = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}.x_j) \qquad (3)$$

Measurements can be made of the $F(\mathbf{h})$, that is the amplitudes only

by irradiating a crystal with monochromatic X-rays and holding an energy-sensitive detector in the appropriate position. Unfortunately no phase-sensitive detector is known and the phases have to be determined in other ways.

Four sources of phase information are available:-

(1) the use of isomorphous replacement

(2) the use of 'anomalous scattering' information

(3) phases calculated from part of the structure which is assumed known

(4) phases refined by the use of equations relating to the phases of different reflexions; these equations are derived from a priori assumptions about the nature of the electron density and may be applied to arbitrary phases, or to those from methods (1), (2) or (3).

Method (3), augmented to some extent by (4), is used in this investigation.

(ii) The Patterson Function

The Fourier transform of the square of the structure factor amplitudes - $F^2(h)$, which can be measured - is called the Patterson function. It has peaks at positions relative to the origin of all the inter-atomic vectors with peak height proportional to the product of the atomic numbers. The Patterson function can be used to solve simple structures, but it is far too complicated, and the peaks too badly resolved, for direct use in protein crystallography. When no direct phase information is available, the only criteria for judging a solution are:-

(1) agreement between observed structure factor amplitudes (F) and those calculated from the proposed structure (F_c)

(2) chemical reasonableness

(3) the support of some physico-chemical properties of the substance - examples are the optical properties of crystals (e.g. Knott), and the nuclear magnetic relaxation times of nuclei in the substance, especially in the presence of paramagnetic ions (e.g. Barry, North, Glassel, Williams and Xavier).

In place of "observed F" above, one could write "observed F^2", which may be replaced by "the Patterson function". The problem is thus equivalent to finding a structure which generates the observed Patterson function. Unfortunately there is no guarantee that any solution will be unique, and indeed Pauling and Shappell in 1930 described the structure determination of bixbyite, where they deduced a small distortion of the metal atoms from cubic close packing, but the direction of this distortion could not be obtained from the F's alone, because a distortion in either direction generated the same Patterson function. They resolved this dilemma by criterion (2) above; into only one of these structures could the oxygen atoms be fitted. However, such "homometric pairs" are likely to be rarely encountered in protein work, although there is no theoretical basis for this hope.

The problem of matching trial structures with observed F's

can thus be carried out in reciprocal space, that is compare F with F_c, or in vector space, that is compare the Patterson function with the Patterson calculated from the trial atomic positions. A particular calculation may thus be described in either or indeed both of these ways.

(iii) The use of a model structure

Method (3) of section (i) of this chapter relies on the use of a known part of the structure - a "partial structure". This may be found either by examining the Patterson function (the "heavy atom" method of small-molecule crystallography), or by postulating that a known molecule or molecular fragment will serve as a partial structure. In the first case, the position of the heavy atoms in the unit cell is known, and the methods of the rest of this chapter are unnecessary. This thesis is concerned, however, with the second case.

(a) The purpose of rotation functions

Although we may guess that a particular atomic arrangement will serve as a partial structure for the crystal we do not know a priori the orientation of this model in the unit cell, and three angular parameters θ_1, θ_2, $\theta_3 \equiv \theta$ must be determined before the vectors h and x in equation (1) refer to the same coordinate system. In space group P_1 the position of the origin has no especial

importance and the model may be used as a partial structure when its orientation is known.

The Patterson function of a structure with more than one molecule in the unit cell may be considered as the sum of two parts: inter-molecular or cross-vectors, and intra-molecular or self-vectors. In a structure of N atoms per molecule and n molecules per unit cell, each atom will make a self-vector to N atoms and a cross-vector to (n-1)N atoms: the self vectors will comprise $\frac{1}{n}^{th}$ of the Patterson function. If we consider only the self-vectors, the resulting reduced calculated Patterson function is independent of the position vector and the problem is temporarily reduced to one of finding three angles. This is the purpose of rotation functions as used in this thesis.

(b) The purpose of translation functions

If there is one molecule per crystallographic asymmetric unit a correctly orientated model will represent the position of one molecule in the unit cell, and the orientation of the other molecules can be found from the space group symmetry; but the positions of these other molecules depend on the positions of the crystallographic symmetry elements with respect to the first molecule. Thus to determine the cross-vectors in one unit cell we must find the vector X representing the position of the origin of our model with respect to the crystallographic origin; this is the purpose of

the translation functions. In space group $P2_1$ the y coordinates of the atoms of both molecules depend in the same way on the value of the y component of the origin fixing vector, and so the Patterson function does not depend on this. For most space groups with two or more symmetry elements all three components of this vector must be found.

If after rotation of its molecule an atom is at $x_j(\theta)$ relative to the local origin, and its position with respect to the crystallographic origin is $x_j(\theta) + X$, there will be (n-1) other atoms at $A_i(x_j(\theta) + X) + d_i$; if there is but one molecule per crystallographic asymmetric unit, all the elements of the point group matrices A_i and translation vectors d_i will be known from the crystallographic symmetry.

$$F_c(h,\theta,X) = \sum_i^n \sum_j^N f_j \exp(2\pi i h(A_i(x_j(\theta)+X) + d_i))$$
$$= \sum_i^n F_i(h,\theta) \exp(2\pi i h A_i X) \qquad (4)$$

Equation (4) is equation (3) expanded to allow for more than one molecule per unit cell. It introduces the molecular transforms F_i which are functions of the rotational parameters θ, but not of the translational parameters X:-

$$F_i(h,\theta) = \sum_j^N f_j \exp(2\pi i h(A_i x_j(\theta)+d_i)) \qquad (5)$$

(c) Non-crystallographic symmetry

If there is more than one molecule per asymmetric unit, these molecules (or protein sub-units) will be related by rotation (or rotation-inversion) matrices A and translation vectors **d** which are approximately valid over part of the unit cell; they are called "non-crystallographic symmetry elements". They may, however, be included in equations (4) and (5), where n is the number of molecules (or sub-units) per unit cell. These non-crystallographic symmetry elements are not necessarily related to the unit cell vectors **a**, **b** and **c** in a simple way, and so are harder to determine than crystallographic symmetry elements. Functions to determine these non-crystallographic rotations A and translations **d** were devised by Rossmann and Blow in 1962 and by Rossmann, Blow, Harding and Coller in 1964, and "the rotation function" often refers to this 1962 non-crystallographic function. But in this thesis it usually has the meaning of section (iii)(a).

(iv) Correlation functions

In the following a Patterson function calculated from the observed structure factor intensities is called the observed Patterson, P, and the calculated Patterson, P_c, is that calculated from a trial model structure.

$$P(\mathbf{x}) = \sum_{\mathbf{h}} F^2 \exp(-2\pi i \mathbf{h}.\mathbf{x})$$

If the orientation of the molecule in the unit cell is partially known, a manual comparison of a few key structure factors can determine the position of the model in the unit cell. Knott in 1940 used the optical properties of naphthalene crystals to reduce the number of parameters, and A. Klug in 1950 used packing considerations for triphenylene.* For a quantitative comparison - a prerequisite of an automatic search method - a correlation function must be defined. A correlation function thus defines how well a set of $F_c(h)$ matches the set of $F(h)$. The specific application of correlation functions as rotation or translation functions is deferred until sections (v) and (vi). As mentioned in section (ii), page 11, a comparison may be made either in vector space or in reciprocal space; we start with vector space methods.

(a) Minimum functions

In "Vector Space", p.242, Buerger pointed out that all the points of a P_c corresponding to N point atoms (N(N-1)/2 independent non-origin points) will lie on a peak in the observed Patterson P when the fit is correct, but that when incorrect, at least some of these N(N-1)/2 points will lie on an empty region in P; this assumes of course that the peaks in P are well resolved. The minimum value of P corresponding to the points of P_c is thus zero when there is no Patterson vector corresponding to one of the vectors in P_c, but large when all the vectors in P_c lie on peaks in P. Nordman

* but see also Pinnock,Taylor & Lipson

and Nakatsu in 1963 used this minimum function as a search or correlation function.

If the model is incomplete, the above discussion will still apply, but if it has wrongly placed atoms, the minimum function will always be small. This, and a reluctance to rely on only one point, led Nordman in 1972 to use a modification in which not the worst fitting, but the average of the 30% or so worst fitting vectors was used as a correlation function. He used several turns of α-helix as a partial structure and searched with this modified minimum function in the Patterson of myoglobin. (His search vectors were not single atom-atom vectors, but overlapping clumps of these.) He successfully located most of the helices.

(b) Huber's function

Huber and Hoppe (e.g. Huber, 1969) have successfully used the function

$$\int_V \{P(X) - P_c(X,\theta,X)\}^2 \, dX \qquad (6)$$

where V is the volume in the unit cell where $P < P_c$. If only part of the molecule is used as a model, at the correct orientation and position $P(X) > P_c(X)$ for all X, and regions where $P < P_c$ represent error. We look, therefore, for a small value of expression (6) at the correct values of θ and X. The inequality $P < P_c$ is more likely

to be satisfied when P_c is large, or when P is small, and so both Huber's and Nordman's functions measure the extent to which large values of P_c fall onto small values of P, but not vice versa. (cf. the equations (17) and (10), where this distinction is not made.)

(c) Residuals

Vand and Pepinsky in 1956 discussed the following reliability factors or residuals

$$R_1 = \sum |F - F_c| \Big/ \sum F \qquad (7)$$

$$R_2 = \sum |F^2 - F_c^2| \Big/ \sum F^2 \qquad (8)$$

$$R_3 = \sum (F-F_c)^2 \Big/ \sum F^2 \qquad (9)$$

$$R_4 = \sum (F^2-F_c^2)^2 \Big/ \sum F^4 \qquad (10)$$

R_1 is the most familiar of these, and perhaps the most useful, because the use of the higher powers of the F's seems to produce noisier functions (see figure (5) on page 53). R_3 and R_4 are progressively easier to discuss theoretically.

(d) Expected values of the residuals

If we have a model consisting of C atoms, each of which has an error Δx_i, the value of the residuals R_1 to R_4 will depend both on the accuracy and on the completeness of the model.

Following Luzzati in 1952 we define a measure of the accuracy of the atomic coordinates as

$$D(\mathbf{h}) = \frac{1}{C} \sum_{i}^{C} \cos 2\pi \mathbf{h}.\Delta x_i$$

D(**h**) is defined for each structure factor, but we may take D to be an average of D(**h**) for a shell of structure factors; D(**h**) will vary strongly with sinθ, but not usually very strongly with direction in reciprocal space. D = 1 for a perfectly accurate model, and D = 0 for a randomly wrong one.

If there are N atoms in the complete structure, we may define as a measure of the completeness of the model

$$\sigma_1^2 = \sum_{i}^{C} f_i^2 \Big/ \sum_{i}^{N} f_i^2 \equiv \Sigma_C / \Sigma$$

If the distribution of atom types is the same in the partial as in the complete structure, σ_1^2 is the ratio of assumed known to total scattering matter.

Parthasarathy and Parthasarathi in 1972 derived expressions for $R_3(D,\sigma_1^2)$ and $R_4(D,\sigma_1^2)$ for the centric and acentric sets of structure factor. Their results are

centric

$$R_3 = 1 + \sigma_1^2 - \frac{4\sigma_1}{\pi} \left[(1-\sigma_1^2 D^2)^{\frac{1}{2}} \sigma_1 D \sin^{-1}(\sigma_1 D) \right] \qquad (11)$$

$$R_4 = 1 + \sigma_1^4 - \frac{2}{3}\sigma_1^2 - \frac{4}{3}\sigma_1^4 D^2 \qquad (12)$$

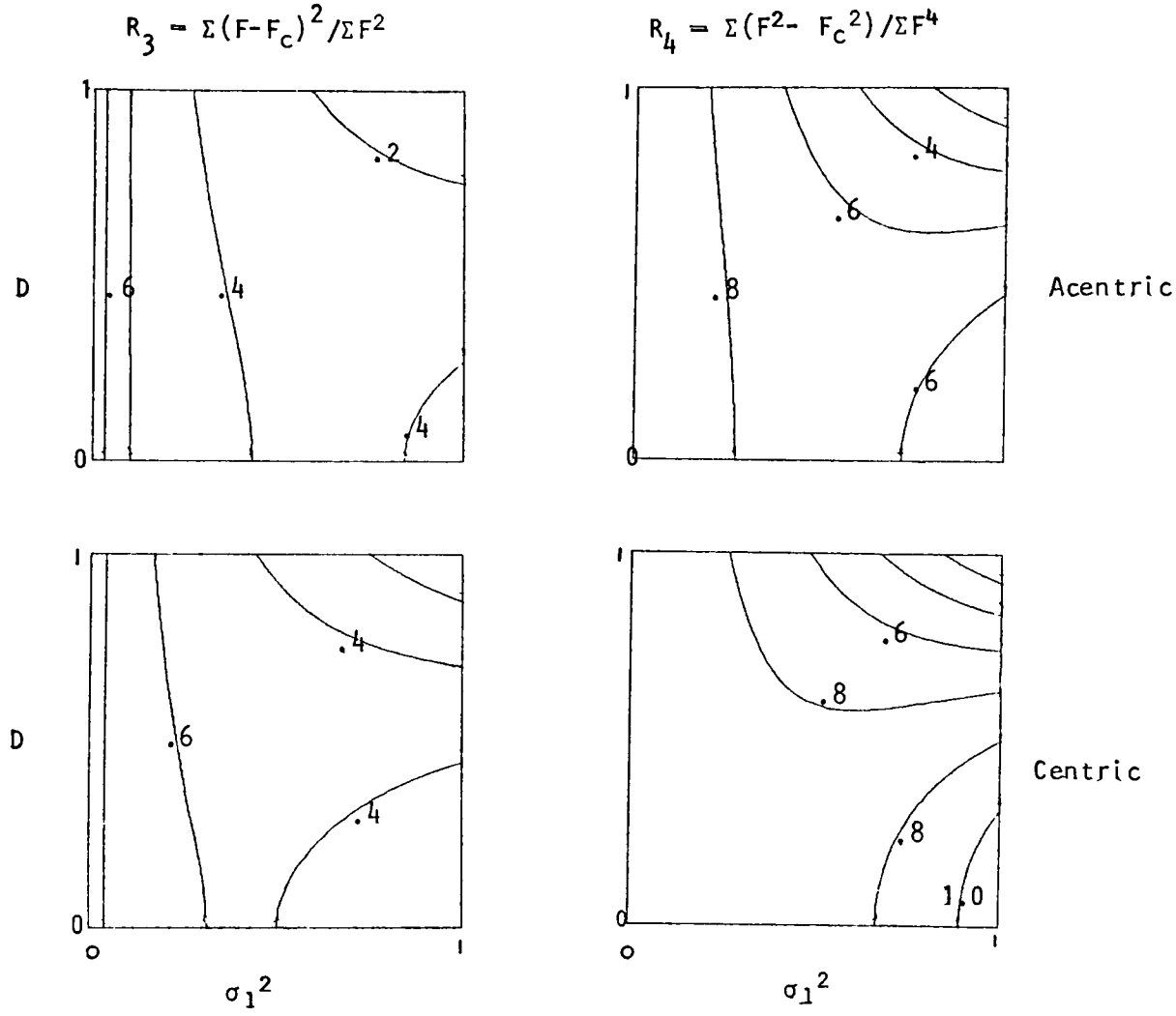


Figure (1)

Theoretical R values for varying values of $\sigma_1^2 = \Sigma_c/\Sigma$ and $D(\mathbf{h}) = \langle \cos 2\pi \mathbf{h}.\Delta\mathbf{x}_j \rangle_c$ where $\Delta\mathbf{x}_j$ is the error in the jth of the C known atoms.

D	R_1 (acentric)	R_1 (centric)
0	0,828	0,586
0,100	0,826	0,583
0,200	0,815	0,576
0,300	0,798	0,563
0,400	0,770	0,544
0,500	0,733	0,516
0,550	0,711	0,500
0,600	0,685	0,479
0,650	0,657	0,458
0,700	0,620	0,434
0,750	0,580	0,404
0,800	0,532	0,369
0,820	0,511	0,353
0,840	0,486	0,336
0,860	0,460	0,318
0,880	0,432	0,297
0,900	0,399	0,273
0,910	0,381	0,261
0,920	0,361	0,246
0,930	0,340	0,231
0,940	0,317	0,213
0,950	0,293	0,195
0,960	0,263	0,175
0,970	0,232	0,152
0,980	0,190	0,124
0,990	0,137	0,089
0,995	0,100	0,064
1,000	0	0

Table (2)

R_1 as a function of D taken from Luzzati (1952)

acentric

$$R_3 = 1 + \sigma_1^2 - 2\,\sigma_1\left[\int_0^{\pi/2} \xi^{\frac{1}{2}} d\phi - \frac{1-\sigma_1^2 D^2}{2}\int_0^{\pi/2} \xi^{-\frac{1}{2}} d\phi\right] \qquad (13)$$

where $\xi = 1 - \sigma_1^2 D^2 \sin^2\phi$

$$R_4 = 1 - \sigma_1^2(1-\sigma_1^2) - \sigma_1^4 D^2 \qquad (14)$$

Equations (11) to (14) are not readily visualised, and so they are plotted at 20% contour intervals in figure (1).

Luzzati in 1952 derived an expression for R_1 as a function of D, but not of σ_1^2: his expression is tabulated in table (2). (All the discussions of likely values of residuals given in this sub-section assume Wilson statistics - see page 73 for a discussion of the validity of this assumption.)

(e) When a residual is likely to be useful

The difference between $R_n(D, \sigma_1^2)$ at the correct value of the rotational and/or translational parameters and at a wrong value is the difference between say $R_n(0.8,\sigma_1^2)$ and $R_n(0,\sigma_1^2)$ if the model has errors corresponding to D = 0.8. Inspection of figure (1) shows that for R_3 and R_4 this difference does not become large until $\sigma_1^2 > 0.75$; that is we should use a partial structure comprising not less than about $\frac{3}{4}$ of the total structure when using R_3 or R_4 as a correlation function. Unfortunately, corresponding expressions for R_1 as a function of σ_1^2 do not seem to have been derived, but we may expect this general

conclusion i.e. that about $\frac{3}{4}$ of the structure should be in the model - to be valid for R_1 also.

Equations (11) to (14) and table (2) can be used to decide whether or not a given value of a residual is less than we would expect from a randomly wrong model. Unfortunately, there seems to be no theory which gives a value for the expected random fluctuations of a residual about its limiting value. The assumptions on which equations (11) to (14) are based lead to these random fluctuations being vanishingly small; however, crystallographic experience seems to show that a departure of more than a few percent from the random value indicates that the model contains some truth, even if the model is systematically misplaced.

(f) The scalar product

This important correlation function may be expressed either in vector space or in reciprocal space. Expanding the numerator of R_4 gives

$$\sum_h (F^2(h) - F_c^2(h))^2$$

$$= \sum_h F^4(h) - 2F^2(h)\, F_c^2(h,\theta,X) + F_c^4(h,\theta,X) \qquad (15)$$

$$= \int_0^1 P^2(x) - 2P(x)\, P_c\,(x,\theta,X) + P_c^2(x,\theta,X)\, dx \qquad (16)$$

$$= \int_0^1 (P(x) - P_c(x,\theta,X))^2 dx \qquad (17)$$

This last expression differs from Huber's function (equation (6)) only in that in (17) the integral is over the whole unit cell.

In the above expressions F and F_c have to be scaled together. If this is such that $\Sigma F^2 = \Sigma F_c^2$, this may be shown to be equivalent to $\int_0^1 P(X)dx = \int_0^1 P_c(x)dx$ in equation (17).

Vand and Pepinsky suggest omitting both the first and third terms of equations (15) and (16), and omitting the factor -2. This leaves a function, R, the scalar product of F^2 and F_c^2.

$$R(\theta,X) = \sum_h F^2(h)\ F_c^2\ (h,\theta,X) \qquad (18)$$

$$= \int_0^1 P(X)\ P_c\ (x,\theta,X)\ dx \qquad (19)$$

When the F^2 and F_c^2 are in agreement, equation (18) will have larger values than when they are uncorrelated: how much larger is considered in sub-section (h).

Considering equation (16) as a translation function, some values of X may cause some of the cross-vectors in $P_c(x)$ to coincide with some of the self-vectors, producing a particularly spikey function P_c. This spikiness may cause equation (19) to have erroneously high values, but in this case $\int_0^1 P_c^2(x)dx$ will also have a higher than average value, which will mean that the use of equation (16) rather than (19), will compensate at least to some extent for this. This

is borne out by figure (5) on page 53, where equation (15) is a less noisy function than equation (18).

(g) The G function

In equation (19) the integral is over the whole unit cell. If the P_c is determined by interpolation in a measured set of structure factors, it may be desirable to restrict the volume of integration to, for example, a sphere centred on the origin, to exclude long vectors which are more likely to be inter-molecular (Rossmann and Blow, rotation function, 1962), or to include only specified inter-molecular vectors (Rossmann et al, translation function, 1964). In reciprocal space this becomes

$$R(\theta,X) = \sum_{h} F^2(h) \sum_{h'} F_c^{\,2}(h,\theta,X)\ G\ (h-h') \qquad (20)$$

where h' refers to h rotated by θ and G is the Fourier transform of the volume of integration in Patterson space.

$$G(x) = 3(\sin 2\pi x - 2\pi x \cos 2\pi x)/(2\pi x)^3$$

for a sphere. $G(0) = 1$, and G is very small when $|x| > 0.725$; the inner summation in equation (20) can thus be over a few terms where $h' \approx h$. If the X's are omitted from equation (20), it becomes the "rotation function" mentioned on page 15.

(h) Expected values of the scalar product

When the partial structure is correctly positioned, the value

Symmetry	$\langle F^4 \rangle / \langle F^2 \rangle^2$
P_1	2 - s
$P\bar{1}$	3 - $1\frac{1}{2}$ s
$P2_1$ acentric	2 - $\frac{1}{2}$ s
$P2_1$ centric zone	3 - $\frac{3}{4}$ s
$P2_12_12_1$ acentric	2 - $\frac{1}{4}$ s
$P2_12_12_1$ centric	3 $\frac{3}{8}$ s

Table (3)

The ratio of the values of the scalar product of $F^2(\mathbf{h})$ and $F_C^2(\mathbf{h})$ for the related and unrelated cases, after Foster and Hargreaves.

$s = \sum_i^N f_i^4 / \left(\sum_i^N f_i^2\right)^2$ which is small if N is large and the atomic weights do not differ very much.

of R (equation (18)) is proportional to $\langle F^2 F_c^2 \rangle$: when the model is incorrectly positioned, F^2 is uncorrelated with F_c^2 and R is proportional to $\langle F^2 \rangle \langle F_c^2 \rangle$. The peak / background ratio when using the scalar product function of F^2 and F_c^2 is thus

$$\frac{\langle F^2 F_c^2 \rangle}{\langle F^2 \rangle \langle F_c^2 \rangle}$$

Foster and Hargreaves have given formulae which depend on the space group for $\langle F^2 \rangle$ and $\langle F^4 \rangle$. When the model is complete and accurate (relative to an arbitrary origin) $\langle F^2 F_c^2 \rangle$ / $\langle F^2 \rangle \langle F_c^2 \rangle$ = $\langle F^4 \rangle$ / $\langle F^2 \rangle^2$, some values of which are given in table (3). From this we see that this peak / background ratio is approximately 2 for acentric data and 3 for centric data. Wilson's distributions can readily be shown to give exactly 2 and 3 respectively: in fact Wilson's distributions implicitly assume that Σf_i^4 is negligible compared to $(\Sigma f_i^2)^2$; this is likely to be true in the examples of chapter 2 because the number of atoms used is large.

More important that these space-group dependent deviations from 2 and 3 is the variation of this peak / background ratio with σ_1^2 and D (see page 19). Parthasarathy and Parthasarathi state that

$$\frac{\langle F^2 F_c^2 \rangle}{\langle F^2 \rangle \langle F_c^2 \rangle} = 1 + D^2 \quad \text{(acentric)}$$

$$= 1 + 2D^2 \quad \text{(centric)}$$

which have the limits of 2 and 3, as above, when the model is perfectly accurate. This does not depend on the value of σ_1^2, and so equation (18) is better than a residual when only a small part of the structure is known. When only a few of the structure factors belong to a centric zone, the peak / background ratio is likely to be of the order of 1.7; see table (5) page 48, and table (6), page 51.

(v) Rotation functions

Rossmann and Blow in 1962 described the use of equation (20) as a rotation function, where $F_c^2(\mathbf{h},\theta)$ was $F^2(\mathbf{h})$ rotated; the G function serves both as a Patterson space cut-off and as an interpolation function (necessary since $F(\mathbf{h})$ is known only at integer values of $\mathbf{h}$.) They describe how a rotation matrix may be set up whose elements are functions of three angles - either an Eulerian or a spherical polar system may be used.

(a) Eulerian angles

When the set of F's has symmetry, it is not necessary to search the whole of the space described by the three angles, and Tollin, Main and Rossmann in 1966 described how the asymmetric unit in Eulerian space may be found and simply described; when spherical polars are used, a simple description of the asymmetric unit is not possible. (Huber, Nordmann, and Rossmann and Blow each use different definitions of Eulerian angles; the definition given by Rossmann and Blow is used

in this thesis; it is, rotate by θ_1 about z, then rotate by θ_2 about the new x, then rotate by θ_3 about the new z. Rossmann and Blow's suggestion for transforming from non-orthogonal to orthogonal axes (y = b, x parallel to a*) has also been used, when necessary, in this thesis.)

A feature of systems of angles defining a three dimensional orientation is that the change in orientation as one angle changes depends on the values of the other angles. An extreme example, in the Eulerian system above, is that when $\theta_2 = 0$, all rotations such that $\theta_1 + \theta_3 =$ constant are equivalent. Lattman in 1972 explained this fully and recommended using parameters $\theta_1 + \theta_3$, θ_2, $\theta_1 - \theta_3$, with steps such that

$$\Delta\theta_+ \cos(\theta_2/2)$$

$$\Delta\theta_2$$

and $$\Delta\theta_- \sin(\theta_2/2)$$

are constant; sections of constant θ_2 should be calculated. The asymmetric unit in this space is, however, a more complicated shape.

(b) Various scalar product rotation functions

Lattman and Love in 1970 pointed out that the G function used by Rossmann and Blow was not necessary if one of the sets of structure factors was calculated from atomic positions, because P_c contained only intra-molecular vectors, and they suggested a linear interpolation

in a stored molecular transform.

Tollin (1964 and private communication) has suggested that it is more accurate (and may be quicker if only a small region of Eulerian space is to be searched) to calculate the $F_c(\mathbf{h},\theta)$ as required by a Fourier transformation of the rotated atomic coordinates (i.e. rotate in real space and Fourier transform, rather than F.T. and rotate in reciprocal space).

Tollin has also suggested that short vectors in Patterson space are characteristic of protein covalent and hydrogen bonds in general, whereas longer vectors (say > 10Å) are characteristic of a particular protein. Unfortunately, whereas each point in the Fourier transform of the whole vector set may be obtained by taking the modulus squared of the Fourier transform of the N atoms, a process involving O(N) operations, the long vector only transform involves $O(N^2)$ operations and a program took 20 seconds per structure factor for distances > 10Å in lysozyme.

(c) Vector space rotation functions

Nordman and Huber et al. have successfully used their correlation functions as rotation functions.

(d) Residual-type rotation functions

In the exceptional case of space group P_1 the position of the origin is of no importance, and Joynson et al. successfully used the residual R_1 (expression (7)) as a rotation function. When there are

4 molecules in the unit cell the self-vectors are only $\frac{1}{4}$ of the total. Figure (1) and the subsequent discussion imply that R_4 and R_3 (and probably by extrapolation R_1 also) will not be useful as rotation functions when only $\frac{1}{2}$ of the atoms are present: this is probably a similar situation to having only $\frac{1}{4}$ of the vectors present, and so residuals will probably not be useful in this context.

(e) An inequality-type rotation function

In equation (4) on page 14 the molecular transforms $F_i(\mathbf{h})$ are functions of the rotational parameters but not of the transitional. Vand and Pepinsky in 1956 suggested (for the centric case, but true also for the acentric) that the mean value of F_c^2 averaged over $\mathbf{X}$ is

$$\langle F_c^2 \rangle_{\mathbf{X}} = \sum_i^n F_i^2$$

since in equation (4) the angles $\mathbf{h}\,\mathbf{A}_i\,\mathbf{X}$ will be random. They suggested that $R^1{}_2(\theta) = \Sigma \left| F^2 - \langle F_c^2(\theta) \rangle_{\mathbf{X}} \right|$ would be a suitable rotation function; figure (5) shows however that the region in $\mathbf{X}$ where R_2 is less than average is quite small, and almost certainly not large enough to cause $R^1{}_2(\theta)$ to have a perceptible minimum, unless n = 1 (when Joynson et al. used $R_1(\theta)$). (Vand and Pepinsky in fact only gave explicit formulae for two centric cases, but the above expression is valid for any space group, including acentric data.)

(f) A better inequality rotation function

From equation (5) we can see that

$$F_c \leqslant \sum_i^n F_i$$

That is, the largest value that F_c can have is the sum of the moduli of the n molecular transforms (n molecules per unit cell). Violations of this inequality may be expressed by

$$R'_1(\theta) = \sum_h (F(h) - \sum_i F_i(h,\theta))\ S(h) \qquad (21)$$

where $S(h) = 0$ if $F < \sum F_i$

$= 1$ if $F > \sum F_i$

Vand and Pepinsky suggested essentially this function for space group $P\bar{1}$ as a special case: I am not sure how original is the generalisation of their idea in equation (21). $R'_1(\theta)$ should be zero at the correct orientation for error-free data; it is reminiscent of Huber's function. The inequality becomes rapidly less likely to be broken as n increases, and may only be useful for n = 2

(g) Computing details

Rotation function studies, whether the Fourier transform/ interpolate type, or the rotate / Fourier transform type, require that sums of the type

$$\sum_{i=1}^{N} f_i \cos(2\pi \mathbf{h}.\mathbf{x}_i)$$

be evaluated many times. In all the calculations described in this thesis, atoms were assumed to be the same shape (that of nitrogen) with the same temperature factor, and their coordinates were stored as integers representing fractions of the unit cell x 8192; all the carbon atoms were stored together, all the nitrogen together, etc. Sums such as

$$\sum_{\substack{\text{all carbon} \\ \text{atoms}}} \cos 2\pi \mathbf{h}.\mathbf{x}$$

were formed, multiplied by the atomic number of the atom type concerned, and then added to the corresponding term for nitrogen, etc., atoms. (Because this process meant that most of the required numbers could be stored in the c.p.u. of the 1906A, this process was nearly twice as fast as the formation of $\Sigma f_i \cos 2\pi \mathbf{h}.\mathbf{x}_i$ by multiplying by f_i each time; this was an idea of David Dover.)

A sine table in intervals of $2\pi/8192$ radians was used. (If the sine table does not contain the number of entries which corresponds to the accuracy of the storage of the atomic coordinates, this causes rounding problems in integer arithmetic; this value of 8192 gave F's agreeing to 1 part in 1000 with those calculated using sines accurate to 1 part in 10^9.) A subroutine in PLAN, the low-level language for the ICL 1906A, was written to form sums of this type; it was 2.5 times quicker than the FORTRAN version; 5000 structure factors with 858 atoms took 90 seconds to calculate.

(vi) Translation functions

(a) in vector space

Huber et al. and Nordman et al. have both used their correlation functions successfully on translation functions.

(b) The scalar product

Tollin, and Crowther and Blow, and Karle have used the scalar product of $F^2(\mathbf{h})$ and $F_c^2(\mathbf{h},\mathbf{X})$ (equation (18) page 24) as a translation function, calling them Q, T and D respectively. From equation (18) it may be seen that since the origin peak is the largest in the Patterson function, R(X) will have large values corresponding merely to the large values of P_c. Tollin first pointed out that these spurious peaks may be removed by removing the Patterson origin peak. Karle has expressed this condition by using

$$P(x) = \sum_{h} (E^2(\mathbf{h}) - 1) \cos 2\pi \mathbf{h}.x$$

Crowther and Blow have suggested that intra-molecular vectors in P may be calculated from the model and subtracted from P; their trial on myoglobin showed that this function (T_1 in their paper) has a better signal to noise ratio than when self-vectors are not removed (i.e. T).

This thesis follows Tollin in seeking the vector **X** required to move from the local to the crystallographic origin: Crowther and Blow, and Karle seek the vector between molecules $(-\mathbf{X}+A_i\mathbf{X}+d_i)$.

(c) The symmetry of $F_c^2(X)$

From equation (4) (page 14) we see that

$$F_c^2(h,\theta,X) = \sum_i \sum_j F_i(h,\theta)\, F_j^*(h,\theta)\, \exp\,(2\pi i h\, (A_i X - A_j X))$$

$$= \sum_i F_i^2 + \sum_i \sum_{j \neq i} F_i F_j^* \exp\,(2\pi i h A_i\, (1 - A_i^{-1} A_j)\, X) \qquad (22)$$

where * means complex conjugate. The A_i form a group, and so $A_i^{-1} A_j$ are members of the group. (It will be recalled, e.g. from Buerger, "Elementary Crystallography", pp493-514, that any combination of the rotation operations (i.e. the A_i) which is allowed in crystallography can be shown to behave as a group(a "point group" because we are not considering the translational symmetry associated with the d_i). It will also be remembered (e.g. from Buerger, p.475) that every element of a group has a unique inverse which is a member of the group (i.e. A_i^{-1}) and any product of elements ($A_i^{-1} A_j$) is also a member of the group.) Since one of the A_i^{-1} is the identity element, the $A_i^{-1} A_j$ include all the rotations in the group; because of the product rule mentioned above, the $A_i^{-1} A_j$ contains no rotations other than those of the group.

In equation (22) the important part as far as determination of the asymmetric unit is concerned, is $(1 - A_i^{-1} A_j)\, X = U$; for we must determine what range of X causes U to assume all possible values in the range 0 to 1. It is convenient to write as a

redefinition

$$(1 - A_i^{-1} A_j)X = (1 - A_k)X'$$

$X' = X$ when A_k is a rotation parallel to the crystallographic axes which define X; when the axis of A_k is not parallel to a crystal axis, X' is X redefined with respect to the axis of A_k and to two orthogonal directions.

For proper rotations it may be seen than an element of the vector $(1 - A_k)X'$ will be zero. (For improper rotations, which are not encountered in protein crystallography, this does not apply; for a mirror plane ($\equiv \bar{2}$) two elements are zero, while for other improper rotations no element is zero.) (A proper rotation is a pure rotation, whilst an improper rotation is the operation of a proper rotation followed by inversion in a centre of symmetry lying on that axis (Buerger, page 23).)

(d) $F_c^2(X)$ is a sum of n functions

From equation (22) and the above discussion we can see that $F_c^2(X)$ is a constant plus a sum of (n-1) two-dimensional functions; the parameters of these functions represent directions orthogonal to the rotation axes. The asymmetric unit of $F_c^2(X)$, and hence of $R(X)$, may be found by an examination of the elements of $(1 - A_k)X'$.

For example, in space group $P2_12_12_1$ if A_k represents a two-fold axis parallel to Z,

$$(1 - A_i^{-1} A_j)\,\mathbf{X} = (X,Y,Z) - (-X,-Y,Z) = (2X,2Y,0)$$

and it is sufficient to search in the range $0 \leqslant X \leqslant \frac{1}{2}$, $0 \leqslant Y \leqslant \frac{1}{2}$, and similarly $0 \leqslant Z \leqslant \frac{1}{2}$. It is probably a coincidence that this is the crystal asymmetric unit too.

In space group $P4_32_12$ the two-fold axes about x and y yield results analogous to those for $P2_12_12_1$ above, but the two-fold about (110) gives

$$(1 - A)\mathbf{X} = (X,Y,Z) - (Y,X,Z) = (X-Y,\ Y-X,\ 2Z)$$

which is a function of (X-Y) and Z; the component of X+Y(in the vector $\mathbf{X}' = (X+Y,\ X-Y,\ Z)$) is zero. The asymmetric unit for this function is thus $0 \leqslant (X+Y) < 1$, $0 \leqslant Z \leqslant \frac{1}{2}$. The n functions which add to form $F_c^2(\mathbf{X})$ thus do not necessarily have compatible asymmetric units. All equivalent values of $\mathbf{X}$ will be covered (it may be seen by considering all the rotations in the group) by allowing $\mathbf{X}$ to vary in the range $0 \leqslant X < 1$, $0 \leqslant Y \leqslant \frac{1}{2}$, $0 \leqslant Z \leqslant \frac{1}{2}$; this includes a duplication in X for (for example) the two-fold rotation axis parallel to Y. This duplication will cause ghost peaks at $R(X_o + \frac{1}{2},\ Y_o,\ Z_o)$ if (X_o,Y_o,Z_o) represents the correct value of $\mathbf{X}$, because at $(X_o+\frac{1}{2},\ Y_o,\ Z_o)$ some of the n functions which add to form F_c^2 will have the same values as they have at $\mathbf{X}_o$, whereas other of these n functions will have different values.

Because $F_c^2(\mathbf{X})$ is a sum of n functions, each representing one

symmetry element, $R(\mathbf{X}) = \sum F^2(\mathbf{h})\ F_c^2(\mathbf{h},\mathbf{X})$ may be factorised into n functions (Tollin, 1966) and each symmetry element may be sought separately. However, the signal to noise ratio will be greater (pace Crowther and Blow) if the two-dimensional functions (R(X,Y), R(Y,Z), R(Z,X)) are added to give the three-dimensional function (R(X,Y,Z)). This latter will reflect its factorisation in having contours in the form of cylinders parallel to the rotation axes; the axes of these cylinders meet at the maximum value of $R(\mathbf{X})$ - see figure (5). (Sabesan tells me that in the comparison of cod insulin with the pig insulin coordinates (see table (4), page 46) the 3 two-dimensional Q functions of Tollin did not unambiguously individually give the correct positions of the molecules, but that a compatibility check on the largest peaks in each function, or a sum of these to give a three-dimensional function, did give an unambiguous answer.)

(e) The asymmetric unit of the residuals

This will be the same as the asymmetric unit of $F_c^2(\mathbf{X})$, since all the residuals may be considered to be functions of F_c^2. These other functions of F_c^2 will not be factorisable as is R, but since R_4 (equation (10)) is closely related to R (equation (18)) it might be expected that R_4 will also show better correlation between F and F_c when only one element of $\mathbf{X}$ is incorrect, rather than when two or three elements of $\mathbf{X}$ are incorrect; furthermore, the other residuals

(equations (7), (8) and (9)) might be expected to show similar properties, as indeed they do in figure (5), page 53.

(f) Computing details

Professor North suggested that I try using the residual R_1 as a translation function, and he pointed out that the F_i in equation (4) may be calculated initially and that the calculation of $R_1(X)$ does not involve a sum over all atoms and all structure factors for each value of X. The F_i were calculated in a way similar to the rotation function (page 32) except that a PLAN routine was written to generate the four F_i for space group $P2_12_12_1$ from one list of atoms and the non-related values of h.

The evaluation of R_1 (equation (7)) for all values of X was performed by using a program specific to point group 222. In fact, two programs were written, one of which evaluated $(hX + kY + \ell Z)$ each time it was encountered and looked up the sine and cosine: the other stored intermediate values of $F_i \exp(2\pi i(hX + kY))$ and $\exp(2\pi i \ell Z_{step})$ etc, etc. They took very nearly the same time to do the same calculation, but the first used much less storage space, and was subsequently always used.

(vii) Refinement of rotational and translational parameters

An approximate value for ϑ, the rotational, and X, the translational parameters may be found by calculating a suitable correlation function

at all unique values of ϑ and X taken at suitable intervals, such as 15° or 2Å for X-ray reflections to a resolution of 6Å. Approximate values of ϑ and X may then be refined either manually by searching at finer intervals around the approximate position, or by an automatic minimisation (or maximisation) procedure. Although the correlation functions R_1 and R_3 (equations (7) and (9)) seem to have similar properties, the latter is more suitable for an automatic search for a minimum; most minimisation procedures assume that the function to be minimised is continuous and that its partial derivatives are continuous, which does not apply to the function R_1 because R_1 uses the modulus function, which is not ordinarily differentiable.

The Nottingham Algorithms Group library of generally useful subroutines is available on the 1906A, and one routine, E04DAF, uses the method described by Fletcher and Power in 1963; a direction of search in the n-dimensional space of n variables is chosen from the gradient at that point and from previous search directions, and the minimum of the function in that direction is found by approximation to a simple polynomial; another direction is then chosen. This may be used on $R_2(\theta,X)$ and on $-R(\theta)$ (equations 8 and (18)). This method has advantages over the usual least squares method in that (i) it is easily adaptable to maximising $R(\theta)$, (i.e. by minimising $-R(\theta)$). (ii) that it requires less storage in the computer, (iii) it is guaranteed to converge, and (iv) it was working properly when I started

using minimising techniques.

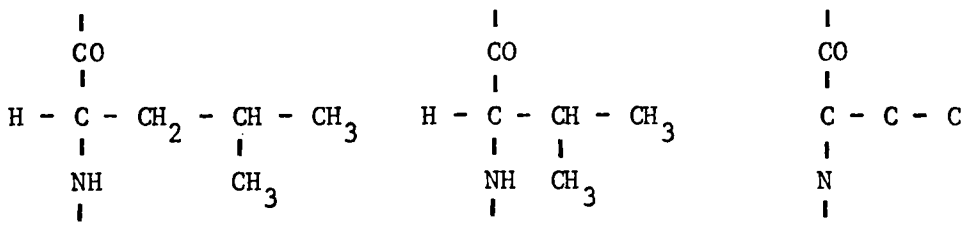
There is of course no guarantee that such a process will converge to a global rather than to a local minimum, but the use of the numerical method to calculate the differential coefficients means at least that a local minimum smaller than the stepping interval (e.g. 0.1^{o}) is not found.

2. APPLICATIONS OF ROTATION AND TRANSLATION FUNCTIONS

(i) Human lysozyme

(a) The partial structure used.

A preliminary model for human lysozyme was chosen by considering the primary sequences of human and hen lysozymes, and using those non-hydrogen atoms which might be expected to be the same in both species. All the backbone atoms were used, except for residues 47 and 48 between which an extra residue is inserted. (See table (1) page 3). Those side chains which were unchanged (N.B. glu ≡ gln, asp ≡ asn) were included completely, those which were different had only the C_β included, though some conservative changes had part of the side chain included. An example of the last case is 129 where leu (hen) → val (human)



and the C_β and C_α atoms of the side chain are used. This process produced a model of 858 non-hydrogen atoms, about 83% of the ordered non-hydrogen scattering matter. (There are 1028 non-hydrogen atoms in human lysozyme.)

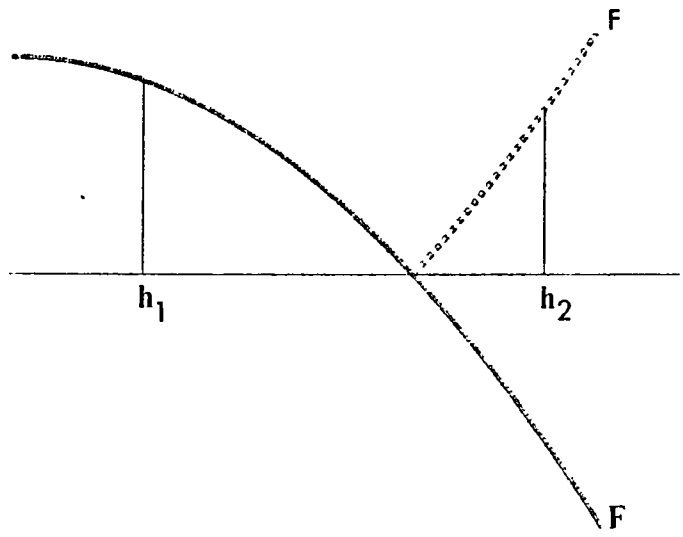


Figure (2)

Interpolation in F(h) in a centric data set.

(b) Rotation search over all Eulerian space

The Fourier transform of this model was calculated by putting one molecule into a hypothetical unit cell (an 80Å cube) and calculating structure factors to a resolution of 6Å. The Rossmann and Blow rotation function to compare these F_c^2 with the F^2 for human lysozyme measured by Stephen Banyard was calculated on the 1906A. The program was modified from one written for the Argus computer by Michael Joynson; it is basically the "fine mesh program" described by Tollin and Rossmann, where a look-up table of values of G (equation (20), page 25) is stored as a function of $|\mathbf{h} - \mathbf{h}'|^2$. The 80Å cube ensures sufficiently fine sampling of reciprocal space since the maximum inter-atomic distance in lysozyme is less than 40Å. Sayre in 1952 (A.C. page 843) pointed out that if for a centro-symmetric crystal we could know $F(\mathbf{h})$ at half-integral values of $\mathbf{h}$, we would know where $F(\mathbf{h})$ changed sign. If this is true, we can see that the wrong interpolation between $F(\mathbf{h}_1)$ and $F(\mathbf{h}_2)$ shown in figure (2) would not occur if $|\mathbf{h}_1 - \mathbf{h}_2| < \frac{1}{2}$. This applies separately to the real and imaginary parts of F(h) for a non-centrosymmetric crystal, but the error on interpolating the modulus is likely to be less severe. See also the last paragraph of Tollin and Rossmann.

The strongest 35 (ex 269) unique observed structure factors with resolution between 10 and 6Å were used in the outer summation in equation (20). Of the calculated structure factors, 27 were used

Source	proteins compared	space group	n	peak-mean in s.d.	peak / backgnd	(next highest peak)-mean	success of trans. search
page 45	human / hen lysozyme	$P2_12_12_1$	4	4.5	1.36	3.0	yes
page 58	orthorh./tetrag. lysozyme	$P2_12_12_1$	4	3.7	1.41	3.4	no
Joynson thesis	α-lactalbumin/ lysozyme	$P2_1$	4	3.1	1.21	3.4	no
Joynson thesis	orthorhomb.(data from Sarma) / tetragonal lysoz.^{m}	$P2_12_12_1$	4	4.1	1.16	3.3	
Lattman & Love	lamprey hemoglobin / myoglobin	$P2_12_12_1$	4	~6.0		~4.8	yes
Lattman et. al.	tuna / spermwhale myoglobin	$P2_12_12_1$	4	~7.1		~6.1	yes
Tollin	seal/sperm whale myoglobin	C2	2	~2.1	1.84	~1.3*	yes
Joynson et. al.	triclinic / tetragonal lysozyme	P_1	1	7.5	1.33	3.7	yes
Sabesan private comm.	cod / pig insulin hexamer	$P2_12_12_1$	4		1.25	nextpk /backgnd 1.10	yes †
"	hagfish /pig insulin monomer	$P4_12_12$	8		1.08	1.06	yes †

Table (4)

A comparison of some rotation searches

* Good results arise from the peak's being near a 2-fold axis in Eulerian space: when this effect is removed the correct & next highest peaks are "comparable".

† The knowledge of the subunit structure was of some assistance here.

The structure factors were to a resolution of ca. 6Å in all these cases.

in each inner summation to interpolate to the rotated position in reciprocal space of the observed values of h. (The G function (page 25) was that of a sphere of radius 30Å). Eulerian angles as defined by Rossmann and Blow were used to describe the various orientations of the model and the asymmetric unit in Eulerian space is 0 to 180° in all three angles. A basic stepping interval of 15° was used in all three angles, but this was increased when sin θ_2 was small. (When $\theta_2 = 0$, $\delta\theta_3$ was 360°; when $\theta_2 = 15°$, $\delta\theta_3$ was 90°; and when $\theta_2 = 30°$, $\delta\theta_3$ was 30°. This is similar to Lattman's treatment (see page 29).)

This search of 1212 orientations took 12 mins on the 1906A, and showed one peak of height 112 against a background of 82, with standard deviation 6.6; the next highest peak had a height of 102. This result is compared with others in table (4) which seems to support Lattmann and Love's observation that peaks more than one standard deviation above all other peaks are probably genuine. A search was also made with all the F's with resolution less than 10Å. There were six peaks between 2 and 3 standard deviations above the mean, one of which was near the correct peak.

(c) Refinement of the rotational parameters

The region around this peak was explored using shells of data with resolution in the ranges ∞ to 6Å, 10 to 6Å, 4 to 6Å, 3 to 4Å and 2.5 to 3Å, the 100 strongest unique structure factors being used in

Resolution Å	method of refinement	ϑ_+	ϑ_2	ϑ_-	Angular distance from 1st line	r.m.s. x from 1st line	no. of function evaluations
finally chosen	manual inter-polation	155.0	159.5	119.8	0	0	
10 to 6	manual interpolat-ion of $R(\vartheta)$	165	158	117	3.64	1.85	482 a
6 to 4	"	151	158.7	119.5	1.12	0.30	64
4 to 3	"	156	159.4	120.5	0.72	0.30	29
3 to2.5	"	152	160.0	119.6	0.73	0.36	74
10 to 6	E04DAF on $-R(\vartheta)$	170	155	122	5.79	2.93	37 b
to 6	"	170	156	122	4.35	2.03	123 b,c
6 to 4	"	158.3	159.5	119.2	1.33	0.44	53 d
to 6	E04DAF on $R_3(\vartheta,X)$	155.6	159.5	120.0	0.46	0.15	24

Table (5)

Refinement of rotational parameters for hen/human lysozyme

E04DAF is a NAG library routine (see section 1(vii))

a This number would have been smaller by about $\frac{1}{2}$ had it been realised sooner that rectangular blocks of $(\vartheta_+,\vartheta_2,\vartheta_-)$ are better in this case than blocks of $(\vartheta_1,\vartheta_2,\vartheta_3)$

b Starting from 180 150 120, the angles found on the first coarse grid

c The last 96 function evaluations caused changes of less than 0.01°

d Starting from 170 156 122

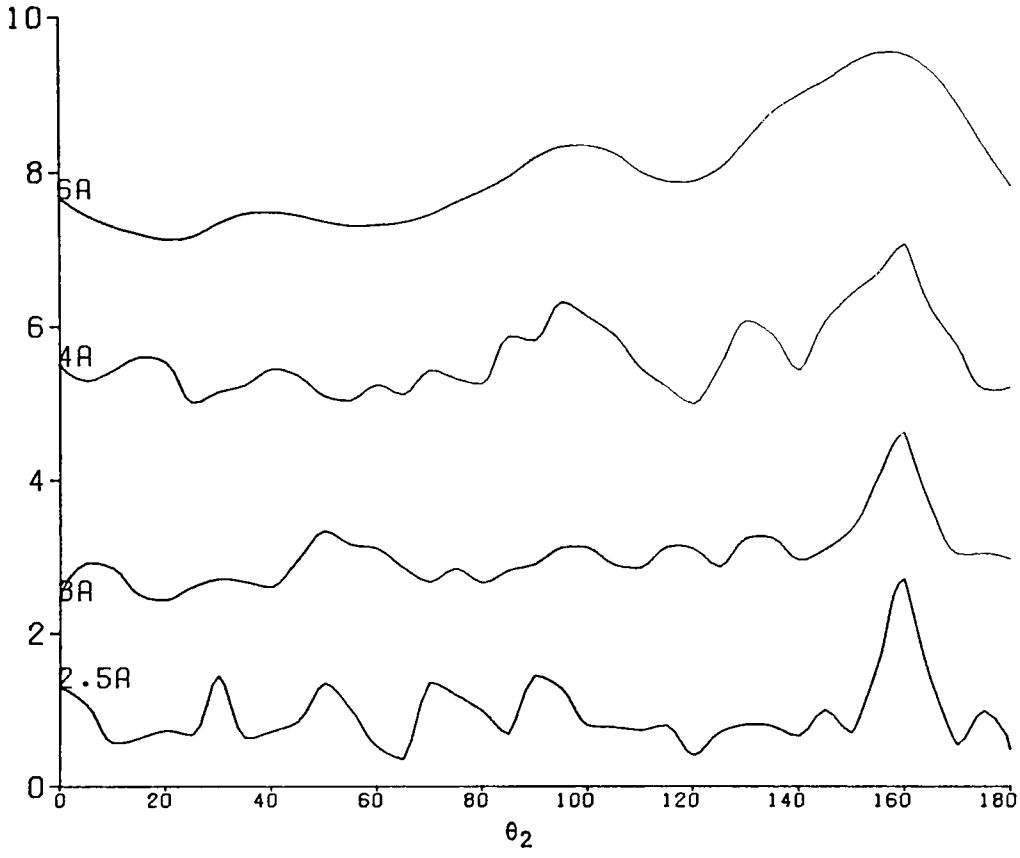


Figure (3)

Sections through the human/hen lysozyme rotation function at constant θ_1 and θ_3. (The vertical scale is arbitrary)

each case. Since only a few orientations are being searched it is quicker to calculate the F_c as they are required rather than to calculate them all and interpolate. Because sin θ_2 is small it is more convenient to search with parameters $(\theta_1 + \theta_3)$, θ_2, $(\theta_1 - \theta_3)$ since the orientation is not very sensitive to the value of $(\theta_1 + \theta_3)$ (see page 29). The three shells of higher resolution structure factors gave orientations in close agreement, but differing slightly from the low resolution position (Table (5)). Automatic refinement was also tried using the minimisation routine mentioned on page 40 on $-R(\theta)$. As may be seen from table (5), automatic peak position refinement works slightly quicker than manual refinement for 3 parameters. Figure (3) shows sections through the finally chosen peak as functions of θ_2 for various shells of structure factors. It may be seen from the sharpness of these peaks that a precision of less than about 0.5° is not meaningful.

(d) Translation search - coarse

A translational search was next made using all the structure factors with resolution greater than 6Å; the residual R_1 and the scalar product R were calculated from the 858 atoms rotated as in the first row of table (5). There are eight equivalent origins in $P2_12_12_1$, and the asymmetric unit for the translation function is 0 to $\frac{1}{2}$ in each direction. This was searched in 20 steps in x and in y and 10 steps in z (about every 1.5Å). A minimum of 54% in the residual was found

Scan	shift in 1/480 of unit cell edge x	y	z	ϑ_+	ϑ_2	ϑ_-	R_1
finally chosen	184	3	252	155	159.5	119.8	see table(9)
coarse grid 10 to 6Å	180	12	240		"		0.54
finegrid							
10 to 6Å	185	4	248		"		0.49
6 to 4Å	184	3	253		"		0.45
4 to 3Å	184	1	252		"		0.45
3 to2.5Å	184	3	250		"		0.48
E04DAF on $R_3(\vartheta, X)$ to 6Å	184.4	3.4	252.1	155.6	159.5	120.0	

Table (6)

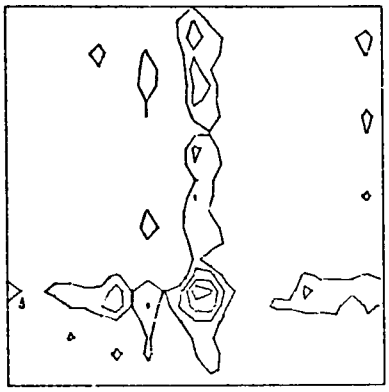
Translation function (residual) results for
hen / human lysozyme
(One unit is ca. 0.12Å in x or y, 0.06Å in z)
E04DAF is a NAG library routine - see page 39.

at this sampling interval; the mean was 69% and the standard deviation 2.9%, Finer sampling reduced the residual to 49%. On the other hand, although R did show a maximum at this point (peak/background = 1.21) it was the sixth highest maximum in the asymmetric unit. (None of the other maxima in R corresponded to a residual of less than 67%.)

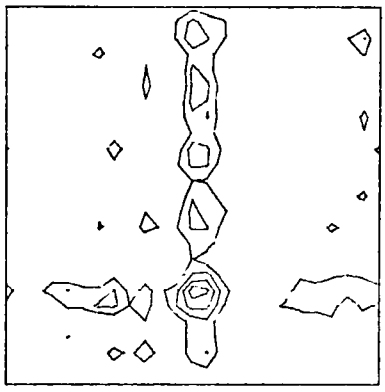
The vector describing the shift necessary to move the rotated model to its correct position with respect to one of the 8 crystallographic origins has now been found. Any of the 8 origins could be used but it is convenient to refer to the one that Ian Swan and Stephen Banyard used in their isomorphous replacement studies. Although the structure factor amplitudes are the same for each of the 8 origins, the phases are different, and the mean phase difference between the calculated and isomorphous replacement phases was 51^{o} for one origin and between 83^{o} and 94^{o} for the other 7. Subsequent work refers to the Swan/Banyard origin.

(e) Translation search - fine

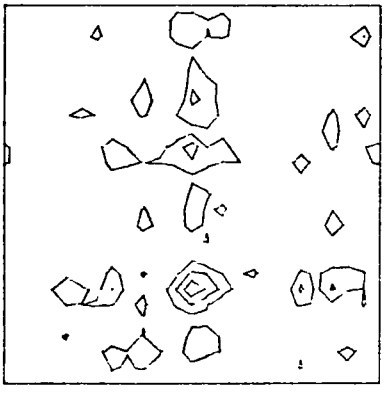
The region in the translation function around the hole in R_1 was searched using shells of higher resolution structure factors, as for the rotational searches, but using all the measured structure factors in each shell; see table (6). The scale factor for R_1 was such that $\sum F_c = \sum F$. (See page 77 for more discussion of scale factors.)



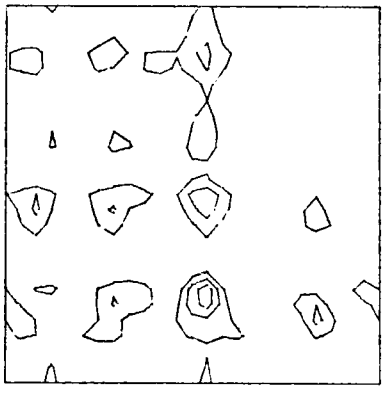
6A TRANSLATION FUNCTION (RESIDUAL)



6A TRANSLATION FUNCTION (SUM (FO-FC) **2)



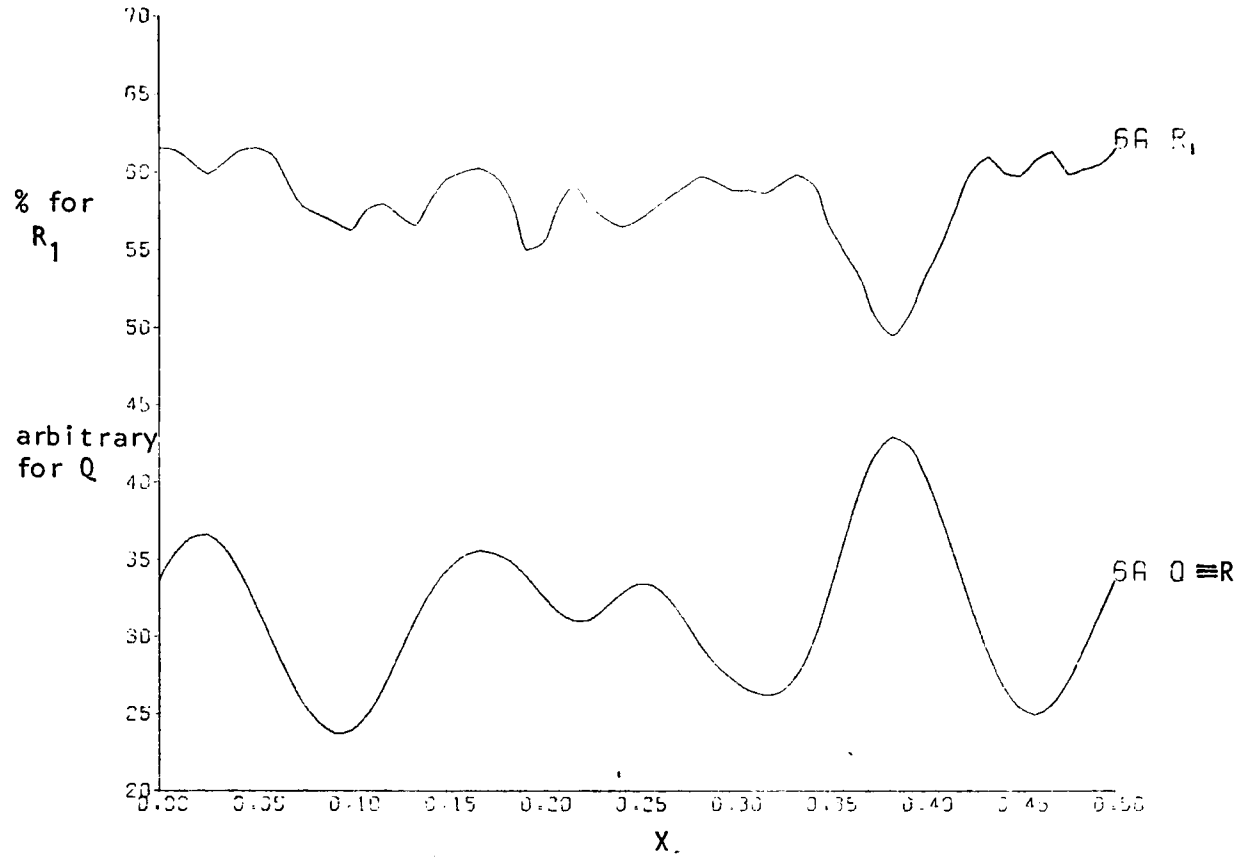
6A TRANSLATION FUNCTION (SUM (FO+FO-FC+FC) **2)



6A TRANSLATION FUNCTION (SUM FO+FC+FC+FC)

Figure (5)

Sections through the main feature of several translation functions. All the 6-10Å F's were used. Contours are in units of 1 standard deviation



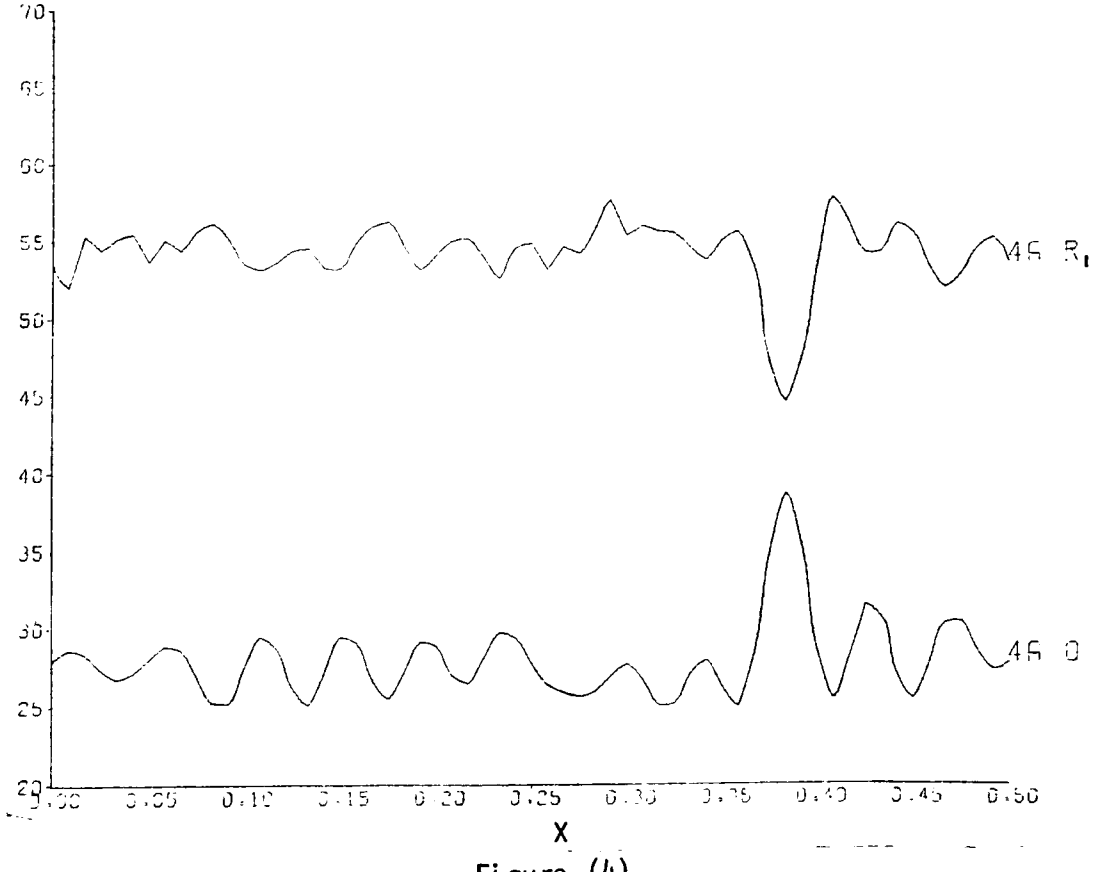
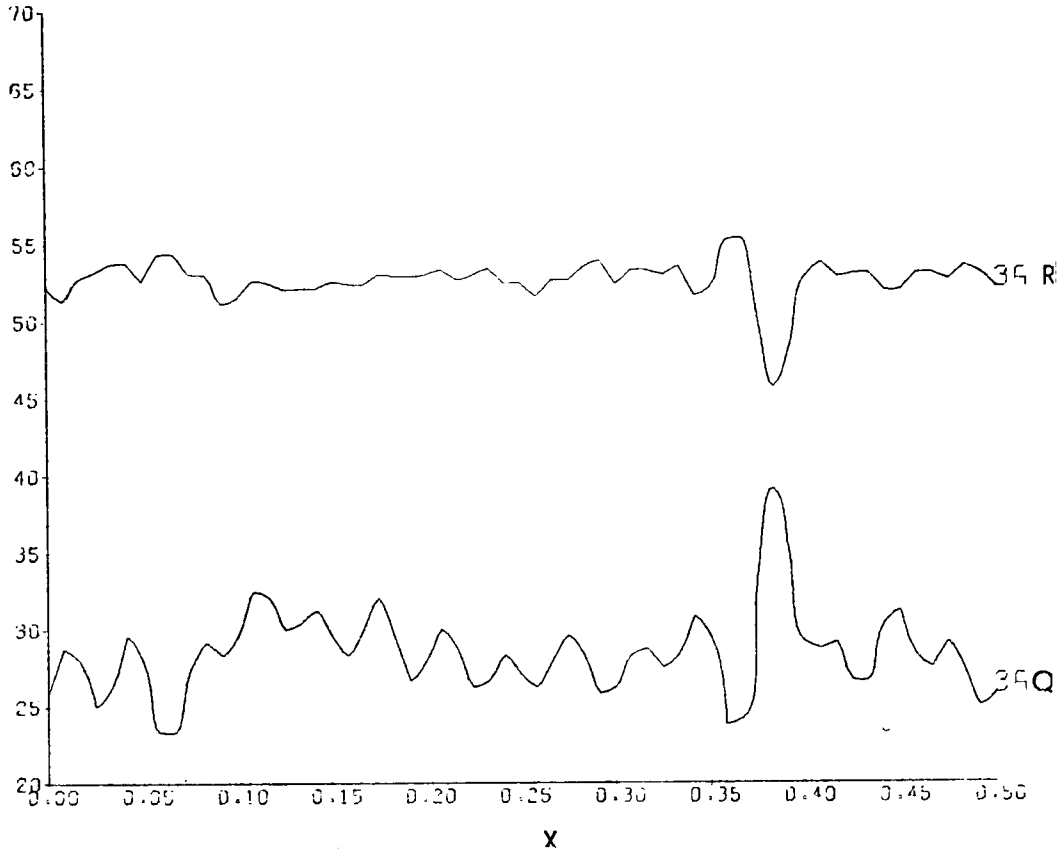


Figure (4)
Sections through $R_1(X)$ and R(X) at various



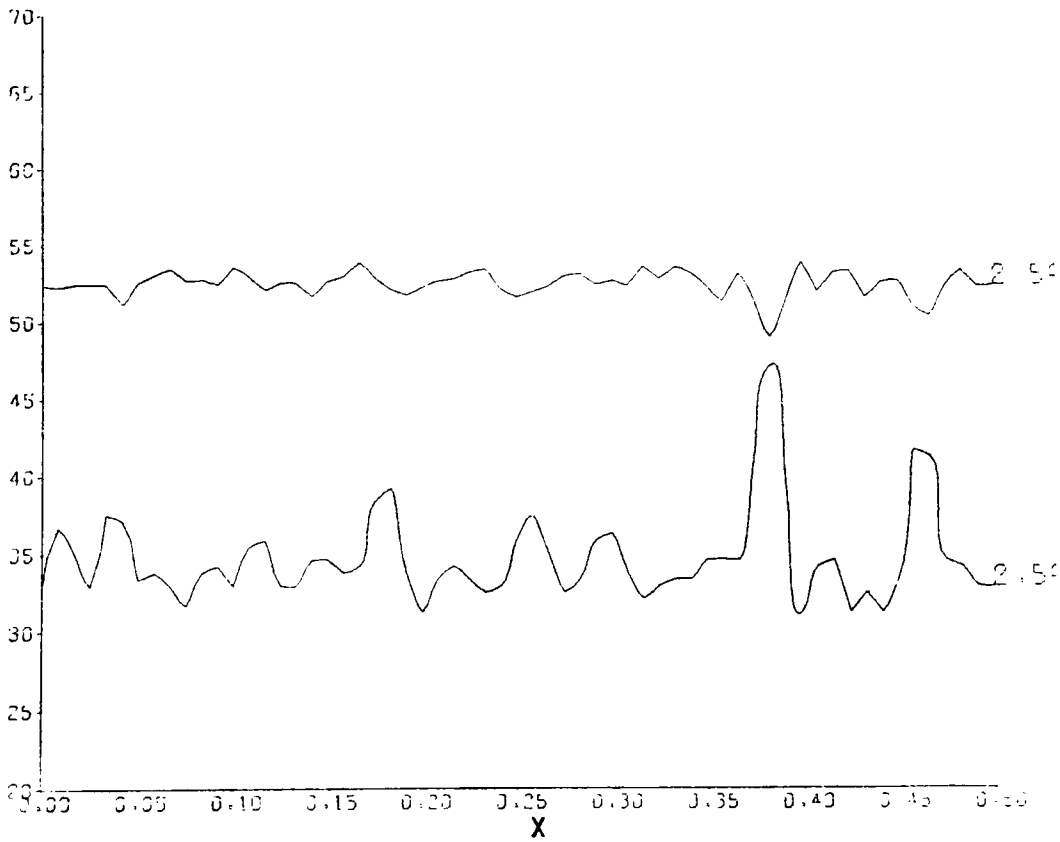


Figure (4) continued

Figure (4) shows sections through the finally chosen position for various resolutions of data for the residual R_1 and the scalar product R; the residual is clearly a less noisy function. Figure (5) shows sections of various residuals and the function R for 6 to 10Å structure factors. They are clearly similar, but those functions using the lower powers of the structure factors are less noisy. The largest false peaks or holes lie in positions where one element of the shift vector is incorrect, as explained on page 38.

(ii) Orthorhombic hen lysozyme

The calculation of the rotation function was performed in exactly the same way as for human lysozyme, except that all 1001 atoms of hen lysozyme were used as the model, and 45 unique structure factors were used because the unit cell is slightly larger. Five peaks were found which were more than 2 standard deviations above background. The three highest of these were investigated at smaller angular intervals, and translation functions (i.e. the residual R_1) were calculated for these using all the 6Å structure factors. No value of the residual less than 66% was found, and a residual as high as 66% is probably not significant.

The results are given in table (7). It is not at all clear why no fit to the data could be found in this case, which should have been better than human lysozyme. Exactly the same programs were

First $\theta_1, \theta_2, \theta_3$ (degrees)	Peak - mean in standard deviat^n	peak/background	refined $\theta_1, \theta_2, \theta_3$	mean R_1	lowest R_1	standard dev. of R_1
30 90 135	3.7	1.41	34 95 135	0.76	0.66	0.028
30 45 60	3.5	1.39	30 48 69	0.77	0.68	0.025
120 120 60	2.9	1.32	121 118 52	0.77	0.68	0.025
165 105 0	2.7	1.30				
75 90 30	2.6	1.29				

Table (7)

Summary of the rotation and translation function studies of orthorhombic lysozyme and the tetragonal lysozyme atoms.

used in both investigations; the two space groups are the same, and the unit cell dimensions comparable.

(iii) Goat α-lactalbumin

(a) Previous work

Michael Joynson had already found a very pronounced non-crystallographic two-fold axis parallel to the crystal axis c by using the rotation function described on page 15. He had calculated rotation functions to compare α-lactalbumin with lysozyme, as described in the section on human lysozyme; he tried using all the atoms of hen lysozyme, and using just the backbone atoms. There were found a number of peaks in these functions which occurred in pairs separated by 180° in θ_3, which corresponds to this non-crystallographic dyad. He had refined the peaks using 4Å structure factors and chosen one of the highest three because of packing considerations. Translation functions to determine the screw component of the non-crystallographic rotation axis (see page 15) could not be clearly interpreted, and neither could Tollin Q functions (see page 34).

(b) Rotational parameters

A model was proposed, consisting of 655 of the hen lysozyme atoms, chosen in a similar manner to those for human lysozyme. (Unfortunately the sequence of goat α-lactalbumin is not known, but that of bovine α-lactalbumin, table (1), was used instead.) The

Source	initial θ_1	θ_2	θ_3	R	refined θ_1	θ_2	θ_3	R	X_1	X_2	Z_1	Z_2	Y	R_1
6	170	40	130	512	170.5	53.1	138.8	629	-	-	60	240	180	47%
									-	-	150	180	0	43%
									384	120	150	180	0	57.1%
									-	-	390	180	240	43%
									360	96	0	-	-	58%
4	140	44	153	507	151.1	43.7	155.2	556	-	-	30	210	120	41%
									-	-	210	210	375	43%
									-	-	390	180	240	43%
6	0	140	150	433	5.1	125.4	156.3	520	-	-	300	30	225	45%
									-	-	360	90	120	45%
10	75	30	45	456	63.9	35.0	51.9	509	NONE < 49%					
6	20	150	170	395	23.1	130.0	161.0	501	NONE < 49%					
10	150	60	45	450	146.0	58.0	41.0	465	-	-	270	90	450	40%
									-	-	270	90	210	39%
									-	-	30	210	210	45%
									-	-	240	30	210	44%
									-	-	390	180	330	45%
10	45	135	0	436	44.7	130.8	4.3	465	-	-	270	90	390	47%
6	60	100	270	320	74.8	125.0	288.1	450						
6	90	80	180	359	90.2	82.7	187.3	394						
6	40	160	10	357	43.4	174.4	3.4	379						
6	90	80	80	323	84.8	89.8	81.0	366						

Table (8)

Summary of the α-lactalbumin / lysozyme rotation and translation function results.

sequence after residue 115 shows many changes and deletions, and was not included. A rotation function was calculated from the largest 30 structure factors with resolution less than 10Å, using the Rossmann and Blow rotation function program. It showed three pairs of peaks higher than 2 standard deviations (σ) above background, two pairs being more than 3σ above background.

These three peaks are shown in table (8) by "10" under "source"; the seven highest peaks from Joynson's 6Å searches are shown by "6" in this column, and his finally chosen peak, refined by 4Å data, by "4" in this column. These eleven peaks were investigated using the 100 largest E's in the 10 to 4Å shell; these were F's sharpened by a temperature factor of $\exp(13.2 \sin^2\theta/\lambda^2)$ and this meant that most of the chosen E's were in the shell 4.5 to 4Å. The two peaks corresponding to the non-crystallographic two-fold were averaged, and the automatic minimisation of $-R(\theta)$ proceeded as summarised in table (8). Most of the peaks were moving by less than 0.1° after about 70 calculations of $R(\theta)$, but the peaks near the highest maximum of R needed more function evaluations; this is consistent with the fact that this peak in Michael Joynson's rotation function spreads lumpily over about 20° in each direction, especially along θ_1, and that his values for the whole molecule, and for main chain only, disagree by 7° in θ_1. The results of these searches are summarised in the first 9 columns of table (8).

Resolution Å	no. of molecules	$R_{1\,true}$	$\frac{\bar{R}_1 - R_{1\,true}}{\sigma}$	$\frac{\bar{R}_1 - R_{1\,true}}{\bar{R}_1 - R_{1\,next}}$	$R_{3\,true}$	$\frac{\bar{R}_3 - R_{3\,true}}{\sigma}$	$\frac{\bar{R}_3 - R_{3\,true}}{\bar{R}_3 - R_{3\,next}}$	$\frac{R_{true}}{\bar{R}}$	$\frac{\bar{R} - R_{true}}{\sigma}$	$\frac{R_{true} - \bar{R}}{R_{next} - \bar{R}}$
10 to 6	8	0.39	7.6	1.8	0.18	7.8	2.2	1.50	4.0	0.98
10 to 6	4	0.49	4.4	1.4	0.30	4.7	1.4	1.27	2.8	0.55
10 to 4	8		9.6	2.4		9.5	2.9	1.93	9.5	2.0
10 to 4	4		4.5	1.8				1.47	7.0	1.4

Table (9)

Test translation functions on tetragonal lysozyme

All the 10 to 6Å structure factors were used, but only the largest 25% of the 10 to 4Å structure factors.

(c) Translational problem

The translational problem is made more difficult by the fact that there are two molecules per asymmetric unit so that the screw component in the non-crystallographic two-fold axis is not known, i.e. that some of the d_i in equation (5) on page 14 are not known. The problem is in fact a 5-dimensional one; it may be expressed as the need to find X_1, X_2, Z_1, Z_2 and Y. X_1 and Z_1 define the relation of one molecule of the non-crystallographic pair to the crystallographic 2_1 axis; X_2 and Z_2 refer to the other of this pair, and Y defines the position along the b axis of this second molecule. (One parameter referring to distances along b is redundant, and taking it to be zero means that the first molecule defines the origin along b.) A complete 5-dimensional search would take too much computing time. One way to proceed is to take one molecule of the pair at a time; another way is to use projections along a and c.

(d) A test on tetragonal lysozyme

In order to see whether or not taking only one molecule of the pair is likely to be useful, a test was made using tetragonal lysozyme structure factors to 6Å. Translation function sections R(X,Y,0), R_1(X,Y,0) and R_3(X,Y,0) were calculated using all 1001 atoms with all 8 molecules in the cell, and again with only 4 molecules (those related by a 4-fold axis). The results summarised in table (9) show (i) that R_1 and R_3 are very similar (R_2 (not included in the table)

was very similar, too), (ii) that leaving out half of the molecules in the unit cell does not prevent the translation function $R_1(X)$ from finding the correct positions of the other molecules, and (iii) that R_1 and R_3 are better than R at discriminating between true and spurious peaks, both when all the molecules are included, and also when half of them are omitted from the calculation. This is, of course, a more favourable case than α-lactalbumin, for the lysozyme atomic coordinates were derived from the structure factors used in this test.

(e) Translational searches for one molecule

As mentioned above, Michael Joynson had already tried to find one of the molecules in the crystal asymmetric unit by using the function R (equation (18)), but the last column of table (9) indicates that this is unlikely to be successful.

The two highest peaks in the rotation function searches (table (8)) were searched by calculating $R_1(X_1,Z_1)$ for each of the orientations $(\theta_1,\theta_2,\theta_3)$ and $(\theta_1,\theta_2,\theta_3+\pi)$, but no value of R_1 less than 86% could be found.

(f) Translational searches in projection for two molecules

Another approach was to use projections along a and c, which reduces the problem to a search for X_1, X_2 and Y (or Z_1, Z_2 and Y) using two-dimensional data. A check is that the two independent calculations should give the same value for Y. The projection

along b was not used because it only reduces the problem to one of four dimensions (X_1, X_2, Z_1, Z_2). For the seven highest peaks in the rotation search (see table (8)), searches for Z_1, Z_2 and Y were made using the 57 structure factors to 6Å resolution with h = 0, and for X_1, X_2 and Y using the 83 structure factors with ℓ = 0.

In the latter search, only one value of R_1 less than 58% could be found; the values of X_1, X_2 and Y at which this occurred is shown by the fifth row of table (8). A two-dimensional search for the parameters Z_1 and Z_2 was performed, holding fixed these values of X_1, X_2 and Y; this search used all the three dimensional structure factors, out to 6Å, but no value of R_1 less than 58% was found. Of the 794 structure factors with resolution out to 6Å, 62 are centric, the rest acentric. Appendix B shows that we should expect the residual R_1 for a randomly wrong structure to be 60.4% with this proportion of centric data, and a residual of 58% is almost certainly not significant.

The results of the searches for Z_1, Z_2 and Y, using the structure factors corresponding to the projection along a are shown in columns 12 to 15 of table (8). Each of the sets of values for Z_1, Z_2 and Y shown in the table were used as constants in a search for X_1 and X_2 using three dimensional structure factors. Only one value of R_1 for the three dimensional data which was less than 58% was found in all these searches; it is indicated in row 3 of table (8). The value of

R_1 = 57.1% is not sufficiently below the random value (60.4%) to be meaningful.

These searches may thus be taken to have failed to account for the measured structure factor amplitudes by postulating a model based on lysozyme.

(g) Nordman's results

Professor C.E. Nordman has also tackled this problem using his Patterson search methods. He has come to no firm conclusion, although the highest of his rotation function peaks is split into two peaks at (in Rossmann and Blow's Eulerian angles)

61, 43, 69 and 87, 27, 31

which are close to the 4th peak in table (8).

(iv) Thiostrepton

(a) Previous work

This antibiotic crystallises in two crystal forms, in space group C_2 (a = 21.757, b = 22.846, c = 23.523 Å, $\beta = 103.94^\circ$) and in space group $P\,4_1 2_1 2$, (a = b = 26.36Å, c = 27.5Å). Its structure had been all but solved by Brian Anderson in the C_2 crystal form; he had found 4 sulphur atoms by examination of the Patterson function and phased from these. He had eventually identified 115 atoms, most of which formed a rigid structure of two nearly parallel rings, but with a long side chain which stuck out into the inter-molecular

liquid; the last few atoms were not clear in the electron density map. He had measured X-ray structure factor amplitudes for the second crystal form ($P4_12_12$) and had calculated a rotation function which showed that the molecule found from the first crystal form needed to be rotated by angles 108, 58, 76° in the Eulerian system described on page 29. Tollin Q functions (see page 34) had been tried in two dimensions to determine the translational parameters of the molecule; the results indicated a most likely solution, but it was not unambiguous.

(b) Translation searches

I calculated a translation function of the form $R_1(X,Y,Z)$ using these 115 atoms and the 112 structure factors with resolution out to 4Å. The mean value of R_1 was 80.6%, standard deviation 4.37% with one hole of 57%, which reduced to 53% on finer sampling. Of this particular set of structure factors, 63% belong to centric zones, and the corresponding value for R_1 for a randomly wrong structure is 73.9% (see Appendix B, page 169). The hole in R_1 is thus well below the level expected for a randomly wrong structure. It may be pointed out that the mean value of R_1 is rather higher than that predicted by appendix B. (This appendix assumes that the distribution functions of F and F_c are the same; this assumption is thus probably not true in this instance. Perhaps the distribution functions of F and F_c are different because no account of the solvent

was made in the calculation of the F_c's; this would cause the model to have too great a contrast at low resolution, making the low resolution F_c's too big. Errors in measurement of the F's may also contribute to the difference in distribution functions.)

(c) False holes

The lowest 2 false holes in this calculation of $R_1(\mathbf{X})$ were 66%, considerably above the value of 53% (or even 57%) which is assumed to represent the correct value of $\mathbf{X}$. This correct value is at (0.656, 0.469, 0.281), and agrees with the value of $\mathbf{X}$ determined from two-dimensional Q functions (Tollin) by Brian Anderson and Wendy Hutcheon (unpublished). One of these false holes was at (0.156, 0.469, 0.281), just half a unit cell away in X; that a false hole might occur in such a position is explained on page 37.

3. REFINEMENT OF HUMAN LYSOZYME

This chapter examines methods by which a partial structure, located in the unit cell as described in chapters 1 and 2, might be completed and refined to form the whole, correct structure. Tests are made on human lysozyme, and comparisons are made with the multiple isomorphous replacement results; the purpose of these comparisons is to check the methods used, rather than to attempt to combine these methods with the multiple isomorphous replacement method. It is hoped that the experience gained in the use of the methods described in this chapter will be generally useful when no parallel isomorphous replacement study is available.

(i) The first approximation to human lysozyme

The 858 atoms selected from the structure of hen egg lysozyme as described in chapter 2 rotated and translated as in table (6) form a first approximation to the structure of human lysozyme, and structure factor phases and amplitudes can be calculated from them. The phases can be compared with those determined by isomorphous replacement by Ian Swan and Stephen Banyard. As mentioned on page 52 the mean difference between the 6Å structure factor phases from isomorphous replacement and from the model was 51°. Figure (12) on page 94 shows how this increases with $2 \sin \theta/\lambda$. Figure (6) shows the mean difference between α_{calc} and α_{iso} of the structure

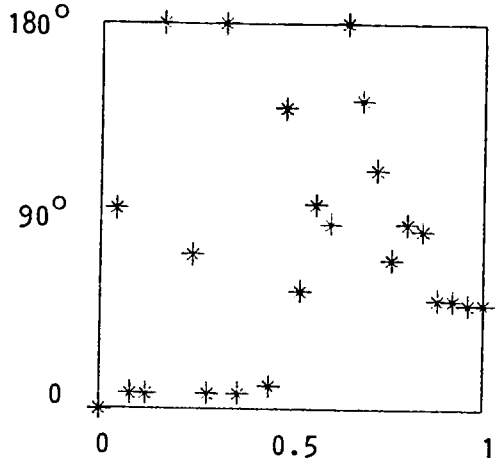
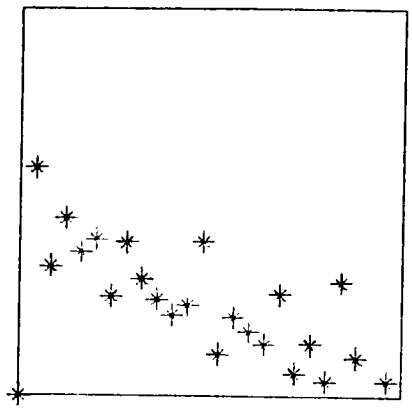


Figure of merit m



F (linear scale)

Figure (6)

Mean difference between M.I.R. phases and calculated phases as functions of M.I.R. Figure of merit and of F. (In the figure of merit graph the four points corresponding to highest m are the average of more than 30 F's: the other points are average of less than 5)

factors with resolution less than 6Å as a function of F; the agreement is better for large F. Figure (6) also shows that structure factors with figure of merit in the isomorphous phase determination less than 0.8 are not correlated with the α_{calc}.

(ii) The average error in the atomic coordinates

(a) Wilson statistics

Much of the theory discussed in this thesis requires that the distribution function of the structure factors be known, or at least assumed. This applies especially to the discussion of the variation of the residuals with D and σ_1^2 (page 19), to the derivation of Sim weighting (page 97), and to the discussion of the likely peak heights of various maps (page 114). Wilson in 1949 derived distribution functions for the centric and acentric cases, assuming a large number of randomly distributed atoms. His results are for acentric data

$$P(F) = \frac{2F}{\Sigma} \exp(-F^2/\Sigma) \qquad (23)$$

and for centric data

$$P(F) = 2(2\pi\Sigma)^{-\frac{1}{2}} \exp(-F^2/2\Sigma) \qquad (24)$$

where $\Sigma = \Sigma f_i^2$

It may be shown from both equations (23) and (24) that the mean value of F^2 is Σ, i.e.

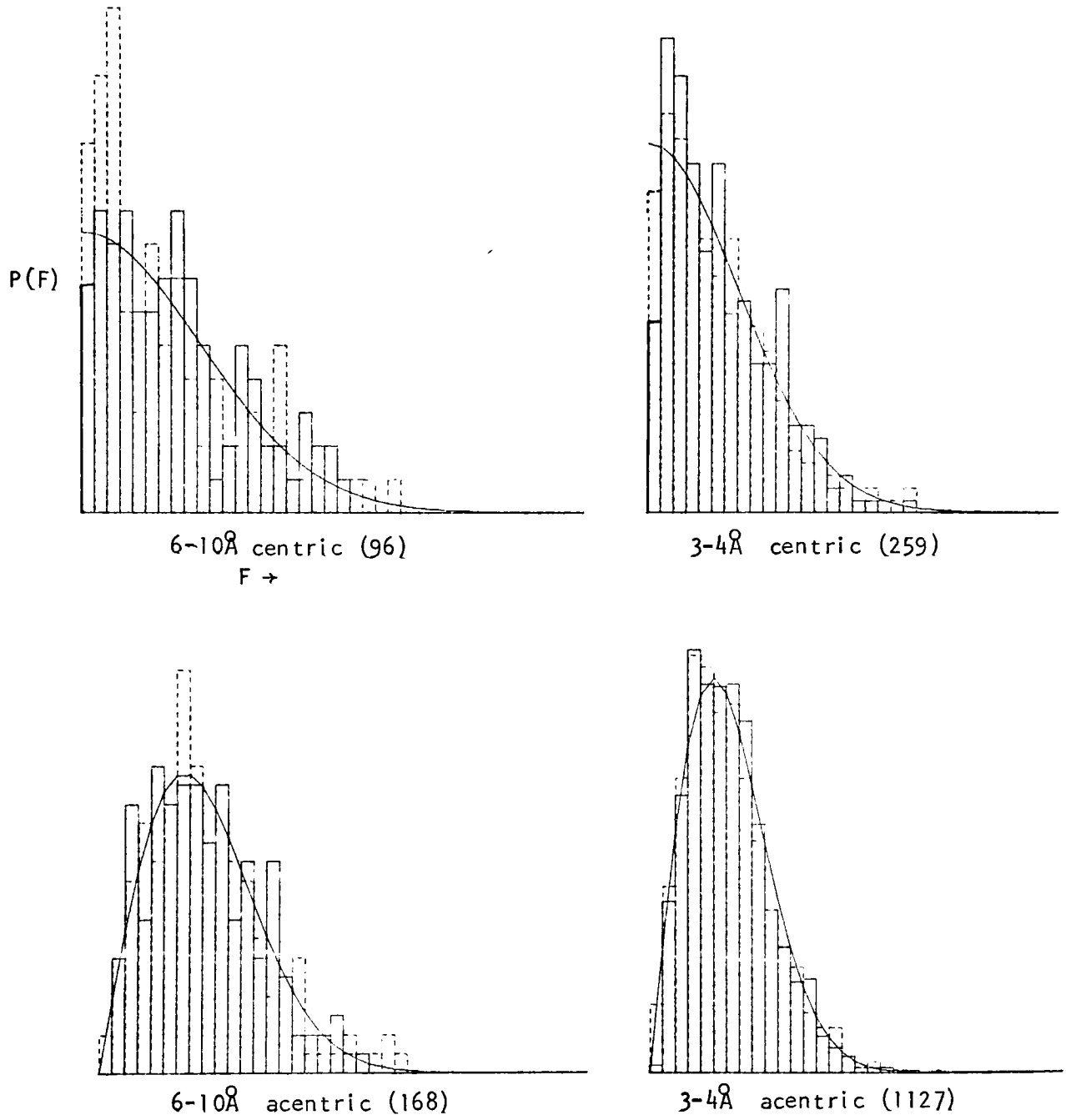


Figure (7)

Probabilities of F as a function of F

Solid histogram is observed F's

Dotted histogram is F_c

Curve is Wilson's (1949) distribution

Number in brackets is number of F's in

$$\langle F^2 \rangle = \int_0^\infty F^2 P(F)\,dF = \Sigma \qquad (25)$$

(see for example Woolfson, page 234).

The opinion that "Wilson statistics do not apply to proteins" is sometimes heard, and so a comparison of equations (23) and (24) with the distribution functions of the measured structure factor amplitudes of human lysozyme was made. A comparison with the set of F_c, determined as in the previous section, was also made. Figure (7) shows that the 6 to 10 Å and 3 to 4 Å shells of structure factors obey these distribution laws quite well; the 6 to 4 Å shell is very similar. The 6 to 10 Å shell of structure factors has more erratic histograms than the 3 to 4 Å shell because there are fewer structure factors in this shell.

Luzzati in 1955 discussed how the distribution functions changed when the atoms were considered to be confined to some regions in the unit cell, rather than randomly distributed over the whole cell; he concluded that this reduced the number of very small F's. The 3-4 Å centric structure factors show to a slight extent that there are less very small F's than Wilson's theory predicts, but random fluctuations is probably a sufficient explanation of this.

The above discussion says nothing, of course, about the distribution of the structure factors corresponding to the unknown atoms and to the errors in the known atoms; this distribution could only be known when the structure was completely solved by looking back,

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 to2.5	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
10 to 6	both	peak/backgnd		rot.fn. 1.36 $\Rightarrow$D=0.60 trans.fn. 1.21 $\Rightarrow$D=0.45			

Table (10)

The values of the residuals R_1, R_3, and R_4 (equations (7),(9),and (10)), and the scalar product R (equation(18)) for the 858 atom model of human lysozyme, and the corresponding values of D (page 19). The F and F_c were scaled so that $\sum F_c^2 / \sum F^2 = 858 / 1028$; σ_1^2 was taken as 858 / 1028 in the derivation of D from R_3 and R_4 (equations (11) to (14)).

and could never be used as an aid to progress. However it probably follows equations (23) and (24) at least as well as the distributions in figure (7).

Table (10) shows the values of R_1, R_3 and R_4 with the corresponding value of D (see table (2) and equations (11) to (14)). When the scaling was such that $\Sigma F_c = \Sigma F$ the values of the R's were larger by about 0.02 for R_1 and R_2, and by 0.05 to 0.4 for R_4. Luzzati in 1952 (his equations (50) and (51)) showed how an r.m.s. Δx could be obtained from D by assuming a normal distribution of Δxj's, but which is unlikely to be found in this case. However, values of an average Δx were calculated according to Luzzati's 1952 theory from these values of D, and were between 0.35 and 0.9 Å; the variations with $\lambda/2 \sin \theta$ and between 2- and 3- dimensional data are less than Luzzati's theory predicts.

(iii) Wrongly Placed Atoms in the Model

Having now a set of 858 atoms as a plausible model for the human lysozyme structure, and the orientation and position of this model in the unit cell having been determined so that the agreement between measured and calculated structure factor amplitudes is greatest, it is possible that some of these atoms are wrongly placed.

One way to determine which, if any, of the atoms in such a model are wrong is to remove each in turn from the model, to calculate a correlation function (page 15) of the remainder, and to regard

as wrong those whose omission improved the correlation. But the atoms are not independent in their effect on a correlation function; having omitted one atom, the effect of omitting a second might be different from that when the first atom was included in the model.

Rather than assigning a weight of either 1 or 0 to each atom, a continuously varying weight might be assigned, thus generating a correclation function of about 1000 independent parameters, whose minimum is to be found. This is similar to a suggestion made by Rollett and Hodgson in 1962 but is clearly an impracticable task in this case. It is plausible that some local regions in the model might be wrong, and so the atoms might be grouped into small regions - such as one or two amino acid residues - and all the atoms in such a region given the same weight. Most of the regions would be expected to have a weight close to unity. An automatic minimisation procedure could thus be employed, with all the weights initially set to 1.

Sixty-four regions were defined, each containing one or, usually, two amino acid residues. An initial value of 1 was assigned to each of the 64 parameters, and the Fletcher and Powell method (page 40) of minimising the correlation function R_3 (equation (9)) was used with all 344 6Å structure factors. That is, $F_c(h) = k \sum_i^{64} w_i F_i(h)$ where $F_i(h)$ is the contribution of the i^{th} region to the h^{th} structure factor, and $\Sigma(F_c - F)^2$ is minimised with respect to the w_i. The scaling factor k between F and F_c was such that $\Sigma F = k\Sigma\left|\sum_i F_i\right|$.

Cycle no.	R_3 arbitrary units	R_1	r.m.s. shift from last cycle
0	498	0.476	
1	332	0.428	0.207
2	302	0.411	0.081
3	274	0.394	0.087
4	267	0.390	0.051
10	254	0.388	0.004

Table (11)

Refinement of weights of parts of the molecule.

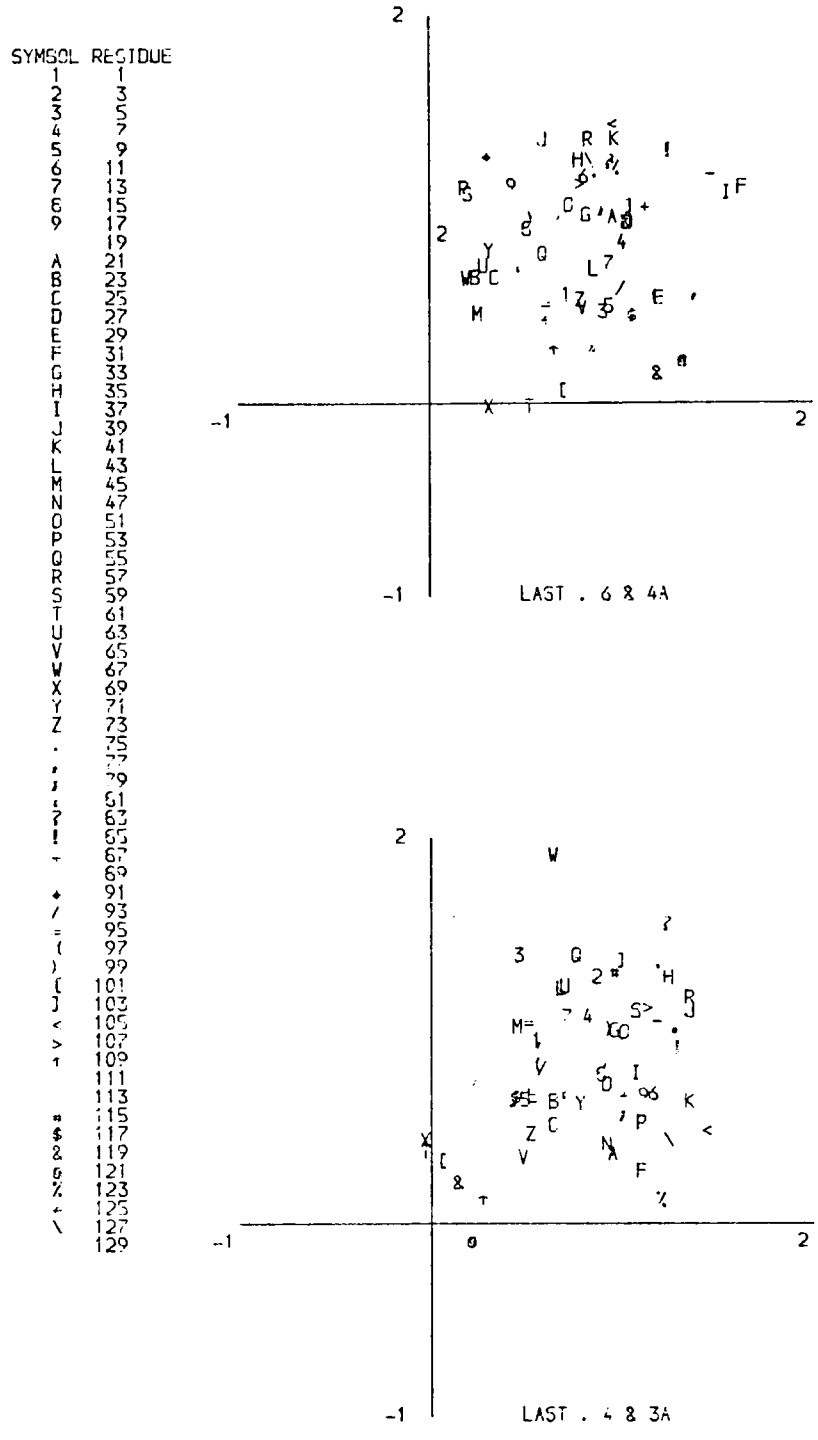


Figure (8)

Correlation between weighting refinements for different shells of data

As table (11) shows the convergence seemed satisfactory, and the values of w_i ranged from 0.07 to 1.63, which is not unsatisfactory. If these values of w_i represent a global rather than a local minimum, it is to be hoped that it can be reached from other starting points, although a finite number of starting points converging to the same minimum would not absolutely prove that minimum to be the deepest. Starting values of all $w_i = 0.5$ were chosen, which gave a minimum in R_3 of 250 (in the same arbitrary units as in table (11)) at values of w_i which differed by 0.15 (r.m.s.) from the first minimum. Starting values of all $w_i = 0.0$ were then chosen, and the minimum was only 423, with values of w_i ranging from -2.3 to 1.9. It is clear that local minima are present, and that the global minimum may not be the first encountered.

If a set of w_i are found from more than one independent set of structure factors, however, some confidence may be placed in their value. Similar minimisations were performed, starting from all $w_i = 1.0$, using all the structure factors in the 6Å to 4Å shell, and using the 700 strongest F's in the 4Å to 3Å shell. If the values of the w_i from two sets of data are plotted on a graph they will fall on a straight line of slope 1 and intercept 0 if perfectly correlated, and if they fall near this line the values of w_i are probably meaningful. Figure (8) shows the comparison of the refinements of the 6Å and the 4Å data sets, and the 4Å and 3Å. The 6Å and 3Å, not shown, display a similar lack of correlation. The first cycle

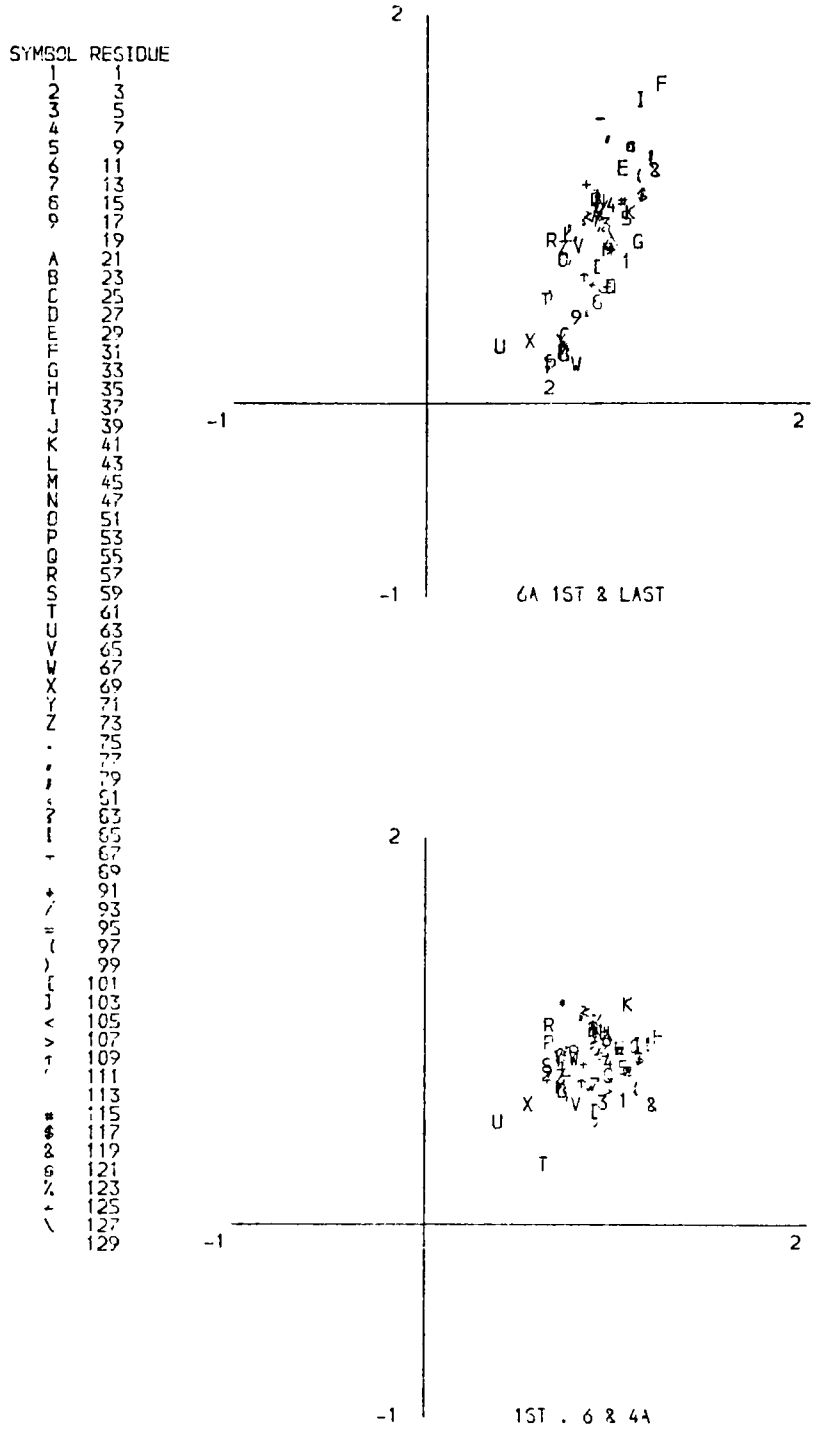


Figure (9)

Correlation between weighting refinements for full matrix and steepest descents, and steepest descents for different shells of data.

of the minimisation is a steepest descents minimisation involving no influence of the effect of varying w_j on w_i. The final values of w_i are correlated with the first cycle values, but have a greater spread; the first cycle values between different data sets are just as uncorrelated as the last cycle values, and these facts are illustrated in figure (9). This method does not seem to be useful.

(iv) Tangent Formula Techniques

(a) Sayre's equation (equation (26))

Although the phases of the structure factors cannot be measured directly, we know that some sets of phases are not permissible; the Fourier transform of the measured structure factor amplitudes with these phases will give an "electron density" not compatible with our prejudices about the structure, such as the non-existence of negative electron density. The formulation of general principles which restrict the permissible phases, and the expression and solution of equations embodying these has occupied many crystallographers for many years, and this sub-chapter is not an exhaustive account.

A structure consisting of a set of discrete equal atoms will have an electron density $\rho(\mathbf{x})$ such that $\rho(\mathbf{x})$ and $\rho^2(\mathbf{x})$ will be very similar (apart from a constant of proportionality) for all $\mathbf{x}$. Sayre in 1952 pointed out that

$$V\, S(\mathbf{h})\, F(\mathbf{h}) = \sum_{\mathbf{k}} F(\mathbf{k})\, F(\mathbf{h}-\mathbf{k}) \qquad (26)$$

where V is the unit cell volume and S(h) is the Fourier transform of the function representing the change in shape of an atom on squaring the structure. S(h) will be spherically symmetrical if the atoms have spherical symmetry, and is thus a function of $2\sin\theta/\lambda$ only. It is often assumed that atoms are points of electron density, when S becomes a constant. Sayre in 1952 showed that if all the F(h) are known to a resolution of 0.7Å, they may all be regenerated from equations (26). If an incomplete and inaccurate set of F(h) is available, approximations to the phases of structure factors both included in and excluded from the starting set may be generated: this is known as tangent formula refinement, and extension, respectively.

Equations (26) predict values for F(h), that is both amplitudes and phases, and the name "tangent formula" arises because the values of tan α(h) are given by dividing the imaginary by the real parts of the right hand side of equation (26); (the name is rather misleading because the angles, and not merely their tangents, are both needed and available.) It is proposed to use the term "tangent formula" to refer to any process which is based on equations (26) and which uses only the angular parts of the left hand side.

This is therefore using less of the available information than for example a process proposed by Sayre in 1972. This is the minimisation in a least squares sense of the difference between the left and right

hand sides of equations (26), i.e. minimise $\sum_{\mathbf{h}} |a\,\mathbf{F}(\mathbf{h}) - \sum_{\mathbf{k}} \mathbf{F}(\mathbf{k})\mathbf{F}(\mathbf{h}-\mathbf{k})|^2$ with respect to the phases $\alpha(\mathbf{h})$.

(b) Probability of correctness of tangent formula phases

J. Karle and I. Karle in 1966 suggested a use for the amplitudes predicted by equations (26); the probability of $\alpha(\mathbf{h})$ having a particular value is

$$P(\alpha) = (2\pi I_o(X))^{-1} \exp\,(X \cos(\alpha - \alpha_{tang})) \qquad (27)$$

where $X = 2N^{-\frac{1}{2}}\, E(\mathbf{h})\, E_{tang}(\mathbf{h})$

E_{tang} is the predicted amplitude and I_o is a Bessel function. Thus X is a measure of the reliability of a phase determination. The $E(\mathbf{h})$ are $F(\mathbf{h})$ sharpened so that they represent point atoms, when the atomic shape factor becomes unity, and the atoms cannot overlap. The above expression is simplified in that it assumes equal atoms; it corresponds to Sim's formula (page 98) for the partial structure case; both are derived from a paper by Cochran in 1955. However, neither Reeke and Lipscomb, nor Weinzierl, Eisenberg and Dickerson, have found a strong correlation of X with discrepancy between isomorphous protein phases and tangent formula refined phases. The latter noted that although subsequent cycles of refinement improved the internal consistency of the phases, the map as judged visually deteriorated after the first or second cycles.

(c) Real space equivalents

Although most workers have used equations (26) or the ones corresponding to point atoms, Barrett and Zwick in 1971 pointed out that direct squaring of the approximate electron density and then calculating the phases by Fourier transforming this new density was at least as convenient and more versatile than the reciprocal space convolution. In fact, squaring is only one of many functions for which ρ new $\approx$ const. X ρ old.

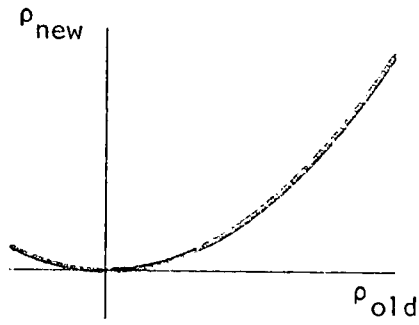
Barrett and Zwick performed a test on myoglobin. They calculated an electron density map of myoglobin using only the structure factors out to 3Å with isomorphous replacement phases. They set the electron density to zero where it fell below a certain level, and then squared the remaining structure; those structure factor phases calculated by this process with a resolution between 3 and 2 Å were compared with the isomorphous replacement phases. They found that the mean $|\alpha$ isomorphous - α tang$|$ was 82^{o}. Gassmann and Zechmeister used minimum and maximum significance levels without squaring; with estriol $((C_{18}H_{24}O_3)_2)$ a starting set of 41 phases produced the correct structure after 30 cycles, but the squaring technique after 15 cycles (starting from the same set of 41 phases) gave one high spurious peak only. Hoppe and Gassmann in 1968 had suggested a reciprocal space approximation to this double truncation.

(d) Justification of some "real space tangent formulae"

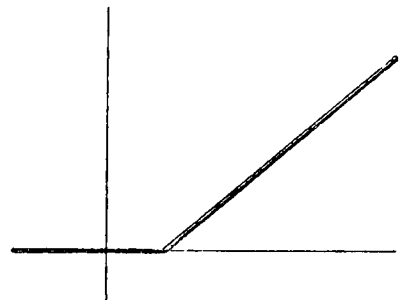
The process of tangent formula refinement may be compared to the well-known technique of Fourier refinement: that is, of calculating electron-density maps, or more usually difference maps with Fourier coefficients $(F-F_c)\exp(i\alpha_c)$, and adjusting the model to fit the clear indications of the difference map. We may think of this process as being equivalent to tidying up an $F\exp(i\alpha_c)$ map, and then Fourier transforming this altered map to produce a new set of phases. It is necessary to alter the map produced from Fourier coefficients $F\exp(i\alpha_c)$, because a mere Fourier transformation of this map, without alterations, reproduces the initial set of phases. If the Fourier refinement is done manually this alteration consists in ignoring small features and those features in a difference map incompatible with our chemical experience; understood features are usually exaggerated. If it is to be done automatically some way of discriminating between wanted and noise features of the difference map must be found.

It is convenient to divide the unit cell conceptually into two sets of regions: those regions which are far from any atoms of the partial structure, and those regions close to atoms in the model.

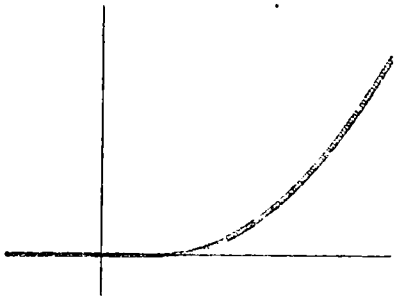
In the empty regions of the model, wanted atoms will appear with nearly half the required height (as explained on page 106) but noise terms will be of lower order. The signal to noise ratio can be



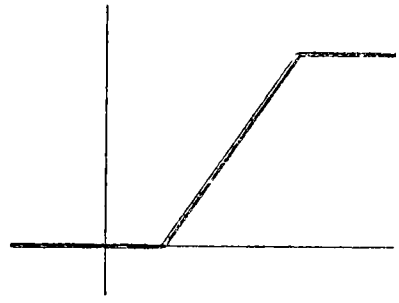
(a) Squaring
(Sayre, Karle, etc, etc.)



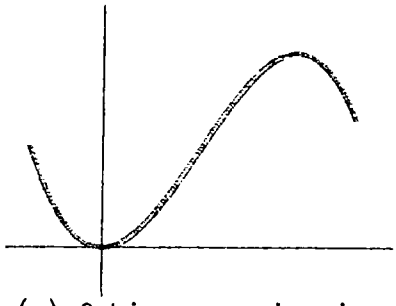
(b) Minimum significance
(Kartha)



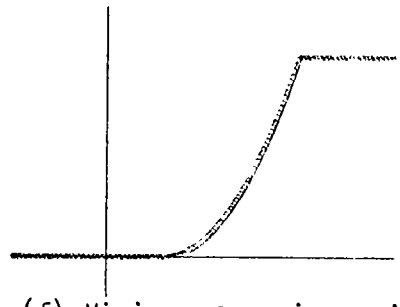
(c) Minimum significance level
and squaring (Barrett & Zwick)



(d) Minimum & maximum significance
levels (Gassmann & Zechmeister)



(e) Cubic approximation to (d)
(Hoppe & Gassmann)



(f) Minimum & maximum significance
levels & squaring

Figure (10)

Electron density modifications used in the
tangent formula

increased therefore by either removing all features in the new map below a certain significance level, (figure (10b)), or by squaring the model, (figure (10a)), or both (10c). However both the lower truncation, and squaring techniques reduce the wanted peak to known peak ratio, since it is always initially less than unity. This may be remedied to some extent by setting all values of the model above a certain level to that level. (Figure (10d) and (10f).) Such an upper level truncation is also useful in preventing one or two peaks from becoming very large after squaring, and artificially dominating the phasing. When atoms are not resolved, their overlap is exaggerated by the squaring, and this can be prevented from becoming too gross by the upper truncation.

Wanted features in the difference map which represent small movements of the atoms may be expected to be small and to be distributed over the model; to distinguish them from the noise peaks in the regions of the model close to atomic positions is not easy manually, and certainly not automatically. One justification for the squaring technique is that, as explained on page 106, wanted features appear with half weight. If initially all the atoms are put in with electron density 1, and one of them should have density $1 + \delta$ (δ may be negative), the map from $F \exp(i\alpha_c)$ will have density at nearly $1 + \delta/2$, but after squaring, it will be approximately $1 + \delta$. However, in the initial stages of refinement it may be better largely

to ignore these regions by applying an upper truncation.

(v) The Application of the Tangent Formula to Human Lysozyme

(a) A test of the Fourier transform procedure

Phases for all 4088 structure factors to 2.5Å resolution for which values of F were available were calculated from the 858 atom model (as described on page 71). An electron density map for the whole unit cell was calculated using these F exp ($i\alpha_c$) as coefficients for the Fourier program in the X-ray 70 system on the 1906A. A program was written which read in this map and Fourier transformed it using a fast Fourier routine of the Cooley-Tukey type, which was kindly given to me by A. Lifschitz. (It was necessary to do the transforms in two dimensions, store these on disc, re-section these, and transform in the remaining direction). The intervals chosen were 64 in x, 64 in y and 32 in z, corresponding to about 1Å in each direction, slightly coarser than the grid used by Barrett and Zwick. Unsharpened F's were used, and the phases from the back-transform were compared with the starting set: they differed by an average of 0.12 degrees. This justifies the grid interval used and shows that series termination errors are negligible with unsharpened F's (N.B. the temperature factor was 21).

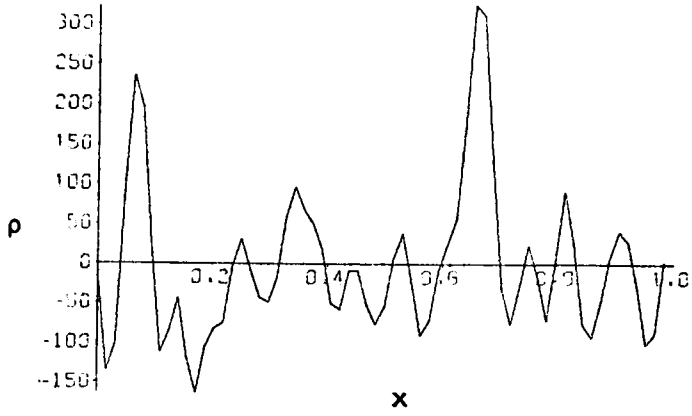
In what follows, my attempts to refine the phases are evaluated by comparing the mean discrepancy between the M.I.R. phases kindly

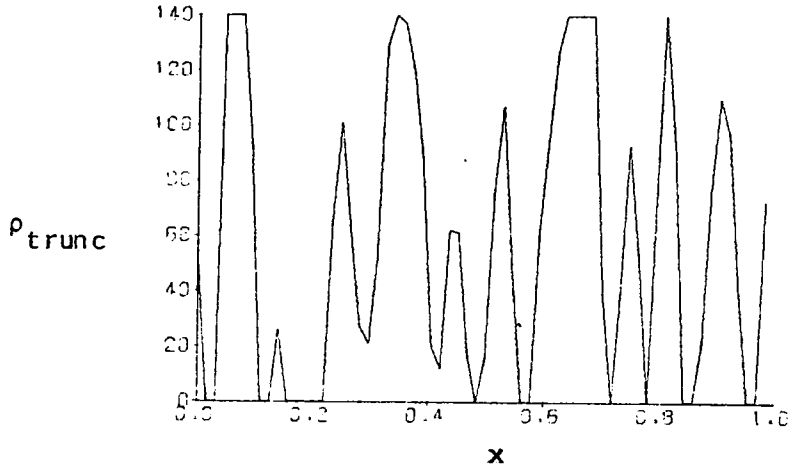
Phase set	mean phase difference in °	
	original α_c's	MIR α's
original α_c after back transformation	0.12	
MIR α's	69.13	
lower cut-off (b)	5.87	69.38
double cut-off (d)	21.58	71.61
squaring & double cut-off (f)	20.41	71.37

Table (12)

Tangent formula refinement results; a comparison with the phases determined by isomorphous replacement, before and after applications of the tangent formula.

b,d,f - see figure (10).





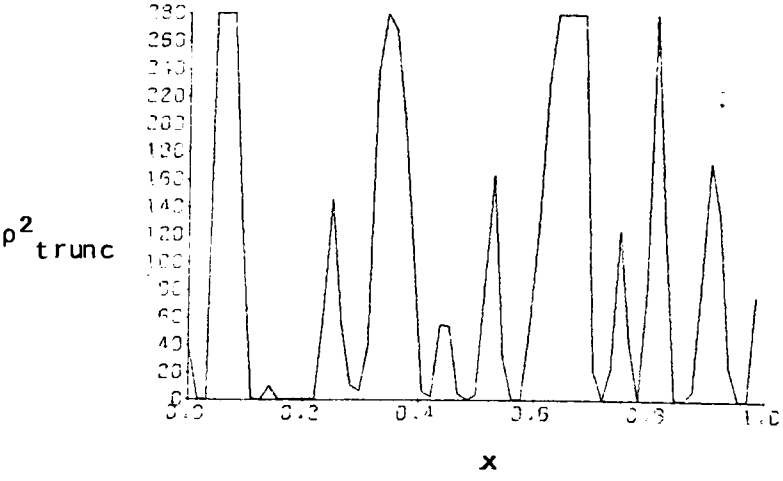


Figure (11)

A typical section through ρ showing changes for tangent formula applications

made available to me by Stephen Banyard, and my phases before and after an attempt at improvement. The M.I.R. phases are not without error, of course; the mean figure of merit is 0.70, indicating an average phase error of 45.5°, except for the set of structure factors out to 6Å, where the mean figure of merit is 0.91, indicating an error of 24.5°. This latter point explains partly why in figure (12) the discrepancy is much less for the low resolution structure factors than for the higher resolution.

(b) Attempts to improve the phases

A lower truncation was applied to the map. The zero level of the map is not known unless F(0,0,0) is accurately known, which it is not. The map was in fact calculated with F(0,0,0) = 0, which makes the mean level of the map zero. A level of -70 was chosen, and it was found that 23.5% of the volume of the unit cell was thereby set to -70. This modified electron density when Fourier transformed gave a set of phases differing by a mean of 5.87° from the starting set; they differed from the isomorphous phases by almost the same amount as the original α_{calc} set (see table (12)).

An upper truncation at 70 was next tried in addition to the lower truncation. The volume thus cut off was 18.6% of the unit cell. A section chosen through this function is shown in figure (11b) together with the unmodified map in figure (11a). This caused a mean phase change of 21.58 degrees, and the resultant phase set

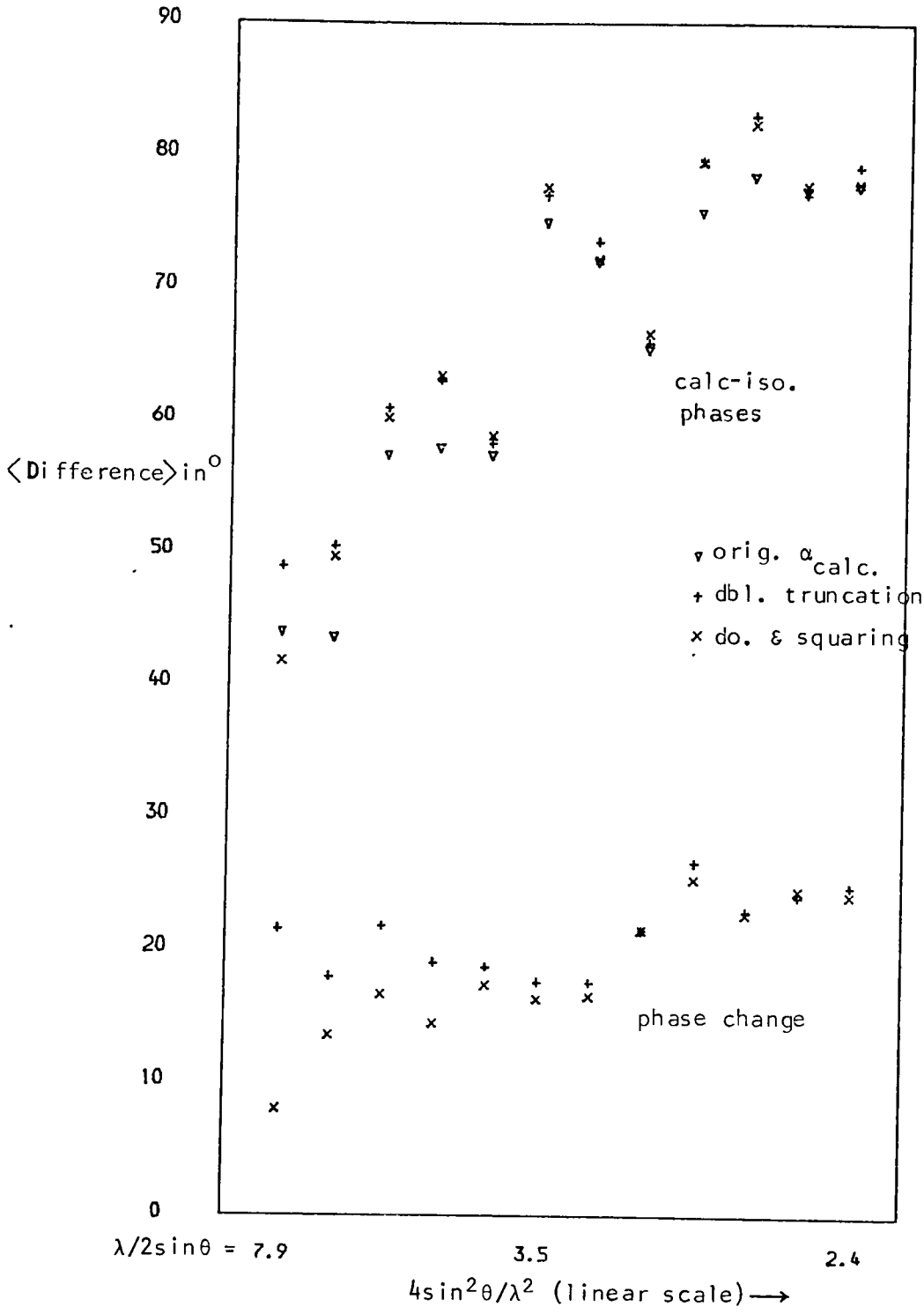


Figure (12)

Tangent formula modifications to the phases as a function of $2\sin\theta/\lambda$

was 2.58^{o} in worse agreement with the isomorphous phases than the original α_{calc}.

A double truncation and squaring was next tried. (Figures (10f) and (11c).) It is of course necessary to set the zero on the ρ new scale to the lower cut-off level before squaring, otherwise the negative regions become positive peaks. (This incidentally emphasizes the importance of including F (0,0,0) in tangent formula applications in reciprocal space.) The mean change in phase was 20.41^{o}, and the new phases were 2.24^{o} in worse agreement with the isomorphous phases than the starting set.

Figure (12) shows the variation of these average phase discrepancies with 2 sin θ/λ; the double truncation and squaring produces a change of only 8^{o} in the structure factors out to $8\mathring{A}$, and in this region an improvement of 2^{o} in the phases is observed. Except for this small success, the automatic phase correction techniques were unsuccessful, and it was decided that although clearly more possibilities could be tried — such as use of sharpened F's, different cut-off levels, more than one cycle of refinement, use of probability functions such as the Karles' (page 85) — effort should be diverted to inspection of difference maps.

(vi) Difference Syntheses (Introduction)

The problem of correcting a partially known structure has

formal similarities to isomorphous derivative theory, and almost every account seems to have its own peculiar nomenclature. In what follows, it is usually assumed that errors in the measured structure factors $F = |\mathbf{F}|$ are negligible and that the F's are on an absolute scale. $\mathbf{F}_c$ is a structure factor calculated from C of the atoms, assumed known, and $\mathbf{F}_Q$ that from the Q missing atoms. Thus

$$\mathbf{F} = \mathbf{F}_c + \mathbf{F}_Q$$

$$\text{or } F \exp(i\alpha) = F_c \exp(i\alpha_c) + F_Q \exp(i\alpha_Q) \qquad (28)$$

The phases α are not known, but the phases α_c will be an approximation to them, an increasingly good approximation as the likely value of F_Q becomes small compared to F_c and F.

We may calculate an electron density map which approximates either to the whole structure, or to the missing part only, since by subtracting $\mathbf{F}_c$ from the structure factors for the whole structure we have the structure factors of the missing part. Whether we use a Fourier synthesis to represent the differences between our approximation and the correct structure, or to represent the correct structure itself is not important to the theories of how to form such syntheses; the same information is carried in either synthesis. Since most of us are not as good as we would like to be at visualising the approximate structure convoluted with the Fourier transform of the region of reciprocal space for which we have

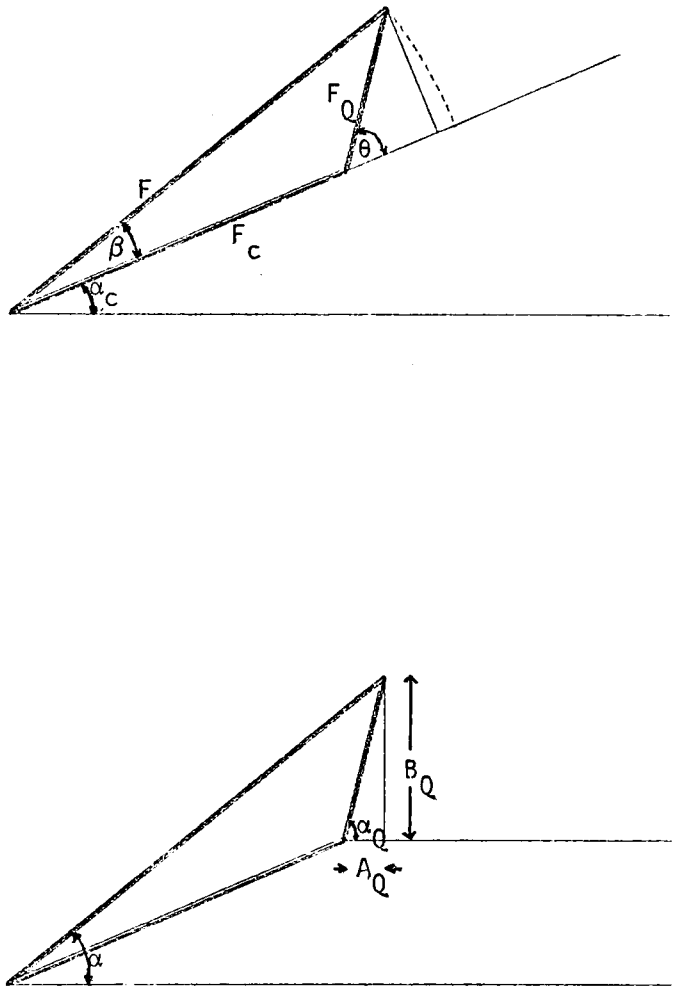


Figure (13)

An example of $\mathbf{F} = \mathbf{F}_c + \mathbf{F}_Q$

measured structure factors, difference maps are usually easier to interpret. The angle β between F and F_c is related to F_Q, and if the probability of F_Q taking any value is assumed to be that proposed by Wilson in 1949, the probability function of β, $P(\beta)$, may be found. α_c is the best estimate we have of α, but this is a better estimate for some structure factors than others because each structure factor has a different $P(\beta)$.

(vii) Sim weighting

Blow and Crick in 1959 suggested that the best map is that in which there is least r.m.s. error. Sim in 1959 and 1960 outlined a scheme for minimising (in an r.m.s. sense) the error in a map, although it does not seem to have been widely used.

(a) Derivation

Sim does not appear to have published his assumptions, but the derivation of his results (equations (31) and (35)) must have been similar to that given below

If $F_Q = A_Q + i\,B_Q$ (see figure (13)), and if the Q atoms are many and are randomly distributed, then (cf. e.g. Woolfson, page 243) if the probability of A_Q having a value A_Q is $P(A_Q)$,

$$P(A_Q)\,P(B_Q)dAdB = (\pi\Sigma_Q)^{-1} \exp(-F_Q^2/\Sigma_Q)\,dAdB \qquad (29)$$

Changing variables as in figure (13) to F_Q and β

$$P(F_Q)\ P(\beta)dFd\beta \ = \ (\pi\Sigma_Q)^{-1}\ \exp\ (-F_Q{}^2/\Sigma_Q)F_Q d\beta dF_Q \qquad (30)$$

Using $F_Q{}^2 = F^2 + F_c{}^2 + 2FF_c \cos\beta$, and integrating equation (30) with respect to β we obtain an expression for $P(F)dF$, and dividing equation (30) by this, we obtain the probability distribution function of β

$$P(\beta) \ = \ \exp\ (X\cos\beta)\ /\ \int_0^{2\pi}\exp\ (X\cos\beta)d\beta \qquad (31)$$

$$\text{where } X = 2FF_c/\Sigma_Q \qquad (32)$$

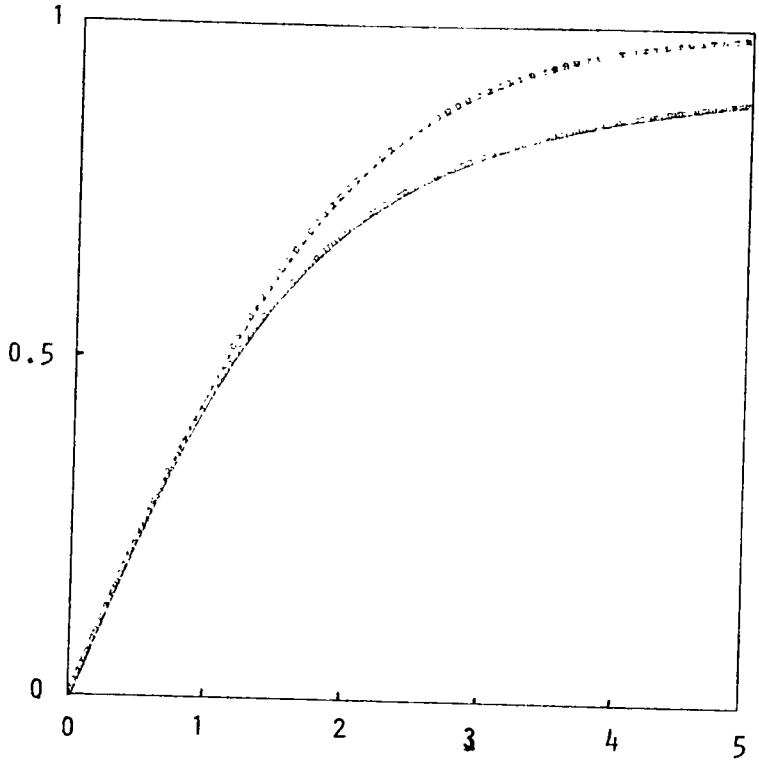
We shall use $W(\mathbf{h})F(\mathbf{h})\ \exp(i\alpha_c(\mathbf{h}))$ as structure factors and investigate the error in the map as a function of the weights W. The error in electron density when $WF\ \exp(i\alpha_c)$ is used rather than $F\ \exp(i\alpha)$ is the Fourier transform of

$$F\ \exp(i\alpha) - W\ F\ \exp(i\alpha_c) = F\ \exp(i\alpha_c)(\exp(i\beta) - W) = F_{err}$$

The best measure of the error in electron density is its modulus, squared and integrated over the whole unit cell. This is the same criterion as Blow and Crick's when ρ_{err} is real, and better in the unlikely case of its being complex.

$$\int_0^1 \left|\rho_{err}\right|^2 d\mathbf{x} \ = \ \int_0^1 \left|\frac{1}{v}\sum F_{err}(\mathbf{h})\ \exp(-2\pi i\mathbf{h}.\mathbf{x})\right|^2 d\mathbf{x}$$

$$= \ \frac{1}{v^2}\sum F_{err}^2 \qquad (33)$$



$$X = 2FF_c/\Sigma_Q$$

Figure (14)

Weighting functions for difference maps

solid line - acentric (Sim)

dotted line - centric (Woolfson)

We do not know the correct value of β, and hence of the mean square error, but knowing $P(\beta)$ we may calculate the expectation value of the mean square error

$$\left\langle \int_0^1 \left| \rho_{err} \right|^2 d\mathbf{x} \right\rangle = \int_0^{2\pi} \frac{1}{V^2} \sum F^2 \left| \exp(i\beta) - W \right|^2 P(\beta) d\beta$$

$$= \frac{1}{V^2} \sum F^2(\mathbf{h}) \left(1 + W^2(\mathbf{h}) - 2W(\mathbf{h}) \int_0^{2\pi} \cos \beta \; P(\beta) d\beta\right) \quad (34)$$

By partially differentiating this with respect to each of the $W(\mathbf{h})$ and equating to zero, we find that $\langle \rho_{err}^2 \rangle$ has a minimum when

$$W(\mathbf{h}) = \int \cos \beta \; P(\beta) d\beta = \langle \cos \beta \rangle \quad (35)$$

$$= \int_0^{2\pi} \cos \beta \; \exp(X \cos \beta) d\beta \; / \int_0^{2\pi} \exp(X \cos \beta) d\beta$$

$$= I_1(X) \; / \; I_0(X) \quad (36)$$

where the I's are Bessel functions; there are routines in the 1906 Libraries NAGF and FSCE for calculating them; X is given by equation (32). Woolfson in 1956 had shown that for structure factors which were limited by symmetry in the unit cell to a phase of 0 or π, the corresponding best value of W is

$$W = \tanh(X/2) \quad (37)$$

Figure (14) shows these weighting functions for the centric and acentric cases.

(b) Values for Σ_Q

Rossmann and Blow in 1961 suggested that Σ_Q should contain a contribution from the errors, Δx_i, in the positions of the supposedly known atoms: if these are in error, the phases α_c will be a poorer approximation to the required phases α. Their expression

$$\Sigma_Q = \sum_{i=1}^{Q} f_i^2 + \frac{4}{3} \pi h^2 \sum_{i=1}^{c} f_i^2 (\Delta x_i)^2$$

is difficult to use in the present context because as section (ii) of this chapter shows (page 77) it is not easy to determine the values of Δx_i^2.

It is easier to derive a value of Σ_Q for use in the weighting functions (see equation (31)) from the $F(\mathbf{h})$, $F_c(\mathbf{h})$ and $P(\beta)$. If we consider, as do Rossmann and Blow, that Σ_Q should contain contributions from the inaccuracies in the supposedly known set of atomic coordinates, we may suppose that the set of Q atoms contains atoms with negative atomic scattering factors at the positions of wrong atoms, and positive scattering factors at the true positions of the atoms which are missing; missing either because they were omitted, or because they were inaccurately placed and were cancelled out by an atom with negative atomic scattering factor. {If too many of the atoms of the partial structure need moving to quite different sites, the analysis of this section breaks down because F_c and F_Q will be

no longer uncorrelated. Small movements however, of the order of less than the apparent diameter of the atoms in an electron density map, do not invalidate the discussion of this section, for this reason:- If

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

whereas the atoms should be at $(\mathbf{x}_j + \Delta\mathbf{x}_j)$, when

$$\mathbf{F} = \sum_j f_j \exp(2\pi i\, \mathbf{h}.(\mathbf{x}_j + \Delta\mathbf{x}_j))$$

then expanding $\exp 2\pi i \mathbf{h}.\Delta\mathbf{x}_j$ as $1 + 2\pi i\ \mathbf{h}.\Delta\mathbf{x}_j$

$$\mathbf{F}_Q = \mathbf{F} - \mathbf{F}_c \approx \sum_j f_j\ 2\ i\pi \mathbf{h}.\Delta\mathbf{x}_j \exp(2\pi i\ \mathbf{h}.\mathbf{x}_j)$$

and unless the $\Delta\mathbf{x}_j$ are correlated the above values for $\mathbf{F}_Q$ and $\mathbf{F}_c$ will not be correlated when $\mathbf{h}.\Delta\mathbf{x}_j \ll 1$ (N.B. $\mathbf{h}.\Delta\mathbf{x}_j$ may take positive or negative values.)}

Equation (31) defines the distribution of values of β we expect when we know F and F_c. We can see from figure (13) that

$$F_Q^2 = F^2 + F_c^2 - 2FF_c \cos\beta$$

The expectation value of F_Q^2 is the mean of the values of F_Q^2 for points in reciprocal space having given values of F and F_c; it may be found by multiplying both sides of the above equation by $P(\beta)$, the distribution function for β, and integrating over all possible values of β:-

$$< F_Q^2 > = \int_0^{2\pi} (F^2 + F_c^2 - 2FF_c \cos \beta)\, P(\beta)\, d\beta$$

$$= F^2 + F_c^2 - 2FF_c \int_0^{2\pi} \cos \beta\, P(\beta)\, d\beta$$

Substituting equation (35) in the above

$$<F_Q^2> = F^2 + F_c^2 - 2FF_c\, W$$

(This equation has been derived (independently) by Srinivasan in another context.) The mean value of F_Q^2 is found by averaging the above equation over all structure factors. But equation (25) shows that the mean value of F_Q^2 is Σ_Q. Thus, if there are H values of **h** (i.e. H structure factors)

$$H\Sigma_Q = \sum(F^2 + F_c^2 - 2FF_cW) \qquad (38)$$

W is given by equations (32) and (36), and so it may be seen that W is a function of Σ_Q; equation (38) is thus an equation of one unknown, Σ_Q, and may be solved numerically.

(c) The value of X

Figure (14) shows that the distribution of values of X might cause three cases to obtain. Firstly, some of the structure factors would have a weight of nearly unity, others nearly zero, and others in between; in this case the general theory outlined above is appropriate.

Secondly, if Σ_Q were very small compared with nearly all the F

and F_c, i.e. if only a small part of the structure were missing, all the X's would be large, and all the weights nearly unity. Many treatments of difference maps consider only this limiting case, when the structure factors become $(F - F_c) \exp(i\alpha_c)$. Maps have often been calculated with these Fourier coefficients, and it is interesting to see what they represent.

(viii) Difference maps - theories

(a) A rule-of-thumb

"The main features of an electron density map are associated more with the phases than the amplitudes of the structure factors." If "main" and "more" are omitted, this rule is simply untrue - it would imply that $F \exp(i\alpha_c)$ maps can never reveal the whole structure. As written above, the rule is too vague to be called a theorem. Its justification is two-fold — firstly the phases determine the signs of the contributions to the structure factors, whereas the amplitudes control merely the sizes of these contributions, and secondly, it seems true in practice — Ramachandran and Srinivasan quote an example on page 66 of their book, where randomly changing the structure factor amplitudes, but retaining the correct phases, caused the main features of the map to be still recognisable.

(b) An extension of the Dodson/Vijayan approach

As is more fully explained in appendix A, Eleanor Dodson and

Vijayan set up an expression for $(F-F_c)\exp(i\alpha_c)$ and consider an approximation for this expression using the Taylor series approximation $(1 + x)^{\frac{1}{2}} = 1 + \frac{1}{2}x$. Taking further terms in the Taylor series gives two formulae: the first applies if $F_Q < F_c$, the second if $F_Q > F_c$. Appendix A on page 163 gives both the algebra in detail, and numerical examples to show that the approximations are valid.

Considering first the case when $F_Q < F_c$, we see from equation (A2) (page 164)

$$(F-F_c)\exp(i\alpha_c) = \tfrac{1}{2}\left(1 - \frac{F_Q^2}{8F_c^2} + \ldots\right) F_Q \exp(i\alpha_Q)$$
$$+ \tfrac{1}{4}\left(1 + \frac{F_Q^2}{16F_c^2} + \ldots\right) \frac{F_Q^2}{F_c} \exp(i\alpha_c)$$
$$+ \tfrac{1}{2} F_Q \exp(i2\alpha_c - i\alpha_Q) + \ldots. \qquad (39)$$

(N.B. $\cos\theta = \frac{1}{2}(\exp(i(\alpha_Q-\alpha_c)) + \exp(i(\alpha_c - \alpha_Q)))$)

The first line on the right hand side of equation (39) represents the correct difference at nearly half weight; applying the rule-of-thumb above, we can see that the second line represents a distorted "ghost" of the known structure; the third line represents noise terms of at least the same size as the wanted part (first term). (The last line is noise only if the F_c and F_Q are uncorrelated, when the terms with phase angles $2\alpha_c - \alpha_Q$, etc, can be given no readily

visualisable meaning.)

Considering secondly the case when $F_Q > F_c$, we see from equation (A3) (page 166)

$$(F-F_c)\exp(i\alpha_c) = \tfrac{1}{2}\left(1 + \frac{F_c^2}{8F_Q^2} + \ldots\right) F_c \exp(i\alpha_Q) + \left(\frac{F_Q}{F_c} - 1 + \frac{F_c}{4F_Q} + \ldots\right) F_c \exp(i\alpha_c) + \tfrac{1}{2} F_c \exp(i2\alpha_c - i\alpha_Q) + \ldots. \qquad (40)$$

To interpret equation (40) we must apply the rule-of-thumb again. The first term on the right hand side is the required difference, somewhat distorted because the amplitudes have been randomly changed (i.e. from F_Q to F_c), and reduced by a factor of nearly $F_c/2F_Q$, which is less than $\frac{1}{2}$ since $F_Q > F_c$; in the second line the dominant term in the bracket is $F_Q/F_c > 1$ and this causes a "ghost" of the known structure to appear: if $F_Q > 2F_c$ the weight of the "ghost" will be greater than that of the wanted term; the third line again represents noise of the same order as the wanted term.

We have seen that the third lines in equations (39) and (40), which have phase angles neither α_c nor α_Q can have amplitudes at least as large as the wanted terms in these equations. This may not be too disastrous, however, because these third lines will produce noise randomly spread over the difference map; if most of

the known atoms are accurately placed and resolved, the missing atoms, the wanted part of the difference maps, will be concentrated in small regions. The total contribution to the structure factors will be similar from each of the noise and the wanted parts, and so in this case the wanted region will be much higher than the noise.

Only in the limiting cases when $F_c >> F_Q$ or $F_Q >> F_c$ for all or nearly all, structure factors can equations (39) or (40) be used alone. In most cases some structure factors will be such that $F_c > F_Q$ and for others the reverse will be true. The correct mixture of equations (39) and (40) and their interpretation becomes rather complicated, and the treatment given in section (x) of this chapter, page 114, is probably more useful. This subsection, and the following one, have been included because first the arguments are easier to follow (in my opinion) than that of section (x), and secondly we can leave the introduction of assumptions of randomness to a later stage of the argument than in section (x), and so the first part of the argument of this sub-section is generally true.

(c) A convolution argument

Ramachandran and Srinivasan have given an alternative explanation of the difference Fourier, which they call the γ' synthesis. A paraphrase of their argument now follows. From equation (3) (page 9) and the truncated Taylor series expansion given on page 106 for $(1 + x)^{\frac{1}{2}}$

$$F^2 = \Sigma(1 + \sum_i \sum_{j=i} \frac{f_i f_j}{\Sigma} \exp(2\pi i \mathbf{h}.(\mathbf{x}_i - \mathbf{x}_j))$$

$$\therefore \quad F \approx \Sigma^{\frac{1}{2}} + \sum_i \sum_{j=i} \frac{f_i f_j}{2\Sigma} \exp(2\pi i \mathbf{h}.(\mathbf{x}_i - \mathbf{x}_j)$$

where $\Sigma = \sum_i^N f_i^2$, the sum over all N atoms in the unit cell. Thus, they say, the Fourier transform of F has peaks in the same places as the Patterson function, but with height reduced by a factor of $\Sigma^{\frac{1}{2}}$ at the origin and by $2\,\Sigma^{\frac{1}{2}}$ elsewhere. (This is using the rule-of-thumb of page 103.) The convolution of this pseudo-Patterson function with the Fourier transform of $\exp(i\alpha)$ (the "phase synthesis") has the correct structure. (N.B. a convolution in real space is equivalent to a multiplication in reciprocal space.) Therefore the phase synthesis can be described to a first — or even a second — approximation by a consideration of this convolution procedure. Having now worked out the interpretation of the phase synthesis, the convolution of the phase synthesis for the calculated phases with the Fourier transform of F (the pseudo-Patterson function) gives us a description of the Fourier transform of F $\exp(i\alpha_c)$. Their result is

Position	Peak height
Known atoms	$\frac{\Sigma_c + \Sigma}{2(\Sigma_c \Sigma)^{\frac{1}{2}}} f_t$
Unknown atoms	$\frac{1}{2} (\frac{\Sigma_c}{\Sigma})^{\frac{1}{2}} f_q$
Pseudo-random ie $\mathbf{x}_i + \mathbf{x}_j - \mathbf{x}_k$	$\frac{f^3}{2(\Sigma_c \Sigma)^{\frac{1}{2}}}$

A comparison of these predicted peak heights with those given by Luzzati's more exact treatment in table (13), page 119— shows that Ramachandran and Srinivasan's treatment slightly overestimates the expected peak height of the unknown atoms, and slightly underestimates the known peak heights. Their conclusion, that when only a small part of the structure is missing and $\Sigma_c \approx \Sigma$ the missing atoms appear with half weight, is correct however. The main criticism of their approach is that their assumption that peak positions in an electron density map are determined by the phases only, and that peak heights are determined by the amplitudes only, is clearly not exactly true, and it is difficult to see under what conditions this is a valid approximation, and when it breaks down.

(ix) The α-synthesis or Sum function

(a) Analogy with difference maps

The third case mentioned on page 104 is when Σ_Q is very large compared to FF_c, and therefore all values of X will be small. By expanding exp(X) as 1 + X it may be shown that

$$I_1(X)/I_0(X) \approx \tanh(X/2) = X/2 + ... \qquad (41)$$

Figure (14) shows that these are valid approximations for the acentric and centric weighting functions up to about X = 1. Thus when only a small part of the structure is known the weighted coefficients

are

$$\frac{FF_c}{\Sigma Q} F \exp(i\alpha_c) = \Sigma_Q^{-1} F^2 F_c$$

If we omit the constant term Σ_Q^{-1}, we have the "α-synthesis", which Raman in 1959 showed could be treated as follows.

$$F = F_c + F_Q$$

we find that

$$F^2F_c = F_c^2F_Q + (F_c^2 + F_Q^2) F_c + F_c^2F_Q^* \qquad (42)$$

where * means complex conjugate. Equation (42), it is to be noted, is not an approximation. The terms on the right hand side may be interpreted by using the approximate rule-of-thumb given on page 105. The three terms will contribute to a map to produce (i) a distorted version of the missing part of the structure, (ii) a distorted version of the original model, with higher weight than the first term and (iii) a noise term of similar magnitude to the wanted part. The interpretation of α-synthesis is thus quite similar to that of equation (39).

(b) Convolution arguments

A multiplication of the Fourier coefficients of two functions is equivalent to the convolution of those functions; the α-synthesis, which has Fourier coefficients F^2F_c is thus the convolution of the

Patterson function of the complete structure with the partial structure. If we assume point atoms, these may be written as,

$$\text{Patterson function} \quad P(\mathbf{u}) = \sum_i^N \sum_j^N f_i f_j \, \delta(\mathbf{x}_i - \mathbf{x}_j, \mathbf{u})$$

and

$$\text{partial structure} \quad \rho_{part}(\mathbf{x}) = \sum_k^C f_k \, \delta(\mathbf{x}_k, \mathbf{x})$$

where $\delta(a,b) = 0$ if $a \neq b$

$= 1$ if $a = b$

and there are N atoms in the complete structure and C in the partial structure. The α-synthesis is thus

$$\rho_\alpha(\mathbf{x}) = \int_0^1 P(\mathbf{u}) \rho_{part}(\mathbf{x}-\mathbf{u}) \, d\mathbf{u}$$

$$= \int_0^1 \sum_i^N \sum_j^N \sum_k^C f_i f_j f_k \, \delta(\mathbf{x}_i - \mathbf{x}_j, \mathbf{u}) \, \delta(\mathbf{x}_k, \mathbf{x}-\mathbf{u}) \, d\mathbf{u}$$

$$= \sum_i^N \sum_j^N \sum_k^C f_i f_j f_k \, \delta(\mathbf{x}, \mathbf{x}_i - \mathbf{x}_j + \mathbf{x}_k)$$

The α-synthesis thus has peaks at positions $\mathbf{x}_i - \mathbf{x}_j + \mathbf{x}_k$ where the $\mathbf{x}_i$ and $\mathbf{x}_j$ are the complete set of atomic positions, and the $\mathbf{x}_k$ are the set of atomic positions in the partial structure. There will be an accumulation of peaks at the positions $\mathbf{x}_i$ - the true atomic positions - each time $(-\mathbf{x}_j + \mathbf{x}_k) = 0$, which if the C atoms are a sub-set of the N atoms in the complete structure, will be

C times. There will be an accumulation of peaks at the x_k positions — i.e. at the positions of the atoms used for phasing — when $x_i - x_j = 0$, i.e. N times. Thus if the partial structure is a sub-set of the complete structure the peak height at the known atom sites x_i will be

$$\sum_{k}^{C} f_k^2 f_i + \sum_{j}^{N} f_j^2 f_i = f_i(\Sigma_c + \Sigma)$$

and at the unknown sites x_k

$$\sum_{j}^{C} f_j^2 f_k = f_k \Sigma_c$$

and the ratio of unknown / known atom peak heights will be the ratio of their atomic numbers multiplied by $\frac{\Sigma_c}{(\Sigma_c+\Sigma)}$. If some of the atoms in the set C are incorrect, it may similarly be shown that the ratio (incorrect atom electron density) / (correct atom) is $\Sigma/(\Sigma_c + \Sigma)$.

The above discussion uses the somewhat unrealistic assumption of point atoms. The next section of this chapter, however, shows that a different argument also derives a ratio of unknown / known electron density peak height of $\Sigma_c/(\Sigma_c + \Sigma)$; these assumptions will be merely that the atoms are numerous and are randomly distributed throughout the unit cell. The results of this sub-section are therefore valid

more generally than the point atom assumption implies.

Returning for a paragraph to the point atom case, we see that noise peaks on the α-synthesis will occur at $x_i - x_j + x_k$; $i \neq j \neq k$, with a height $f_i f_j f_k$, whereas the wanted peaks have height $f_k \Sigma_c$. If all the atoms are the same, the ratio of wanted / noise peak height is $\Sigma_c / f^2 = C$. Buerger discusses the α-synthesis under the name "image seeking sum function", but dismisses it because in his examples this ratio is often only two. In the example of human lysozyme discussed later there are hundreds of atoms in the partial structure, and so the signal / noise ratio should be of the order of several hundred. The point atom approximation however, does not apply in many real situations, and the signal / noise ratio expected for a more realistic situation is discussed in the next section; it is much lower than this figure because the numerous noise peaks overlap to form larger noise peaks.

(x) Expectation values of signal to noise ratios

Equations (39) and (40) are generally true, but to interpret them we need to make some assumptions, for example that $\mathbf{F}_c$ and $\mathbf{F}_Q$ are uncorrelated. Luzzati in 1953 gave an explanation of maps with Fourier coefficients $F \exp(i\alpha_c)$ which assumes that Wilson statistics apply to both F_c and F_Q, and that F_c and F_Q are uncorrelated; that is, he built these assumptions into his theory at the beginning.

He derived formulae firstly for χ, the expected height of an unknown atom in a map with Fourier coefficients $F \exp(i\alpha_c)$, secondly for τ, the expected height of the known atoms in this map, and thirdly for μ, the ratio of χ to the r.m.s. value of electron density not accounted for in either of these categories. His argument is summarised below.

(a) Peak height at known atom site τ

The electron density at the t'th of the C known atoms in such a map is

$$\rho(\mathbf{x}_t) = \sum F(\mathbf{h}) \cos(2\pi \mathbf{x}_t.\mathbf{h} - \alpha_c) \qquad (43)$$

$P(F/\mathbf{F}_c)dF$ is the probability of a structure factor having values between F and F + dF when the known C atoms have corresponding structure factor $\mathbf{F}_c$. From equation (29) by changing variables to F and β and integrating over all values of β we obtain

$$P(F/\mathbf{F}_c)dF = \int_0^{2\pi} \frac{F}{\Sigma_Q} \exp\left(-(F^2 + F_c^2 - 2FF_c\cos\beta)/\Sigma_Q\right) d\beta dF$$

$$= 2\pi\, x \exp(-x^2-y^2)\, I_0(2xy)dx \qquad (44)$$

where $x = F\Sigma_Q^{-\frac{1}{2}}$, $y = F_c\Sigma_Q^{-\frac{1}{2}}$

$P(\mathbf{F}_c/\mathbf{F}_t)\, F_c dF_c d\alpha_c$ is the probability of finding $\mathbf{F}_c$ within the element of area $F_c dF_c d\alpha_c$ on an Argand diagram, where $\mathbf{F}_t$, the Fourier transform

of the known atom we are considering, is known; F_t has phase $2\pi x_t.h$. By a similar argument we obtain

$$P(F_c/F_t)F_c dF_c d\alpha_c = \frac{F_c}{\Sigma_c} \exp(- (F_c^2 + F_t^2 - 2F_c F_t \cos(2\pi x_t.h - \alpha_c))/\Sigma_c) dF_c d\alpha_c$$

$$= y\ \psi \exp(- y^2\psi - F_t^2\psi + 2y\psi F_t \cos(\ \ )) dy d\alpha_c \qquad (45)$$

where $\psi = \Sigma_Q/\Sigma_c$

The expectation value of $\rho(x_t)$, $\rho_m(x_t)$, is given by multiplying equation (43) by the probability of F having a given value, given that we know F_t, and integrating.

$$\rho_m(x_t) = \sum \int_0^\infty F\ P(F/F_t) \cos(2\pi h.x_t - \alpha_c)\ dF \qquad (46)$$

Luzzati perhaps ought to have mentioned that he uses the following theorem of probability theory:

$$P(F/F_t) = \int P(F/F_c)\ P(F_c/F_t)\ dF_c \qquad (47)$$

We consider only the case where many atoms are known, when F_t^2/Σ_c is negligible, and expanding (45) in terms of F_t and integrating over α_c we obtain the first order term in Ft as

$$\rho_m(x_t) = \tau f_t$$

where $\tau = \int_0^\infty \int_0^\infty 4\ \psi^2 x^2 y^2 \exp(- x^2 - y^2(1 + \psi))\ I_0\ (2xy) dx dy \qquad (48)$

(b) Peak height at unknown atom site χ

By a similar argument we obtain the expectation value of the density at one of the unknown atoms, q, assuming F_q^2/Σ_Q to be negligible, as:-

$$\rho_m(x_q) = \chi f_q$$

where $\chi = \int_0^\infty \int_0^\infty 4\psi x^2 y(x I_1(2xy) - y I_0(2xy)) \exp(-x^2-y^2(1+\psi)) dxdy$ (49)

Luzzati expanded the Bessel functions I_0 and I_1 as power series and integrated to obtain slowly converging power series in $\psi = \Sigma_Q/\Sigma_C$.

(c) The signal to noise ratio μ

The mean square electron density in the map with coefficients $F \exp(i\alpha_c)$ is $< F^2 > = \Sigma$. The electron density squared accounted for by the above is

$$\chi^2 \sum_{q=1}^{Q} f_q^2 + \tau^2 \sum_{t=1}^{C} f_t^2 = \Sigma_Q(\chi^2 + \tau^2/\psi) \qquad (50)$$

The signal / r.m.s. noise ratio is thus

$$\mu = \chi\Sigma_Q^{\frac{1}{2}} / (< F^2 > - \Sigma_Q(\chi^2 + \tau^2/\psi)^{\frac{1}{2}}$$

$$\mu = \chi/(< F^2 > / \Sigma_Q - \chi^2 - \tau^2/\psi)^{\frac{1}{2}} \qquad (51)$$

where in this case $< F^2 > = \Sigma = \Sigma_C + \Sigma_Q = (1 + 1/\psi)\Sigma_Q$

The denominator of equation (51) is the r.m.s. electron density in a map from Fourier coefficients $F \exp(i\alpha_c)$ not accounted for by the

known or unknown atom sites. Thus the fact that unknown atoms do not appear with their full height, for example, does not introduce anything into this expression for the noise in such a map. This is not the case in Blow and Crick's definition of error, where the r.m.s. error is averaged over the whole unit cell (see page 98).

(d) Extension of Luzzati's treatment to maps with weighted coefficients

The above argument may be used when other Fourier coefficients are used; if Sim weighting is used, equations (43), the corresponding one for $\rho(x_q)$, and (51) have F replaced by $F\, I_1(2xy)/I_0(2xy)$; if the α-synthesis is used these equations have F replaced by $F^2 F_c$. The results are summarised below. {$E = \exp(-x^2-y^2(1+\psi))$, $\psi = \Sigma_Q/\Sigma_C$, $z = 2xy$, and the I's are Bessel functions.}

$$\tau = \int_0^\infty \int 4\psi^2 x^2 y^2 E\, I_0(z)\, dxdy \qquad (52)$$

$$\chi = \int_0^\infty \int 4\psi\, x^2 y\, E(x\, I_1(z) - y\, I_0(z))\, dxdy \qquad (53)$$

$$\langle F^2\rangle/\Sigma_Q = \int_0^\infty \int 4\psi\, x^3 y\, E\, I_0(z)\, dxdy = 1 + 1/\psi \qquad (54)$$

$$\tau_{sim} = \int_0^\infty \int 4\psi^2 x^2 y^2 E\, I_1(z)\, dxdy = 1 \qquad (55)$$

$$\chi_{sim} = \int_0^\infty \int 4\psi\, x^2 y\, E\, (x\, I_1^2(z)/I_0(z) - y\, I_1(z))\, dxdy \qquad (56)$$

$$\langle F_{sim}^2\rangle/\Sigma_Q = \int_0^\infty \int 4\psi\, x^3 y\, E\, I_1^2(z)/I_0(z)\, dxdy \qquad (57)$$

$$\tau_\alpha = \int_0^\infty \int \Sigma_Q 4\psi^2 x^3 y^3 E\, I_0(z)\, dxdy = (1 + 2/\psi)\Sigma_Q \qquad (58)$$

$\frac{\Sigma_c}{\Sigma}$	τ	χ	$\frac{\langle W^2F^2\rangle}{\Sigma_Q}$	χ_{Sim}	χ/τ	χ_α/τ_α	μ	μ_{Sim}	μ_α
0.1	2.5461	0.1258	0.1924	0.0813	0.049	0.091	0.205	0.297	0.286
0.2	1.8452	0.1804	0.3929	0.1429	0.098	0.167	0.298	0.408	0.365
0.3	1.5437	0.2242	0.6234	0.1949	0.145	0.231	0.375	0.492	0.402
0.4	1.3695	0.2632	0.9077	0.2410	0.192	0.286	0.447	0.564	0.414
0.5	1.2546	0.2995	1.2855	0.2835	0.239	0.333	0.517	0.629	0.408
0.6	1.1729	0.3348	1.8238	0.3238	0.285	0.375	0.588	0.692	0.387
0.7	1.1120	0.3700	2.6964	0.3631	0.333	0.412	0.663	0.755	0.352
0.8	1.0652	0.4064	4.4027	0.4027	0.382	0.444	0.747	0.821	0.298
0.9	1.0286	0.4464	9.4449	0.4450	0.434	0.474	0.847	0.895	0.218

Table (13)

Expected peak heights χ and τ, unknown/known peak height ratios and signal to noise ratios μ for maps with coefficients $Fexp(i\alpha_c)$, $WFexp(i\alpha_c)$ (Sim), and $F^2F_c exp(i\alpha_c)$ (α-synthesis). See equations (52) to (61).

$$\chi_\alpha = \int\int_0^\infty \Sigma_Q 4\psi\, x^3y^2\, E\,(x\, I_1(z) - y\, I_0(z))\, dxdy = {}^{\Sigma Q}/_{\psi} \qquad (59)$$

$$\langle F_\alpha{}^2 \rangle = \langle F^4 F_c{}^2 \rangle = \sum^3 \int\int 4\psi\, x^5y^3\, E\, I_0(z)\, dxdy = \frac{2\sum_Q^3}{\psi}\,(1+\tfrac{1}{\psi})(1+\tfrac{3}{\psi}) \qquad (60)$$

$$\mu_\alpha = (2/\psi + 3 + \psi)^{-\frac{1}{2}} \qquad (61)$$

The above nine integrals were evaluated numerically using the routine QB01A in the library OXLIB1900 on the 1906A computer. (The Chebyshchev method was used; each integral took about 1200 function evaluations (ca. 6 secs) to yield an accuracy of 1 part in 10^4; the limits were zero to where the function was less than 10^{-30} its maximum value). Those integrals for which an analytic expression is not known are listed in table (13) as functions of Σ_c/Σ, the fraction of the structure known, together with the signal / r.m.s. noise ratios μ. {Careful comparison of table (13) with table (1) in Luzzati's 1953 paper will show that my values of τ and χ, and hence μ, are larger than Luzzati's. A calculation of his series for the integrals for τ and χ (his equations (35) and (42)) with up to 350 terms yields the values given in table (13).}

(e) Conclusions

There are several interesting points arising from equations (52) to (61) and table (13). Firstly equation (55) shows that the peak heights on a Sim weighted map are correct, and so a Sim weighted difference map, with Fourier coefficients $(WF - F_c)\exp(i\alpha_c)$ will

have no "ghost" of the partial structure.

Secondly, the unknown to known peak height ratio is χ/τ; this is given in table (13) for the unweighted case: for Sim weighting it is χ_{sim}: for the α-synthesis it is from equations (58) and (59)

$$\frac{1/\psi}{1 + 2/\psi} = \frac{\sum_c}{\sum + \sum_c}$$

which is the same as the expression derived on page 113, and is also listed in table (13). The ratio χ/τ can be seen to be greatest for the α-synthesis and least for the unweighted maps.

Thirdly, the signal to noise ratio μ is always greater for Sim weighted maps than for unweighted maps. For the α-synthesis this ratio μ is nearly equal to that for Sim weighting — and greater than μ for unweighted maps — when only a small proportion of the structure is known, but is low when most of the structure is known. (The reason for these poor results when most of the structure is known may be seen from equation (42); the distortions mentioned in connexion with the term $(F_c^2 + F_Q^2)\mathbf{F}_c$ are not only distortions of peak heights, but also represent noise peaks, which swamp the term $F_c^2\mathbf{F}_Q$ when F_Q is small.)

Sim weighting was derived by minimising the r.m.s. error in the map calculated; this error is composed of changes in the atom peak heights from their true values, and of extraneous noise peaks.

Both of these types of contribution to the error are reduced by the application of Sim weighting: the α-synthesis is better at correcting the peak heights to their correct value, but at the expense of introducing extraneous noise peaks.

(xi) Human lysozyme difference maps

(a) The maps calculated

As explained in section (i) of this chapter, a preliminary set of 858 atoms was obtained, and a set of $F_c(\mathbf{h})$ was calculated from them by a method similar to that described on page 32. Maps were calculated with the following sets of Fourier coefficients.

1. $F \exp(i\alpha_c)$
2. $WF \exp(i\alpha_c)$ W is given by equations (36) and (37) on page 101.
3. $F^2F_c \exp(i\alpha_c)$
4. $(WF-F_c) \exp(i\alpha_c)$
5. $(W'F-F_c') \exp(i\alpha_c')$
6. $mF \exp(i\alpha_{mir})$

According to the theory of Sim (see page 98) map 2 should be that map which represents the whole structure with the least possible r.m.s. error. Maps 1 and 3 also represent the whole structure (see pages 96 and 111), and were calculated for comparison with map 2. According to the analysis of Sim-weighted maps of the previous

section — see especially equation (55) — map 2 should contain the partial structure with its proper weight, and so map 4 should represent the differences between the whole and the partial structure, with no "ghost" of the partial structure.

In map 5 the primes refer to the fact that the F_c were calculated from only 327 atoms, as described later. Sections of map 6 were kindly lent to me by Drs. Colin Blake and Stephen Banyard. (m is their figure of merit, the expectation value of the cosine of the error in their phase determination.)

(b) The value of Σ_Q

The value of Σ_Q required for the calculation of Sim and ~~Cochran's~~ Woolfson's weighting functions was calculated by solving equation (38) numerically (using subroutine C05AAF in the NAG library). The structure factors were divided into shells with $\lambda/2\sin\theta$ limits of 6, 4, 3 and 2.5 Å, as in section (ii) (page 76) of this chapter; a value of Σ_Q was found for each shell. The scaling between F and F_c was such that $0.83\Sigma F^2 = \Sigma F_c^2$ for each shell, since (page 43) the 858 atoms represent 83% of the ordered non-hydrogen scattering matter.

A set of values of $X = 2FF_c/\Sigma_Q$ was thus obtained. Over all structure factors the median value of X was 1.90, and reference to figure (14) on page 100 shows that this is near the most curved part of the weighting function curves. Therefore the weighting used neither applies a weight of unity to nearly all F's, nor a weight

nearly always proportional to X, but a mixture of these. Map 2 is thus intermediate between maps 1 and 3.

(c) The mean value of $\cos^2\theta$

We can see, either from figure (13) or from equation (A2), that when $F_Q << F_c$

$$F-F_c \approx F_Q \cos\theta \tag{62}$$

$$(F-F_c)^2 \approx F_Q^2 \cos^2\theta \tag{63}$$

If F_c and F_Q are uncorrelated, all values of the angle $\theta = \alpha_Q - \alpha_c$ are equally probable; averaging equation (63) over all structure factors, we see that

$$\langle (F-F_c)^2 \rangle \approx \tfrac{1}{2} \langle F_Q^2 \rangle \tag{64}$$

and equation (25) shows that the right hand side of equation (64) is (by Wilson's theory) $\frac{1}{2}\Sigma_Q$, and so

$$\langle (F-F_c)^2 \rangle \approx \tfrac{1}{2}\Sigma_Q \tag{65}$$

For shells of structure factors with resolution out to 6, 4, 3 and 2.5 Å the ratio of $\langle (F-F_c)^2 \rangle / \Sigma_Q \approx \langle \cos^2\theta \rangle$ was 0.54, 0.50, 0.48 and 0.47 respectively. If these values are considered to be significantly different from 0.5, an explanation of their variation with the resolution of the data could be that at low resolution the missing atoms appear to be in much the same places as the known atoms (since the missing atoms are often side chains) and so there

is a tendency for $|\theta| = |\alpha_Q - \alpha_c|$ to be small more often than a random theory would predict; this would make $\langle \cos^2\theta \rangle > \frac{1}{2}$. On the other hand, at higher resolution, this similarity of position of the C and Q atoms will not be apparent. Contributions to Σ_Q from inaccurately placed known atoms may cause an opposite effect. On page 103 it was stated that the terms

$$\sum_{t=1}^{C} f_t \exp(i\alpha_t) \quad \text{and} \quad \sum_{t=1}^{C} 2\pi i \mathbf{h}.\Delta\mathbf{x}_t \, f_t \exp(i\alpha_t)$$

would not be correlated, but perhaps this is not true; the factor i in the second term may cause F_Q and F_c to tend to be 90° apart in the Argand diagram, and so $\langle \cos^2\theta \rangle < \frac{1}{2}$.

(d) Theory of what the maps represent

In order to apply the results of section (x) of this chapter, we need to know how much of the structure may be assumed to be correctly known: the figure of 83% mentioned earlier assumed that the intermolecular liquid and hydrogen atoms were negligible, and that the known atoms were accurate. If the known and unknown parts of the structure may be considered to be independent

$$\Sigma = \Sigma_c + \Sigma_Q \tag{66}$$

(N.B. We are considering inaccuracies in the partial structure to be contained in the "unknown" part in this discussion.)

We know Σ from $\Sigma = \langle F^2 \rangle$ and we have just discussed the value of Σ_Q. Hence Σ_c may be found on the same scale as Σ and Σ_Q. The ratio of known to total structure is Σ_c/Σ and for the shells of structure factors with resolution out to 6, 4, 3 and 2.5 Å this ratio is 0.54, 0.55, 0.49 and 0.40 respectively. (Perhaps these values of Σ_c/Σ should be used instead of 0.83 in the scaling of ΣF^2 and ΣF_c^2 in the determination of Σ_Q (subsection (b)), and this process continued until the values of Σ_c/Σ converge; equation (66) is probably not sufficiently dependable, however, for this process to be exact, and anyway we shall only use the results in a semi-quantitative way.)

Taking then Σ_c/Σ as 0.5, this gives $\psi = \Sigma_Q/\Sigma_c = 1.0$, and reference to equations (52) to (61) and table (13), page 119, we see that

1 For map 1 we would expect to see the known structure reappear with $\tau = 1.25$ times its proper weight, whilst the unknown atoms will have only $\chi = 0.30$ of their proper weight; the signal to noise ratio μ will be 0.52, and $\chi/\tau = 0.24$

2 For map 2 we would expect the known atoms to reappear with their proper weight, while the unknown have a weight of 0.28. The ratio of χ/τ is thus 0.28, better than with map 1, as is the signal to noise ratio $\mu = 0.63$.

3 The α-synthesis, map 3, has a ratio of unknown/known peak

heights of $\chi/\tau = 0.33$, better than for maps 1 or 2, but the signal to noise ratio μ is 0.41, worse than for maps 1 or 2.

4 As already mentioned, a useful consequence of Sim weighting is that map 4 should contain no "ghost" of the known structure; the signal to noise ratio μ is 0.63 as for map 2.

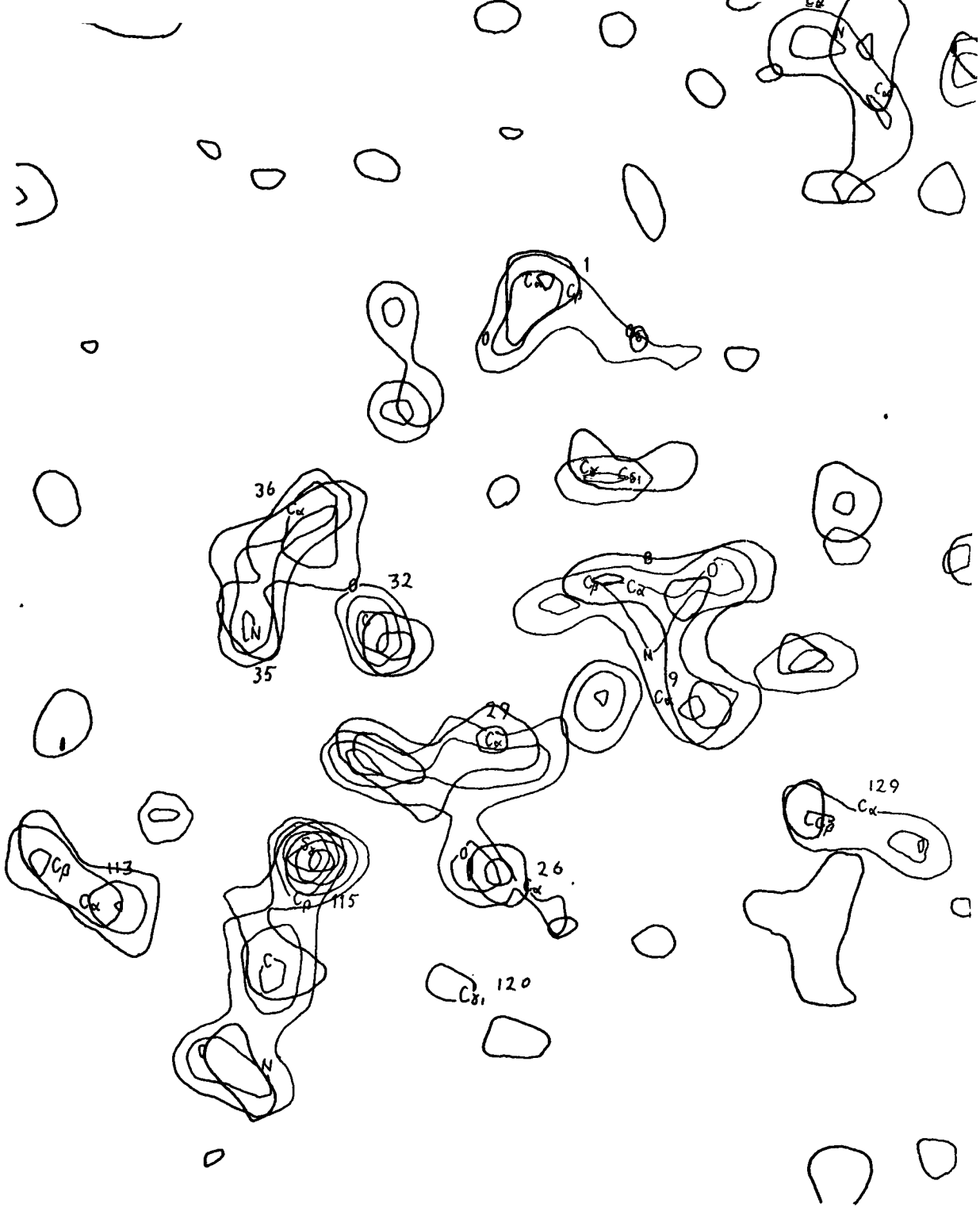
(e) The Fourier calculations

The calculations of maps 1 to 5 were performed using the Fourier program in the X-Ray 70 system on the 1906A. The electron density was calculated at intervals of 1/80, 1/80, 1/40 of the unit cell vectors **a**, **b**, **c**; i.e. about every 0.7Å. Forty sections of constant z were calculated (i.e. the whole unit cell in this direction) with x from 205/240 to 376/240, and y from -4/80 to 50/80, which includes almost a whole molecule. (This is more than an asymmetric unit, and so every part of a molecule is present.) These limits were suggested to me by Stephen Banyard.

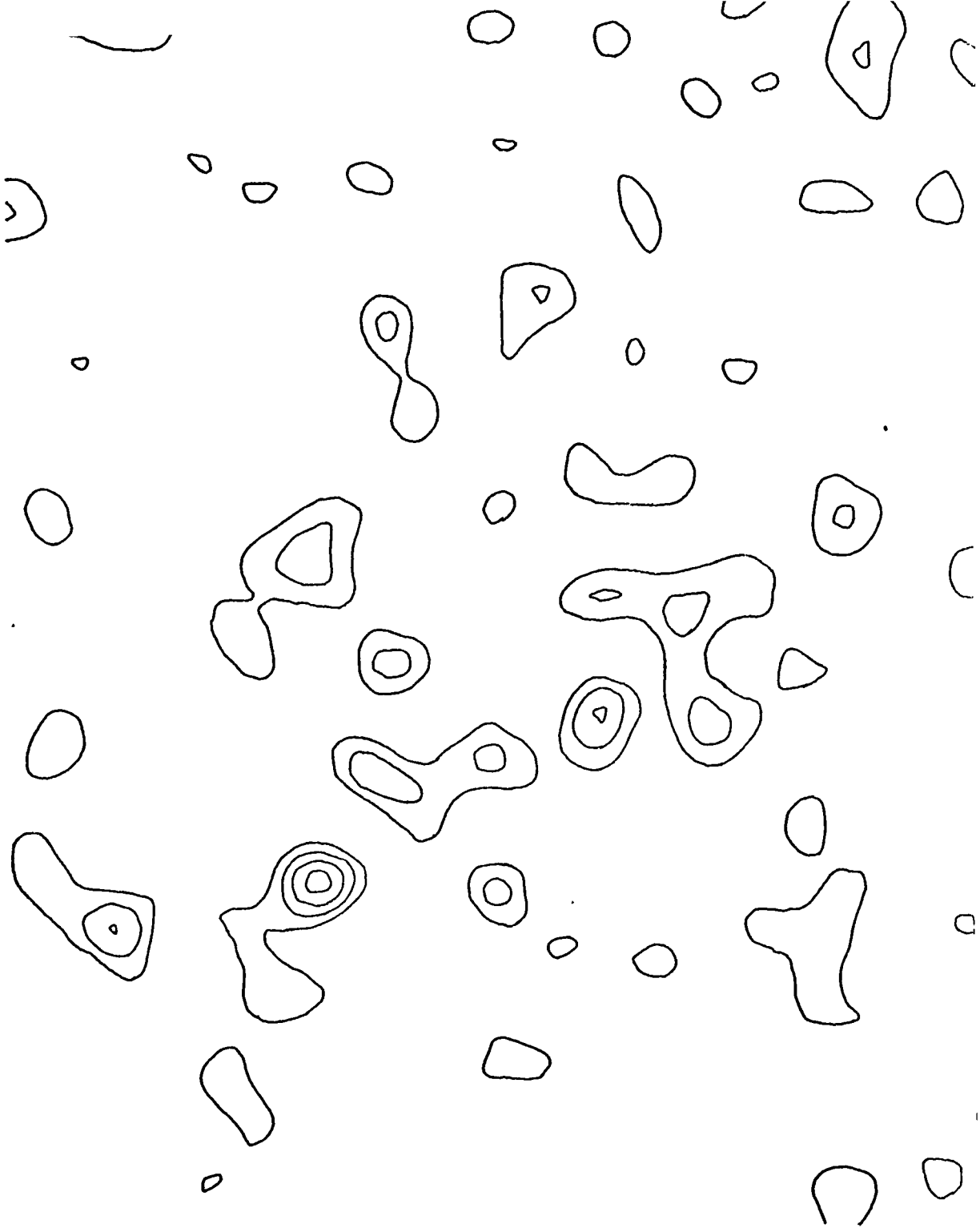
These maps were calculated with F(0,0,0) as zero, because its true value is not known accurately. This means that the average level of the maps is zero. Contours were drawn to represent the electron density at equal but arbitrary intervals, with only levels greater than zero being drawn on the maps which represent electron density (rather than differences) i.e. maps 1, 2, 3 and 6. The contour levels on the difference maps (4 and 5) were five times as close together as those on maps 1 and 2, and negative contours were

Figure (15)

Sections $z = 156/240$ of the Sim weighted map WF exp $(i\alpha_c)$ (top, transparent) and the M.I.R. map (bottom).



map 6
map 2



map 6

shown dotted on the difference maps. The zero level of the difference maps was omitted for clarity. The contours were drawn on a pen and ink plotter on the 1906A using a program adapted by myself from one written for the Argus computer by Stephen Banyard. This program also rotated and translated a set of atoms, generated the equivalent atoms in all the molecules in the unit cell, and plotted the appropriate atoms on the contoured electron density sections.

(f) A preliminary look at the maps

Figure (15) shows a section through maps 2 and 6. It is apparent that there is a good deal of similarity between these maps. Secondly, in the inter-molecular liquid regions — the top left-hand corner and the bottom right-hand corner — there are more features on the M.I.R. map than the α_{calc} map. Thirdly it may be seen that the electron density, while usually occurring at atom sites, does not always follow the atoms rigidly - see the C_{δ} of residue 1 (near top right-hand corner) for example. Fourthly, there is a fairly important peak just to the right of the centre of map 6 which has no counterpart in map 2: this is discussed later.

I next looked at the difference map corresponding to map 2, i.e. map 4, without comparing it with the M.I.R. map; examples are given in figures (17) to (19). In some parts of this map it appeared that the atoms were approximately correct — they were in

blank regions on the difference map. Other regions showed fairly clearly that movements to the atoms should be made — see for example figure (17). Many regions, however, showed atoms in holes in the difference map, with positive peaks nearby, but it was not clear towards which positive regions in the difference map the atoms had moved — see figure (18) for an example. Many of the ends of the side chains were in negative regions with no very clear corresponding positive region; this may be explained by the fact that the same temperature factor was applied to all atoms, whereas side chains nearly always have higher temperature factors — i.e. a smaller electron density at the atomic centres — see for example table 3 of Watenpaugh, Sieker, Herriott and Jensen, where the temperature factors are given individually for all the atoms in rubredoxin.

There are two — and only two — places where a collision occurs between different molecules in their rotated and translated positions. These both involve the guanidium groups of arginine residues 14 and 61, which collide with the main chain of residue 126, and with the side chains of valine 109 and asparagine 113 respectively. These are the locations of two deep holes in the difference map.

(g) 327 atoms

One way of dealing with the situation of uncertainty of interpretation of the difference map is to omit all atoms which fall on

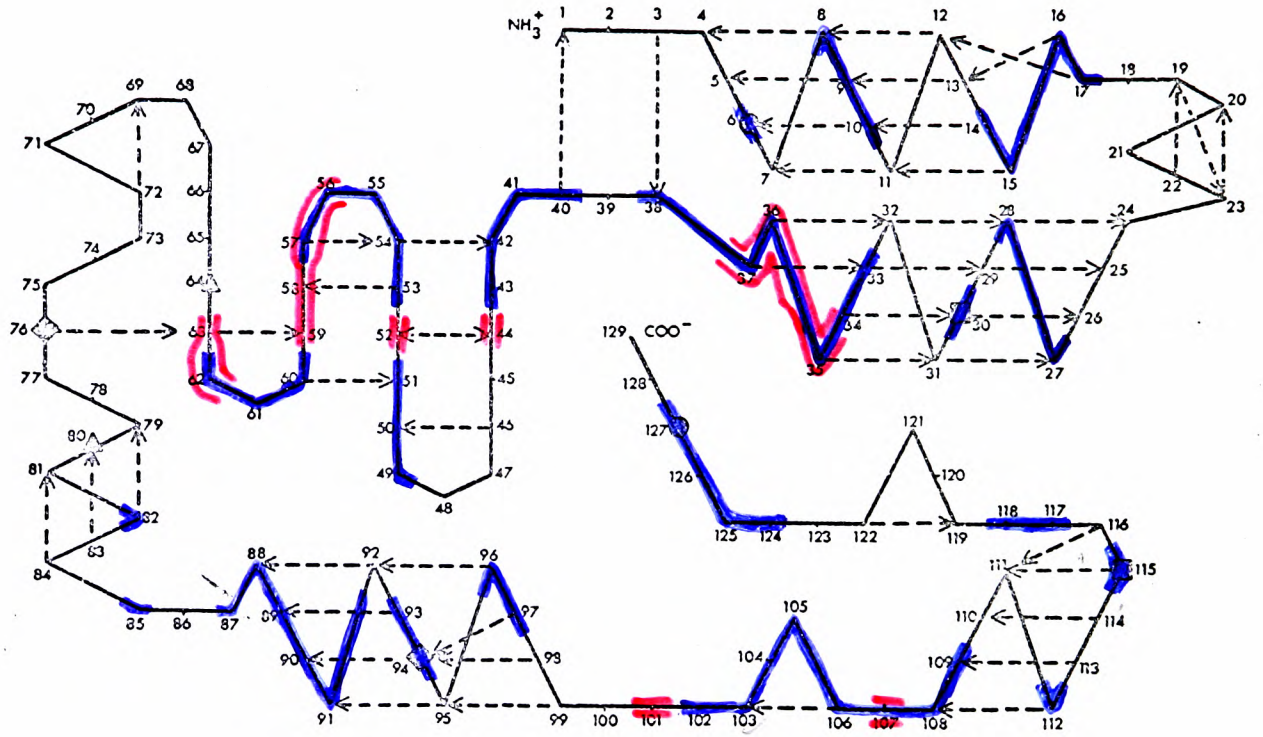


Figure (16)

A diagram showing main-chain NH→CO hydrogen bondsin hen lysozyme. The blue regions are those residues which appeared not to have moved according to map 4. The red regions are concerned with the active site of lysozyme

or near negative regions of the difference map. Each residue was treated as a whole, and if more than one of the four main-chain atoms fell within a negative region on map 4 the whole residue was omitted; if the main-chain appeared to be in the right place but the side-chain lay in a negative region, the side chain atoms in the negative region were omitted.

Figure (16) is adapted from a diagram in the review by Imoto et al. The blue regions are those parts of the main chain which seemed fairly clearly not to have moved, according to map 4. The red regions are amino acid residues associated with the active site of lysozyme (see Imoto et al). Residues numbers 82 and 112 seem to be isolated in the middle of parts of the chain which have moved, and it was thought that their non-appearance in negative regions of map 4 might be due to a coincidental movement of atoms from, and other atoms into, their positions; since caution was the word for this section, these two residues were omitted also.

This selection gave a total of 327 atoms, which as figure (16) shows, were in regions which might be expected to be unchanged in the transition from hen to human lysozyme; that is the parts of the molecule where the conformation of the chain is held either by cystine bridges, or by hydrogen bonds into a definite secondary structure, or in the region of the active site. Nearly all of the blue regions in figure (16) fall into one of these three categories.

Resolution Å	R_1 (327 atoms)		R_1 (858 atoms)
	$\sum F_c^2 = \sum F^2$	$\sum F_c^2 = \frac{327}{1028} \sum F^2$	$\sum F_c^2 = \frac{858}{1028} \sum F^2$
∞ to 6	0.58	0.62	0.47
6 to 4	0.54	0.57	0.43
4 to 3	0.51	0.55	0.44
3 to2.5	0.49	0.53	0.47

Table (14)

Residuals for the 327 and 858 atom sets as functions of sinθ

Structure factors were calculated from these atoms as before — i.e. by assuming that each atom had the shape of nitrogen but its own atomic number, by applying a temperature factor to each atom of 21.46, ignoring the inter-molecular liquid, and scaling separately within the usual four shells.

(h) Statistics of the 327 atom phases

This new set of F_c was treated similarly to the 858 atom model. The variation of the residual R_1 with sin θ is shown in table (14); I am not sure what is the "fairest" way to scale F and F_c in order to compare the values of R_1 with those of the 858 atom model; both the ways given in table (14) show that R_1 is larger for the fewer atoms — as expected — but that it decreases with increasing resolution. This decrease may be explained by the fact that the 327 atoms are not randomly chosen from the structure but rather have large chunks of the molecule been omitted; the low resolution view will be grossly distorted, but the high resolution view will see parts of the molecule quite accurately.

The value of Σ_Q was evaluated as described in sub-section (b) of this section, page 123, with $\Sigma F_c^2 = \frac{327}{1028} \Sigma F^2$. The value of the ratio of known to total scattering matter is, as before, $(\Sigma - \Sigma_Q)/\Sigma$, which for the shells of structure factors with resolution out to 6, 4, 3 and 2.5 Å was 0.40, 0.30, 0.30, 0.28 (N.B. 327/1028 = 0.32). Comparing these numbers with the corresponding ones for the 858

atom model (page 126), we see that although these values of Σ_c/Σ are smaller than for the 858 atom model, the reduction is by no means as great as the ratio 327/858, which would presumably apply to a set of 327 atoms randomly chosen from the 858 atoms. This indicates that contributions to Σ_Q from inaccuracies in the 327 atoms are comparatively smaller than for the 858 atom model. Why the ratio of $(\Sigma - \Sigma_Q)/\Sigma$ should be larger than 327/1028 for the 6Å structure factors is not clear.

Taking Σ_c/Σ to be 0.30, we see from table (13) that we would expect the signal to noise ratio μ to be 0.49, which is rather lower than μ for map 2 (0.63).

The median value of $X = 2FF_c/\Sigma_Q$ was 0.91, and reference to figure (12) shows that the weighting in the 327 atom model is again intermediate between no weighting and the α-synthesis (section (ix)).

(i) Contouring map 5

Map 5 is a difference map, but because the F_c represent only a small proportion of the total atoms, we would expect there to be more positive regions than negative. We know that F(0,0,0) should be larger than $F_c(0,0,0)$, but F(0,0,0) is not known accurately, so $F_c(0,0,0)$ and F(0,0,0) were both omitted; this makes the mean value of the map zero, rather than positive. An electron density equal to one contour level on map 4 was added to all points on map 5, which was then contoured in the same way as map 4. This somewhat arbitrary choice of $(F(0,0,0) - F_c(0,0,0))$ gave a map which

Set of atoms	r.m.s. Δx Å
858 atoms	1.75
backbone only of same (508 atoms)	1.52
327 atoms	1.42
backbone only of same (225 atoms)	1.35

Table (15)

Root mean square distances between rotated & translated hen lysozyme atoms and the preliminary set of atomic positions for human lysozyme.

sometimes reached the third positive level, but rarely reached the second negative level, which is what one might expect: see page 143 for an example.

None of the 327 atoms fall into a negative region on this map.

(j) Comparison with measured coordinates

As mentioned earlier, a preliminary set of atomic coordinates has been measured by Dr. Colin Blake, Simon French and Adrian Butement from the wire model built from map 6. Values of the r.m.s. difference between these atomic positions and those corresponding to the models derived from the hen lysozyme atoms were calculated. As table (15) shows, the 327 atoms are closer to those of human lysozyme than the remainder of the 858 atom set, and backbone atoms are more similar in the two molecules than side chain atoms. A more detailed discussion of these sets of coordinates is left to Adrian Butement.

(xii) Interpretation of human lysozyme maps

(a) How far to move atoms

Before proceeding to look at actual maps in detail, we need to know how to tell from a difference map how far inaccurately placed atoms should be moved. If an atom lies in a hole in a difference map, and we think that its real position is indicated by a peak some distance away, the problem is trivial. If, however, the electron

densities of the atom at its true and at its inaccurate positions overlap, the situation is more complicated.

A well-known approach — see for example Stout and Jensen page 377 — is to expand $\rho(\mathbf{X} + \Delta\mathbf{X})$ as a Taylor series up to the term in Δx^2 and then it may be shown that the required shift of an atom's position, Δx, is given by

$$\Delta x = - \frac{\text{slope}}{\text{curvature}} = - \frac{\partial \Delta\rho / \partial x}{\partial^2 \rho / \partial x^2} \qquad (66)$$

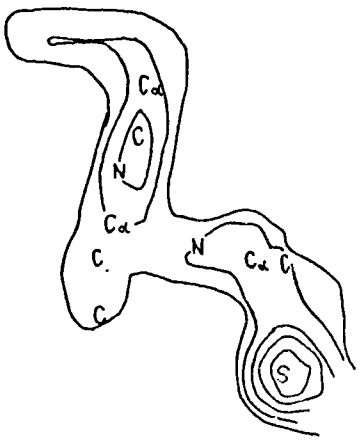
The slope is here the gradient of the difference map, and the curvature is the curvature of the corresponding electron density map. Stout and Jensen (on page 382) point out that a factor of 2 should be inserted in the right hand side of equation (66), when the structure has no centre of symmetry, because features on the $\Delta\rho$ map usually appear with only half weight compared to those on the electron density (ρ) map; see the $\frac{1}{2}$ in equation (39), page 106, for an explanation of this. (This factor of $\frac{1}{2}$ was first pointed out by Cruikshank in 1950.) However, as was shown in section (x) of this chapter, this reduction by a factor of $\frac{1}{2}$ applies only in the limiting case when nearly all the structure is accurately known. It might be better to use the appropriate value of χ or χ_{sim} given in table (13) and multiply the right hand side of equation (66) by $1/\chi$.

The curvature of the electron density near an atomic centre was found from map 2. Four atoms in different stretches of main chain and well defined side chain were chosen, and about 5 values of x and $\rho(x)$ were fitted in a least squares sense to a parabola, from which the curvature was readily obtained. Of the four values of the curvature, none differed by more than 10% from the mean, and so equation (67) is probably applicable to all atoms to within an accuracy of about 10%, except for those atoms near the ends of side chains where the curvature is less than in well defined regions of the molecule. If d is the distance in $\mathring{A}$ between the first negative contour level on map 4 and the first positive level the other side of an atom which seems to have moved, the shift to apply, according to equation (66) is

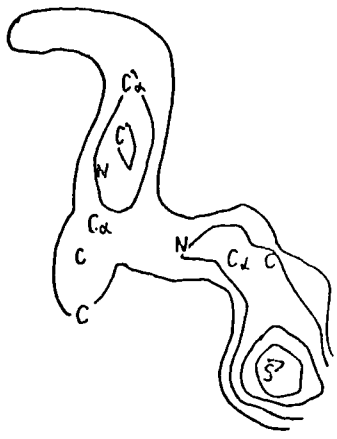
$$\Delta x = \frac{0.55}{d} \mathring{A} \qquad (67)$$

The factor $0.55\mathring{A}^2$ takes into account the five-fold difference in contour levels between maps 2 and 4, the Luzzati factor 1/0.28, and the fact that d is the distance between 2 contour levels on map 4, the zero level having been omitted.

Equation (67) shows that when an atom has moved a short distance, the value of Δx is proportional to d^{-1}. However, when it has moved a large distance the approximations implied in equation (66) break down, and equation (67) no longer applies; Δx is simply the distance



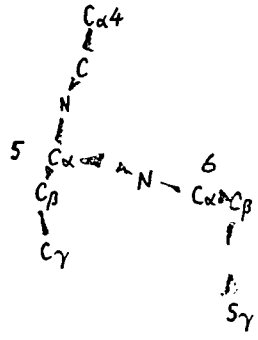
map 1
Fexp($i\alpha_c$)



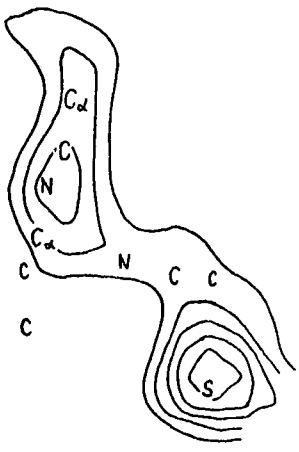
map 2
WF exp($i\alpha_c$)

Figure (17)

A superimposition of features on part of the sections z = 120/240 and 114/240 for maps 1 to 6.

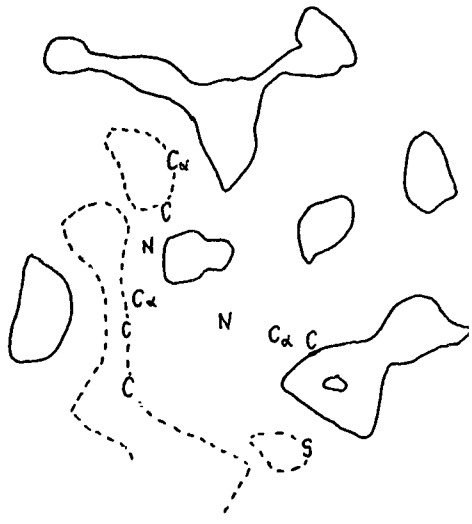


The atoms used in phasing maps 1 to 4. Map 5 was phased with



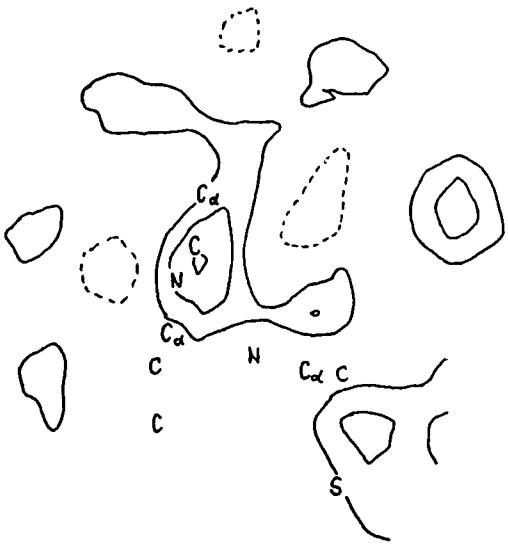
map 3

$F^2F_c\exp(i\alpha_c)$



map 4

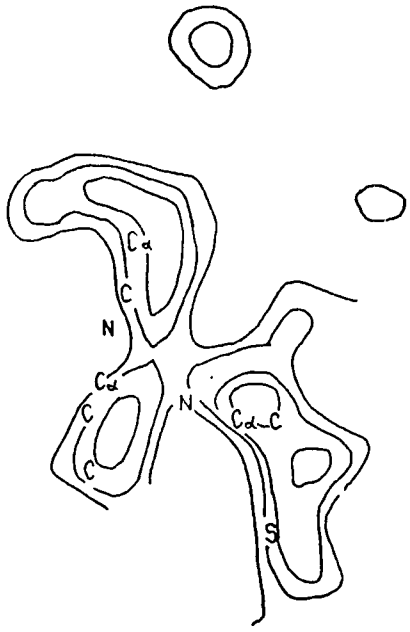
$(WF-F_c)\ \exp(i\alpha_c)$



map 5

$(W'F-F'_c)\ \exp(i\alpha'_c)$

(327 atoms)



map 6

isomorphous replacement

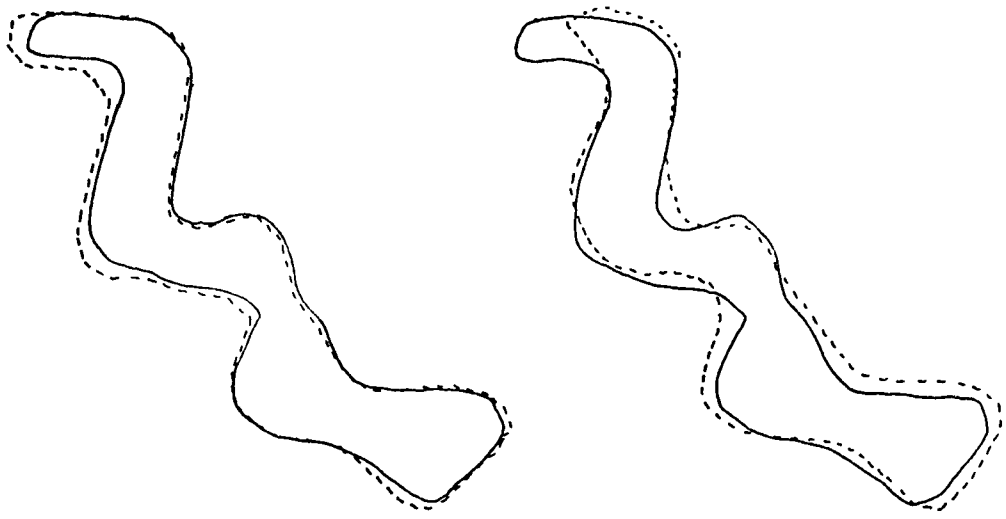
to the appropriate peak on the difference map, which increases as d increases. Experience will obviously be an important guide in deciding when these two limiting cases apply, and in deciding how far to move the atom when neither seems exclusively applicable.

(b) Residues 4, 5 and 6

Most of the main-chain atoms of these residues lie in a plane perpendicular to z, and figure (17) shows the section z = 120/240 for the 6 maps discussed on page 122, together with some features on the section z = 114/240, and the atoms falling within 0.3Å of those sections.

Maps 1, 2 and 3 look quite similar, with the top of the ridge of electron density corresponding to atoms Cα4 to Cα5 slightly to the right of the atomic positions used for the phasing. Map 2 is apparently contoured at a slightly higher level than map 1 because the weights of the Fourier coefficients are always less than 1. Map 3 is at an effectively lower resolution than map 1 because the weighting function FF_c tends to decrease with increasing sin θ. Map 6 confirms that the atoms are approximately correct, but that they should be moved to the right by about 0.7Å (N.B. 1Å ≡ 5 mm.); this is a larger distance than a simple interpretation of any of the maps 1, 2 or 3 would indicate (about 0.3Å).

Figure (18) shows the lowest contour level on the z = 120/240 section for map 2 (solid) superimposed on the corresponding contour



map 1 dotted
map 2 solid

map 3 dotted
map 2 solid

Figure (18)

The lowest contour level of the section
z 120/240 to compare maps 1,2 and 3.

for map 1 (left, dotted) and map 3 (right, dotted). This shows more clearly than figure (17) that the Sim weighted map shows a larger shift to the right than the unweighted map — and reference to map 6 in figure (17) shows that a shift to the right larger than either map 1, 2 or 3 indicates, is required. Sim weighting is thus an improvement in this example (and others, not discussed). The comparison of maps 2 and 3 in figure (18) shows that map 3 is of rather lower resolution, but there is no other apparent systematic difference.

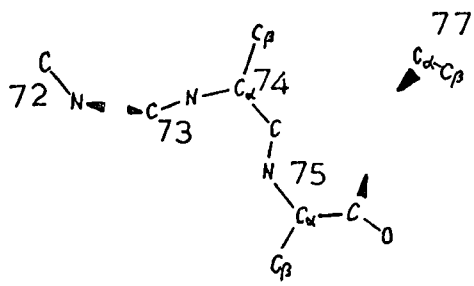
Returning to figure (17), we see that the difference map, map 4, shows that the C_α and N atoms of residue 6 are not near a negative region, but that the atoms $C_\alpha 4$ to $C_\alpha 5$ are near negative regions. The distance between the negative and positive contours across atom $C_\alpha 4$, for example, is about 0.8Å (4 mm.), and so equation (67) indicates that it should be moved by about 0.7Å, in agreement with the shift predicted by map 6.

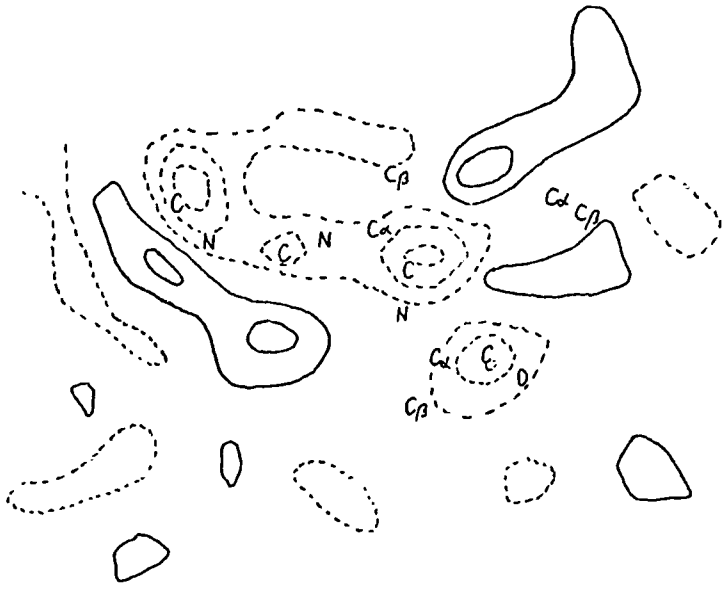
The atoms N6, $C_\alpha 6$, $C_\beta 6$ and $S_\delta 6$ were included in the 327 atom set, but the others shown in figure (17) were omitted. As may be seen, a region of electron density appears in map 5 about 0.5Å to the right of the original position of the omitted main chain atoms; $C_\beta 5$ and $C_\gamma 5$ do not reappear on map 5. The 4 atoms used for phasing which appear in this region are near to positive regions, which indicate that they should be moved up and to the right.

Figure (19)

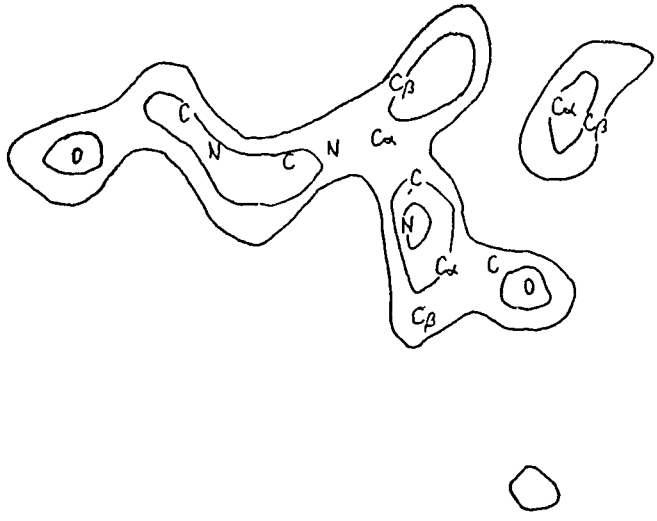
A superimposition of parts of the sections z = 72/240 and 78/240, with the atoms lying near those sections.

The atoms are

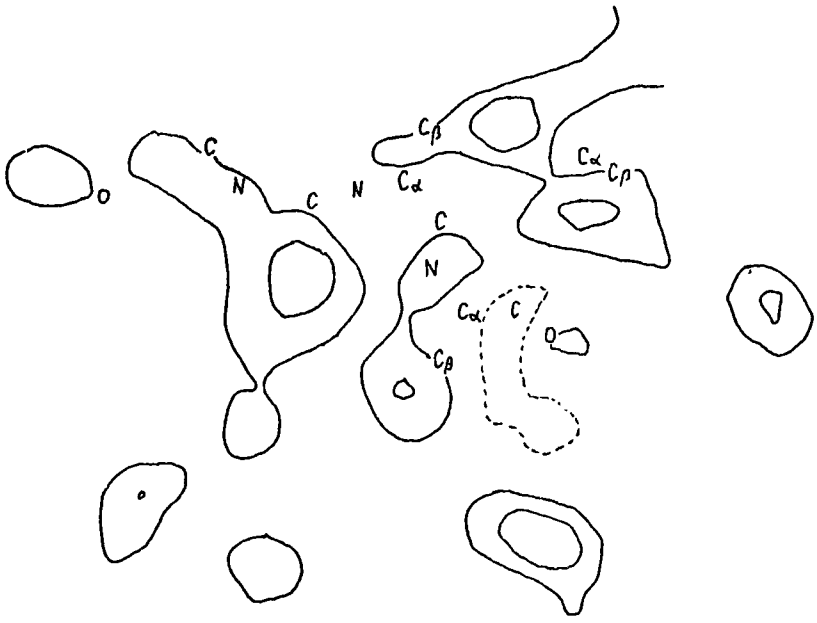




map 4



map 2



map 5



map 6

We see then that any of maps 1 to 5 would have told us fairly clearly to move the main chain atoms towards the position indicated by map 6. The directions of the shifts of individual atoms, however, are not usually such as to preserve the bond lengths and angles: we would have to average the shifts in some way over a group of atoms in order to preserve the bond lengths.

Residue 4 changes from glycine in hen to glutamic acid in human lysozyme, and so no side chain atoms were included in residue 4. Maps 1, 2 and 3 show very little, if any, trace of electron density for the side chain, but maps 4 and 5 both have a spur from the main chain at $C_\alpha 5$ (not quite continuous on map 5); a wire model of an ℓ-amino acid shows that with N5 above and to the left of $C_\alpha 5$, that $C_\beta 5$ must be in the spur shown. Map 6 shows no electron density connecting to the main chain at this point, and so in this example the α_{calc} phases are better than the M.I.R. phases. (A minor binding site of uranyl nitrate in the M.I.R. study was near glu_4, which may be why the electron density is poor in this region.)

(c) Residues 72, 73, 74, 75, 77

Figure (19) shows a superimposition of part of sections z = 72 and 78/240, with the atoms that lie nearer to those sections than to any other. Many of the main-chain atoms of the above residues form a roughly planar group which is fairly easily drawn.

Map 4 shows that many of these atoms lie near the bottom of

holes in the difference map, and it is not clear to which positive regions these atoms should be moved — especially the atoms of residues 74 and 75. In the choice of atoms to calculate map 5, all of these were omitted.

If map 2 had been more closely examined in conjunction with map 4, most of the ambiguity would have been resolved; residues 72 and 73 obviously move down the page, 77 is approximately correctly placed, and 74 and 75 are probably also in nearly the correct position: their appearance in holes in map 4 is probably due to their high temperature factor, since this region of the molecule is a section on the outside of the molecule with no very well defined secondary structure. (As figure (16) shows, these residues make few hydrogen bonds to the rest of the molecule.) These conclusions are verified by reference to map 6.

Map 5 has perhaps been contoured at too high a level. It clearly has features in common with map 6, and a poorly connected band of electron density runs from C72 to N75. (N.B. none of the atoms drawn in figure (19) was used in the phasing of map 5.) There is also a superimposed noise, which it is not easy (and probably not possible) to distinguish from wanted features — the hole into which C75 falls is an example.

(d) Side chains

There are 18 residues where there are more than two extra side

chain atoms to be found. In about half of these, some at least of the side chain could be seen on map 4 unambiguously in the same place as on map 6; in these residues the rest of the side chain usually appeared clearly in map 5 too. An example is residue 77, where C_α, C_β and N were used for the phasing of maps 2 and 4, but not for map 5; the missing atoms of the histidine ring were clearly visible on map 4, just noticeable on map 2, and the whole ring was a definite blob (showing flattening perpendicular to the ring) on map 5.

On the other hand when the side chains were indistinct on map 6 they were usually indistinct on the other maps too. One example is tyrosine 62, which is unclear on both M.I.R. and α_{calc} maps. However, this is not always the case — in the example of glutamic acid 4 given earlier map 4 was better than map 6. Another exception — in the opposite sense — is arginine 41, where on map 4 any of three isolated blobs could represent the guanidium group, and map 5 showed no suitable electron density at all, while a low level but continuous ribbon of electron density on map 6 connected a blob near one of those on map 4 to the main chain.

Return to figure (15), page 128, for another example of side chain completion. The large blob near the centre of the page on map 6, which has no real counterpart on map 2, is the sulphur atom of methionine 29 (which is valine in hen lysozyme). This sulphur

atom is less dense than that of cystine 115, probably because it has a higher temperature factor. There is some indication on map 2 of an extension of the side chain towards the sulphur atom position, but it is not very definite: map 4 shows no peak at the methionine 29 sulphur atom position.

Another point to make about side chains is that they often have higher temperature factors than main chain atoms, as mentioned on page 132. Aspartic acid 18, for example has its carboxylate group in a quite neat hole in map 4; the whole residue was omitted from the 327 atoms set for this reason. Map 2 does not indicate that it has moved, for the atoms lie at the centre of the blob of electron density on this map. Map 5 shows the reappearance of the carboxylate group, with separate lobes for each oxygen atom — though this is probably a coincidence — and the $O_{\delta_1} - O_{\delta_2}$ vector is in the same direction as on map 6.

(e) Rearrangements (17, 23, 28, 105, 111)

The regions discussed so far have shown that small movements of the main chain can be detected by the maps 2, 4 and 5, and that omitted side chains can often be placed with the aid of these maps. The M.I.R. study has shown that linked conformational changes involving up to five side chains occur, and it is important to see if difference maps can unravel what is happening. One example discussed by Stephen Banyard is the set of residues 17, 23, 28, 105, 111.

The two central residues of this cluster, 23 and 105 become smaller in changing from hen to human — changing from tyrosine and methionine respectively to isoleucine. Tryptophan 28 twists inwards, partially to fill in this hole, and 17 expands from leucine to methionine, the extra sulphur atom occupying part of the space occupied by tryptophan 28 in hen lysozyme. On the other side the benzene ring of tryptophan 111 moves inwards too, while the bottom of the pyrrole ring of 111 hinges outwards slightly.

Can these movements of some atoms to positions close to those vacated by other atoms be found by difference maps? When one knows what one is looking for, features in map 4 indicative of most of the changes mentioned in the previous paragraph can be found. The benzene ring of trp 28 lies in a hole in map 4, with a peak between it and met 17, which represents the position of the new sulphur atom of met 17, rather than a movement of trp 28; map 2 seems to show that no change is necessary, while map 5 shows a small peak for the sulphur of met 17 and a very small peak where the benzene ring of trp 28 is in hen lysozyme. Although no other peak is a possible site for the new sulphur atom of met 17, this is not strong enough evidence to force on us the correct interpretation of these difference maps.

The other moved tryptophan, 111, also lies in a hole in map 4,

with a peak on map 4 near the pyrrole ring to indicate its outward movement. There is a peak to represent the inward movement of the benzene ring, but this movement is about 5Å, and the positive peak would not necessarily be associated with a negative one 5Å away. Map 2 clearly shows that trp 111 has moved (N.B. map 2 does not always show changes clearly), and map 5 has a fairly large blob of density where the benzene ring of trp 111 should be, but poorly connected to the main chain. In short, the whole region is recognisably disturbed, but difference maps do not tell precisely how.

(f) Rearrangements (1, 85, 86)

The other region which the M.I.R. study has shown to be different in hen and human lysozymes is residues 1, 85 and 86. In hen lysozyme the side chain of serine 85 points outwards, and the atoms $C_\alpha 85$ to $C_\alpha 87$ (when viewed along x with z vertical) adopt a V formation with $C_\alpha 86$ at the point of the V; the side chain of lysine 1 sticks out into the inter-molecular liquid. In human lysozyme, glutamine 85 moves into the body of the molecule by about 1.7Å with the side chain pointing inwards, displacing two water molecules; asparagine 86 now remains in a roughly horizontal plane joining 85 and 87, which move apart slightly; the side chain of lysine 1 now points upwards towards the position occupied by ser 86 in hen lysozyme.

It must be admitted that the difference maps of this region are rather disappointing. The inwards motion of 85 is apparent, and 86 seems to have moved slightly, but not very dramatically, according to maps 2 and 4. The main chain atoms of 85 and 87 were included in the 327 atom set, and map 5 shows electron density at the site of the omitted 86. We know from the M.I.R. studies that this in fact represents the side chain of lys 1, but this is not clear from map 5. Lys 1 was omitted from the 327 atom set because it showed movements on map 4, and the main chain reappears — as it should — on map 5. However the old position of the side chain has no electron density on map 5; instead there is a blob of electron density in the direction which map 6 tells us the side chain is now pointing, but it is not continuous with the electron density at the site of 86 in hen lysozyme.

Thus these difference maps do not indicate that an important rearrangement has taken place in this region.

4. CONCLUSIONS

Inter-species comparisons of proteins have now been made at low resolution for three proteins: seal and spermwhale myoglobin (Tollin), hagfish and pig insulin (Sabesan and Dodson, unpublished), and recently pig and dogfish lactate dehydrogenase (Hackert, Ford and Rossmann). However, no very precise conclusions could be drawn from these low resolution studies, apart from one of the general similarity of the proteins from different species. The refinement of rubredoxin from an inaccurate set of atomic positions and 1.5Å structure factor amplitudes by Watenpaugh et al. has shown that refinement of protein structures is possible. The work in this thesis on human lysozyme firstly adds to the list of low resolution studies above, and proceeds to explore the differences between the proteins at high resolution.

(i) Locating the model

When rotation function and translation function searches fail to find a set of parameters which allow a postulated model to account for the observed X-ray diffraction intensities, we do not know if this is because the model is basically wrong, or if our rotation or translation search methods are inadequate.

(a) Orthorhombic lysozyme

The failure to locate the lysozyme molecule in the high temperature crystal of hen lysozyme is surprising, especially so in view of the success with human lysozyme. Imoto et al. note that "lysozyme is an unusually stable protein"; the general similarity of hen and human lysozymes in different space groups suggests that crystal forces are not major determinants of the lysozyme molecules' conformation. Thus although there is a reversible change in the conformation of lysozyme in solution (see Imoto et al. page 772), the transition temperature never seems to be below about 70°C in the absence of chemical denaturants, and the effect of this thermal denaturation process at 25°C would not be expected to be large. There is a report (see Imoto et al. page 785) that lysines 33 and 96 can be covalently linked in solution, implying that there can be conformations of lysozyme in solution very different from that in the crystal. If this report is true, which Imoto et al. seem to doubt, such a conformation may be the basis of the high temperature crystal form of lysozyme. The failure of the rotation and translation searches does not prove however, or even in my opinion strongly indicate, that the conformation in the high temperature crystal is very different.

(b) α-lactalbumin

The failure in the case of goat α-lactalbumin is less surprising. There is a number of physical, chemical, and physiological comparisons

of bovine α-lactalbumin and hen lysozyme, reviewed by Arnon and Maron. They seem roughly equally divided on whether there is or is not a substantial similarity in the conformations of these two proteins in solution. α-lactalbumins always seem to crystallise as a dimer — see Fenna — whereas although lysozyme often crystallises as a dimer, this is less common than as a monomer — see Imoto et al. page 684.

Whereas the successful rotation function studies showed one clear-cut peak, the rotation function studies on α-lactalbumin and lysozyme showed no definite monolithic peak. Michael Joynson's studies with all the lysozyme atoms and with just the backbone atoms, and my studies with some of the side-chain atoms showed that the position of the largest peak on the rotation function varied by up to 10^{o} depending on which atoms were chosen. This peak was not smooth, but had several shoulders, which may be caused by there being regions in α-lactalbumin which are similar to corresponding regions in lysozyme, but these regions may be assembled in a slightly different way in the two molecules.

(c) The methods

Most of the methods used in chapter 2 can be described as matching observed and calculated Patterson functions. The assumptions of Wilson statistics — i.e. random atomic positions — imply that the Patterson function will be composed of inter-atomic vectors

which are random too, and so the Patterson function of lysozyme, which obeys Wilson statistics quite well, will not have a few outstanding features, but a larger number of small features. Lysozyme does not have very much regular secondary structure, compared to myoglobin which is almost entirely helical, or lactate dehydrogenase, which has a large sheet region; these latter two proteins have larger features on their Patterson functions than lysozyme because the inter-atomic vectors from these regular secondary structure features will tend to add rather than be uncorrelated.

A lack of strong features on the Patterson functions will mean that there is less difference between correctly and incorrectly aligned Pattersons rather than when strong features are present. Referring to table (4) on page 46, we see that the proteins lysozyme and insulin, which have comparatively little secondary structure, presented more difficulty than the myoglobins (and lactate dehydrogenase, published too late to be included) which have substantial secondary structure.

When rotation and translation searches fail, it is not possible to know whether the rotational search has succeeded or not. However, I know of no cases where the rotation function showed a peak outstanding above the others and where the translation searches failed, and so it is probable that the rotation searches are failing, rather than the translation searches.

(ii) Completing and refining the model

The determination of the differences between hen and human lysozyme differs from the refinement of rubredoxin (Watenpaugh et al.) firstly in that the atomic displacements are larger, secondly they are linked changes rather than being essentially random errors, thirdly the human lysozyme structure factors only extend to 2.5Å resolution, rather than 1.5Å for rubredoxin, and fourthly that only 83% of the non-hydrogen protein atoms are known for human lysozyme, rather than almost all the rubredoxin atoms.

The poorer resolution in the human lysozyme does not seem to be troublesome where the refinement is making small movements to the main chain, or adding atoms onto side chains. When a rearrangement of the atoms occurs however, this may be the root of the problem. If a particular region in the molecule has electron density features which may be assigned to the amino acid residues in various ways, the precise electron density in the connections between the blobs of electron density is crucially important, and perhaps the failure to show that residues 1, 85 and 86 had rearranged was because these connections are too imprecise at this resolution.

When the least squares refinement of amino acid residue occupancies was done it was thought that some groups might have moved by several Ångstroms, and that this would be revealed by a low occupancy at the original site. It now appears that these movements are much less, and perhaps a refinement of temperature factors would have

been more successful. Temperature factors are a measure of the precision of the locations of the atoms; they can reveal inaccuracies in the positions, or by reaching very high values, that atoms are completely wrong — see the paper by Rossmann, Jacobson, Hirshfeld and Lipscomb, for example. It may be necessary to treat each atom separately, since temperature factors do vary considerably within one residue (see Watenpaugh et al.); an attempt to refine 1028 temperature factors with 4000 data would probably have a sufficient excess of data to parameters, for Watenpaugh et al. refined 2000 parameters (x,y,z and B of each atom of rubredoxin) with 5000 data. However, the model of rubredoxin which was refined was more complete and more accurate than that of human lysozyme.

A complete least squares refinement of 4000 parameters (x, y, z and B for each atom) with 4000 data is not possible: higher resolution would obviously be necessary. However, I see no reason why an improved model of human lysozyme could not be obtained by the methods begun at the end of chapter 3 of this thesis (page 131 onwards); the determination of which atoms needed rearranging could probably be found by least squares refinement of temperature factors, and those atoms with high temperature factors would be omitted. Genuinely imprecise atoms would then reappear on a map phased from the remaining atoms, and the rearranged atoms should be then visible.

APPENDIX A

The Difference Fourier

Referring to figure (13) on page 97, we see that

$$F^2 = F_c^2 + F_Q^2 + 2F_cF_Q \cos\theta$$

where $\theta = \alpha_Q - \alpha_c$

Defining $x = F_Q/F_c$, we develop an expression for difference Fourier coefficients as a series in ascending powers of x.

$$F^2 = F_c^2(1 + x)^2 \left(1 + \frac{2x(\cos\theta - 1)}{(1 + x)^2}\right)$$

from the above equation for F . A Taylor's series expansion for $(1 + a)^{\frac{1}{2}}$ is

$$(1 + a)^{\frac{1}{2}} = 1 + \tfrac{1}{2}a - \tfrac{1}{8}a^2 + \tfrac{1}{16}a^3 - \tfrac{5}{128}a^4 + \ldots \qquad (A1)$$

Now $F = F_c(1 + x)\left(1 + \frac{2x(\cos\theta - 1)}{(1 + x)^2}\right)^{\frac{1}{2}}$

and using equation (A1) in this gives

$$F - F_c = F_c\{x + \frac{x}{1+x}(\cos\theta - 1) - \tfrac{1}{2}\frac{x^2(\cos\theta - 1)^2}{(1 + x)^3}$$

$$+ \tfrac{1}{2}\frac{x^3(\cos\theta - 1)^3}{(1 + x)^5} - \tfrac{5}{8}\frac{x^4(\cos\theta - 1)^4}{(1 + x)^7} + \ldots\ldots\}$$

Expanding the $(1 + x)^{-n}$ as a Taylor's series gives

$$
\begin{aligned}
F - F_c = F_c\{ & x \\
& + (x - x^2 + x^3 - x^4 + \ldots)\ (\cos\ - 1) \\
& - \tfrac{1}{2}(\ \ x^2 - 3x^3 + 6x^4 - \ldots)\ (\tfrac{3}{2} - 2\cos\theta + \tfrac{1}{2}\cos 2\theta) \\
& + \tfrac{1}{2}(\ \ \ \ x^3 - 5x^4 + \ldots)\ (\tfrac{5}{2} + \tfrac{15}{4}\cos\theta - \tfrac{3}{2}\cos 2\theta + \tfrac{1}{4}\cos 3\theta) \\
& - \tfrac{5}{8}(\ \ \ \ \ \ x^4 - \ldots)\ (\tfrac{35}{8} - 7\cos\theta + \tfrac{7}{2}\cos 2\theta - \cos 3\theta \\
& \qquad + \tfrac{1}{8}\cos 4\theta)\}
\end{aligned}
$$

$$
\begin{aligned}
F - F_c = F_c\{ & \tfrac{1}{4}x^2 + \tfrac{1}{64}x^4 + \ldots \\
& + \cos\theta(1 - \tfrac{1}{8}x^3 + \ldots) + \cos 2\theta(-\tfrac{1}{4}x^2 + \tfrac{1}{16}x^4 - \ldots) \\
& + \cos 3\theta(\ \ \tfrac{1}{8}x^3 - \ldots) + \cos 4\theta(\ \ \ \ -\tfrac{5}{64}x\ + \ldots)
\end{aligned}
$$

$$
\begin{aligned}
F - F_c = {} & F_Q \cos\theta + \frac{F_Q^2}{4F_c}(1 - \cos 2\theta) + \frac{F_Q^3}{8F_c^2}(\cos 3\theta - \cos\theta) \\
& + \frac{F_Q^4}{64F_c^3}(1 + 4\cos 2\theta - 5\cos 4\theta) + \ldots . \qquad \text{(A2)}
\end{aligned}
$$

Equation (A2) is the beginning of a power series in $x = F_Q/F_c$; it is tested in table (A1) where column 3 is the exact value of $F-F_c$ and column 5 is the value according to equation (A2). It is seen to be a quite good approximation when $x < 1$ and a moderately good one when $x = 1$. For the case when $x > 1$ we define

$$z = F_c/F_Q$$

and derive an expression for $F-F_c$ as a power series in z. The

$\frac{F_Q}{F_c}$	θ cycles	$\lvert F-F_c \rvert$	Eq.(A4)	Eq.(A2)	Eq.(A3)
0.2500	0.0000	0.2500	0.2500	0.2500	0.2500
0.2500	0.1250	0.1900	0.1914	0.1900	4.1287
0.2500	0.2500	0.0308	0.0500	0.0308	-6.7501
0.2500	0.3750	-0.1580	-0.0914	-0.1580	8.3713
0.2500	0.5000	-0.2500	-0.1500	-0.2500	-1.7500
0.5000	0.0000	0.5000	0.5000	0.5000	0.5000
0.5000	0.1250	0.3990	0.4024	0.3998	0.7500
0.5000	0.2500	0.1180	0.1667	0.1172	-0.5000
0.5000	0.3750	-0.2632	-0.0690	-0.2631	0.7500
0.5000	0.5000	-0.5000	-0.1667	-0.5000	-1.5000
0.7500	0.0000	0.7500	0.7500	0.7500	0.7500
0.7500	0.1250	0.6196	0.6245	0.6260	0.6984
0.7500	0.2500	0.2500	0.3214	0.2417	0.1204
0.7500	0.3750	-0.2916	0.0184	-0.2855	-0.0873
0.7500	0.5000	-0.7500	-0.1071	-0.7500	-1.2500
1.0000	0.0000	1.0000	1.0000	1.0000	1.0000
1.0000	0.1250	0.8478	0.8536	0.8741	0.8741
1.0000	0.2500	0.4142	0.5000	0.3750	0.3750
1.0000	0.3750	-0.2346	0.1465	-0.1866	-0.1866
1.0000	0.5000	-1.0000	0.0000	-1.0000	-1.0000
1.2500	0.0000	1.2500	1.2500	1.2500	1.2500
1.2500	0.1250	1.0809	1.0873	1.1581	1.0920
1.2500	0.2500	0.6008	0.6944	0.4761	0.5860
1.2500	0.3750	-0.1085	0.3016	0.0809	-0.0960
1.2500	0.5000	-0.7500	0.1389	-1.2500	-0.7500
1.5000	0.0000	1.5000	1.5000	1.5000	1.5000
1.5000	0.1250	1.3176	1.3243	1.5012	1.3230
1.5000	0.2500	0.8028	0.9000	0.4922	0.7963
1.5000	0.3750	0.0624	0.4757	0.5731	0.0659
1.5000	0.5000	-0.5000	0.3000	-1.5000	-0.5000
1.7500	0.0000	1.7500	1.7500	1.7500	1.7500
1.7500	0.1250	1.5568	1.5636	1.9349	1.5597
1.7500	0.2500	1.0156	1.1136	0.3589	1.0124
1.7500	0.3750	0.2600	0.6637	1.3549	0.2610
1.7500	0.5000	-0.2500	0.4773	-1.7500	-0.2500
2.0000	0.0000	2.0000	2.0000	2.0000	2.0000
2.0000	0.1250	1.7979	1.8047	2.5000	1.7996
2.0000	0.2500	1.2361	1.3333	-0.0000	1.2344
2.0000	0.3750	0.4736	0.8619	2.5000	0.4738
2.0000	0.5000	0.0000	0.6667	-2.0000	0.0000

Table (A1)

Some approximations to $\lvert F-F_c \rvert$

steps are similar to the previous case, and are given somewhat abbreviated.

$$F = F_Q^2(1+z)^2\left(1+\frac{2z(\cos\theta-1)}{(1+z)^2}\right.$$

$$F = F_Q(1+z)\left(1+\frac{z(\cos\theta-1)}{(1+z)^2} - \frac{1}{2}\frac{z^2(\cos\theta-1)^2}{(1+z)^4} + \text{etc.}\right)$$

$$F-F_c = F - zF_Q$$

$$F-F_c = F_Q\left(1+\frac{z(\cos\theta-1)}{1+z} - \frac{1}{2}\frac{z^2(\cos\theta-1)^2}{(1+z)^3} + \text{etc.}\right)$$

$$\begin{aligned} F-F_c &= F_Q\{1+(z-z^2+z^3-z^4+\ldots)(\cos\theta-1) \\ &\quad -\tfrac{1}{2}(z^2-3z^3+6z^4-\ldots)(\tfrac{3}{2}-2\cos\theta+\tfrac{1}{2}\cos2\theta) \\ &\quad +\ldots.\} \\ &= F_Q + F_c(\cos\theta-1) + \frac{F_c^2}{4F_Q}(1-\cos2\theta) + \frac{F_c^3}{8F_Q^2}(\cos3\theta-\cos\theta) \\ &\quad + \frac{F_c^4}{64F_Q^3}(1+4\cos2\theta-5\cos4\theta)+\ldots \end{aligned} \qquad \text{(A3)}$$

In table (A1), column 6 gives the value of $F-F_c$ given by equation (A3) for various values of F_Q/F_c and θ It may be seen that equation (A3) is a good approximation when $F_Q > F_c$ and no doubt a better still approximation would result from taking more terms in equation (a3) - and (A2) too.

Eleanor Dodson and Vijayan take only the first two terms in equation (A1), and they deduce (for any value of F_Q/F_c)

$$F - F_c = \frac{F_Q^2}{F_Q + F_c} + \frac{F_Q F_c}{F_Q + F_c} \cos\theta \qquad (A4)$$

Table (A1) shows in column 4 the values of $F-F_c$ predicted by equation (A4); we see that it is usually a poorer approximation to the true value of $F-F_c$ than that given by equation (A2) when $F_Q < F_c$, or by equation (A3) when $F_Q > F_c$.

In the limit of $F_Q << F_c$, the right hand side of equation (A4) becomes

$$F_Q \cos\theta + \frac{F_Q^2}{F_c} \qquad (A5)$$

whereas equation (A2) becomes

$$F_Q \cos\theta + \frac{F_Q^2}{4F_c}(1 - \cos 2\theta) \qquad (A6)$$

As is discussed on page 106, the term with no dependence on θ represents a "ghost" of the model structure used to calculate F_c; a comparison of equations (A5) and (A6) shows that this "ghost" has only one quarter of the strength attributed to it by Eleanor Dodson and Vijayan.

APPENDIX B

Mixed centric and acentric data

The entries in table (2) for D = 0 give the expected value of the residual R_1 for the centric and acentric probability distributions of Wilson (1949). In 1950 Wilson had derived these values of $2\sqrt{2} - 2$ and $2 - \sqrt{2}$ (his equations (8) and (10)) for the centric and acentric probability distributions respectively. From equations (1) and (5) of Wilson's 1950 paper

$$R_{random} = \frac{2\langle F\rangle - 4\langle G(F)\rangle}{\langle F\rangle} \qquad \text{(B1)}$$

where $G(F) = \int_0^F F\,P(F)\,dF$ (his equation (4))

If a proportion c of the structure factors obey the centric distribution, and (1-c) obeys the acentric, equation (B1) becomes

$$R_{random} = \frac{2c\langle F\rangle_C + 2(1-c)\langle F\rangle_A - 4c\langle G\rangle_C - 4(1-c)\langle G\rangle_A}{C\langle F\rangle_C + (1-c)\langle F\rangle_A} \qquad \text{(B2)}$$

where the subscripts C and A refer to the centric and acentric probability distributions used to find the mean values of F and G. Values for $\langle F\rangle_C$, $\langle F\rangle_A$, $\langle G\rangle_C$ and $\langle G\rangle_A$ are given by Wilson. Substituting these in equation (B2) gives

$$R_{random} = \frac{(4-2\sqrt{2})c + \pi(1 - \frac{\sqrt{2}}{2})(1-c)}{\sqrt{2}c + \frac{\pi}{2}(1-c)} \qquad \text{(B3)}$$

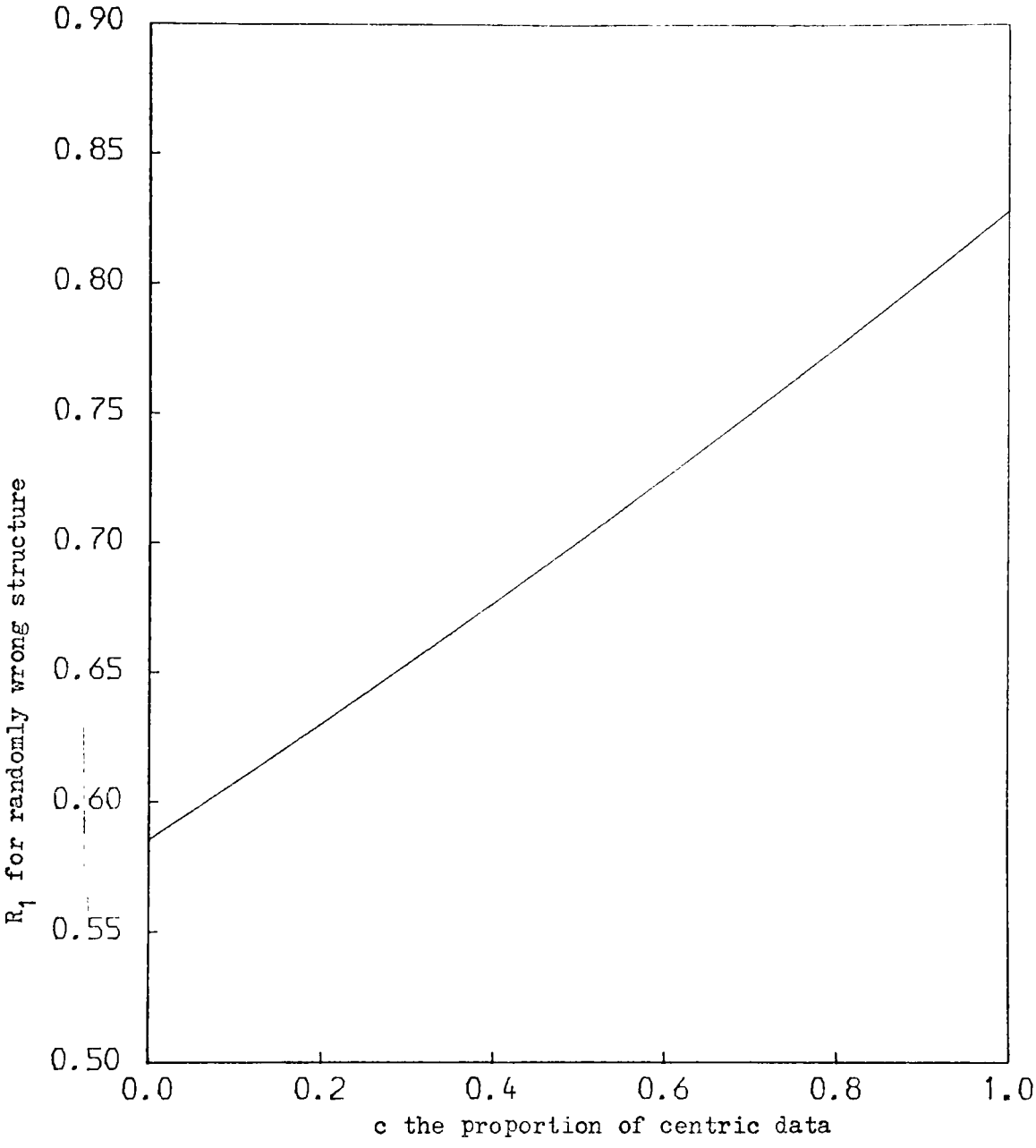

0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
R_1 for randomly wrong structure
0.0
0.2
0.4
0.6
0.8
1.0
c the proportion of centric data

Figure (B1)

This reduces to Wilson's equations (8) and (10) when c = 1 and c = 0 respectively. Figure (B1) shows equation (B3) as a function of c; it may be seen that to approximate this to a straight line will never produce an error in the expected value of R_{random} greater than 0.006. The value of R_1 corresponding to a randomly wrong structure for data which is partly centric and partly acentric may thus be safely approximated by linear interpolation between the extreme values of 0.586 (all acentric) and 0.828 (all centric).

REFERENCES

Anderson, Hodgkin and Viswamitra, "The structure of thiostrepton" Nature (1970) 225, 233.

Arnon and Maron, "An immunological approach to the structural relationship between hen egg-white lysozyme and bovine α-lactalbumin". J.Mol.Biol. (1971) 61, 225

Aschaffenburg, Fenna, Handford and Phillips, "Crystallography of α-lactalbumin. I Low-salt crystals of goat α-lactalbumin" and "II High-salt crystals of goat α-lactalbumin". J.Mol.Biol. (1972) 67, 525 and 529.

Banyard and Blake, see Banyard, D.Phil thesis, Oxford, 1973.

Barrett and Zwick, "A method for the extension and refinement of crystallographic protein phases utilizing the fast Fourier transform". A.C. (1971) A27, 6.

Barry, North, Glassel, Williams and Xavier, "Quantitative determination of mononucleotide conformations in solution using lanthanide ion shift and broadening N.M.R. probes". Nature (1971) 232, 236.

Blake, Koenig, Mair, North, Phillips and Sarma, "Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2Å resolution", Nature (1965) 206, 757.

Blake and Swan, "X-ray analysis of structure of human lysozyme at 6Å resolution", Nature (1971) 232, 12.

Blow and Crick, "The treatment of errors in the isomorphous replacement method", A.C. (1959) 12, 794.

Brew, "Lactose synthetose: evolutionary origins, structure and control", in Essays in biochemistry (ed. Campbell and Dickens) (1970) 6, 93.

Browne, North, Phillips, Brew, Vanaman and Hill, "A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme", J.Mol.Biol. (1969) 42, 65.

Buerger, Vector Space, Wiley (1959).

Cochran, "Relations between the Phases of Structure Factors", A.C. (1955) 8, 473.

Crowther and Blow, "A method of positioning a known molecule in an unknown crystal structure", A.C. (1967) 23, 544.

Cruickshank, "The convergence of the least-square and Fourier refinement methods", A.C. (1950) 3, 10.

Fenna, D.Phil thesis, Oxford, 1973.

Fletcher and Powell, "A rapidly converging descent method for minimization", Comp.J. (1963) 6, 163.

Foster and Hargreaves, "The use of moments of X-ray intensity in space group determination", A.C. (1963) 16, 1124 and 1133. See table I, page 1134.

Gassmann and Zechmeister, "Limits of phase expansion in direct methods", A.C. (1972) A28, 270.

Hackert, Ford and Rossmann, "Molecular orientation and position of the pig M_4 and H_4 isoenzymes of lactate dehydrogenase in their crystal cells". J.Mol.Biol. (1973) 78, 665.

Hoppe and Gassmann, "Phase Correction, a New Method to solve Partially Known Structures". A.C. (1968) B24, 97.

Huber, "Programmed 'Faltmolekül" method" in Crystallographic computing, Proc.. 1969, Ed. F.R. Ahmed, p.96.

Imoto, Johnson, North, Phillips and Rupley, "Vertebrate lysozymes", in The Enzymes (1972), vol. VII, Academic Press.

Jollès and Berthou, "High temperature crystallisation of lysozyme: an example of phase transition." FEBS Letters (1972), 23 21

Joynson, D.Phil. Thesis, Oxford, 1970.

Joynson, North, Sarma, Dickeson and Steinrauf, "Low resolution studies on the relationship between the triclinic and tetragonal forms of lysozyme", J.Mol.Biol. (1970) 50, 137.

Karle, "Translation functions and direct methods", A.C. (1972) B28, 820.

Karle and Karle, "The symbolic addition procedure for phase determination for centrosymmetric and noncentrosymmetric crystals", A.C. (1966) 21 849.

Kartha, referred to in Barrett and Zwick.

Klug, "The application of the Fourier transform method to the analysis of the structure of triphenylene, $C_{18}H_{12}$", A.C. (1950) 3, 176.

Knott, "Molecular structure factors and their application to the solution of the structures of complex organic crystals" (i.e. naphthalene and aliphatic chains), Proc.Phys. Soc.Lond. (1940) 52, 229.

Lattman, "Optimal sampling of the rotation function", A.C. (1972) B28, 1065.

Lattman and Love, "A rotational search procedure for detecting a known molecule in a crystal", A.C. (1970) B26, 1854.

Lattman, Nockolds, Kretsinger, Love, "Structure of Yellow Fin Tuna Metmyoglobin at 6Å Resolution", J.Mol.Biol. (1971) 60, 271.

Luzzati, "Traitement Statistique des Erreurs dans la determination des structures cristallines", A.C. (1952) 5, 802.

Luzzati, "Resolution d'une structure cristalline lorsque les positions d'une partie des atomes sont connues; traitement statistique", A.C. (1953) 6, 142.

Luzzati, "Sur l'emploi des méthodes statistiques dans l'étude de la structure cristalline des protéines", A.C. (1955) 8, 795.

Nordman, "An application of vector space search methods to the Patterson function of myoglobin", A.C. (1972) A28, 134.

Nordman and Nakatsu, "Interpretation of the Patterson function of crystals containing a known molecular fragment. The structure of an Alstonia alkaloid", J.A.C.S. (1963) 85, 353.

North and Phillips, "X-ray studies of crystalline proteins", Prog. in biophysics and molecular biology (1969)

Parthasarathy and Parthasarathi, "Discrepancy factors for use in crystal structure analysis", A.C. (1972) A28, 426.

Pauling and Shappell, "The crystal structure of bixbyite and the C-modification of the sesquioxides", Z.Krist. (1930) (A) 75, 128-142.

Phillips, "Advances in protein crystallography", Adv. in research by diffraction methods (1966) 2, 75.

Ramachandran and Srinivasan, Fourier methods in Crystallography, Wiley (1970)

Raman, "Syntheses for the deconvolution of the Patterson function. Part II. Detailed theory for non-centro-symmetric crystals", A.C. (1959) 12, 964.

Reeke and Lipscomb, "On the Use of the Tangent Formula to Extend the Resolution of Protein Phases", A.C. (1969) B25, 2614. This paper is also on Tangent Formula refinement.

Rollett and Hodgson, Symposium 50 Jahre Röntgeninterferenzen, München (1962). See Hoppe and Gassmann, 1968.

Rossmann and Blow, "The refinement of structures partially determined by the isomorphous replacement method", A.C. (1961) 14, 641.

Rossmann and Blow, "The detection of sub-units within the crystallographic asymmetric unit", A.C. (1962) 15, 24.

Rossmann, Blow, Harding, Coller, "The relative positions of independent molecules within the same asymmetric unit". A.C. (1964) 17, 338.

Rossmann, Jacobson, Hirshfeld and Lipscomb, "An account of some computing experiences", A.C. (1959) 12, 530.

Sayre, "The squaring method: a new method for phase determination", A.C. (1952) 5, 60.

Sayre, "Some implications of a theorem due to Shannon", A.C. (1952) 5, 843.

Sayre, "On least-square refinement of the phases of crystallographic structure factors", A.C. (1972) A28, 210.

Srinivasan, "Weighting functions in crystal-structure analysis" Z.Krist. (1968) 126, 175. This paper is actually about expectation values of difference Patterson Fourier coefficients.

Stout and Jensen, 'X-ray structure determination, a practical guide, Macmillan (1968).

Stryer, Kendrew and Watson, "The mode of attachment of the azide ion to sperm whale metmyoglobin", J.Mol.Biol. (1964) 8, 96.

Tollin, "On the determination of molecular location", A.C. (1966) 21, 613.

Tollin, "Determination of the orientation and position of the myoglobin molecule in the crystal of seal myoglobin", J.Mol.Biol. (1969) 45, 481.

Tollin and Cochran, "Patterson function interpretation for molecules containing planar groups." A.C. (1964) 17, 1322.

Tollin, Main and Rossmann, "The Symmetry of the Rotation Function", A.C. (1966) 20, 404.

Tollin and Rossmann, "A description of various rotation function programs", A.C. (1966) 21, 872.

Vand and Pepinsky, "Determination of molecular translation and orientation using modified disagreement functions", Z. Krist. (1956) 108, 1.

Watenpaugh, Sieker, Herriot and Jensen, "Refinement of the model of a protein: rubredoxin at 1.5Å resolution", A.C. (1973) B29, 943.

Weinzierl, Eisenberg and Dickerson, "Refinement of protein phases with the Karle-Hauptman tangent formula", A.C. (1969) B25, 380.

Wilson, "The probability distribution of X-ray intensities", A.C. (1949) 2, 318.

Wilson, "Largest likely values for the reliability index", A.C. (1950) 3, 397.

Woolfson, "An improvement of the 'heavy-atom' method of solving crystal structures", A.C. (1956) 9, 804.

Woolfson, An introduction to X-ray crystallography, C.U.P. (1970).

Dodson and Vijayan "The determination and refinement of heavy atom parameters in protein heavy-atom derivatives. Some model calculations using acentric reflexions." A.C. (1971) B27 2402

Pinnock, Taylor and Lipson, "A re-determination of the structure of triphenylene." A.C. (1956) 9 173

Qurashi, "An extension of the inequalities method of sign determination by means of negative density transforms." A.C. (1953) 6 103

Sim, "The distribution of phase angles for structures containing heavy atoms." A.C. (1959) 12 813

Sim, "A note on the heavy atom method." A.C. (1960) 13 511