NEW TECHNIQUES FOR DIRECT METHODS IN X-RAY CRYSTALLOGRAPHY

BY

ALLAN NISBET HENDERSON

A thesis submitted to the University of Glasgow for the degree of Doctor of Philosophy in the Faculty of Science

Chemistry Department

February 1994

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

This Thesis is Dedicated to

Lilias W. Henderson

ACKNOWLEDGEMENTS

Many colleagues in the chemistry department have been very generous in their sharing of crystallographic knowledge and experience during the course of this research. In particular my supervisor Chris Gilmore, was a constant source of encouragement, technical advice and inspiration. I am indebted to others who also contributed to theories in particular Keith Henderson and Colin Bannister, and to those who maintain the computers on which all this work was performed, particularly Stuart McKay and latterly Chris Edwards.

I would also like to take this opportunity of thanking Professor M.M. Woolfson and Yao Jia-Xing for kindly providing the data that was used in the APP experiments.

Thanks are also due to Jonathan Silsby for providing the leave of absence from Electronic Data Systems that finally allowed this thesis to be produced.

Finally thanks are due to the S.E.R.C. for funding this research, and also to The British Crystallographic Association for funding attendance at conferences in the U.K.

ABSTRACT

The principal aim of this thesis is the further development of the methods of solution of crystal structures using the techniques of direct methods. All the research undertaken has used either a Bayesian approach to the statistics or has used the more specific maximum entropy technique. Much of the work results in an implementation to the MITHRIL program which is then used as part of the testing strategy, and is mentioned throughout this thesis.

The first chapter is an introduction to direct methods to give the reader an overview in the techniques which will be expanded upon later in the thesis. In addition to explaining the major techniques used in the field a section on the maximum entropy method is included along with a brief explanation of the function of the maximum entropy program MICE.

The second chapter details the maximum entropy method and presents the results from the entropy maximisation of maps produced using phase sets generated from random starting phases by the SAYTAN program. A small protein Avian Pancreatic Polypeptide (App) was used as the test structure. No conventional figure of merit was able to discriminate between the phase sets yet by applying standard maximum entropy procedures using the MICE program and examining the log likelihood gain (LLG) the correct phase sets were identified.

The third chapter details a Bayesian method of obtaining temperature factors, scale factors and estimated standard deviations on these figures for use in the normalisation of structure factors to normalised structure factors. A full derivation of the new formula using Bayesian methods and the Wilson statistics is provided along with details of the implementation into the MITHRIL90 program. A full set of test results based on selections of x-ray diffraction data for seventeen test structures is given. The results show that this is a perfectly adequate method that provides reasonable standard deviations of the normalised structure factors. The greatest advantage of this new theory is that it has the ability to be extended to use Bayesian priors to generate better normalisation equations.

The fourth chapter details a likelihood based figure of merit, LOGLIK, designed to compare observed and calculated E-magnitudes for the reflections that are not

involved in the direct methods phasing procedure. This gives a measure of the internal consistency of the three phase invariants used in phasing. A derivation of the formula that yields the calculated E-magnitudes is given. The results are given for twenty two test structures and correlations between LOGLIK and conventional figures of merit. The results show that while LOGLIK contains new information it has no new advantage over conventional figures of merit, and indeed can only be used for ranking phase sets into a preferred order not the determination of correctness.

Also included in the thesis is an appendix that contains the manual for the use of the MITHRIL90 program, that incorporates the new normalisation method and the LOGLIK figure of merit.

CONTENTS

LIST OF FIGURES

Chapter 1 INTRODUCTION TO DIRECT METHODS

Chapter 2 THE USE OF LIKELIHOOD IN THE SOLUTION OF THE STRUCTURE OF AVIAN PANCREATIC POLYPEPTIDE

LIST OF TABLES

CHAPTER 1

INTRODUCTION TO DIRECT METHODS

1.0 THE PHASE PROBLEM

1.1 Problem Definition

An X-ray diffraction pattern contains much information including; the shape and size of the unit cell, the symmetry within the crystal (or space group) and the relative atomic positions within the cell. The aim of the vast majority of X-ray diffraction experiments is the elucidation of these relative atomic positions.

The results obtained from a diffraction experiment is normally the diffraction pattern. A set of intensities can be routinely acquired, to high precision, using an automatic diffractometer, followed by mathematical correction for geometry and scaling. The intensity of each reflection is dependant on the positions of the atoms in the unit cell and the scattering power of each atom. These intensities can be related to structure factor amplitudes by EQ 1.1.1

$$|F_{\underline{h}}| = \sqrt{\frac{I_{\underline{h}}}{L_p}} \qquad \text{(EQ 1.1.1)}$$

Where

$|F_{\underline{h}}|$ is the structure factor amplitude of reflection $\underline{h}$

$I_{\underline{h}}$ is the measured intensity of reflection $\underline{h}$

L_p is the combined Lorentz factor and polarisation factor

To obtain an electron density map of our structure we require the parameters $\underline{x},\rho$ which are the Fourier transform of $\underline{h},\varphi$. The structure factors with which we will determine the atomic positions are complex numbers and are defined by EQ 1.1.2

$$F_{\underline{h}} = |F_{\underline{h}}| e^{-i\varphi_{\underline{h}}} \qquad \text{(EQ 1.1.2)}$$

Where

$F_{\underline{h}}$ is the structure factor of reflection $\underline{h}$

$\varphi_{\underline{h}}$ is the phase of reflection $\underline{h}$

As can be seen from EQ 1.1.2 it is necessary to determine the phase of the reflections before we can proceed to atomic co-ordinates. This phase cannot be measured and must therefore be calculated. This problem would normally be insoluble, but as we frequently have many more observed reflections than degrees of freedom to define the atomic positions, the problem becomes soluble in practice. This is the Crystallographic Phase Problem.

1.2 Practical Solutions

Of the several methods that have been devised to solve the phase problem, the principal three methods in common use at the moment are:

1. Patterson (heavy atom) method (Patterson, 1934) which relies on the structure containing a heavy atom which will dominate the scattering. The maxima in the Patterson function, which does not depend on the phases, yield the atomic positions of the heavy atoms from the vector peaks. Once the position of this powerfully diffracting atom is known then phases can be calculated for the structure factors and the remaining structure determined by Fourier methods. In general this technique is not applicable to structures that contain no heavy atoms.

2. Isomorphous Replacement, which relies on being able to soak into the crystal a solution containing a heavy atom without altering the crystal or molecular structure in any way. This heavy atom is then located by Patterson methods. This technique is used most frequently in the solution of protein structures.

3. Direct methods, so called, because an attempt is made to determine phases directly from the structure factor amplitudes using a series of equations that relate the two. This technique is the most commonly used for structure determination as there is no requirement for a heavy atom. This makes the technique applicable to the majority of small molecules that contain approximately all atoms of equivalent scattering power.

2.0 PHASE RELATIONSHIPS

Phases may be determined by mathematically relating structure factors. These relationships being purely mathematical in nature are perfectly suited to solution by the use of computers. It is the power of the technique coupled with recent advances in computing technology that has led to the proliferation of structure solution programs in the last decade.

The first of these structure factor relationships to be used was the Cauchy-Schwartz inequality relationships by Harker and Kasper (Harker & Kasper, 1948). This states that, for a centrosymmetric crystal, if $|F_{\underline{h}}|$ and $|F_{2\underline{h}}|$ are both large then $F_{2\underline{h}}$ is probably positive (i.e. $\varphi_{2\underline{h}} = 0$). This inequality is a consequence of the electron density being positive at all points within the crystal. As this equation is only true for centrosymmetric reflections with a large structure factor amplitude, these equations are inadequate for non-centrosymmetrics and are of very limited use for large, or complex structures that diffract more weakly.

A more generally applicable relationship is the Sayre Equation (Sayre, 1952) given as EQ 2.0.1

$$F_{\underline{h}} = \frac{\Theta_{\underline{h}}}{V} \sum_{\underline{k}} F_{\underline{k}} F_{\underline{h}-\underline{k}} \qquad \text{(EQ 2.0.1)}$$

Where

$$\Theta_{\underline{h}} = \frac{f_{\underline{h}}}{\gamma_{\underline{h}}} \qquad \text{(EQ 2.0.2)}$$

Where

$f_{\underline{h}}$ is the scattering factor for each atom

$\gamma_{\underline{h}}$ is the Fourier transform of the squared electron density

V is the volume of the unit cell

This equation makes no assumptions about the nature or shape of atoms within the unit cell. In order to take advantage of the knowledge that atoms are discrete points in

space we must express the structure factors as the normalised structure factor $E_{\underline{h}}$. Although the Sayre equation continues to have a practical use in modern direct methods its algebraic nature limits this usefulness.

New relationships, derived from probability theory, that were more generally applicable were developed by Karle and Hauptman (Karle & Hauptman, 1950) in the early fifties and quickly found practical application in the solution of naphthalene (Karle & Hauptman, 1953). These probabilistic relationships are still the basis of modern conventional direct methods.

3.0 NORMALISATION

3.1 Determination of $E_{\underline{h}}$

It is convenient to work with structure factors from which the effect of variation in atomic scattering factor with variation in $(sin^2\theta)/\lambda^2$, and the effect of thermal vibration has been eliminated as far as possible. This is called a normalised structure factor and for a reflection h is given the symbol $E_{\underline{h}}$. It may be defined in its simplest form as 3.1.1

$$E_{\underline{h}} = \frac{F_{\underline{h}}^{observed}}{\left|F_{\underline{h}}^{expected}\right|} \qquad \text{(EQ 3.1.1)}$$

$$\left|E_{\underline{h}}\right|^2 = \frac{K\left|F_{\underline{h}}\right|^2}{\varepsilon \sum_{j=1}^{N} f_j^2} \qquad \text{(EQ 3.1.2)}$$

Where

K is the scale factor required to place $F_{\underline{h}}$ onto an absolute scale.

ε is the epsilon factor which is dependant on the point group and reflection indices.

Equation 3.1.2 by the use of f_j assumes a point atom at rest. As an atom vibrates about a point its electron density will appear smeared and this lower density electron cloud will have a correspondingly reduced ability to diffract X-rays, so a correction must be made using equation 3.1.3

$$[f_j]_T = f_j e^{-Bsin^2\theta/\lambda^2} \qquad \text{(EQ 3.1.3)}$$

Where

$[f_j]_T$ is the scattering factor of the atom at temperature T

B can be shown to be defined by equation 3.1.4

$$B = 8\pi^2 U_j \qquad \text{(EQ 3.1.4)}$$

Where U_j is the mean squared amplitude of atomic vibration

Wilson proved (Wilson,1942) that a plot of $ln\left(\frac{\langle I_{rel}\rangle}{\sum_{j=1}^{N}[f_j]_T^2}\right)$ versus $(sin^2\theta)/\lambda^2$ will

give an approximate straight line, which when least squares fitted gives B, the isotropic temperature factor, from the slope, and also yields K, the scale factor, from the intercept at $\theta = 0^o$. If the plot of the points on the Wilson plot deviate significantly from the straight line then it is recommended that the K-curve technique devised by Karle and Hauptman (Karle & Hauptman,1953) is used to determine B and K.

Another method of determination of B and K has been developed based on a Bayesian statistical technique. This technique will be discussed fully in Chapter 2.

With both B and K determined then $|E_{\underline{h}}|$ may be calculated using equation 3.1.5

$$|E_{\underline{h}}|^2 = \frac{K|F_{\underline{h}}|^2}{\varepsilon\sum_{j=1}^{N} f_j e^{-2Bsin^2\theta/\lambda^2}} \qquad \text{(EQ 3.1.5)}$$

It is important to note that the reflection $\underline{h}$ does not change phase during the normalisation procedure.

3.2 Probability Distributions

The equations that define the probability distributions (equations 3.2.1 and 3.2.2) of normalised structure factors have no dependency on the atomic scattering factors, thus making the distribution independent of structural complexity. This is a major difference from the probability distributions of the structure factors themselves, which are structurally dependant.

$$P(|E|) = \frac{\sqrt{2}}{\sqrt{\pi}} e^{\frac{-|E|^2}{2}} \qquad \text{(EQ 3.2.1)}$$

for centrosymmetrics

$$P(|E|) = 2|E|e^{-|E|^2} \qquad \text{(EQ 3.2.2)}$$

for non-centrosymmetrics

The normalisation of the structure factors is a crucial part of the direct methods structure determination process, as all procedures that follow are dependant on the E-magnitudes. All the factors that can effect these magnitudes must be very carefully considered, as the normalisation procedure may determine whether a structure will be solved or not.

Using the fact that the fraction of the E's lying in the range a,b can be defined by equation 3.2.3

$$\int_a^b P(E)\,dE \qquad \text{(EQ 3.2.3)}$$

We may then predict ideal distributions for each type of reflection. These results are given in Table 1. By comparison with the statistics of our real data and the results in Table 1, we may determine if the data set, or a projection of the data is centric or acentric.

	Centric	**Acentric**
$\langle E \rangle$	0.798	0.886
$\langle E^2 \rangle$	1.000	1.000
$\langle E^3 \rangle$	1.596	1.329
$\langle E^4 \rangle$	3.000	2.000
$\langle (E^2-1) \rangle$	0.968	0.736
$\langle (E^2-1)^2 \rangle$	2.000	1.000
$\langle (E^2-1)^3 \rangle$	8.691	2.415

TABLE 1. Theoretical distribution of E-magnitudes

3.3 Unitary Structure Factors

Like normalised structure factors the unitary structure factor, $U_{\underline{h}}$, is corrected for fall off in scattering power with increasing $(sin^2\theta)/\lambda^2$. The unitary structure factor is defined as:

$$U_{\underline{h}} = \frac{F_{\underline{h}}}{\sum_{j=1}^{N} f_j} \qquad \text{(EQ 3.3.1)}$$

Since

$$|F_{\underline{h}}| \leq \sum_{j=1}^{N} f_j$$

Then $|U_{\underline{h}}| \leq 1$

While the need for unitary structure factors is conventional direct methods is near non-existent, the new structure solution technique of Maximum Entropy is more efficiently expressed in terms of $U_{\underline{h}}$.

4.0 ORIGIN AND ENANTIOMORPH DEFINITION

4.1 Origin Definition

In order to fully define a structure in real space we must define an origin to enable us to obtain a fixed reference frame for atomic co-ordinates. This is done in reciprocal space by assigning phases to a limited number of $|F_{\underline{h}}|$ or $|E_{\underline{h}}|$. In general there is a choice of origins for each space group since the space group defines the direction of the symmetry axes but not their absolute positions. The selection of origin defining reflections and the phases to be assigned to them, is determined by the International Tables for X-ray Crystallography. The origin defining reflections must contain no seminvariants and be linearly independent of each other, i.e. the combinations of 2 or 3 reflections must not be structure invariants or seminvariant.

The effect of changing the origin in real space will obviously have an effect on the phases of the reflection in reciprocal space, as the two are related by a Fourier transform. The change in phase due to a shift in origin is given by equation 4.1.1

$$2\pi\varphi'_{\underline{h}} = 2\pi\varphi_{\underline{h}} - 2\pi\,(\underline{h} \cdot \Delta\underline{x}) \qquad \text{(EQ 4.1.1)}$$

Where

$\Delta\underline{x} = (\Delta x, \Delta y, \Delta z)$

$\varphi'_{\underline{h}}$ is the phase at the new origin

$\varphi_{\underline{h}}$ is the phase at the old origin

4.2 Enantiomorph Definition

In non-centrosymmetric structures we lack a centre of symmetry and so only one enantiomorph can exist. In the absence of anomalous scatterers, measured intensities are insensitive to the enantiomorph and thus it must be defined. If an enantiomorph is not fixed at the beginning of a structure solution we may determine a solution where some reflections define one enantiomorph while the others define the alternate enantiomorph, resulting in both enantiomorphs appearing in the electron density map.

To define the enantiomorph we must assign a phase to a structure invariant, or seminvariant. In practice we would assign a phase to a single reflection with large E-magnitude that participated in an invariant whose phases sum was not 0 or π.

A change of enantiomorph entails the reversal of all phases i.e. $\varphi_{\underline{h}}$ becomes $-\varphi_{\underline{h}}$. This change results in a reversal of all invariants and seminvariants also. It may be thought of as a reflection of the atomic positions in the origin.

5.0 STRUCTURE INVARIANTS AND SEMINVARIANTS

5.1 Structure Seminvariants

Seminvariants are a consequence of space group symmetry and origin choice. They can be structure factors, a product of structure factors, a structure factor phase, or a sum of phases. These are all invariant with regard to origin shift, provided the origin choices are permissible, as defined in The International Tables for X-ray Crystallography.

For a seminvariant an origin shift of $\Delta\underline{x}$, from one permissible origin to another, should leave $\varphi_{\underline{h}}$ unchanged, i.e. $\Delta\varphi_{\underline{h}}$ should equal zero (or modulus 2π). This leads to the requirement that equation 5.1.1 must be true for a structure seminvariant.

$$\underline{h} \cdot \Delta\underline{x} = 0, or \ldots n = integer \qquad \text{(EQ 5.1.1)}$$

e.g. in space group $P2_1$ allowable origins lie at (0,y,0), (1/2,y,0), (0,y,1/2), (1/2,y,1/2)

$$\begin{bmatrix} h & k & l \end{bmatrix} \times \begin{bmatrix} 0 \\ y \\ 0 \end{bmatrix} = ky$$

$$\begin{bmatrix} h & k & l \end{bmatrix} \times \begin{bmatrix} 1/2 \\ y \\ 0 \end{bmatrix} = ky + \frac{h}{2}$$

$$\begin{bmatrix} h & k & l \end{bmatrix} \times \begin{bmatrix} 0 \\ y \\ 1/2 \end{bmatrix} = ky + \frac{l}{2}$$

$$\begin{bmatrix} h & k & l \end{bmatrix} \times \begin{bmatrix} 1/2 \\ y \\ 1/2 \end{bmatrix} = ky + \frac{h}{2} + \frac{l}{2}$$

The only type of reflection that can make all the above equations zero or an integer is; h is an even indice, k is zero and l is an even index e.g. (2,0,4). Any linear combination of phases that yields a total that is of the form (e,0,e) will also be a seminvariant e.g.

$$\varphi_{eko} + \varphi_{e\bar{k}o}$$

$$\varphi_{e\bar{k}o} + \varphi_{eke} + \varphi_{e0o}$$

Both the above linear combinations of phases are structure seminvariants in the space group $P2_1$.

Of particular use in direct methods are seminvariants of only one reflection as shown in the above examples. These are called Σ_1 relationships and form the basis of some figures of merit, they are also used in the phase determination process.

5.2 Structure Invariants

A structure invariant is a quantity the value of which depends only on the structure and not on the choice of origin, a simple example being that of E-magnitudes. As can be seen from equation 4.1.1 the phase of a reflection is frequently dependent on both the structure and on the position of the origin, however certain linear combinations of reflections are structure invariants.

Let us examine the product of three reflections:

$$E'_{\underline{h}}E'_{\underline{k}}E'_{-\underline{h}-\underline{k}} = E_{\underline{h}}e^{2\pi i(\underline{h}\cdot\Delta\underline{x})} \times E_{\underline{k}}e^{2\pi i(\underline{k}\cdot\Delta\underline{x})} \times E_{-\underline{h}-\underline{k}}e^{2\pi i(-\underline{h}-\underline{k}\cdot\Delta\underline{x})} \quad \text{(EQ 5.2.1)}$$

$$E'_{\underline{h}}E'_{\underline{k}}E'_{-\underline{h}-\underline{k}} = E_{\underline{h}}E_{\underline{k}}E_{-\underline{h}-\underline{k}}e^{2\pi i(\underline{Q}\cdot\Delta\underline{x})} \quad \text{(EQ 5.2.2)}$$

Therefore

$$E'_{\underline{h}}E'_{\underline{k}}E'_{-\underline{h}-\underline{k}} = E_{\underline{h}}E_{\underline{k}}E_{-\underline{h}-\underline{k}} \quad \text{(EQ 5.2.3)}$$

Since $|E|$ is a structure invariant then the sum $\varphi_{\underline{h}} + \varphi_{\underline{k}} + \varphi_{\underline{h}-\underline{k}}$ must also be an invariant.

This shows that the product of any three reflections for which the sum of the indices are zero, is a structure invariant. These three phase invariants are called triplets or Σ_2 relationships.

6.0 THE NEIGHBOURHOOD PRINCIPLE

It has been shown that for a fixed enantiomorph any invariant is uniquely determined by the observed E-magnitudes. For an invariant, in favourable circumstances, the phase is determined by a small set of E-magnitudes and is relatively insensitive to the remaining bulk of observed $|E_{\underline{h}}|$. These small sets of influential reflections are called the neighbourhoods of the invariant. These neighbourhoods are nested, so all reflections contained in the first neighbourhood are also within the second neighbourhood. The value of the invariant is more dependent on the reflections within the first neighbourhood than on the second etc.

Provided the E-magnitude of each of the reflections within the neighbourhood is known then a good estimate is possible for the invariant from the conditional probability distribution. The estimate for the invariant is particularly good in the favourable case when the variance of the distribution is small.

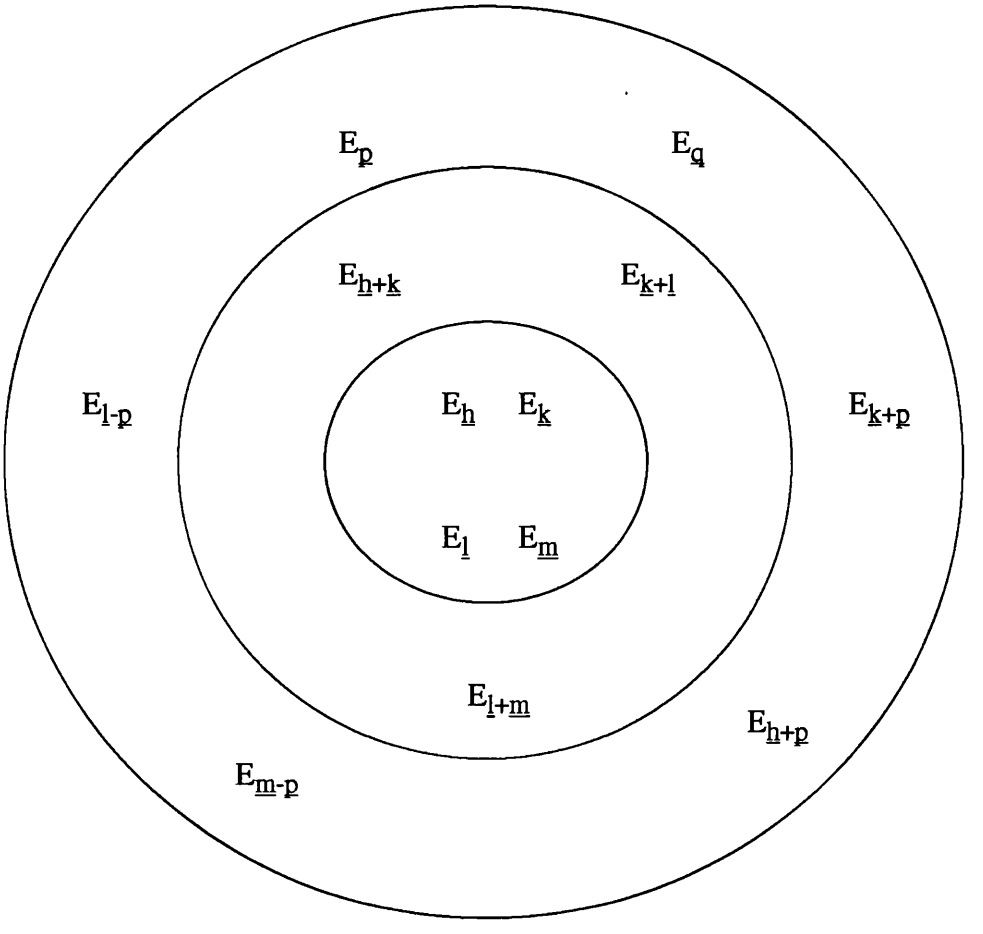


FIGURE 1. The first, second and third neighbourhoods for a quartet.

7.0 TRIPLETS

7.1 Three Phase Structure Invariants

A triplet or three phase structure invariant relationship is an invariant of the form:

$$\Phi_3 = \varphi_{\underline{h}} + \varphi_{\underline{k}} + \varphi_{\underline{l}} \qquad \text{(EQ 7.1.1)}$$

when

$$\underline{h} + \underline{k} + \underline{l} = \underline{O} \qquad \text{(EQ 7.1.2)}$$

Where

$\underline{O}$ is the null vector

i.e. the sum of the indices of the three reflections involved are all zero.

For centrosymmetric structures this phase sum can only have values of 0 or π, normally called the sign of the relationship, where a phase sum of zero is said to be positive and a phase sum of π is negative. When the E-magnitudes of the reflections are large a triplet is much more accurate and can be used to determine the phase of a third reflection, providing the phase of the other two reflections are known (Sayre, 1952).

$$S_{\underline{h}} S_{\underline{k}} S_{\underline{l}} \approx 1 \qquad \text{(EQ 7.1.3)}$$

The neighbourhoods of the triplet are as follows;

First neighbourhood: Composed of the principal terms $|E_{\underline{h}}|$, $|E_{\underline{k}}|$ and $|E_{\underline{l}}|$.

Second neighbourhood: Is formed by consideration of the quintet (five phase invariant)

$$\Phi_5 = \varphi_{\underline{h}} + \varphi_{\underline{k}} + \varphi_{\underline{l}} + \varphi_{\underline{p}} + \varphi_{-\underline{p}} \qquad \text{(EQ 7.1.4)}$$

an so is composed of $|E_{\underline{p}}|$, $|E_{\underline{h}+\underline{p}}|$, $|E_{\underline{h}-\underline{p}}|$, $|E_{\underline{k}+\underline{p}}|$, $|E_{\underline{k}-\underline{p}}|$, $|E_{\underline{l}+\underline{p}}|$ and $|E_{\underline{l}-\underline{p}}|$.

Where

$\underline{p}$ is a floating vector

7.2 Cochran Distributions

If the magnitudes of the three structure factors involved in a triplet relationship and the number and type of atoms in the unit cell are known, then it has been shown by Cochran (Cochran, 1955) that the probability of the triplet being equal to zero is given by:

$$P(\Phi_3(\underline{h},\underline{k})) = \frac{1}{2\pi I_o(\kappa)} e^{\kappa \cos \Phi_3(\underline{h},\underline{k})} \qquad \text{(EQ 7.2.1)}$$

Where

$$\kappa = 2\frac{\sigma_3}{\sigma_2^{3/2}} |E_{\underline{h}} E_{\underline{k}} E_{\underline{h}-\underline{k}}| \qquad \text{(EQ 7.2.2)}$$

$$\sigma^n = \sum_{i=1}^{N} Z_i^n \qquad \text{(EQ 7.2.3)}$$

I_0 is a zero order modified Bessel function

It should be noted that in an equal atom structure the following is true:

$$\frac{\sigma_3}{\sigma_2^{3/2}} = \frac{1}{\sqrt{N}} \qquad \text{(EQ 7.2.4)}$$

If we plot the probability of Φ_3 versus the phase angle then we can see that for a higher value of κ then the greater the probability that the three phase relationship is equal to zero (see Figure 2). As κ is dependant on the E-magnitudes of the reflections being used, it is obviously preferable to construct our triplets from reflections with large E-magnitudes.

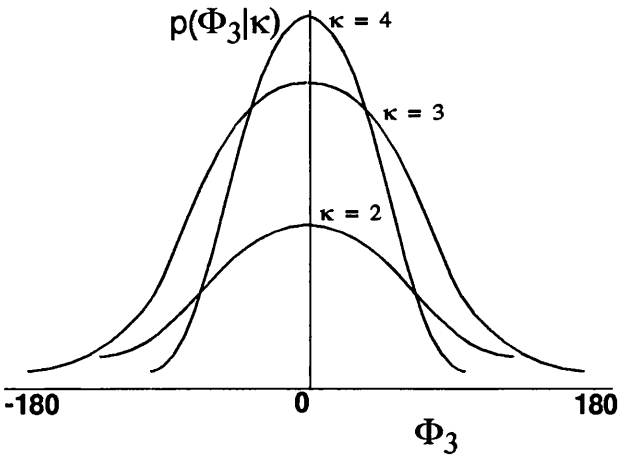


FIGURE 2. Cochran distributions for κ=2,3,4

It can be seen from Figure 2 that the variance of the distribution greatly decreases with increasing κ. This variance (or square of the standard deviation) is given by the summation of the infinite series of Bessel functions (Karle & Karle, 1966):

$$V(\kappa) = \frac{\pi^2}{3} + 4\sum_{t=1}^{\infty} \frac{(-1)^t I_t(\kappa)}{t^2 I_0(\kappa)} \qquad \text{(EQ 7.2.5)}$$

A plot of EQ 7.2.5 is shown in Figure 3

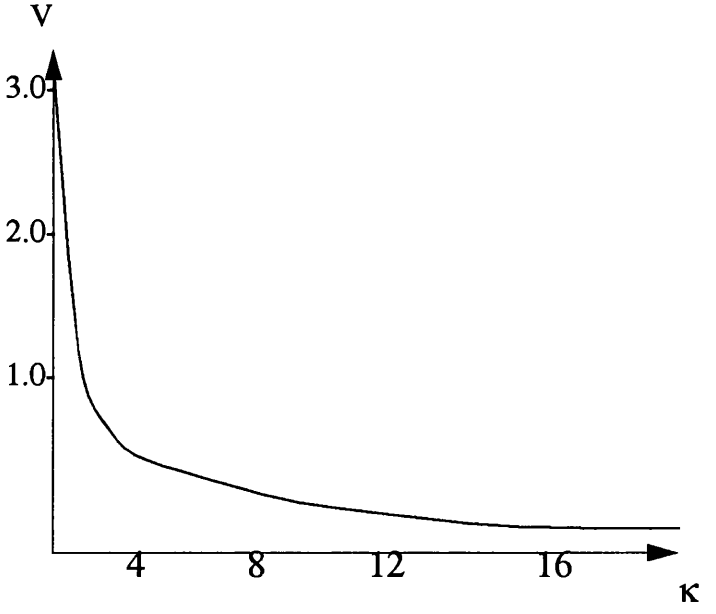


FIGURE 3. A plot of the variance of a triple phase relationship as a function of κ

8.0 QUARTETS

A quartet relationship is an invariant of the form:

$$\Phi_4 = \varphi_{\underline{h}} + \varphi_{\underline{k}} + \varphi_{\underline{l}} + \varphi_{\underline{m}} \qquad \text{(EQ 8.0.1)}$$

when

$$\underline{h} + \underline{k} + \underline{l} + \underline{m} = \underline{O} \qquad \text{(EQ 8.0.2)}$$

Where

$\underline{O}$ is the null vector

In practical use this relationship is not as important in the determination of phases as the triplet relationship, but is of extreme importance in the solution of structures that crystalise in a space group that contains no translational symmetry.

As can be seen in Figure 1, the first neighbourhood consists of the principal terms $|E_{\underline{h}}|$, $|E_{\underline{k}}|$, $|E_{\underline{l}}|$ and $|E_{\underline{m}}|$. The second neighbourhood is composed of the quartet cross terms:

$|E_{\underline{h}+\underline{k}}|$, $|E_{\underline{k}+\underline{l}}|$, $|E_{\underline{l}+\underline{m}}|$. The third neighbourhood can be constructed by the addition of an arbitrary reflection $\underline{p}$ and its associated reflection $\underline{q}$, such that

$$\underline{h} + \underline{k} + \underline{p} + \underline{q} = \underline{O} \qquad \text{(EQ 8.0.3)}$$

To ensure that the probability of this quartet being true it is best if both $|E_{\underline{p}}|$ and $|E_{\underline{q}}|$ are large. The addition of these two new reflections has produced two new quartets.

$$\Phi_4 = \varphi_{\underline{h}} + \varphi_{\underline{k}} + \varphi_{\underline{p}} + \varphi_{\underline{q}} \qquad \text{(EQ 8.0.4)}$$

$$\Phi_4 = \varphi_{\underline{h}} + \varphi_{\underline{k}} + \varphi_{-\underline{p}} + \varphi_{-\underline{q}} \qquad \text{(EQ 8.0.5)}$$

Consideration of these two invariants gives rise to the third neighbourhood composed of the two new principal terms $|E_{\underline{p}}|$, $|E_{\underline{q}}|$ and the new unique cross terms $|E_{\underline{m}-\underline{p}}|$, $|E_{\underline{h}+\underline{p}}|$, $|E_{\underline{l}-\underline{p}}|$ and $|E_{\underline{k}+\underline{p}}|$.

It is possible to calculate a probability function for Φ_4 given the E-magnitudes and cross terms using EQ 8.0.6

$$P(\Phi_4 \mid |E_{\underline{h}}||E_{\underline{k}}||E_{\underline{l}}||E_{\underline{m}}||E_{\underline{h}+\underline{k}}||E_{\underline{k}+\underline{l}}||E_{\underline{l}+\underline{m}}|) = \qquad \text{(EQ 8.0.6)}$$

$$\frac{1}{L} e^{-2B\cos\Phi_4} I_o\left(\frac{2}{\sqrt{N}}|E_{\underline{h}+\underline{k}}|Y_{12}\right) I_o\left(\frac{2}{\sqrt{N}}|E_{\underline{k}+\underline{l}}|Y_{23}\right) I_o\left(\frac{2}{\sqrt{N}}|E_{\underline{l}+\underline{m}}|Y_{31}\right)$$

Where

I_0 is the zero order Bessel function

L is a normalising constant

$$B = \frac{2}{N}|E_{\underline{h}}E_{\underline{k}}E_{\underline{l}}E_{\underline{m}}|$$

$$Y_{12} = \sqrt{|E_{\underline{h}}|^2|E_{\underline{k}}|^2 + |E_{\underline{l}}|^2|E_{\underline{m}}|^2 + 2|E_{\underline{h}}||E_{\underline{k}}||E_{\underline{l}}||E_{\underline{m}}|\cos\Phi_4}$$

$$Y_{23} = \sqrt{|E_{\underline{k}}|^2|E_{\underline{l}}|^2 + |E_{\underline{h}}|^2|E_{\underline{m}}|^2 + 2|E_{\underline{h}}||E_{\underline{k}}||E_{\underline{l}}||E_{\underline{m}}|\cos\Phi_4}$$

$$Y_{31} = \sqrt{|E_{\underline{l}}|^2|E_{\underline{h}}|^2 + |E_{\underline{k}}|^2|E_{\underline{m}}|^2 + 2|E_{\underline{h}}||E_{\underline{k}}||E_{\underline{l}}||E_{\underline{m}}|\cos\Phi_4}$$

There are three types of quartet all distinguished by the value of $\cos(\Phi_4)$.

1. Positive Quartets: These are the quartets for which the E-magnitudes of the cross terms are large then the probability of the value of $\cos(\Phi_4)$ being +1 is large (Schenk, 1973). These quartets are highly correlated with triplets and so are rarely used in phase determination. The probability distribution for such quartets are shown in Figure 4.

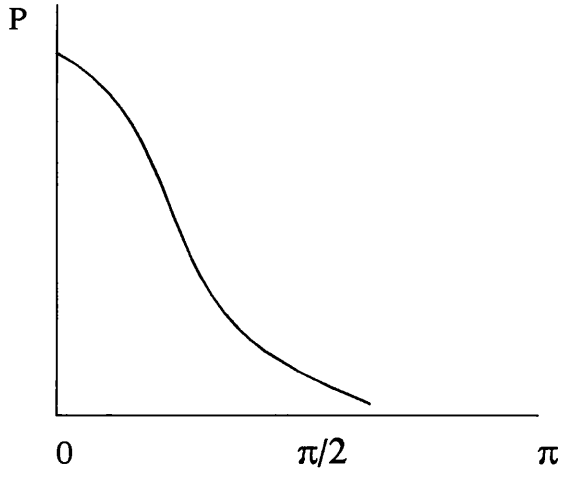


FIGURE 4. The probability distribution for a positive quartet.

2. Negative Quartets: These quartets are those which have small E-magnitudes on the cross terms of the second neighbourhood. For these quartets the probability is that $\cos(\Phi_4) = -1$ i.e $\Phi_4 = \pi$ (Hauptman, 1974). These quartets are more readily used in direct methods as they are not as closely correlated to the triplets as positive quartets are. They can be used actively in phase determination and also in the figure of merit NQEST (DeTitta, Edmonds, Langs & Hauptman, 1975). The probability distribution for the values of such quartets are shown in Figure 5.

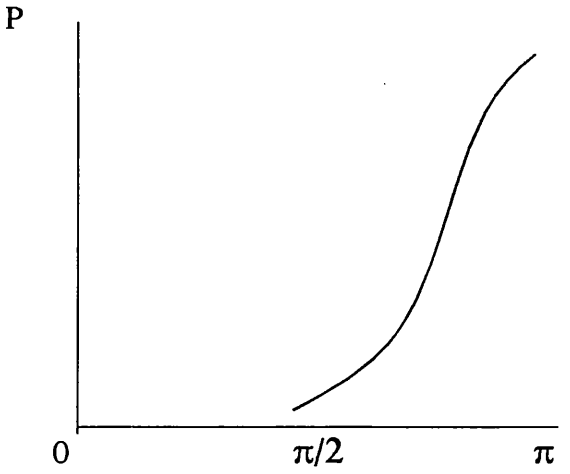


FIGURE 5. The probability distribution of a negative quartet

3. Enantiomorph Sensitive Quartets: These are the quartets that have a mix of large and small E-magnitudes or a collection of intermediate E-magnitudes in the cross terms (Hauptman, 1975). These are less reliable than positive or negative quartets and are infrequently used in structure determination. They most likely have $\Phi_4 = \pm\frac{\pi}{2}$. The probability distribution for such quartets is shown in Figure 6.

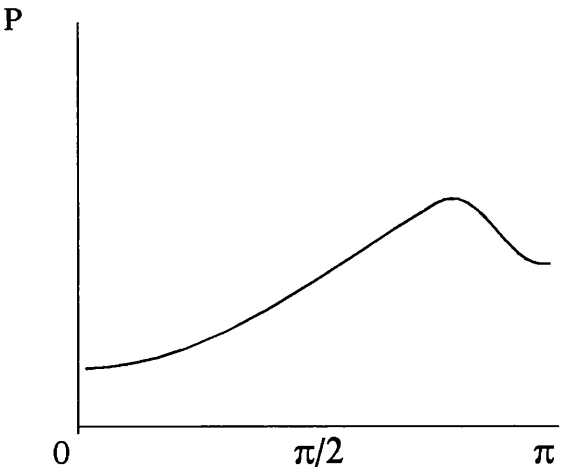


FIGURE 6. The probability distribution for an enantiomorph sensitive quartet

9.0 STARTING SETS

In no structure solution do we start in complete ignorance about the phases of the reflections. A set of reflections can be phased immediately, composed of the origin defining reflections, an enantiomorph defining reflection (if appropriate) and Σ_1 determined reflections. Also in our starting set we can include a small number of reflections that are deemed suitable for phase permutation, or symbolic addition with assigned phase values.

The selection procedure of the best reflections for origin definition and permutation is intrinsically linked with the convergence procedure (Germain, Main & Woolfson, 1970). The procedure starts with the calculation of $\alpha(\underline{h})_{est}$ for each reflection. This figure is a measure of each reflections connectivity to all other reflections through the invariants.

$$\alpha(\underline{h})_{est} = \sum_{\underline{k}} \kappa(\underline{h},\underline{k})^2 + \sum_{\underline{k}}\sum_{\underline{l}} \kappa(\underline{h},\underline{k})\,\kappa(\underline{h},\underline{l})\,\frac{I_1\{\kappa(\underline{h},\underline{k})\}}{I_0\{\kappa(\underline{h},\underline{k})\}}\,\frac{I_1\{\kappa(\underline{h},\underline{l})\}}{I_0\{\kappa(\underline{h},\underline{l})\}} \quad \text{(EQ 9.0.1)}$$

where $\kappa(\underline{h},\underline{k})$ is defined in EQ 7.2.2

From this estimation of $\alpha(\underline{h})_{est}$ it can be seen that a good choice of reflections to have in our initially phased set will be those with maximum connectivity through the invariants. This selection is done iteratively where the reflection with the lowest value of $\alpha(\underline{h})_{est}$ is removed from the list of reflections to be phased and the value of $\alpha(\underline{h})_{est}$ recalculated for all reflections. A reflection will not be dropped from the list if that reflection is the last of its type that is required for origin or enantiomorph definition. If a reflection has a value $\alpha(\underline{h})_{est} = 0$ then its phase cannot be calculated from the phases of the remaining reflections, this reflection is then selected for permutation.

The convergence procedure is designed in such a way that it leads to strong phase development and multiple interactions. It should ideally avoid weak links in the convergence map (a weak link is a phase indication from one invariant only). Convergence mapping is not a robust procedure and marked differences can appear in a map following small changes in the E-magnitudes of the reflections contributing to the map.

10.0 INITIAL PHASING TECHNIQUES

10.1 Phase Permutation

Following convergence we have some reflections for which we know the phase, and others which we will permute, i.e. assign values to the phase. These assigned phases will then be propagated through the convergence map to determine the values of the phase of all other reflections in that phase set. For even a small number of permuted reflections this can lead to a very large number of possible phase sets. In early versions of MULTAN (Main, 1985) the phase was permuted through the values $\frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}$ and $\frac{7\pi}{4}$. This gives rise to a 4^n number of phase sets for n permuted reflections. As the calculation and refinement of each phase set is not insignificant it is important to try to minimise the number of permuted reflections. Obviously centro-symmetric phases need only be permuted over the two possible values π and 0.

10.2 Magic Integer Permutation

To reduce the number of phase sets generated and refined the idea of magic integers (White & Woolfson, 1975) was introduced into direct methods. Now almost every direct methods computer program that deals with phase permutation uses the magic integer approach. It can be shown that a set of n integers $m_1, m_2, m_3, \ldots, m_n$ can be approximated to a set of n phases that an be represented by a single variable x (in the range $0 < x < 2\pi$), using the equation;

$$\varphi_i = m_i x \qquad \text{(EQ 10.2.1)}$$

As the sets of integers can only approximately represent the set of n phases it is important to reduce the root mean square (r.m.s.) errors to a minimum by using the most efficient sets of magic integers (Main, 1977). These are shown in Figure 7.

n	Sequence								Number of sets	r.m.s. error
2	2	3							12	26
3	3	4	5						20	29
4	5	7	8	9					32	37
5	8	11	13	14	15				50	42
6	13	18	21	23	24	25			80	47
7	21	29	34	37	39	40	41		128	48
8	34	47	55	60	63	65	66	67	206	49

FIGURE 7. The most efficient sets of magic integers for up to eight phase permutations

Although the r.m.s errors are large they are acceptable, and the reduction in the number of sets generated allows larger numbers of phases to be permuted.

10.3 Random Phasing

During the study of phase refinement processes, tests were done on the radius of convergence (Baggio, Woolfson, Declercq & Germain, 1978). The tests involved producing a set of phased reflections with random errors in the phase of a known r.m.s. error and testing if the refinement process could reduce this error to below 30^{o}. With a starting r.m.s. error of 75^{o} six out of ten tests produced a correct phase set. With an r.m.s. error of 90^{o} four out of ten phase sets were correctly refined. When totally random phases were used it was still found that a small fraction of the phase sets could be refined to correctness. This fact was the basis for the random methods of phasing the intitial set.

There are two principal methods of random phasing

1. Assigning random phases to a reduced starting set, say, 100 reflections. These 100 reflections are then refined and used to extrapolate the remaining phases and then these in turn are refined. This is the procedure used in the YZARC (Baggio, Woolfson, Declercq & Germain, 1978) program.

2. Random phases are assigned to all reflections and then refined using carefully calculated weights assigned on the basis of whether the phase is known absolutely i.e. origin defining reflections, known confidently i.e. Σ_1 reflections, or a randomly assigned phase. This procedure is the one used in the RANTAN program (Yao Jia-Xing, 1981).

Random phasing methods are now responsible for the majority of routine structure solutions using direct methods. These techniques are more easily programmed than the phase permutation techniques and in many cases are more robust tools for structure solution.

It is unlikely that the starting phases assigned in either phase permutation or random assignment will be correct, even after refinement. Therefore it is important to generate a number of phases sets. The number of sets generated is dependant on whether phase permutation or random assignment is used, whether the structure is centro-symmetric or non-centrosymmetric, and on the number of atoms in the asymmetric unit. Regardless of how many phase sets have to be generated, it is only necessary that one correct phase set, identifiable from its Figure of Merit is produced, this will allow the determination of the crystal structure.

11.0 PHASE REFINEMENT

11.1 Tangent Refinement

There are many different methods of phase refinement but the most common is tangent refinement (Karle & Hauptman, 1956). It is a means of relating all of the indications of an individual phase from all the associated invariant relationships. It can be written in its weighted form as:

$$\tan\phi_{\underline{h}} \approx \frac{\sum_{\underline{k}} W_{\underline{k}} W_{\underline{h}-\underline{k}} |E_{\underline{k}}||E_{\underline{h}-\underline{k}}| \sin(\phi_{\underline{k}} + \phi_{\underline{h}-\underline{k}})}{\sum_{\underline{k}} W_{\underline{k}} W_{\underline{h}-\underline{k}} |E_{\underline{k}}||E_{\underline{h}-\underline{k}}| \cos(\phi_{\underline{k}} + \phi_{\underline{h}-\underline{k}})} \qquad \text{(EQ 11.1.1)}$$

Where $W_{\underline{k}}$, $W_{\underline{h}-\underline{k}}$ are weights

The above formula is derived on the assumption that all invariants are independent of each other. This can lead to the over refinement of phases.

The tangent formula is used to refine phases in an iterative manner, with refinement continuing until there are minimal shifts in the calculated phases from one cycle to the next.

The tangent formula is now very rarely used without a weighting scheme. As a consequence of the convergence procedure, the reflections at the bottom of the convergence map will have a much greater influence on the phase estimate than those further up the map. Reflections near the bottom of the map therefore have a weight much closer to one while those further up are correspondingly downweighted. Weighting schemes are important in the early stages of phasing, but as refinement increases all the weights tend to unity. The weighting scheme that works well with many structures can be written as:

$$W_{\underline{h}} = \min\left[\frac{\alpha(\underline{h})}{5}, 1.0\right] \qquad \text{(EQ 11.1.2)}$$

A further development in weighting schemes to be used with the tangent formula comes from Hull and Irwin (Hull & Irwin, 1978). This helps prevent the rapid movement of the weight to unity.

$$W_{\underline{h}} = \min\left[\frac{\alpha(\underline{h})}{5}, 1.0, \frac{\{\alpha(\underline{h})_{est}+5\}}{\alpha(\underline{h})}\right] \qquad \text{(EQ 11.1.3)}$$

Where

$$\alpha(\underline{h}) = \sum_{\underline{k}} \kappa_{\underline{hk}} \cos(\varphi_{\underline{k}} + \varphi_{\underline{h}-\underline{k}}) \qquad \text{(EQ 11.1.4)}$$

$\kappa_{\underline{hk}}$ is given by EQ 7.2.2

$\alpha(\underline{h})_{est}$ is given in EQ 9.0.1

This weighing scheme also prevents $\alpha(\underline{h})$ (Karle & Karle, 1966) exceeding $\alpha(\underline{h})_{est}$, as the reflection will be downweighted if it does so. This prevents the overconsistancy of phases associated with the tangent formula. This weighting scheme is also used preferentially in solving structures that contain heavy atoms, have pseudo-symmetry, or belong to a symmorphic space group.

The shape of the Hull & Irwin weighting scheme is shown in Figure 8.

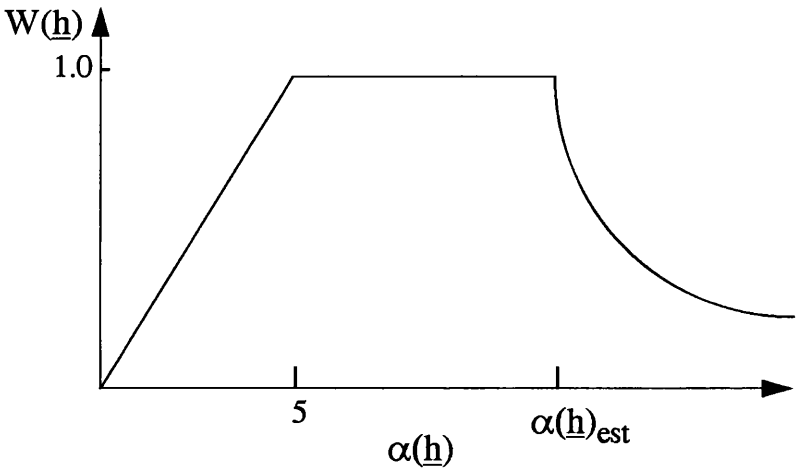


FIGURE 8. The Hull & Irwin weighting scheme

A further development in tangent formula weighting has been suggested by Giacovazzo (Giacovazzo, 1979) based on an improved estimate for $\alpha(\underline{h})$ that takes the uncertainty of known phases into account.

11.2 Linear Equations

An alternative to the tangent formula for refinement is to set up the phase relationships as a set of linear equations with least-squares refinement to obtain a solution (Woolfson, 1977).

A three phase invariant can be written as

$$W\varphi_{\underline{h}} + W\varphi_{\underline{k}} + W\varphi_{\underline{l}} \approx Wn \qquad \text{(EQ 11.2.1)}$$

Where

W is the weight of the invariant

n is an integer

If the full set of integers are known and a full set of phases is also known then the complete set of relationships can be written in matrix form.

$$\mathbf{A}\,\mathbf{x} \approx \mathbf{c} \qquad \text{(EQ 11.2.2)}$$

By replacing ≈ by = we can obtain a least squares solution for the integers n.

$$\mathbf{x} = (\mathbf{A}^{\mathbf{T}}\mathbf{A})^{-1}\,\mathbf{A}^{\mathbf{T}}\,\mathbf{c} \qquad \text{(EQ 11.2.3)}$$

If approximate phases are available then EQ 11.2.2 and EQ 11.2.3 provide a cyclic method of refinement.

This linear equations method has been used in the program YZARC. In this program random phases are assigned to the top 100 reflections. These phases are then refined by linear equations. Only the top 100 reflections are refined to reduce the size of the matrix that has to be inverted for the procedure. The 100 reflections then have further refinement and extension using the tangent formula.

11.3 Phase Annealing

The theory of simulated annealing (Kirkpatrick, Gelatt & Vecchi, 1983) is a generally applicable tool for the determination of a global minimum within a system that contains many local minima. It is usually used to determine minima from random starting positions. It can be thought of as an analogy of statistical thermodynamics.

For phase annealing we will frequently start from a random position and then by slow "cooling" of the phase set reach a global minimum in phase space. To simulate this random movement within the phases we need a Boltzman like probability distribution, which is defined as EQ 11.3.1 for centrosymmetric structures.

$$P_{positive} = \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{\alpha}{2}\right) \qquad \text{(EQ 11.3.1)}$$

$$P_{negative} = 1 - P_{positive} \qquad \text{(EQ 11.3.2)}$$

This gives the ratio of

$$\frac{P_{positive}}{P_{negative}} = e^{-\alpha} \qquad \text{(EQ 11.3.3)}$$

If EQ 11.3.3 has α divided by kT then we have the Boltzman equation.

We now can use a Metropolis (Metropolis, Rosenbluth, Rosenbluth, Teller & Teller, 1953) algorithm to flip the phases. After each change in the phases the value of α is evaluated. If α is greater then we test the probability ratio against a random number in the range 0-1. Consequently if the ratio is greater than the random number then we move to the higher α state.

This sort of iterative process should end with the global minimum for phase space, however the structure solution may not be at this minimum, as the function we are minimising may not define accurately enough the phase space. Because of this it is best to approach the problem with a multisolution approach, with many random starting positions chosen.

This technique has been successfully implemented into the crystallographic package SHELX-90, and applied to a number of test structures in a variety of space groups (Sheldrick, 1990).

12.0 FIGURES OF MERIT

Following phase refinement all phase sets must be evaluated to determine the presence of structural information. In a multisolution environment this is essential as it is not feasible to generate E-maps for each phase set and check each map individually. There are a great many figures of merit (FOMs) and frequently direct methods programs will combine these FOMs into a single measure and rank all phase sets according to this FOM.

Ideally any FOM should use phase relationships that have not been used as a phasing technique to make the FOM independent of the technique used. This however leads to the situation that if a relationship is strong enough for a FOM then it should also be good enough to derive phases. FOMs are generally a compromise of these two facts.

There are several FOMs in use in the computing packages that are available. Presented in the following sections is a small sample of the most commonly used FOMs. In Chapter 3, a new figure of merit based on the log-likelihood gain is presented in detail.

12.1 ABSFOM

ABSFOM (Germain, Main & Woolfson, 1971) is a measure of the internal consistency of the triplet invariants used in estimating the phases. It is defined as:

$$ABSFOM = \frac{\sum_{\underline{h}} \alpha_{\underline{h}} - \sum_{\underline{h}} \langle \alpha_{\underline{h}}^2 \rangle_r^{1/2}}{\sum_{\underline{h}} \langle \alpha_{\underline{h}}^2 \rangle_{est}^{1/2} - \sum_{\underline{h}} \langle \alpha_{\underline{h}}^2 \rangle_r^{1/2}} \qquad \text{(EQ 12.1.1)}$$

Where

$$\langle \alpha_{\underline{h}}^2 \rangle_r = \sum_{\underline{k}} \kappa_{\underline{h}\underline{k}}^2 \qquad \text{(EQ 12.1.2)}$$

$\langle \alpha_{\underline{h}} \rangle_{est}$ is defined in EQ 9.0.1

$\alpha_{\underline{h}}$ is defined in EQ 11.1.4

α_{est} is the estimated value of α, and α_r is the expected value of α for random phases. Therefore ABSFOM is zero for random phases, one if there is a perfect match between the estimated and measured values of α and greater than one for an overconsistent phase set. Typically ABSFOM is considered to be a good indication if it lies in the range 0.9-1.3.

As ABSFOM is closely related to Σ_2 relationships and is also closely tied to the tangent refinement and extension method. When the tangent formula has refined the phases to overconsistancy ABSFOM becomes large, and is small with underconsistent phase sets.

12.2 RESID

RESID can defined in terms of α by the following:

$$R_\alpha = \frac{\sum_{\underline{h}} \left| \alpha_{\underline{h}} - \langle \alpha_{\underline{h}}^2 \rangle_{est}^{1/2} \right|}{\sum_{\underline{h}} \langle \alpha_{\underline{h}}^2 \rangle_{est}^{1/2}} \qquad \text{(EQ 12.2.1)}$$

Where

$\langle \alpha_{\underline{h}} \rangle_{est}$ is defined in EQ 9.0.1

$\alpha_{\underline{h}}$ is defined in EQ 11.1.4

R_α is the residual between the estimated and measured α. As the FOM is dependant on α it is tied to the tangent formula phasing methods, as is ABSFOM. A correct phase set should have a residual value less than 20%, although structures that contain large numbers of atoms in the unit cell may have correct phase sets with R_α in excess of this figure.

12.3 PSI-ZERO

PSI-ZERO (Cochran & Douglas, 1955) is a measure of fit of the small magnitude E's. It is defined as:

$$\psi_o = \frac{\sum_{\underline{h}} \left| \sum_{\underline{k}} E_{\underline{k}} E_{\underline{h}-\underline{k}} \right|}{\sum_{\underline{h}} \sqrt{\left(\sum_{\underline{k}} |E_{\underline{k}} E_{\underline{h}-\underline{k}}|^2 \right)}} \qquad \text{(EQ 12.3.1)}$$

ψ_o is largely independent of the phasing method and is sensitive to atomic positions. A value of ψ_o less than one is usually indicative of a correct phase set. Although if all other figures of merit are good but a value of ψ_o greater than one may indicate a correct fragment in the wrong position.

12.4 NQEST

NQEST uses negative quartets. It is defined as:

$$NQEST = \frac{\sum_{\underline{h},\underline{k},\underline{l},\underline{m}} W_{\underline{h}\underline{k}\underline{l}\underline{m}} \cos\left(\varphi_{\underline{h}} + \varphi_{\underline{k}} + \varphi_{\underline{l}} + \varphi_{\underline{m}}\right)}{\sum_{\underline{h},\underline{k},\underline{l},\underline{m}} W_{\underline{h}\underline{k}\underline{l}\underline{m}}} \qquad \text{(EQ 12.4.1)}$$

Where

$$W_{\underline{h}\underline{k}\underline{l}\underline{m}} = \left|1 - 2P^{+}\right| \text{ for centrosymmetrics} \qquad \text{(EQ 12.4.2)}$$

$$W_{\underline{h}\underline{k}\underline{l}\underline{m}} = \frac{1}{\sigma^2} \text{ for non-centrosymmetrics} \qquad \text{(EQ 12.4.3)}$$

Where

σ^2 is the variance of the quartet probability distribution

Note that the summation, $\sum_{\underline{h},\underline{k},\underline{l},\underline{m}}$, is over all the negative quartets.

Like ψ_o, NQEST uses the information contained in the small E-magnitudes, although the information is used in a different way, thus making the two figures of merit independent of each other. This dependence on small E-magnitudes also makes NQEST largely independent of the phasing procedure. NQEST can lie anywhere in the range -1.0 to +1.0, with the most negative values likely to correspond to correct phase sets.

12.5 Combined Figure of Merit (CFOM)

The majority of available direct methods programs contain a combined figure of merit. As the array of figures of merit is very large most packages have different FOMs and thus a different CFOM. Presented here is the CFOM from MITHRIL90 (see Appendix A). Regardless of which package and which FOMs are used, CFOM is usually the weighted sum of the FOM difference divided by the FOM range.

$$CFOM = W_1 \frac{ABSFOM - ABSFOM_{min}}{ABSFOM_{max} - ABSFOM_{min}} + W_2 \frac{(\psi_o)_{max} - \psi_o}{(\psi_o)_{max} - (\psi_o)_{min}} + W_3 \frac{(R_\alpha)_{max} - R_\alpha}{(R_\alpha)_{max} - (R_\alpha)_{min}} + W_4 \frac{NQEST_{max} - NQEST}{NQEST_{max} - NQEST_{min}} \qquad \text{(EQ 12.5.1)}$$

For CFOM the larger the value the more likely the correctness of the phase set. The maximum value of CFOM is $W_1 + W_2 + W_3 + W_4$. These weights are normally set to unity but can be changed to favour specific FOMs. If a FOM has not been calculated then the weight associated with that FOM is set to zero. In MULTAN (Woolfson, 1991) the weights for CFOM are set to 0.6, 1.2, 1.2 for ABSFOM, ψ_o and R_α respectively.

CFOM is used to rank phase sets generated by the multisolution technique for selection of the phase set most likely to contain structural information.

13.0 MAPS

13.1 Electron Density Maps

The ultimate aim of the vast majority of X-ray crystal studies is the elucidation of the molecular structure of the crystal. This is done by the examination of the electron density map. In any structure the majority of the electrons in the molecule will be at or close to the atomic centres, so the peaks in the electron density map will correspond to the atomic sites.

The electron density map $\rho(\underline{x})$ is a Fourier transform of the phased reflections $(\underline{h}, F^{obs})$. It is defined as

$$\rho(\underline{x}) = \frac{1}{V}\sum_{\underline{h}} |F^{obs}|_{\underline{h}} e^{-2\pi i(\underline{h}\cdot\underline{x}) + i\varphi_{\underline{h}}} \qquad \text{(EQ 13.1.1)}$$

where

V is the volume of the unit cell in $Å^3$

$|F^{obs}|$ is the amplitude of the observed structure factor

φ is the phase of the calculated structure factor

13.2 E-maps

The reflections that are phased most reliably are those with large E-magnitudes. However, study of the normalisation procedure shows that reflections with large $|E_{\underline{h}}|$ are also frequently those with high values of $\sin\theta$, and not those with large $|F_{\underline{h}}|$. This means that the reflections that we have determined phases for need not be those that will give the best electron density map. For this reason we calculate an E-map (Karle, Hauptman, Karle & Wing, 1958), a Fourier synthesis using EQ 13.1.1, but replacing $|F^{obs}_{\underline{h}}|$ with $|E_{\underline{h}}|$.

Since E's correspond to point atoms at rest the peaks of an E-map are very sharp, and must be sampled at small grid spacings, typically 0.333Å. As the method of selecting reflections for phasing produce highly interrelated reflections, spurious peaks can appear in E-maps, i.e. benzene rings frequently have a peak in the centre of the ring. The relative heights of the peaks should be used to assign chemical types to peaks only

with extreme caution. The best methods of assigning chemical type is to have some idea of the overall structure and using this to assign atomic type to each of the peaks.

Once a peak list has been obtained then an attempt must be made to identify chemically reasonable fragments. This can be done manually by examining print out or more conveniently by the automatic interpretation of the peak list using the computer (Main & Hull, 1978). The auto-interpretation algorithm is based on maximum and minimum bond lengths and angles for chemically acceptable fragments.

The interpretation of the map is perhaps the most important phase of the structure solution, for it is here that the chemical structure will become apparent. To aid in the task of examining maps the greater availability of computer graphics is proving to be invaluable. With an option to examine either interpreted peak lists for fragments graphically or contoured E-maps for molecular positions, the time taken for the successful identification of a correct solution decreases dramatically.

14.0 THE FAILURES OF CONVENTIONAL DIRECT METHODS

Modern direct methods are an enormously powerful tool in the elucidation of molecular structure, with the majority of crystal structures being solved using this technique. Some computer packages are now so complex and reliable that from normalisation to assignment of atom types and isotropic refinement is completely automatic.

Occasionally a structure proves impossible to solve with no interpretable E-maps being produced. These failures can be attributed to many things, the most obvious being poor quality diffraction data, although, with modern diffractometers this is becoming less of a problem.

Problem structures are frequently large, containing many atoms in the unit cell, which reduce the reliability of the phase relationships. A further factor that affects the reliability of the phase relationship is that the probability formulae are based on randomly distributed atoms. This is obviously a false assumption as atomic positions are determined by strict rules of bond lengths and angles. The breakdown of the random atom model is a very serious problem in the solution of protein structures.

The most common problem is the choice of starting sets as the convergence map is very sensitive to the starting position. A balance must be made between generating few phase sets with a weak link in the map and the computer time required to permute all the reflections required to remove all the weak links. This is only a problem with the multi-solution method.

Even when reliable phase relationships have been used to determine phases, the tangent formula, used for phase refinement, can be prove to be unstable. This instability arises from the assumption that all phase relationships are independent and can lead to over-refinement of the phases, which result in over-correlation and meaningless E-maps, with correspondingly unreasonable figures of merit. The figures of merit themselves can prevent the elucidation of a structure by failing to provide sufficient discrimination to indicate a suitable phase set for examination.

The other major problem category are the large structures i.e those with a molecular weight in excess of 1,500 Daltons, in addition to reducing the reliability of the phase relationships, these structures produce cluttered E-maps that are very difficult to interpret and fragments may be missed by the crystallographer when checking such maps.

15.0 MAXIMUM ENTROPY

15.1 Bayes Theorem

The majority of statistics encountered in crystallography are classical or frequentist statistics. There is however a more natural type of statistics based on Bayes Theorem (Bayes, 1763).

$$P(F|E) \propto P(F)\,P(E|F) \qquad \text{(EQ 15.1.1)}$$

Where

$P(F|E)$ is the *posterior* and is calculated after measurement. It is the probability of F occurring given that event E has already occurred

$P(F)$ is the *prior* and is our knowledge before measurement. It is the probability of event F occurring.

$P(E|F)$ is the *likelihood* and correspond to our measured data. It is the probability of event E being true if event F has happened.

In image processing it is possible to write Bayes Theorem as

$$p(image|data) \propto p(image)\,p(data|image) \qquad \text{(EQ 15.1.2)}$$

The full power of the Bayesian method is realised when the posterior is used as the prior in a new calculation of the posterior and the process repeated in a cyclic method until the likelihood is unchanged.

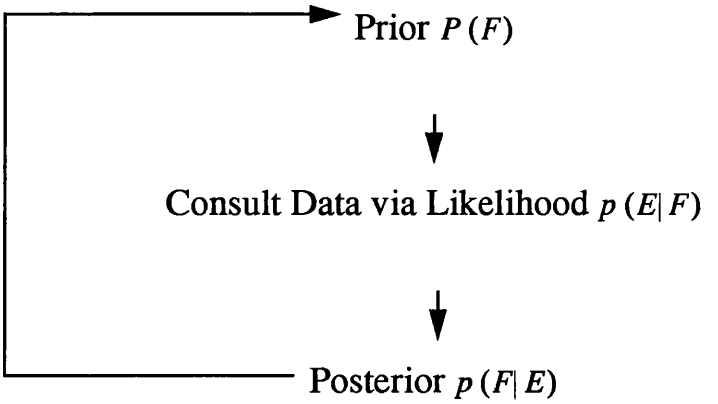


FIGURE 9. The cyclic approach to Bayesian refinement.

The obvious weakness of the method is that it is simple to bias results toward the solutions that are sought. This bias is reduced by placing constraints upon the Bayesian prior (Shannon & Weaver, 1949), these constraints are:

(i) The prior must always be capable of reproducing what we know to within experimental error.

(ii) The prior must be maximally non-committal towards that which we do not know. i.e. It should not assume that missing data are zero.

To make our prior maximally non-committal we can say that it has minimal information content about which we have no knowledge. A probability function has an entropy (or information content) that is defined as:

$$S = -I = -\sum_i p_i \log p_i \qquad \text{(EQ 15.1.3)}$$

Where

S is the entropy of the probability function

I is the information content of the function

p_i is the probability of event i

EQ 15.1.3 is for the discrete case, e.g. the rolling of dice, for the continuous case, e.g. an electron density, it may be written as an integral

$$S = -\int p(\underline{x}) \log p(\underline{x})\, dx \qquad \text{(EQ 15.1.4)}$$

To produce our maximally non-committal prior we now need only to maximise the function given in EQ 15.1.4

This technique is called the maximum entropy (ME) technique and is used extensively in under-determined problems, such as image processing.

15.2 Maximum Entropy in Crystallography

The above concepts have been shown to be applicable to the phase problem in x-ray crystallography (Bricogne 1984, Bricogne 1988).

The power of conventional (probabilistic) direct methods are limited by two principal factors:

1. The random atom model which has been used to derive so much of modern conventional direct methods theory is flawed. The model cannot represent the ordered structure of atoms imposed by the laws of chemistry.

2. The use of the Edgeworth series in the estimation of the joint probability distribution of the structure factors. The Edgeworth series is only accurate for small values (ideally zero) of E-magnitudes. This means that in practice we need a full set of atomic resolution data.

There are two principal ways to avoid the limitations given above.

The first is to use a maximum entropy approach. This is not based on the assumption that atoms have a random distribution. This theory has been applied to problems similar in nature to the crystallographic phase problem previously with considerable success i.e. statistical theory of communications (Rice, 1944; Rice 1945). Almost all theories of conventional direct methods are derivable using the maximum entropy approach, confirming the validity of the underlying theory (Bricogne, 1984).

The second is to eliminate the asymptotic convergence problems of the Edgeworth series using the saddlepoint approximation. This technique allows the evaluation of the joint probability distribution of the structure factors when some of the E-magnitudes are large.

It has been shown that the two techniques described above are essentially equivalent (Bricogne, 1988).

15.3 Normalising data for ME Calculations

It has been found that for ME calculations it is better to use unitary structure factors, which for equal point atoms of unit weight are given by

$$|U_{\underline{h}}| = \frac{|E_{\underline{h}}|}{\sqrt{N}} \qquad \text{(EQ 15.3.1)}$$

Where

$|E_{\underline{h}}|$ is the normalised structure factor magnitude

N is the number of atoms in the unit cell

From the above equation it can be seen that the maximum value for $|U_{\underline{h}}|$ is 1.0. $|U_{\underline{h}}|$ similarly to $|E_{\underline{h}}|$ is independent of the scattering angle θ. The values for $|U_{\underline{h}}|$ may have a variance estimated using the technique of Hall and Subramanian (Hall & Subramanian, 1982).

$$\sigma^2_{\underline{h}} = |U_{\underline{h}}|^2 \left(\frac{\sigma^2(|F_{\underline{h}}|)}{|F_{\underline{h}}|^2} + \frac{\sigma^2(K)}{K^2} + \left(\frac{\sin\theta}{\lambda}\right)^4 \sigma^2(B) + \frac{2\left(\frac{\sin\theta}{\lambda}\right)^2 r(K,B)\,\sigma(K)\,\sigma(B)}{K} \right) \quad \text{(EQ 15.3.2)}$$

Where

$\sigma(|F_{\underline{h}}|)$ is the standard deviation on the magnitude of the structure factor

$\sigma(K)$ is the standard deviation on the scale factor

$\sigma(B)$ is the standard deviation on the temperature factor

$r(K,B)$ is the correlation coefficient between the scale and the temperature factors.

These variances are necessary for the maximisation of the entropy calculations as they are used as weights for each individual reflection. They are also used in the calculation of Σ, a measure of the structure complexity, although the technique is not sensitive to the errors on the reflections.

15.4 The Basis Set

In determining a solution to the phase problem for a specific set of data we do not start in a position of complete ignorance. The phases of the origin reflections must be set to locate an origin in real space, this will result in the phases of 1-3 reflections being known before we begin. In the case of non-centrosymmetric structures we must also define an enantiomorph which generally results in a further reflection being assigned

an arbitrary phase. These 1-4 phased reflections constitute our basis set {H}, and will be used in the calculation of our first non-uniform prior, $q(\underline{x})$. The origin is defined by selecting a subset of reflections, {L}, with the largest U-magnitudes that lie within a user defined resolution range, typically > 2.5Å. The reflection which, when temporarily removed from {L}, causes the smallest reduction in summation 15.4.1 is then removed permanently from the subset, provided that the ability to define a legal origin is retained within the new smaller subset. This process continues in an iterative manner until only origin and enantiomorph defining reflections remain, at which point the set {L} has become the set {H}.

$$\sum_{\underline{i}} |U_{\underline{h}}||U_{\underline{k}}||U_{\underline{l}}||U_{\underline{m}}||E_{\underline{i}}^2 - 1| \qquad \text{(EQ 15.4.1)}$$

Where

$\underline{h}, \underline{k}, \underline{l}, \underline{m} \in \{L\}$

$\underline{i}$ is a reflection in the second neighbourhood of the set {L}, that has at least two sets of contributors.

The selection of origin defining reflections that maximise 15.4.1 has the effect of maximising the sensitivity of the log likelihood gain.

This preference for low resolution reflections is made as the maximisation of the likelihood is best carried out while moving from low to high resolution. It is believed that this preference for a low resolution basis set is caused by the conditional distribution of the phases being more non-committal towards the future build up of electron density within the unit cell. This allows for less committed phase extrapolation and a smoother building of the $q^{ME}(\underline{x})$ map.

The first basis set need not always start with only the origin and enantiomorph reflections (although this is the normal case for single crystal work). In electron diffraction work there may be additional phased reflections from the Fourier transform of the lattice image and for proteins there may be a set of phases that have been derived from isomorphous replacement.

15.5 Obtaining a $q^{ME}(\underline{x})$ Map

Let $U_{\underline{h} \in H}$ be the phased reflections of the basis set, and $U_{\underline{h} \in K}$ the remaining unphased reflections of the data set. We now use the reflections in the basis set to calculate our initial electron density map $\rho(\underline{x})$ by Fourier transform. This is not a true electron density map but should be thought of as a "probability distribution of atoms" that can be treated in the same way as, and appears very similar to, a conventional electron density map. This map contains only the information that we input and so is heavily biased. As in conventional direct methods, areas of negative electron density are forbidden in the ME formalism and so all negative portions of the map $\rho(\underline{x})$ are set to zero to form a new map $\rho'(\underline{x})$. The new map is still highly biased and so the entropy of the image $\rho'(\underline{x})$ is maximised subject to the constraint that the Fourier transform of the electron density map $\rho'(\underline{x})$ must contain the basis set reflections to within experimental error of those input to the original calculation. This maximum entropy map is denoted as $q^{ME}(\underline{x})$.

It can be shown (Bricogne, 1984) that when our electron density distribution has maximum entropy relative to our prior distribution $m(\underline{x})$ and, while continuing to obey the constraints, then our new distribution can be described exactly using the exponential model:

$$q^{ME}(\underline{x}) = \frac{m(\underline{x})}{Z(m, \omega)} e^{\omega(\underline{x})} \qquad \text{(EQ 15.5.1)}$$

Where

$$Z(m, \omega) = \int_V m(\underline{x}) e^{\omega(\underline{x})} d^3\underline{x} \qquad \text{(EQ 15.5.2)}$$

Note: $Z(m, \omega)$ is a normalising constant

$$\omega(\underline{x}) = \sum_{\underline{h} \in H} \zeta_{\underline{h}} e^{-2\pi i (\underline{h} \cdot \underline{x})} \qquad \text{(EQ 15.5.3)}$$

Where the complex Lagrange multipliers $\zeta_{\underline{h}}$ are obtained from the condition that $q^{ME}(\underline{x})$ reproduces the experimental constraints. This can be expressed algebraically as

$$\int_V q^{ME}(\underline{x})\, e^{2\pi i(\underline{h}\cdot\underline{x})} d^3\underline{x} = U_{\underline{h}\in H} \qquad \text{(EQ 15.5.4)}$$

The Fourier transform of the $q^{ME}(\underline{x})$ map contains not only the phases of the basis set reflections used in its' initial calculation but now also contains phase information on $U_{\underline{h}\in K}$. The map also contains information on reflections that were not measured experimentally. This phase extrapolation is shown graphically in Figure 10.

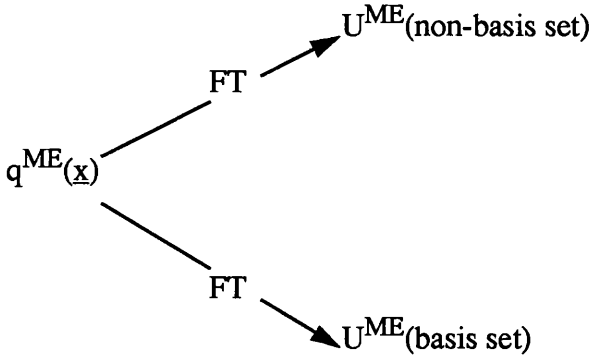


FIGURE 10. The reflections produced from the Fourier transform of the $q^{ME}(\underline{x})$ map.

The reliability of the extrapolated phases from this $q^{ME}(\underline{x})$ map is dependent on the product $|U_{\underline{h}}|^{obs}|U_{\underline{h}}|^{ME}$. The larger the product the greater the accuracy of the extrapolated phase (Gilmore, Bricogne & Bannister, 1990). Obviously when the basis set is very small, say the origin and enantiomorph only, then the extrapolation will be very weak for all reflections. While it is possible to add the strong extrapolates to the basis set this causes difficulties as it traps the entropy maximisation into a local maximum. As the extrapolate method traps the entropy, we require some other technique to move reflections from the non-basis set {K} to the basis set {H}.

15.6 The Phasing Tree

To alleviate the problem of adding weakly extrapolated reflections to the basis set, reflections with large U-magnitude are given permuted phases just as in conventional direct methods. These permuted phases can either be assigned by the full factorial or magic integer (White & Woolfson, 1975) methods. These new permuted reflections are now added to the basis set and are treated as additional constraints for the entropy maximisation of the $q^{ME}(\underline{x})$ map. The reflections are chosen in such a manner as to ensure optimal enlargement of the second neighbourhood of the basis set, while favouring reflections for which $|U_{\underline{h}}|^{ME}$ is small i.e. the maximum entropy

extrapolation is telling us little or nothing about these reflections. This implies that our next $q^{ME}(\underline{x})$ map should have a maximum information gain relative to the previous map. These reflections about which the $q^{ME}(\underline{x})$ map know least are also the reflections least coupled to those that are already known and this should prevent islands of highly correlated reflections being determined. Using this method of phase permutation gives rise to a node for each choice of phase for each permuted reflection, and so builds our phasing tree as shown in figure 11.

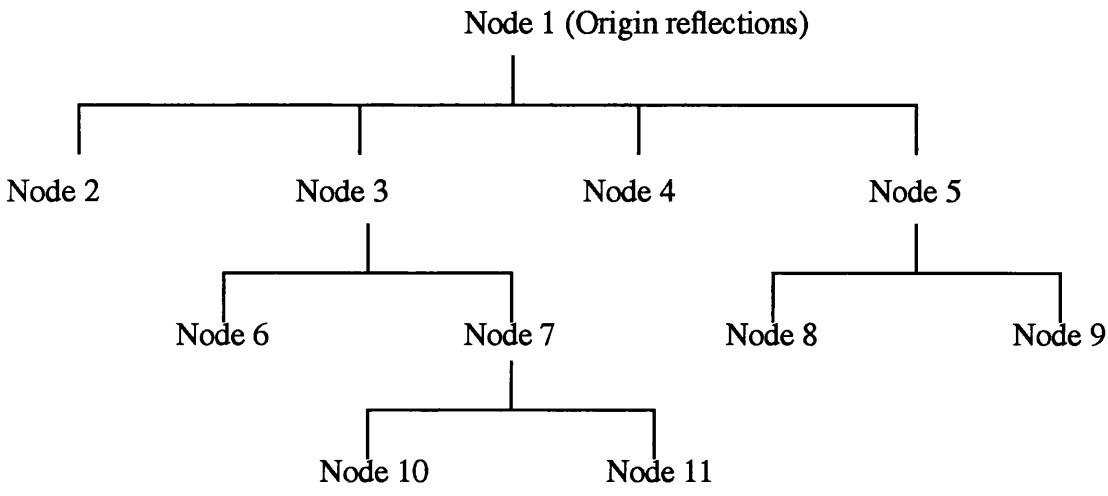


FIGURE 11. The early stages of a centrosymmetric phasing tree.

Such a phasing tree could become computationally unwieldy if each node was to be extended, as the maximisation of the entropy is computationally intensive. To determine which of the nodes is the most promising to continue the phasing procedure with, we use two figures of merit, entropy and likelihood.

15.7 Entropy

This figure is used to determine the current "strength" of the node. This loss of entropy $S_m(q^{ME}(\underline{x}))$ in moving from a uniform distribution of atoms to the now non-uniform atomic distribution is the measure of the reduction of the population of probable structures when the constraints of the basis set are enforced (Shannon & Weaver, 1949).

The figure of entropy can be thought of as the girth of the branches on the tree that have lead to this point in the phasing procedure.

Mathematically entropy is defined as:

$$S_m(q^{ME}(\underline{x})) = -\int q(\underline{x}) \log\left[\frac{q(\underline{x})}{m(\underline{x})}\right] d^3\underline{x} \qquad \text{(EQ 15.7.1)}$$

When used on its own entropy appears to be a poor figure of merit (Lemaréchal & Navaza, 1991; Shankland, Gilmore, Bricogne & Hashizume, 1993). However one research group claims to have solved the structure of recombinant bovine chymosin (Gilliland, Windborne, Nachman & Wlodawer, 1990) ab initio using entropy as the figure of merit to determine phase choices (Sjölin, Prince, Svensson & Gilliland, 1991). This result disagrees with the observations made by others (Gilmore, Henderson & Bricogne, 1991) and goes against the theory (Bricogne, 1988), who stated that entropy is best used in conjunction with likelihood in the formula:

$$NS_m(q^{ME}(\underline{x})) + L \qquad \text{(EQ 15.7.2)}$$

Where

N is the effective number of atoms in the unit cell

L is the likelihood of the system at best match of U-magnitudes

$S_m(q^{ME}(\underline{x}))$ is the entropy of the $q^{ME}(\underline{x})$ map

The above formula gives the theoretical optimum figure of merit for the discrimination of phase sets. However in the early stages of phase determination likelihood has proved to be a reliable figure of merit, with entropy becoming more relevant as more phased reflections are added to the basis set. This results in likelihood alone being used as the criteria for phase set discrimination during the early stages of phase determination.

15.8 Likelihood

Just as entropy is the figure used to determine the current and past power of a node, likelihood is the figure used to determine the future viability of the node under examination. It can be thought of as the girth of the branches of the tree that will radiate away from this node.

In direct methods we do not use the likelihood in its native form but rather calculated it relative to the null hypothesis L_0, the likelihood that for all extrapolated reflections $|U_{\underline{h}}|^{ME} = 0$. This is called the likelihood ratio, or if we take logs of the equation it is called the log likelihood gain.

Mathematically the formulae for the acentric log likelihood gain is defined as

$$L = \sum_{\underline{h} \in K} \left[\log I_0 \left(\frac{2N}{\varepsilon_{\underline{h}}} |U_{\underline{h}}|^{obs} |U_{\underline{h}}|^{ME} \right) - \frac{N}{\varepsilon_{\underline{h}}} \left(|U_{\underline{h}}|^{ME} \right)^2 \right] \qquad \text{(EQ 15.8.1)}$$

Where

I_0 is a zero order Bessel function

For centric reflections the log likelihood gain is defined as

$$L = \sum_{\underline{h} \in K} \left[\log \cosh \left(\frac{N}{\varepsilon_{\underline{h}}} |U_{\underline{h}}|^{obs} |U_{\underline{h}}|^{ME} \right) - \frac{N}{2\varepsilon_{\underline{h}}} \left(|U_{\underline{h}}|^{ME} \right)^2 \right] \qquad \text{(EQ 15.8.2)}$$

The above versions of the likelihood equations were derived using the diagonal approximation to the likelihood and so are only sensitive to the extrapolated moduli $|U_{\underline{h}}|^{ME}$, and not the associated extrapolated phases.

According to a fundamental theorem of classical statistics (Neyman & Pearson, 1933), likelihood is the best possible figure of merit for making optimal decisions. In the context of direct methods this means that although a great many figures of merit have been developed and used successfully, likelihood must be the best (Bricogne, 1991).

By combining the prior distribution of the unitary structure factors with the likelihood in Bayes theorem EQ15.1.1 we can use the maximisation of likelihood as a means to refine phases (Bricogne, 1988; Gilmore, Bricogne & Bannister, 1991).

15.9 Centroid Maps

A $q^{ME}(\underline{x})$ map is not a traditional electron density map although it is very similar and displays many of the features normally associated with the electron density map for a specific compound. The $q^{ME}(\underline{x})$ map is put through a Sim filter (Sim, 1959; Sim,

1960) to produce a centroid map. In the centroid map both reflections from the basis set {H} and the extrapolated reflections {K} are used and each are assigned weights $w_{\underline{h}}$. Phase angles are those extrapolated by the ME process.

For centric reflections:

$$w_{\underline{h}} = \tanh\left[\frac{N}{\varepsilon_{\underline{h}}}|U_{\underline{h}}|^{obs}|U_{\underline{h}}|^{ME}\right] \qquad \text{(EQ 15.9.1)}$$

For acentric reflections

$$w_{\underline{h}} = \frac{I_1\left[\frac{2N}{\varepsilon_{\underline{h}}}|U_{\underline{h}}|^{obs}|U_{\underline{h}}|^{ME}\right]}{I_0\left[\frac{2N}{\varepsilon_{\underline{h}}}|U_{\underline{h}}|^{obs}|U_{\underline{h}}|^{ME}\right]} \qquad \text{(EQ 15.9.2)}$$

Where

I_1 and I_0 are first and zero order Bessel functions respectively

Maps calculated from this new weighted extended set of structure factors will show enhanced resolution without the need for any new phase information to be added. This feature of the maximum entropy technique results in very sharp and clean electron density maps being produced (Gilmore, Bricogne & Bannister, 1991).

16.0 MITHRIL90

In the early 1980s it was decided to bring together all the important theoretical advances made in the 1970s into a single computer program. The result of this NATO funded project was the MITHRIL package (Gilmore, 1984; Gilmore & Brown, 1988). The name MITHRIL is an acronym for **M**ultan with **I**nteractive facilities, **T**riplet checking, **H**igher invariants, **R**andom phasing, **I**ntelligent control of flow and options and **L**inear equations phasing. In 1988 it was decided to upgrade the MITHRIL package to reflect the theory developed during the 1980s, also to introduce some minor bug fixes and to make the package more user friendly. MITHRIL90 was never released in this form but was integrated into the commercially available package CRYSTAN.

The program itself is written in standard Fortran 77 in a modular form with each module having its own menu and user options. It is designed to be run in either batch or interactive modes with a great deal of flexibility in the user options for structure solution. The flowchart of the modules through the program is shown in Figure 12. The basic framework is based around a highly modified version of MULTAN80, that incorporates the major features from the following programs:

1. MAGEX (Hull, Viterbo, Woolfson & Shao-Hui, 1981; Shao-Hui & Woolfson, 1982) a magic-integer based program that uses the primary-secondary method for the magic integer representation of the phases. This allows the use of a large number of reflections and relationships to be used from the start in a structure solution.

2. YZARC (Baggio, Woolfson, Declercq & Germain, 1978) is a procedure for the random assignment of phases to the starting set of up to 100 reflections. They are then refined and used for phase extension followed by further phase refinement. It is essentially a forerunner of the RANTAN procedure.

3. RANTAN (Yao Jia-Xing, 1981) is a program that assigns random phases to all reflections and then refines them using the tangent formula with a carefully controlled weighting scheme.

4. LSAM (Germain & Woolfson, 1968) a program to solve structures using a symbolic addition method.

The program has a large number of options and techniques available for structure solution and has proved to be an efficient and powerful tool in the elucidation of many structures. The majority of structures can be solved by MITHRIL using the default options and it is only the most obstinate of data sets that have proved impossible to solve.

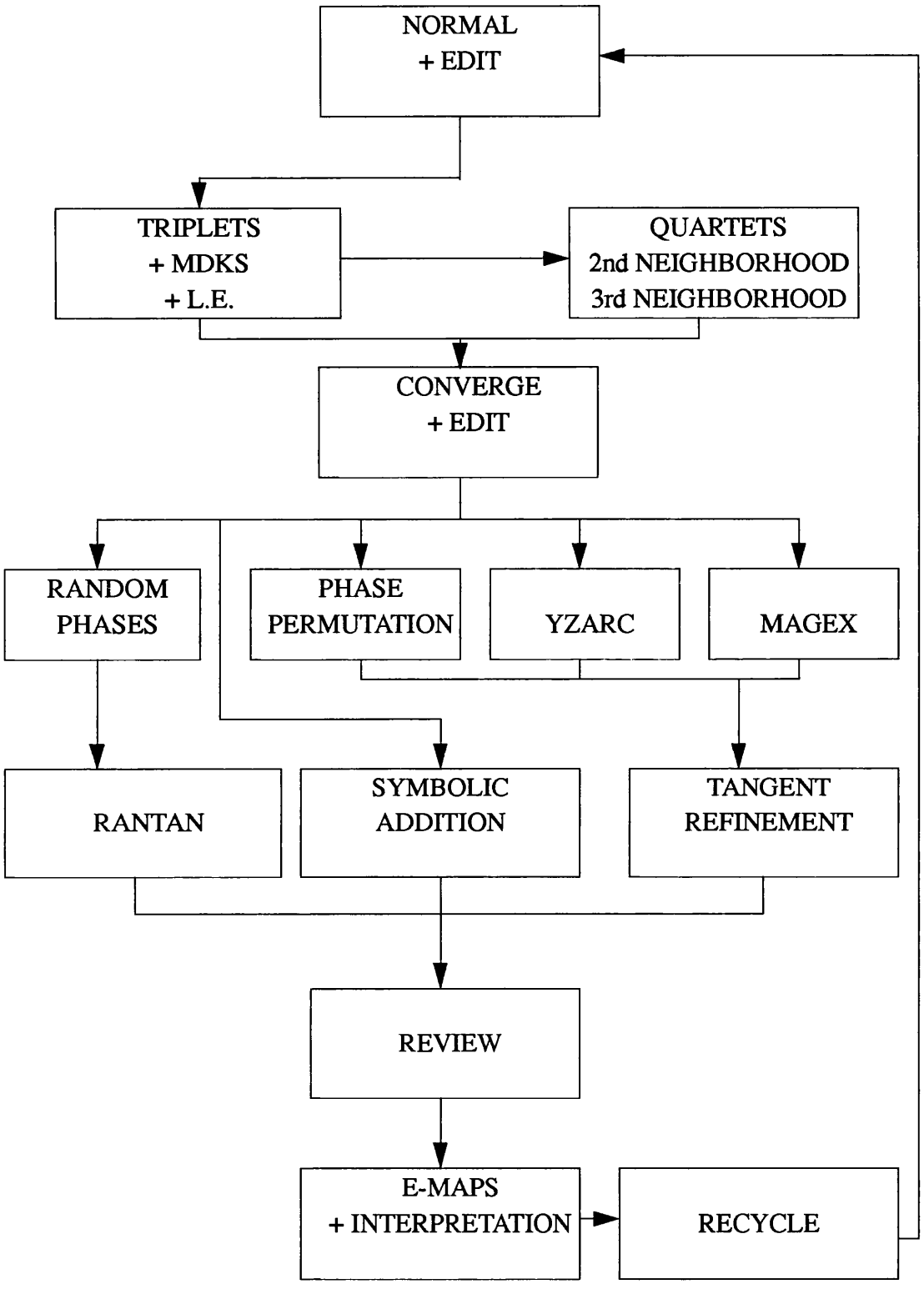


FIGURE 12. Flowchart of the modules of the MITHRIL package

17.0 MICE

In the mid 1980s the work of Bricogne in the area of Maximum Entropy as applied to crystallography (Bricogne, 1984; Bricogne, 1988) was sufficiently advanced to produce a computer program based on these theories. A collaboration of C.J Gilmore and G. Bricogne at LURE resulted in the computer program MICE (Bricogne & Gilmore, 1990), **M**aximum entropy **I**n a **C**rystallographic **E**nvironment. The program is written in standard Fortran 77 in a modular form similar to MITHRIL. Some of the tasks required for a full ab initio solution of a structure using MICE are done using the MITHRIL program with an interface to MICE.

The maximum entropy technique is computationally very intensive but with modern computers and the new network processor distributed version of the program (Shankland, Gilmore, Bricogne & Hashizume, 1993) being used it is now feasible to use MICE almost routinely. Many structures have been solved using MICE including single crystal X-ray of both centro and non-centro symmetric structures (Gilmore, Bricogne & Bannister, 1990), powder structures (Tremayne et al., 1992) and has been used for phase extension in protein crystallography (Xiang, Carter, Bricogne & Gilmore, 1993). The program has also found application in the solution of electron and fibre diffraction patterns. The program is not commercially available but is being used at several academic sites around the world.

The program is unique in direct methods because it uses entropy maximisation and likelihood for structure solution. It also generates no invariants explicitly but the procedure uses all invariants implicitly. The radical difference between MICE and other conventional direct methods packages make comparisons difficult. However based on the success of MICE with very difficult structures at low resolution indicate that it represents a useful addition to the armoury of direct methods tools, finding greater application solving problems that conventional direct methods cannot tackle.

References

Baggio, R., Woolfson, M.M., Declercq, J.P. & Germain, G. (1978). *Acta. Cryst.* **A34**, 883-892

Bayes, T. (1763). *Phil. Trans. Roy. Soc.* **53**, 370-418

Bricogne, G. (1984). *Acta. Cryst.* **A40**, 410-445

Bricogne, G. (1988). *Acta. Cryst.* **A44**, 517-545

Bricogne, G. (1991). *Maximum Entropy in Action*, Edited by B. Buck & V.A. MacAulay, pp 187-216. Clarendon Press, Oxford

Bricogne, G. & Gilmore, C.J. (1990). *Acta. Cryst.* **A46**, 284-297

Cochran, W. (1955). *Acta. Cryst.* **8,** 473-478

Cochran, W. & Douglas, A.S. (1955). *Proc. Roy. Soc.* **A227**, 486-500

DeTitta, G.T., Edmonds, J.W., Langs, D.A. & Hauptman, H. (1975). *Acta. Cryst.* **A31**, 472-479

Germain, G., Main, P. & Woolfson, M.M. (1970). *Acta. Cryst.* **B26**, 274-285

Germain, G., Main, P. & Woolfson, M.M. (1971). *Acta. Cryst.* **A27**, 368-376

Germain, G., & Woolfson, M.M. (1968). *Acta. Cryst.* **B24**, 91-96

Giacovazzo, C. (1979). *Acta. Cryst.* **A35**, 296-305

Gilliland, G.L., Windborne, E.L., Nachman, J. & Wlodawer, A. (1990). *Proteins: Struct. Funct. Genet.* **8**, 82-101

Gilmore, C.J. (1984). *J. Appl. Cryst.* **17**, 42-46

Gilmore, C.J., Bricogne, G. & Bannister C. (1990). *Acta. Cryst.* **A46**, 297-308

Gilmore, C.J. & Brown, S.R. (1988). *J. Appl. Cryst.* **21**, 571-572

Gilmore, C.J., Henderson, A.N. & Bricogne, G. (1991). *Acta. Cryst.* **A47**, 842-846

Hall, S.R. & Subramanian, V. (1982). *Acta. Cryst.* **A38**, 598-600

Harker, D. & Kasper, J.S. (1948). *Acta. Cryst.* **1**, 70-75

Hauptman, H. (1974). *Acta. Cryst.* **A30**, 472-476

Hauptman, H. (1975). *Acta. Cryst.* **A31**, 680-687

Hull, S.E. & Irwin, M.J. (1978). *Acta. Cryst.* **A34**, 863-870

Hull, S.E., Viterbo, V., Woolfson, M.M. & Shao-Hui, Z. (1981). *Acta. Cryst.* **A37**, 566-572

Karle, I.L., Hauptman, H., Karle, J. & Wing, A.B. (1958). *Acta. Cryst.* **11**, 257-263

Karle, J. & Hauptman, H. (1950). *Acta. Cryst.* **3**, 181-187

Karle, J. & Hauptman, H. (1953). *Acta. Cryst.* **6**, 473-476

Karle, J. & Hauptman, H. (1956). *Acta. Cryst.* **9**, 635-651

Karle, J. & Karle, I.L. (1966). *Acta. Cryst.* **21**, 849-859

Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. (1983). *Science* **220**, 671-680

Lemaréchal, C. & Navaza, J. (1991). *Acta. Cryst.* **A47**, 631-632

Main, P. (1977). *Acta. Cryst.* **A33**, 750-757

Main, P. (1985). *Crystallographic Computing 3,* Edited by G.M. Sheldrick, C. Kruger & R. Goddard, pp. 206-215, Clarendon Press, Oxford

Main, P. & Hull, S.E. (1978). *Acta. Cryst.* **A34**, 353-361

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087-1092

Neyman, J. & Pearson, E. S. (1933). *Phil. Trans. Roy. Soc.* **A231**, 289-337

Patterson, A.L. (1934). *Phys. Rev.* **46**, 372-376

Parthé, E. & Gelato, L.M. (1984). *Acta. Cryst.* **A40**, 169-183

Rice, S.O. (1944). *Bell System Tech. J.* **23**, 283-332

Rice, S.O. (1945). *Bell System Tech. J.* **24**, 46-156

Sayre, D. (1952). *Acta. Cryst.* **5**, 60-65

Schenk, H. (1973). *Acta. Cryst.* **A29**, 77-82

Shannon, C.E. & Weaver, W. (1949). *The Mathematical Theory of Communications*, Urbana, Univ. of Illinois Press

Shankland, K., Gilmore, C.J., Bricogne, G. & Hashizume, H. (1993). *Acta. Cryst.* **A49**, 493-501

Shao-Hui, Z. & Woolfson, M.M. (1982). *Acta. Cryst.* **A38**, 683-685

Sheldrick, G.M. (1990). *Acta. Cryst.* **A46**, 467-473

Sim, G.A. (1959). *Acta. Cryst.* **12**, 813-815

Sim, G.A. (1960). *Acta. Cryst.* **13**, 511-512

Sjölin, L., Prince, E., Svensson, L.A. & Gilliland, G.L. (1991). *Acta. Cryst.* **A47**, 216-223

Tremayne, M., Lightfoot, P., Mehta, M.A., Bruce, P.G., Harris, K.D.M., Shankland, K., Gilmore, C.J. & Bricogne, G. (1992) *Journal of Solid State Chemistry* **100,** 191-196

White, P.S. & Woolfson, M.M. (1975). *Acta. Cryst.* **A31**, 53-56

Wilson, A.J.C. (1942). *Nature* **150**, 151-152

Woolfson, M.M. (1977). *Acta. Cryst.* **A33**, 219-225

Woolfson, M.M. (1991). *Direct Methods of Solving Crystal Structures*, Edited by H. Schenk, pp. 87-103, Plenum Press, New York

Xiang, S., Carter, C.W., Bricogne, G. & Gilmore, C.J. (1993). *Acta. Cryst.* **D49**, 193-212

Yao Jia-Xing (1981). *Acta. Cryst.* **A37**, 642-644

CHAPTER 2

THE USE OF LIKELIHOOD IN THE SOLUTION OF THE STRUCTURE OF AVIAN PANCREATIC POLYPEPTIDE

1.0 INTRODUCTION

1.1 Avian Pancreatic Polypeptide

Avian Pancreatic Polypeptide (App) is one of a series of closely related pancreatic polypeptides found in mammals and avians. It is a 36-amino acid polypeptide of considerable biological activity (Lonovics, Devit, Watson, Rayford & Thomson, 1981). It acts as part of a physiological feedback system inhibiting pancreatic secretion after a meal. Elevated levels of pancreatic polypeptide are seen in humans suffering from duodenal ulcer. It has also been suggested as a screening method for the early detection of pancreatic tumours.

The crystals have the following crystallographic parameters

Monoclinic Space group C2

$\underline{a}$= 34.18Å $\underline{b}$=32.92Å $\underline{c}$=28.45Å β= 105.26°

It has a crystal formula of $C_{109}N_{53}O_{58}Zn + 80H_2O$, giving 301 non-hydrogen atoms in the asymmetric unit and an approximate molecular weight of 4,800 daltons.

The structure was initially solved and refined to a resolution of 1.4Å (Blundell, Pitts, Tickle, Wood, 1981). It was solved to 2.1Å by single isomorphous replacement, and anomalous scattering from Hg in $HgCl_2$. The phases were refined using a Hull and Irwin weighted tangent formula.

In 1983 the structure was solved to 0.98Å (Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell, 1983) by collecting 53,000 reflections and merging to 17,058 unique reflections with R_{merge} = 6.0%. By a refinement of the 1.4Å structure with the new diffraction data a final R factor for anisotropic non-hydrogens of 15.6% was obtained. It was these merged reflections that were used as the data for the experiments that follow.

The basic crystal structure is that of symmetrical dimers each linked through the zinc atoms. Three crystallographically related App molecules contribute ligands to the metal atoms. Each zinc is bonded to atoms N and O of residue GLY1 in the first App molecule. It is bonded to atom OD1 of residue ASN23 of the second molecule, atom NE2 of residue HIS34 in the third and the O of a water molecule. This gives a zinc co-ordination of a distorted trigonal bi-pyramid. The molecular structure is that of a small globular protein with a hydrophobic core. The residues 2-8, shown in Figure 1, form a polyproline II - like helix closely packed by hydrophobic interactions against an α-helix comprised of residues 14-32.

GLY PRO SER GLN PRO THR TYR PRO GLY ASP ASP ALA PRO VAL GLU ASP LEU ILE ARG PHE TYR ASP ASN LEU GLN GLN TYR LEU ASN VAL VAL THR ARG HIS ARG TYR

FIGURE 1. The 36 residues of Avian Pancreatic Polypeptide

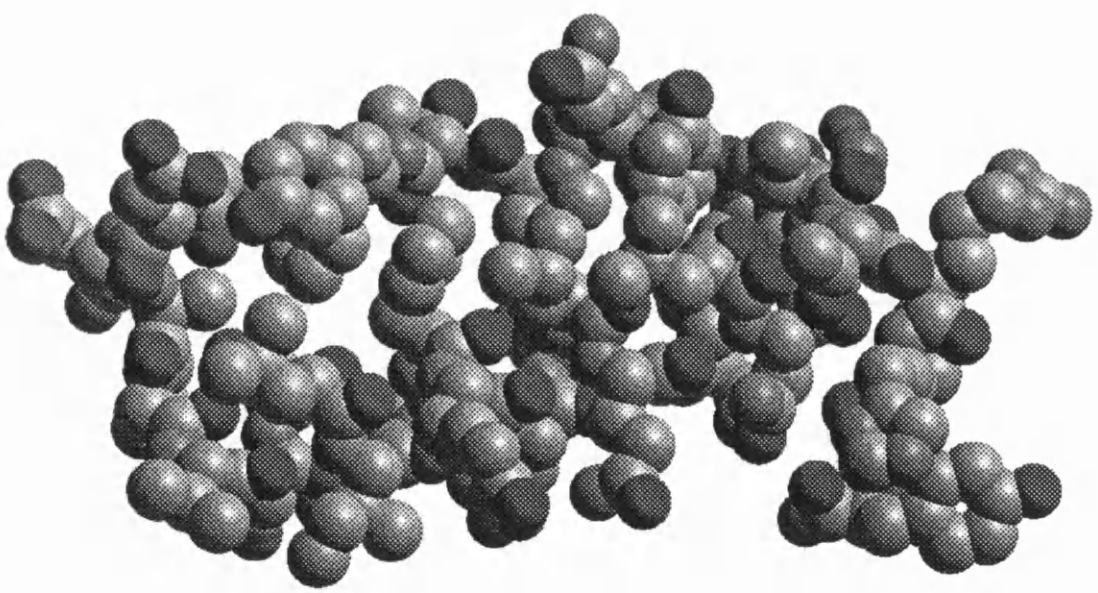


FIGURE 2. The structure of Avian Pancreatic Polypeptide, a shaded sphere image

1.2 Difficulties in the Solution of App

The data supplied for the experiments detailed in this chapter were supplied by Michael M. Woolfson and contained only a fraction of those observed by the initial diffraction experiment. We further reduced the supplied data to a smaller resolution shell to speed up the computation. This very small amount of data should make the solution or identification of solutions impossible for conventional direct methods for the following reasons.

i. The phase problem is only soluble as more data is gathered than there are parameters required to define the system. This has led to the general rule that it is necessary to gather at least five times the number of reflections as there are atoms in the asymmetric unit, or else the structure will prove to be very difficult to solve. In the case of App we have approximately 2.7 reflections per atom and so the phase problem will prove very difficult for this structure.

ii. We have no high resolution reflections in our data set, conventional direct methods depend on these high resolution reflections for phasing and structures cannot be solved without them. This is expressed best in Sheldricks' rule:

"If less than 50% of the theoretically observable reflections are observed in the resolution range 1.1-1.2Å are observed ($|F_{\underline{h}}| > 4.0\sigma(|F_{\underline{h}}|)$) then the structure will be difficult to solve by conventional direct methods."

The determination of the crystal structure of App has been attempted with conventional direct methods techniques using the program SAYTAN (Woolfson & Jia-Xing, 1988, Woolfson & Jia-Xing, 1990). This program generated 1000 trial phase sets, from the largest 727 reflections and successfully refined some phase sets to correctness. The program failed in its attempt to discriminate between the good and bad phase sets using conventional figures of merit. In an attempt to solve an unknown structure this failure would prevent the structure being determined without the laborious task of examining the E-map from each of the phase sets.

2.0 THEORY

2.1 The Sayre Equation & Tangent Formula

The Sayre equation (Sayre, 1952) is an algebraically derived formula that can be used to relate the phases of three reflections, thus determining the phase of one reflection from the known phases of two others. It is only true under strict conditions and contains no probabilistic information and so is now of limited use when compared with later work done on triplet theory.

$$|F_{\underline{h}}|e^{i\varphi_{\underline{h}}} = \frac{\Theta_{\underline{h}}}{V}\sum_{\underline{k}}|F_{\underline{k}}||F_{\underline{h}-\underline{k}}|e^{i(\varphi_{\underline{k}}+\varphi_{\underline{h}-\underline{k}})} \qquad \text{(EQ 2.1.1)}$$

Where:

$$\Theta_{\underline{h}} = \frac{f_{\underline{h}}}{\gamma_{\underline{h}}} \qquad \text{(EQ 2.1.2)}$$

and:

- $f_{\underline{h}}$ is the scattering factor for each atom
- $\gamma_{\underline{h}}$ is the Fourier coefficient of the squared electron density of each atom
- V is the volume of the unit cell in Å^3

While this equation is only true under special circumstances, it can be seen through inspection, that for large $|F_{\underline{k}}|$, $|F_{\underline{h}-\underline{k}}|$, that the sign, S, of $|F_{\underline{h}}|$ is probably equal to the sign of the product of the signs of $F_{\underline{k}}$, $F_{\underline{h}-\underline{k}}$, for centric reflections.

$$S_{\underline{h}}S_{\underline{k}}S_{\underline{h}-\underline{k}} \approx 1 \qquad \text{(EQ 2.1.3)}$$

The Sayre equation is only true for structures containing identical, spherically symmetrical atoms, whose electron densities do not overlap, and which are at rest or possess identical isotropic temperature factors. It should also be noted that the summation over all reflections $\underline{k}$ implies an infinite summation. Assuming that all these conditions are met then the Sayre equation provides a route whereby diffracted amplitudes and phases can be related by exact equations.

The tangent formula (Karle & Hauptman, 1956) which is most often used for the determination and refinement of phases can be derived from the Sayre equation, but is more correctly derived by other methods. In its weighted form it may be written:

$$\tan\phi_{\underline{h}} \approx \frac{\sum_{\underline{k}} W_{\underline{k}} W_{\underline{h}-\underline{k}} |E_{\underline{k}}||E_{\underline{h}-\underline{k}}| \sin(\phi_{\underline{k}} + \phi_{\underline{h}-\underline{k}})}{\sum_{\underline{k}} W_{\underline{k}} W_{\underline{h}-\underline{k}} |E_{\underline{k}}||E_{\underline{h}-\underline{k}}| \cos(\phi_{\underline{k}} + \phi_{\underline{h}-\underline{k}})} \qquad \text{(EQ 2.1.4)}$$

Where $W_{\underline{k}}$, $W_{\underline{h}-\underline{k}}$ are weights.

This form of the tangent formula treats all invariants as independent of each other and this assumption, being false, can lead to overrefinement of phases. This is especially true when only invariants that use reflections with large E-magnitudes are involved. As the probability of the tangent formula being true greatly increases with increasing $|E_{\underline{h}}||E_{\underline{k}}||E_{\underline{h}-\underline{k}}|$ there are good practical reasons for selecting phase relationships that involve large E-magnitudes.

Where overconsistency of phases is a problem there are two approaches used to overcome this. The first is to use a new weighting scheme in the tangent formula, such a scheme is that devised by Hull and Irwin (Hull & Irwin, 1978). The alternative is to use the Sayre equation directly in an iterative refinement method. Using triplets that involve two strong and one weak E-magnitude prevents the overconsistancy caused by only using the largest E-values.

2.2 SAYTAN

SAYTAN (SAYre TANgent refinement) is a program that can be used as an alternative to the traditional tangent refinement of direct methods. At its core it uses the Sayre-equation tangent formula (Debaerdemaeker, Tate & Woolfson, 1985; Debaerdemaeker, Tate & Woolfson, 1988)(EQ 2.2.1). Using the Sayre-equation tangent formula gives some improvement over the traditional tangent refinement, particularly with structures that lack translational symmetry (Debaerdemaeker, Tate & Woolfson, 1984).

$$\varphi_{\underline{h}} = \text{phase of } \{t(\underline{h}) - 2Kq(\underline{h})\} \qquad \text{(EQ 2.2.1)}$$

Where

$$t(\underline{h}) = \sum_{\underline{k}} \left[\frac{1}{g(\underline{h})} + \frac{1}{g(\underline{k})} + \frac{1}{g(\underline{h}-\underline{k})} \right] E_{\underline{k}} E_{\underline{h}-\underline{k}} \qquad \text{(EQ 2.2.2)}$$

$$q(\underline{h}) = \sum_{\underline{l}} \left[\frac{1}{g(\underline{l})^2} \right] \sum_{\underline{k}} E_{\underline{k}} E_{\underline{l}-\underline{k}} E_{\underline{h}-\underline{l}} \qquad \text{(EQ 2.2.3)}$$

$$K = \frac{\sum_{\underline{k}} E_{-\underline{k}} t(\underline{k})}{\sum_{\underline{k}} E_{-\underline{k}} q(\underline{k})} \qquad \text{(EQ 2.2.4)}$$

g(h) is the scattering factor of atoms with squared electron density

This equation tends to return phases for reflections with large E-magnitudes which will satisfy Sayres equation (EQ 2.1.1) for reflections with both large and small E-magnitudes. This is an important method of avoiding the overconsistancy problems that are associated with the traditional tangent refinement. The addition of these small E-magnitude reflections gives additional phasing power to this method, although the small E-magnitude reflections are not themselves phased.

The SAYTAN program can generate many hundreds of random phase sets which it will then refine to a minimum in phase space. The determination of correct phase sets is then done by conventional figures of merit (FOMs) (see chapter 1), and also with the TFOM (Refaat, Tate, Yao Jia-Xing, Woolfson, 1990) figure of merit.

$$TFOM = \sum_{\underline{h},\underline{k}} K(\underline{h},\underline{k}) \left\{ \cos(\varphi_{\underline{h}} + \varphi_{\underline{k}} + \varphi_{\underline{h}-\underline{k}}) - \frac{I_1(K(\underline{h},\underline{k}))}{I_0(K(\underline{h},\underline{k}))} \right\}^2 \qquad \text{(EQ 2.2.5)}$$

Where

$$K(\underline{h},\underline{k}) = \frac{2}{\sqrt{N}} \left| E_{\underline{h}} E_{\underline{k}} E_{\underline{h}-\underline{k}} \right| \qquad \text{(EQ 2.2.6)}$$

There have been some considerable difficulties with these FOMs giving false indications of correct phase sets in the case of App. It is not unusual in a large or complex structure for the traditional figures of merit to fail or give false indications. The TFOM was the most successful figure of merit, but even this could not discriminate in any conclusive way, resulting in only 4 correct sets in the top 20

favoured sets. It would be possible to examine each electron density map to determine if any structural information was present, however this would be very time consuming, and obviously this is far from a satisfactory solution to this problem. This FOM problem appears to be the only real problem in the SAYTAN solution of App.

The SAYTAN algorithm is sensitive to the quality and resolution of data (Debaerdemaeker, Tate & Woolfson, 1985) which is not a problem with a maximum entropy approach to solving crystal structures. This may also be why the FOMs in SAYTAN have such difficulty in pointing to the correct phase sets.

2.3 MICE

MICE (**M**aximum entropy **I**n a **C**rystallographic **E**nvironment) (Gilmore, Bricogne & Bannister, 1990) is a computer program written in standard Fortran 77 on a UNIX platform. Its structure follows very closely that of MITHRIL (Gilmore, 1984; Gilmore & Brown, 1988), with a collection of independent modules linked by a central control program. The structure and control of the program with both batch and interactive modes of operation is very similar to MITHRIL and indeed, there is an interface between the two programs.

MICE is an a generalised computing tool for the elucidation of crystal structures using a maximum entropy and maximum likelihood approach as suggested by Bricogne (Bricogne, 1984;Bricogne 1988). Although the maximisation of the entropy is very computationally intensive, improvements in hardware technology and greater refinement of the code have produced a much faster program capable of solving large structures, to high resolution, in a relatively short period of time.

The following is a breakdown of the processing of the data through the MICE program:

i. The normalisation of the structure factors $|F_{\underline{h}}|^{obs}$ is done in the program MITHRIL, using the Wilson plot method, to give normalised structure factors $|E_{\underline{h}}|^{obs}$. These normalised structure factors are then converted to unitary structure factors $|U_{\underline{h}}|^{obs}$, and have a variance estimated using the technique of Hall and Subramanian (Hall & Subramanian, 1982).

ii. The determination of origin and enantiomorph, if appropriate, is performed in a process described in Chapter 1 section 15.4

iii. The origin and enantiomorph reflections and the remaining phased reflections are now used as constraints (the basis set $\{H\}$) in the calculation of a non-uniform maximum entropy prior or $q^{ME}(\underline{x})$ map.

iv. The $q^{ME}(\underline{x})$ map is now updated and has its new entropy maximised using the exponential modelling technique (Bricogne, 1984). The exponential modelling algorithm is unstable and is only stabilised by a plane search technique that relies on a variety of damping factors, bumpers and other checks to prevent the plane search from finding false maxima (Bricogne & Gilmore, 1990). In early cycles of entropy maximisation a more computationally economic line search may be used without loss of robustness.

v. At each cycle of entropy maximisation the log likelihood gain is determined.

vi. Steps (iv) and (v) are repeated until the reduced χ^2 statistic (EQ 2.3.1) has reached unity. The χ^2 statistic is a measure of the fit between $\left|U^{ME}_{\underline{h}\in H}\right|$ and $\left|U^{obs}_{\underline{h}\in H}\right|$.

$$\chi^2 = \frac{1}{(2n_a + n_c)} \sum_{\underline{h}\in H} \frac{\left|U^{obs}_{\underline{h}} - U^{ME}_{\underline{h}}\right|^2}{S^2_{\underline{h}}} \qquad \text{(EQ 2.3.1)}$$

Where

n_c is the number of centric reflections in the basis set $\{H\}$

n_a is the number of acentric reflections in the basis set $\{H\}$

The term $(2n_a + n_c)$ is the number of degrees of freedom for the system.

The variable $S_{\underline{h}}$ is a measure of the variance and is given by EQ 2.3.2

$$S^2_{\underline{h}} = \sigma^2_{\underline{h}} + p\varepsilon_{\underline{h}}\Sigma \qquad \text{(EQ 2.3.2)}$$

Where

$\varepsilon_{\underline{h}}$ is the standard epsilon factor

$\sigma_{\underline{h}}^2$ is the error on $\left|U_{\underline{h} \in K}^{OBS}\right|$

p is an empirical parameter set to unity for this work.

Σ is related to the effective number of atoms within the unit cell, and so can be thought of as a measure of structural complexity. It is a refinable parameter in the likelihood calculation.

2.4 The Role of Σ_a and Σ_c Criteria

In the calculation of χ^2 a quantity Σ must be evaluated (see EQ 2.3.2). This quantity can be broken down into two component parts Σ_a and Σ_c, these two parts relating to acentric and centric reflections respectively. These figures are used in both the calculation of χ^2 and N_{eff}, a parameter that reflects both the quality of phase information and molecular complexity, it is defined as the effective number of atoms in the unit cell. N_{eff} can be calculated as the weighted mean of $\frac{1}{\Sigma_c} + \frac{2}{\Sigma_a}$.

The correct sets all show lowered values of both Σ_a and Σ_c. While these give an indication of correct phase sets they are highly correlated to χ^2 and do not give substantial discriminative powers when compared to log-likelihood gain. These parameters also give false indication as to the correctness of set 15 shown in Table 6.

2.5 The Role of Entropy

We maximise the entropy of our system only to minimise bias in our results i.e be maximally non-committal to what we do not know. Entropy is a function of χ^2 and this dependence on χ^2 prevents entropy being used as a figure of merit (see Table 2). In cases where two phase sets have opposing indications of correctness based on entropy and likelihood then a formula that has been used to select the favoured phase set is

$$FOM = NS + L \qquad \text{(EQ 2.5.1)}$$

Where

N is the effective number of atoms in the unit cell, a refinable parameter in the likelihood calculation.

S is the entropy

L is the log likelihood gain

The set is preferred if FOM is a maximum. Normally there is a strong indication from likelihood and there is no need to refer to the entropy for set indication.

3.0 EXPERIMENTAL & RESULTS

3.1 Preparation of Data

All data processed through the MICE program was kindly provided by Michael Woolfson and Yao Jia-Xing. All MICE jobs were run on a cluster of three MASSCOMP computers, a 5400, 5450 and a 6300, using the UNIX operating system. It had been pre-processed in the following way (Gilmore, Henderson & Bricogne, 1991).

1. The 16,538 data were normalised to give E-magnitudes using a Wilson plot (Wilson, 1942).

2. The largest 800 and smallest 200 were selected for further processing. 73 of the largest and 18 of the smallest were lost during convergence due to insufficient connectivity. This leaves 727 large E's connected by 9726 triplets and 182 weak reflections connected by 6106 triplets.

3. Random phases were assigned to the 909 reflections and refined using the SAYTAN program until convergence. 1000 trial sets were generated and tested by SAYTAN.

4. Three subsets of the initial 1000 trials were extracted and were used as input to the MICE program. The selection of the subsets was made by M.M. Woolfson and attempt to simulate a method that could be used to solve unknown structures using a SAYTAN/MICE combination.

   (i) A collection of 10 phases sets containing one set with an absolute phase error of $<40^{o}$, all others having an absolute phase error $> 79^{o}$

   (ii) A collection of 50 phase sets containing six sets with an absolute phase error of $< 50^{o}$ all others having an absolute phase error $> 80^{o}$

   (iii) A collection of 20 phase sets that were the preferred sets filtered by the TFOM figure of merit.

Each phase set that was received was processed by the MICE program as follows:

1. The data received contained only h,k,l,E and φ. Errors on the E's are necessary for the weighting of reflections in their contribution to the likelihood, so these errors were simulated. This was done by assigning:

$$\sigma(|E_{\underline{h}}|) = 0.1|E_{\underline{h}}| + 0.01 \qquad \text{(EQ 3.1.1)}$$

This proved to be quite adequate for our use and does reflect the normal level of error found in measured data.

2. These E-values, $|E_{\underline{h}}|$,were than converted to Unitary Structure Factors, $|U_{\underline{h}}|$, using the equation

$$|U_{\underline{h}}| \cong \frac{|E_{\underline{h}}|}{\sqrt{N}} \qquad \text{(EQ 3.1.2)}$$

Where
N is the number of atoms in the unit cell

3. Unlike a normal maximum entropy solution, we start with the set $\{H\}$, the set of phased reflections, which is composed of 117 unique reflections at 1.9Å that have been phased by the SAYTAN program. The subset of reflections at 1.9Å was chosen to speed up the calculation of the Fourier transforms in MICE. As all $q^{ME}(\underline{x})$ maps must be oversampled, to reduce aliasing errors, producing a map of 0.98Å data would require a grid of 0.3Å. This would give rise to approximately a 1.2 million point Fourier map. As each cycle of entropy maximisation requires a maximum of 14 Fourier transforms, the computing power required to perform such a calculation would have been beyond the computing capabilities of our laboratory at the time that this work was performed. The final choice of sampling the map at 0.6Å gives rise to a 172,800 point Fourier map.

4. The initial 117 phased reflections $\underline{h} \in H$ are used as constraints in the calculation of the non-uniform prior. The reflections $\underline{h} \in H$ are duly returned from the Fourier transform, at the end of each cycle of entropy maximisation from the final $q^{ME}(\underline{x})$ map. Information is now returned from the final $q^{ME}(\underline{x})$ map in the form $|U^{ME}_{\underline{h} \in K}|$ and also $\varphi^{ME}_{\underline{h} \in K}$, where $\{K\}$ is the set of all reflections not in the basis set $\{H\}$.

5. The log-likelihood gain (LLG) (EQ 3.1.3) is calculated to compare $\left|U^{ME}_{\underline{h}\in K}\right|$ and $\left|U^{obs}_{\underline{h}\in K}\right|$.

$$L = L(H) - L(H_o) \qquad \text{(EQ 3.1.3)}$$

Where

$L(H)$ is the log likelihood based on the agreement of observed and extrapolated U-magnitudes for all reflections in the set $\{K\}$.

$L(H_o)$ is the log likelihood null hypothesis (i.e. all extrapolated magnitudes are set to zero)

Steps 4 and 5 are repeated until the fit of $\left|U^{ME}_{\underline{h}\in H}\right|$ and $\left|U^{OBS}_{\underline{h}\in H}\right|$ is such that the reduced χ^2 statistic (EQ 2.3.6) is equal to 1.

For all phase sets the reduced χ^2 statistic was allowed to go to unity before entropy maximisation was terminated. At $\chi^2 = 1$ we should have an optimum match of $\left|U^{OBS}_{\underline{h}\in H}\right|$ and $\left|U^{ME}_{\underline{h}\in H}\right|$. In practice it was found that the LLG was maximised at approximately $\chi^2 = 1.9$. It was important to follow the change in LLG for all χ^2, for all phase sets, in order to determine the shape of the log-likelihood gain graph for this set of diffraction data.

3.2 Experiment 1

A set of ten phase sets, one of which has $\langle|\Delta\varphi|\rangle < 40^o$, and all others have $\langle|\Delta\varphi|\rangle > 79^o$ were provided by M.M. Woolfson. The data were processed through the MICE program and the results are shown in Table 1.

This shows that the LLG very strongly indicates the set with the lowest $\langle|\Delta\varphi|\rangle$. While the figures may appear to be small it must be remembered that this is a $\log_e$ scale. It is interesting to note that the correct phase set is the one that conventional figures of merit point to as being the least likely of all ten sets to contain any relevant structural information (see Table 2). Note that the ψ_o figure of merit strongly indicates that there is overconsistancy of phases. Note also that for a structure of this size, the traditional figures of merit indicate that all sets with the exception of the correct one are good candidates for structure solutions. In Table 2 we also show that the entropy of the map is not a good figure of merit.

Set no.	Absolute Phase Error	Maximum $L(H) - L(H_0)$	Corr χ^2	$\Sigma_a \times 10^{-3}$	$\Sigma_c \times 10^{-3}$
1	79.77	-0.013	2.47	0.68	1.02
2	39.19	11.269	1.81	0.60	0.86
3	84.54	0.193	1.58	0.67	0.99
4	82.91	0.078	2.25	0.68	0.98
5	84.83	-0.026	3.44	0.71	1.10
6	84.38	0.004	2.52	0.68	1.02
7	84.51	0.369	1.40	0.67	0.97
8	81.09	3.479	0.97	0.66	0.94
9	83.34	0.006	2.98	0.69	1.02
10	84.95	0.097	2.45	0.68	0.99

TABLE 1. Log-likelihood gain for 10 trial phase sets generated by SAYTAN

Set no.	Absolute Phase Error	Maximum $L(H) - L(H_0)$	Entropy S	Ψ_0	R_α	ABS FOM
1	79.77	-0.013	-0.329	0.87	29.95	0.66
2	39.19	11.269	-0.518	2.73	29.24	1.70
3	84.54	0.193	-0.842	0.77	27.21	0.67
4	82.91	0.078	-0.413	0.84	26.33	0.69
5	84.83	-0.026	-0.632	0.90	27.53	0.70
6	84.38	0.004	-0.312	0.83	26.94	0.67
7	84.51	0.369	-0.948	0.89	26.18	0.72
8	81.09	3.479	-0.153	0.82	26.04	0.72
9	83.34	0.006	-0.165	0.90	26.15	0.70
10	84.95	0.097	-0.331	0.80	27.24	0.69

TABLE 2. A comparison of maximum log-likelihood gain and corresponding entropy with normal figures of merit for 10 trial phase sets.

In Figure 3 we show the shape of the graph of log-likelihood gain against χ^2 for phase set 2 from Table 1, the correct set. This characteristic maximum at $\chi^2 \approx 1.9$ is repeated in all subsequent experiments and can also be used as an indication of a correct phase set, for this structure only. Compare the graph in Figure 3 with that in Figure 4, a graph showing the same plot but for phase set 5 from Table 1. Notice how there is now no meaningful likelihood maximum and this is characteristic for all incorrect phase sets.

FIGURE 3. A plot of log-likelihood gain vs Reduced χ^2 for phase set 2 $\langle|\Delta\varphi|\rangle = 39.2^o$ Note that this shows the characteristic log-likelihood maximum at $\chi^2 = 1.9$

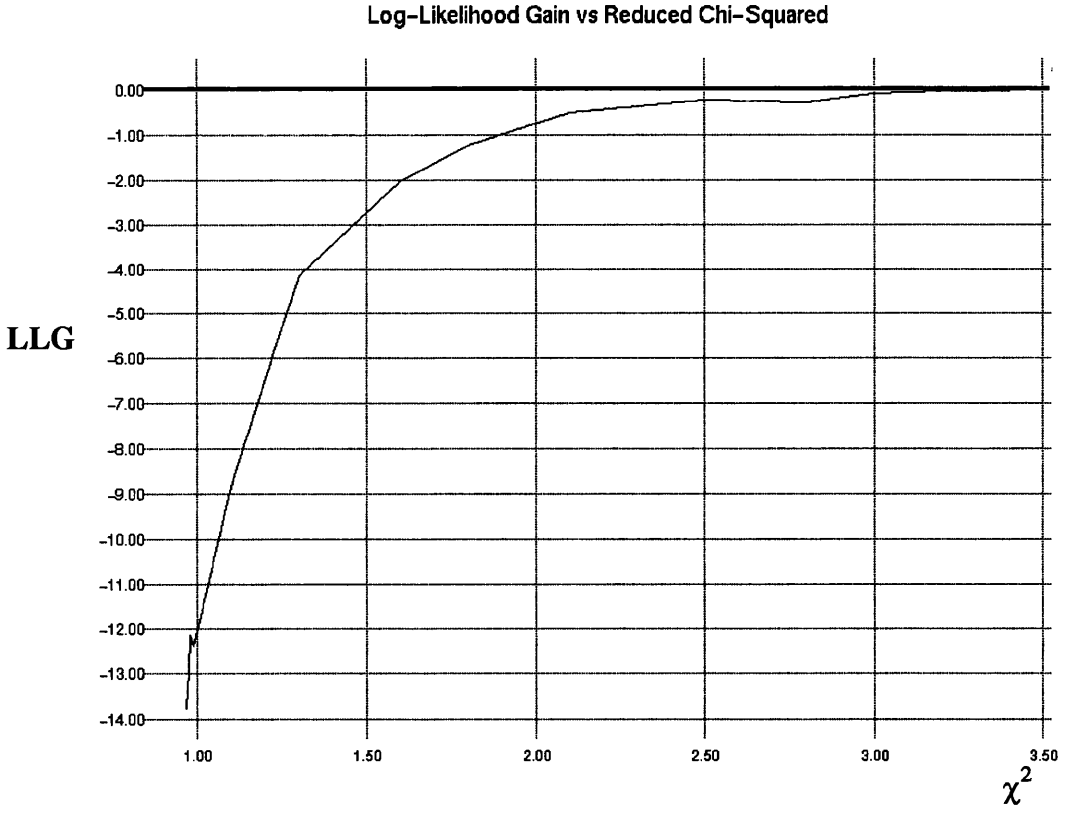


FIGURE 4. A plot of log-likelihood gain vs Reduced χ^2 for phase set 5 $\langle|\Delta\varphi|\rangle = 84.83^o$ Note that this does not show the characteristic log-likelihood maximum at $\chi^2 = 1.9$

3.3 Experiment 2

A set of 50 phase sets, six of which have $\langle|\Delta\varphi|\rangle < 50^o$, the remaining 44 have $\langle|\Delta\varphi|\rangle > 80^o$ were provided by M.M. Woolfson. As in experiment 1, the data were processed through the program MICE and the results of this experiment is shown in Table 3.

The largest log-likelihood gain for an incorrect phase set is 3.60 while the minimum log-likelihood gain for a correct set is 4.58. The log-likelihood of 4.58 does not occur at $\chi^2 = 1.9$ but at $\chi^2 = 2.4$. In a normal structure solution this set would have to be classified as marginal and any better basis set used to proceed with the structure determination.

The remaining five sets are all very strongly indicated with the maximum LLG occurring at the characteristic $\chi^2 \approx 1.9$.

Set no.	Absolute Phase Error	Maximum $L(H) - L(H_0)$	Corr χ^2	$\Sigma_a \times 10^{-3}$	$\Sigma_c \times 10^{-3}$
1	85.44	3.594	0.98	0.66	0.96
2	84.45	1.366	1.22	0.67	0.96
3	84.71	0.069	2.41	0.68	0.95
4	82.57	1.923	0.98	0.66	0.95
5	83.71	0.333	1.49	0.67	0/97
6	83.36	2.612	1.25	0.67	0.95
7	42.75	10.521	1.93	0.62	0.89
8	84.33	-0.014	3.07	0.69	1.06
9	85.68	0.146	1.77	0.68	0.99
10	85.01	0.223	1.51	0.67	0.96
11	47.11	11.104	1.93	0.62	0.89
12	85.02	0.188	2.05	0.68	0.99
13	83.58	0.457	1.53	0.67	0.97
14	80.24	3.605	0.97	0.65	0.92
15	48.64	8.470	2.03	0.63	0.91
16	85.09	1.288	1.58	0.66	0.95
17	84.16	0.052	3.05	0.69	1.03
18	48.19	10.496	1.95	0.62	0.88
19	85.60	0.002	3.15	0.69	0.98
20	82.01	1.905	1.52	0.66	0.95
21	85.70	1.802	0.99	0.66	0.94
22	49.18	4.584	2.35	0.66	0.95
23	84.04	0.174	2.04	0.68	0.98
24	85.96	1.856	1.30	0.66	0.95
25	83.38	1.095	0.98	0.66	0.94

TABLE 3. Log-likelihood gain for 50 phase sets generated by SAYTAN

Set no.	Absolute Phase Error	Maximum $L(H) - L(H_0)$	Corr χ^2	$\Sigma_a \times 10^{-3}$	$\Sigma_c \times 10^{-3}$
26	85.45	2.093	1.30	0.66	0.94
27	82.24	0.099	2.81	0.68	1.02
28	83.01	-56.686	1.80	0.68	0.97
29	85.79	0.487	1.99	0.67	0.96
30	84.81	0.089	2.62	0.68	1.01
31	80.62	-0.019	3.30	0.70	1.12
32	43.28	10.546	1.94	0.63	0.89
33	85.17	0.121	1.49	0.67	0.99
34	84.52	0.078	1.76	0.68	0.97
35	83.36	-0.018	3.20	0.69	1.06
36	85.59	0.075	1.78	0.68	0.99
37	85.22	0.289	1.54	0.67	0.98
38	82.48	3.420	0.98	0.66	0.93
39	84.74	0.019	2.61	0.68	0.98
40	85.40	0.008	1.74	0.68	0.99
41	85.47	-4.793	3.43	0.70	1.14
42	83.28	-15.918	2.83	0.69	1.04
43	82.18	-41.973	3.28	0.68	1.00
44	81.10	-18.551	3.48	0.69	1.12
45	83.90	-31.611	3.35	0.70	1.12
46	84.13	-0.043	3.10	0.69	1.07
47	85.43	0.119	1.82	0.68	0.97
48	84.62	2.975	0.97	0.66	0.95
49	86.17	0.689	1.55	0.67	0.98
50	81.92	0.005	3.37	0.71	1.08

TABLE 3. (continued): Log-likelihood gain for 50 phase sets generated by SAYTAN

This maximum log-likelihood gain at $\chi^2 \approx 1.9$ is not the expected result. From the theory we would expect an LLG maximum at $\chi^2 = 1.0$, which is the point of best fit between $\left|U^{ME}_{\underline{h}\in H}\right|$ and $\left|U^{obs}_{\underline{h}\in H}\right|$. It is believed that this behaviour is caused by the Zn atoms forming a pseudo-centrosymmetric lattice within the cell. This effect would be most pronounced at 2.0Å where the Zn atoms are the dominant feature of the electron density maps. This will cause a severe enantiomorph definition problem, and it was observed that extrapolated phases $\varphi^{ME}_{\underline{h}\in K}$ and phases input as constraints $\varphi^{ME}_{\underline{h}\in H}$ both showed a dependence on the value of the reflection index k.

Once we have detected our six correct sets from the sample of fifty sets we then calculate an electron density map. This differs from the $q^{ME}(\underline{x})$ map in that it is Sim filtered to generate a centroid map. In this map reflections belonging to both $\{H\}$ and $\{K\}$ are used and assigned weights $W_{\underline{h}}$, that are dependent on whether the reflections are centric or acentric

$$W^{centric}_{\underline{h}} = \tanh\left(\frac{N_{eff}}{\varepsilon_{\underline{h}}\left|U^{ME}_{\underline{h}}\right|\left|U^{obs}_{\underline{h}}\right|}\right) \qquad \text{(EQ 3.3.1)}$$

$$W^{acentric}_{\underline{h}} = \frac{I_1\left(\left(\frac{N_{eff}}{\varepsilon_{\underline{h}}}\right)\left|U^{ME}_{\underline{h}}\right|\left|U^{obs}_{\underline{h}}\right|\right)}{I_0\left(\left(\frac{N_{eff}}{\varepsilon_{\underline{h}}}\right)\left|U^{ME}_{\underline{h}}\right|\left|U^{obs}_{\underline{h}}\right|\right)} \qquad \text{(EQ 3.3.2)}$$

Where

N_{eff} is the effective number of atoms in the unit cell, and is a refinable parameter in the likelihood calculation.

I_1, I_0 are the first and zero order Bessel functions respectively

$\varepsilon_{\underline{h}}$ is the crystallographic epsilon factor

3.4 Experiment 3

The last set of data to be tested were the 20 phase sets preferred by the TFOM figure of merit. This group of phase sets contains four sets which have $\langle|\Delta\varphi|\rangle < 50^o$ the remaining sixteen sets having $\langle|\Delta\varphi|\rangle > 70^o$. The test is to determine whether LLG is sensitive enough to discriminate between a correct phase set containing relevant structural information with $\langle|\Delta\varphi|\rangle = 49.5^o$, and an incorrect phase set containing no information with $\langle|\Delta\varphi|\rangle = 70.2^o$. This experiment was also done to simulate a method which could be performed in a practical situation to solve difficult structures. As it is very time consuming to perform a full ab initio solution of a crystal structure using maximum entropy techniques, it may be possible to generate many thousands of phase sets using the SAYTAN program and then pass only those preferred by the TFOM figure of merit to the log-likelihood calculation. There are problems with this however, as SAYTAN takes the phases to its preferred minimum in phase space. When the SAYTAN refined phases are taken into MICE this is not the phase minimum that would be expected by the maximum entropy technique.

The results are shown in Table 4 and again the correct sets are indicated most strongly, with a log-likelihood gain > 10. Fifteen of the remaining sixteen sets all have a maximum log-likelihood gain < 1.5. One set, set 15, has a maximum log-likelihood gain of 10.22, yet has a $\langle|\Delta\varphi|\rangle = 80.2^o$. By all previous tests a log-likelihood in excess of 10 is considered to be a strong indication that the phase set is correct. An examination of the electron density map for set 15 yielded no interpretable information, while the maps of the four correct sets all produced the four Zinc atoms quite clearly in Figures 5 and 6. This anomaly may be a local entropy maximum only applicable to the 1.9Å resolution subset because phases are being taken from a different phasing technique which has determined a different optimum point.

Set no.	Absolute Phase Error	Maximum $L(H) - L(H_0)$	Corr χ^2	$\Sigma_a \times 10^{-3}$	$\Sigma_c \times 10^{-3}$
1	80.29	0.110	3.40	0.68	1.01
2	80.55	0.455	1.39	0.67	1.04
3	76.83	0.066	2.87	0.68	1.05
4	76.55	0.082	3.41	0.68	1.01
5	76.56	1.313	2.46	0.67	1.00
6	71.77	0.037	3.22	0.69	1.14
7	49.46	10.019	2.00	0.63	0.92
8	46.11	10.013	1.73	0.58	0.85
9	79.06	0.075	3.57	0.68	1.01
10	40.59	10.566	1.79	0.60	0.88
11	76.50	0.054	3.57	0.68	1.01
12	41.10	11.896	1.83	0.61	0.89
13	74.23	-0.002	3.32	0.70	1.22
14	80.61	0.070	3.42	0.68	1.01
15	80.16	10.226	2.01	0.63	0.92
16	74.37	0.097	2.68	0.68	1.06
17	82.27	0.065	3.58	0.68	1.01
18	70.15	1.084	1.41	0.66	0.97
19	75.52	0.560	1.31	0.66	0.99
20	76.98	0.533	1.61	0.67	0.99

TABLE 4. Log-likelihood gain for the 20 phases sets most favoured by the TFOM figure of merit.

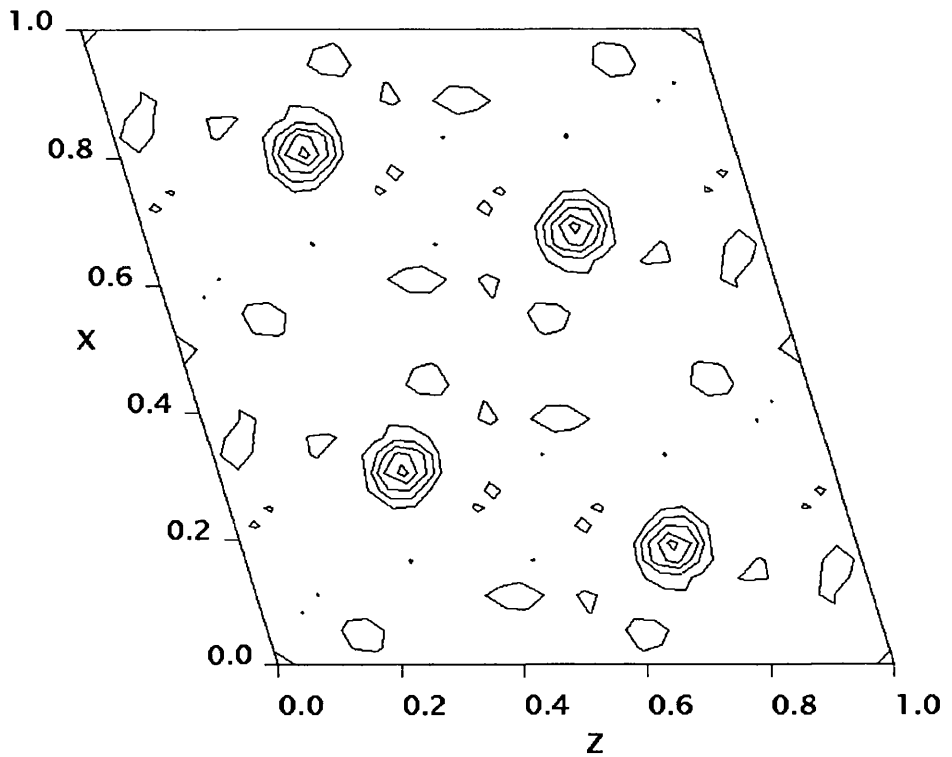


FIGURE 5. The electron density map projected down the Y-axis showing the position of the four zinc atoms.

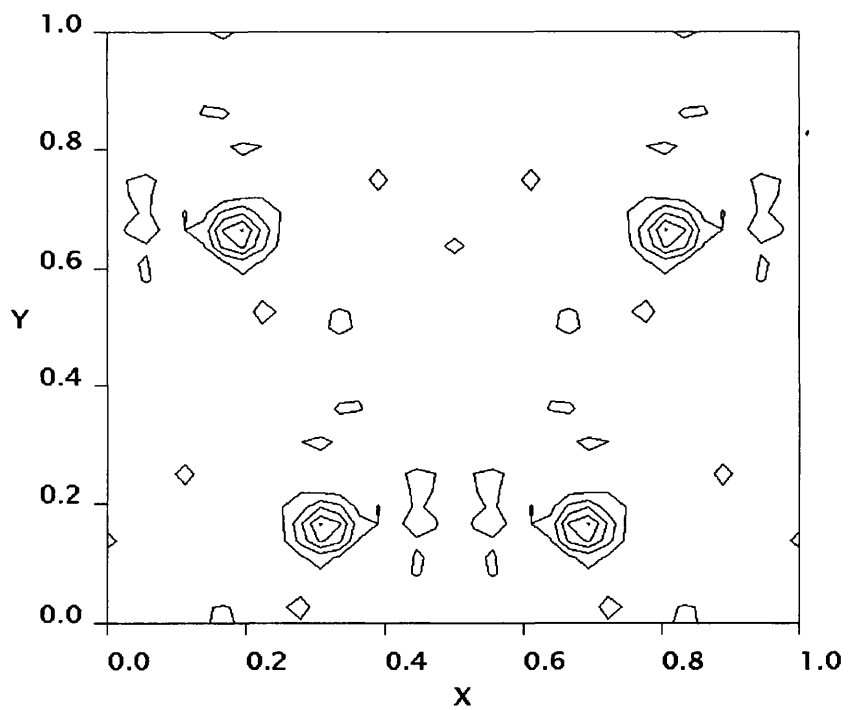


FIGURE 6. The electron density map projected down the Z-axis showing the position of the four zinc atoms.

3.5 Summary

The previous experiments have quite clearly shown that log-likelihood gain (LLG) is a figure of merit of unique discriminatory powers. The experiments have also shown that LLG can be used in situations where the phase problem is severely underdetermined and continue to produce valid results.

Under the conditions that LLG determined the correct phase sets (i.e. 117 reflections at 1.9Å) no conventional figure of merit would have been able to make any conclusive indication of correctness. The results given in the previous sections are further proof of the usefulness of the maximum entropy method in situations where conventional direct methods fail.

While we have not shown a full ab initio solution for App in this work, but we have demonstrated a very powerful technique for the solution of large or difficult structures. The program SAYTAN has proven itself to be of considerable power and speed at producing refined correct phase sets, but because it uses only conventional figures of merit it cannot discriminate between good and bad phase sets. The MICE program has demonstrated the ability to very clearly discriminate between phase sets. If these two programs are used together then the possibility of being able to solve even larger structures using direct methods increases.

4.0 FUTURE WORK

The computer limitations on the maximum size of a Fourier map are being reduced and new techniques that allow distributed processing, coupled with advanced computer hardware are allowing much faster computation times. These enhancements are all being implemented to be used in conjunction with the MICE program. These advances should allow a full resolution calculation at 0.98Å resolution for the data. This should alleviate the enantiomorph problem and also hopefully remove the anomalous set 15 in Table 6.

As our data set is large consisting as it does of 117 unique reflections it should be possible to refine the phases of these reflections even further than is possible using SAYTAN (Bricogne, 1984). These further refined phases would provide better quality maps for fragment refinement and also provide a better starting point for phase extrapolations in MICE.

It should be possible to take our 117 reflections and move onward through the phasing tree to give a full solution for the structure, not just the Zinc positions. If this proves successful then a full ab initio solution of the protein starting from origin and enantiomorph is required. This ab initio work is currently being undertaken by C.J. Gilmore. This work involves the determination of the zinc positions at 3Å, which has already been done. This has allowed the MICE program to break the enantiomorph problem at an earlier stage and also removed the problem of LLG peaking at $\chi^2 \approx 1.9$. It has worked well, and with a minimum number of reflections, likelihood strongly indicates the presence of an optimal phase set, selected from 4000, with which to proceed.

References

Blundell, T.L., Pitts, J.E., Tickle, I.J. & Wood, S.P. (1981). *Proc. Natl. Acad. Sci. USA* **78,** 4175-4179

Bricogne, G. (1984). *Acta. Cryst.* **A40**, 410-445

Bricogne, G. (1988). *Acta. Cryst.* **A44**, 517-545

Bricogne, G. & Gilmore, C.J. (1990). *Acta. Cryst.* **A46**, 284-297

Debaerdemaeker, T., Tate, C. & Woolfson, M.M. (1984). *Acta. Cryst. Suppl.* **A40**, PS17.2.18

Debaerdemaeker, T., Tate, C. & Woolfson, M.M. (1985). *Acta. Cryst.* **A41**, 286-290

Debaerdemaeker, T., Tate, C. & Woolfson, M.M. (1988). *Acta. Cryst.* **A44**, 353-357

Gilmore, C.J. (1984). *J. Appl. Cryst.* **17**, 42-46

Gilmore, C.J. & Brown, S.R. (1988). *J. Appl. Cryst.* **22**, 571-572

Gilmore, C.J., Bricogne, G. & Bannister, C. (1990). *Acta. Cryst.* **A46**, 297-308

Gilmore, C.J., Henderson, A.N. & Bricogne, G. (1991). *Acta. Cryst.* **A47**, 842-846

Glover, I., Haneef, I., Pitts, J.E., Wood, S.P., Moss, D., Tickle, I.J. & Blundell, T.L. (1983). *Biopolymers* **22,** 293-304

Hall, S.R. & Subramanian, V. (1982). *Acta. Cryst.* **A38**, 598-600

Hull, S.E. & Irwin, M.J. (1978). *Acta. Cryst.* **A34**, 863-870

Karle, J. & Hauptman, H. (1956). *Acta. Cryst.* **9**, 635-651

Lonovics, J., Devit, P., Watson, L.C., Rayford, P.C. & Thomson, J.C. (1981). *Arch. Surg.* **116**, 1256-1264

Refaat, L., Tate, C., Yao Jia-Xing & Woolfson, M.M. (1990). *Acta. Cryst. Suppl.* **A46**, PS 02.08.13

Sayre, D. (1952). *Acta. Cryst.* **5**, 60-65

Wilson, A.J.C. (1942). *Nature* **150**, 151-152

Woolfson, M.M. & Yao Jia-Xing (1988). *Acta. Cryst.* **A44**, 410-413

Woolfson, M.M. & Yao Jia-Xing (1990). *Acta. Cryst.* **A46**, 409-413

CHAPTER 3

A BAYESIAN METHOD OF NORMALISATION

1.0 INTRODUCTION

1.1 The Discrete Atom Constraint

Correct electron density must always conform to certain constraints i.e the electron density must always be positive. It is these constraints that provide the mathematical basis for the theories of direct methods. One of the most powerful of these constraints is that the electron density must contain discrete atoms when the observed data is of atomic resolution, and as this is the very nature of matter this restriction on the electron density is always true.

To make use of this restriction mathematically we remove the shape of the atom from $F_{\underline{h}}$, converting the structure factor to a normalised structure factor, $E_{\underline{h}}$. The removal of the shape of the atom is done by dividing the structure factor by the scattering factors of the atoms in the unit cell (see Chapter 1 EQ 3.1.2). This reduces the electron density at the atoms to points, and thus prevents any decrease in scattering power with increasing $(sin^2\theta)/\lambda^2$. When normalised structure factors are used in phase determining formulae the phase constraints are strengthened so that the electron density map should contain well resolved atomic peaks.

All the above mathematical quantities can be related to real space quantities using Fourier transforms:

$E_{\underline{h}}$	x	atomic scattering factor	=	$F_{\underline{h}}$
↕ F.T.		↕ F.T.		↕ F.T.
point atom structure	*	real atom	=	electron density

1.2 Normalisation

To obtain a value for $E_{\underline{h}}$ we are first required to calculate the following variables:

1. A scale factor, K, required to place $|F_{\underline{h}}^{obs}|$ and $|F_{\underline{h}}^{calc}|$ onto the same absolute scale.
2. An overall isotropic temperature factor, B, to remove the effect of thermal vibration from the atomic shape.

These variables can be calculated from the intensity statistics using a Wilson Plot (Wilson, 1942) as detailed in Chapter 1, Section 3.1. An alternative method of normalising data has been suggested using the Patterson map (Nielsen, 1975). The Wilson plot method is the most common normalisation technique used and has been modified to be more accurate under certain circumstances e.g. the K-curve technique (Karle & Hauptman, 1953). These modifications are improvements, but are not a major departure from the basic technique or theoretical assumptions of the Wilson plot.

Normalisation is one of the most important steps in any structure solution as the majority of the probability methods used for determining phases are sensitive to the E-magnitudes of the reflections (Hall & Subramanian, 1982a). The E-magnitudes have a profound effect on the convergence mapping procedure (Germain, Main & Woolfson, 1970) and this will determine whether a structure can be solved. Direct methods only solve so many structures with such a high success rate due to the multisolution method, finding alternative phasing paths through the convergence map, some of which lead to an acceptable structure solution.

The major failures of the Wilson plot are encountered when sparse or systematically missing data is used to determine the scale and temperature factors. To aid in the normalisation of such data and also to provide a new alternative normalisation method, a Bayesian technique has been developed and integrated into the MITHRIL90 program (Gilmore, 1984; Gilmore & Brown, 1988). The remainder of this chapter will detail the theory and experimental results of this new technique.

2.0 THEORY

2.1 The Wilson Plot

As can be seen from Chapter 1 Section 3.1, it is possible to obtain the scale and temperature factors from a plot of $ln\left(\frac{\langle I_{rel}\rangle}{\sum_{j=1}^{N}[f_j]_T^2}\right)$ versus $(sin^2\theta)/\lambda^2$.

To make this plot $(sin^2\theta)/\lambda^2$ is split into resolution shells, each having the same volume and thus should contain approximately equal numbers of reflections. The selection of the size of the shell is important as each shell must contain enough reflections to be statistically representative but must also be small enough to display minimal variation in $(sin^2\theta)/\lambda^2$ (Rogers, 1965). Shells that contain too few observed reflections cannot be used statistically and this is a difficulty with organic structures that diffract weakly at high $(sin^2\theta)/\lambda^2$ due to large thermal effects.

The theory predicts that this plot should ideally be a straight line but in practice is frequently a curve. This deviation from ideal behaviour can be caused by systematic errors in the measured intensities. These errors can either come from the technique used in the measurement of the intensities or may come from the non-random distribution of atoms in the unit cell. The random distribution of atoms in the unit cell is central to the derivation of the Wilson statistics, but this assumption can be wrong in any of four categories of atomic distribution:

i. If the number of independent atoms in the unit cell is low (e.g. <8) then this cannot be a uniform distribution. The converse of this may also be true in the case of a large number of atoms that are concentrated within a relatively small volume of the unit cell, this atomic distribution can occur with proteins.

ii. When a number of atoms lie on symmetry elements then these atoms will contribute to a subset of reflections only (Hauptman & Karle, 1953). The experimental curves become more anomalous the heavier the atoms in the special position (Shmueli, Rabinovich & Weiss, 1990).

iii. When the relative number of heavy atoms is large. This can be thought of as a special case of (i), as the contribution of the light atoms will be negligible relative to the heavy atoms present.

iv. Pseudosymmetry can cause the crystal to diffract as though it possessed higher symmetry than is truly present. This affects the normalisation procedure by making the choice of space group ambiguous.

To reduce deviations from the idealised straight line it is common practice to overlap contiguous resolution shells. This has the effect of reducing scatter of reflections in the shells and also removes the effects of anomalous shells. An example of this overlapping shells technique is given for diamantane (Rogers & Kennard, unpublished) in Table 1.

Set Number	Range of $(sin^2\theta)/\lambda^2$	Mean $(sin^2\theta)/\lambda^2$	Number of Refs. in Range	Average Intensity
1	0.000 - 0.042	0.026	285	1663
2	0.021 - 0.063	0.044	422	1234
3	0.042 - 0.084	0.065	547	997
4	0.063 - 0.105	0.085	615	506
5	0.084 - 0.126	0.106	686	299
6	0.105 - 0.147	0.126	767	239
7	0.126 - 0.168	0.149	825	288
8	0.147 - 0.189	0.169	882	285
9	0.168 - 0.210	0.190	910	202
10	0.189 - 0.231	0.211	992	146
11	0.210 - 0.252	0.231	1056	106
12	0.231 - 0.274	0.253	1093	65
13	0.252 - 0.295	0.272	1047	40
14	0.274 - 0.316	0.294	1066	34
15	0.295 - 0.316	0.304	580	29

TABLE 1. Distribution of Reflections within Resolution shells for Diamantane

It is normal practice to perform a least squares fit of the points on the Wilson plot to a straight line. In an attempt to improve the fitting of this straight line to the curve Hall and Subramanian (Hall & Subramanian, 1982a) tested weighting the least squares procedure. This weighting procedure was found to be unreliable as the weights were based on $|F_{\underline{h}}^2|$ and $\sigma(|F_{\underline{h}}^2|)$ which are highly correlated to each other, and this introduces a systematic bias into the refinement.

2.2 K-Curve Fitting

The basis of the Wilson plot is the assumption of random distributions of atoms in the unit cell. In reality structural regularities cause the Wilson plot to deviate from the idealised straight line. To take these regularities into account it is possible to fit a monotonically decreasing function to the experimentally derived curve (Karle & Hauptman, 1953). This is the K-curve technique and is normally recommended when the Wilson plot is significantly removed from the straight line.

Extensive studies were carried out on normalisation in the early eighties by Hall and Subramanian (Subramanian & Hall, 1982) into the optimum choice for a normalisation technique. These studies have indicated that the Wilson plot / least squares fitting technique is in fact better than the K-curve technique.

The difference between the scale and temperature factors derived by the K-curve and Wilson Plot technique is small, however these small inaccuracies give rise to less accurate invariants and therefore an increase in the potential for phasing errors. It is this sensitivity to error in the phasing process that make determination of the best possible estimates for B and K essential.

2.3 Errors on the Scale and Temperature Factors

A method for estimating the errors on the E-magnitudes based on the precision of the intensity data and the Wilson plot parameters was introduced in the early eighties (Hall & Subramanian, 1982b). Values for these errors may be obtained by using the following equation:

$$\sigma^2_{\underline{h}}(|E_{\underline{h}}|) = |E_{\underline{h}}|^2\left(\frac{\sigma^2(|F_{\underline{h}}|)}{|F_{\underline{h}}|^2} + \frac{\sigma^2(K)}{K^2} + \frac{sin^4\theta}{\lambda^4}\sigma^2(B)\right) + |E_{\underline{h}}|^2\left(\frac{2\left(\frac{sin^2\theta}{\lambda^2}\right)^2 r(K,B)\,\sigma(K)\,\sigma(B)}{K}\right) \qquad \text{(EQ 2.3.1)}$$

Where

$|F_{\underline{h}}|$ and $\sigma(|F_{\underline{h}}|)$ can both be obtained from the experimental observations

B,K and $|E_{\underline{h}}|$ can all be obtained from the Wilson plot

$r(K,B)$, $\sigma(K)$ and $\sigma(B)$ are all parameters that can be determined from the Wilson plot least-squares fitting procedure

2.4 A New Bayesian Method

The largest errors in normalisation occur, not in the fitting of the straight line, but when the data is very sparse i.e. where a subset of reflections are absent, this is often caused by the threshold of unobserved reflections being set too high when the data are being collected, thus omitting weak reflections.

While it is possible to insert these missing weak reflections back into the data set using a Bayesian technique (French & Wilson, 1978), this is inappropriate in the case of missing reflections that are strong. The process of using the full Bayesian insertion technique is very complex involving additional data measurement and computational requirements, although a partial application of the technique has been suggested for routine use (Hall & Subramanian, 1982a).

The strength of a new fully Bayesian method of the determination of K and B would be the ability to normalise the data without the need for such a complex method to alter the intensity data prior to the normalisation. Ideally we would use the measured data in the best way possible to yield optimum values of K and B.

2.5 Derivation of the Formula

Let us define the Bayesian temperature factor as β which is related to the conventional temperature factor B, by the equation

$$\beta = 2B \qquad \text{(EQ 2.5.1)}$$

and also define the Bayesian Scale factor κ which is related to the conventional scale factor K by the equation

$$\kappa = 1/K \qquad \text{(EQ 2.5.2)}$$

With application of Bayes Theorem (Bayes, 1763) we obtain:

$$P(\kappa, \beta | \{R_i\}) \propto \prod_{i=1}^{N} P(R_i | \kappa, \beta) \cdot P(\kappa, \beta) \qquad \text{(EQ 2.5.3)}$$

Where

N is the number of observed reflections

κ is the Bayesian scale factor

β is the Bayesian temperature factor

$P(\kappa, \beta | \{R_i\})$ is the probability of κ, β given the set of measured intensities $\{R_i\}$

$P(R_i | \kappa, \beta)$ is the probability of an individual intensity given κ, β

$P(\kappa, \beta)$ is our prior knowledge of κ, β

If we substitute Wilson statistics for $P(R_i | \kappa, \beta)$ and substitute unity for $P(\kappa, \beta)$ then we obtain for N reflections:

$$\log(P(\kappa, \beta | \{R_i\})) = \sum_{i=1}^{N} \left(\frac{-R_i^2}{\kappa exp\left(\frac{-\beta sin^2\theta}{\lambda^2}\right)} - \log\kappa + \beta\frac{sin^2\theta}{\lambda^2} + \log 2R_i \right) \quad \text{(EQ 2.5.4)}$$

Where

$$R_i = \frac{|F_i|}{\varepsilon \sum_{j=1} f_j^2} \quad \text{(EQ 2.5.5)}$$

Where

N is the number of atoms in the unit cell

f_j is the atomic scattering factor

ε is the epsilon factor which is dependant on the point group and reflection indices

As $\log 2R_i$ does not vary with κ, β it is only a constant added to the probability and can be ignored for the purposes of the determination of maxima, also $\log\kappa$ is not dependant on i and so can be taken outside the summation.

Therefore EQ 2.5.4 can be rewritten as:

$$\log\left(P\left(\kappa,\beta|\ \{R_i\}\right)\right) = \sum_{i=1}^{N}\left[\frac{-R_i^2 exp\left(\frac{\beta sin^2\theta}{\lambda^2}\right)}{\kappa} + \beta\frac{sin^2\theta}{\lambda^2}\right] - N\log\kappa \quad \text{(EQ 2.5.6)}$$

EQ 2.5.6 allows the calculation of a surface of probability, whose global maximum is at the most probable values for κ and β. The profile of this peak will yield errors on the values of κ and β.

In the derivation of EQ 2.5.6 we have used no prior information on κ, β. In Bayesian terminology we have used a uniform prior i.e. we have multiplied our probability surface by a plane. This need not be the case, and any suitable prior could be substituted into $P(\kappa, \beta)$ for the derivation of another equation. The ability to change the prior gives an easily extendable theory which can be tailored to the type of data used for the normalisation. This technique is especially applicable to protein diffraction data.

As we have used a uniform prior for $P(\kappa, \beta)$ and Wilson statistics for $P(R_i|\ \kappa, \beta)$ the formula EQ 2.5.6 is mathematically equivalent to the Wilson plot, although the method used to determine κ, β in the Bayesian technique will differ from that used in the Wilson plot. This difference in the method of determination of most probable values of κ, β should result in very small differences in the values of κ, β found by the two techniques. Differences greater than the esd of the values of κ, β would not be expected.

2.6 Obtaining Optimal Values of κ, β

In the version of the program implemented in to MITHRIL90 two techniques for obtaining the maximum from the probability surface were coded.

(1) A grid search where an area of probability map of course grain is calculated and searched for a maximum. Once a new maximum value for $\log\left(P\left(\kappa,\beta|\ \{R_i\}\right)\right)$ is obtained another probability map is calculated around this point with a finer grain, which is itself searched for a maximum. This process is repeated until the user is satisfied that κ, β have been determined accurately enough. A version of this procedure was implemented with the program making the decision as to when the values of κ, β were accurate enough for normal use. Using the grid search technique is

only recommended when the eigenvector/eigenvalue system becomes ill conditioned, as it cannot provide esd estimates for κ or β.

(2) An eigenvector/eigenvalue system was set up to speed the searching process and also to provide accurate esd estimates on the values of κ, β.

Matrices must be set up to determine shifts toward the maximum of the probability surface as in EQ 2.6.1

$$\mathbf{Ax = -b} \qquad \text{(EQ 2.6.1)}$$

Where

$$A = \begin{bmatrix} \frac{d^2f(\kappa,\beta)}{d\beta^2} & \frac{d^2f(\kappa,\beta)}{d\beta d\kappa} \\ \frac{d^2f(\kappa,\beta)}{d\kappa d\beta} & \frac{d^2f(\kappa,\beta)}{d\kappa^2} \end{bmatrix} \qquad \text{(EQ 2.6.2)}$$

A Hessian matrix

$$b = \begin{bmatrix} \frac{df(\kappa,\beta)}{d\beta} \\ \frac{df(\kappa,\beta)}{d\kappa} \end{bmatrix} \qquad \text{(EQ 2.6.3)}$$

$$x = \begin{bmatrix} \delta\beta & \delta\kappa \end{bmatrix} \qquad \text{(EQ 2.6.4)}$$

The solution to EQ 2.6.1 is determined by performing a Householder reduction of **A** to give the tridiagonal form of the matrix, followed by a tridiagonal QL implicit solution to yield the eigenvalues, eigenvectors and shifts (Press, Flannery, Teukolsky & Vetterling, 1992). The shifts are then applied to κ, β and the values of the elements of matrix **A** are re-evaluated. This process continues until the shifts are small. The derivatives of EQ 2.5.6 follow:

$$\frac{df(\kappa,\beta)}{d\beta} = -\frac{1}{\kappa}\sum_{i=1}^{N}\frac{R_i^2 sin^2\theta}{\lambda^2} exp\left(\frac{\beta sin^2\theta}{\lambda^2}\right) + \sum_{i=1}^{N}\frac{sin^2\theta}{\lambda^2} \qquad \text{(EQ 2.6.5)}$$

$$\frac{df(\kappa,\beta)}{d\kappa} = \frac{1}{\kappa^2}\sum_{i=1}^{N} R_i^2 exp\left(\frac{\beta sin^2\theta}{\lambda^2}\right) - \frac{N}{\kappa} \qquad \text{(EQ 2.6.6)}$$

$$\frac{d^2f(\kappa,\beta)}{d\kappa^2} = \frac{-2}{\kappa^3}\sum_{i=1}^{N} R_i^2 exp\left(\frac{\beta sin^2\theta}{\lambda^2}\right) + \frac{N}{\kappa^2} \qquad \text{(EQ 2.6.7)}$$

$$\frac{d^2f(\kappa,\beta)}{d\beta^2} = -\frac{1}{\kappa}\sum_{i=1}^{N} \frac{R_i^2 sin^4\theta}{\lambda^4} exp\left(\frac{\beta sin^2\theta}{\lambda^2}\right) \qquad \text{(EQ 2.6.8)}$$

$$\frac{d^2f(\kappa,\beta)}{d\kappa d\beta} = \frac{d^2f(\kappa,\beta)}{d\beta d\kappa} = \frac{1}{\kappa^2}\sum_{i=1}^{N} \frac{R_i^2 sin^2\theta}{\lambda^2} exp\left(\frac{\beta sin^2\theta}{\lambda^2}\right) \qquad \text{(EQ 2.6.9)}$$

2.7 Calculation of Errors on κ, β

The matrix technique for the determination of the probability maximum yields two eigenvectors u_1 and u_2, it also yields two eigenvalues λ_1 and λ_2. These eigenvectors may be thought of as the axes of an equal probability ellipse, whose length is the eigenvalue, from the centre of the ellipse, see Figure 1.

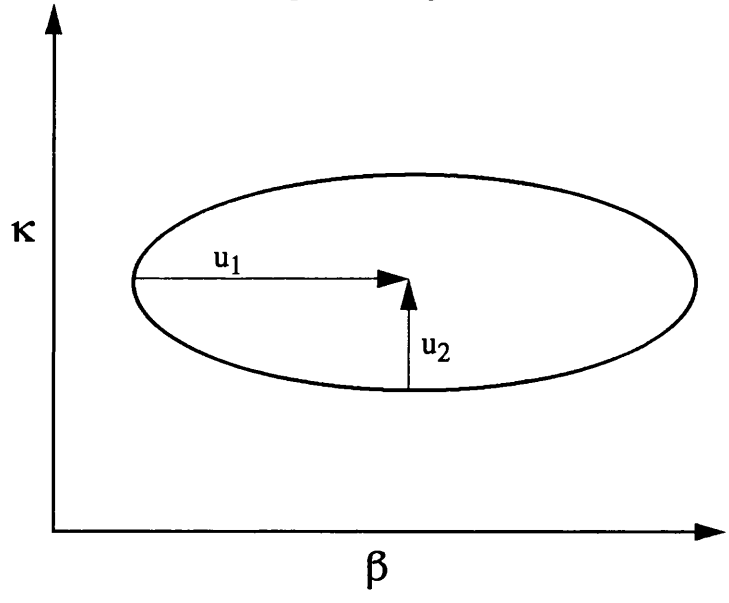


FIGURE 1. The eigenvectors u_1 and u_2 on an ellipse of equal probability for β, κ.

This allows the calculation of the errors on β, κ from the following equations (R.K. Henderson, private communication):

$$\sigma(\kappa) = \sqrt{\sum_{i=1}^{2} \frac{(u_i e_1)^2}{\lambda_i^2}} \qquad \text{(EQ 2.7.1)}$$

$$\sigma(\beta) = \sqrt{\sum_{i=1}^{2} \frac{(u_i e_2)^2}{\lambda_i^2}} \qquad \text{(EQ 2.7.2)}$$

Where $e_1 = [0\ 1]$ $e_2 = [1\ 0]$

2.8 Implementation into MITHRIL90

The source code for the Bayesian normalisation method was written to be flexible and easily extended as the theory developed. All subroutines were written in standard FORTRAN 77 to be run on a UNIX workstation.

The subroutines that performed the Householder reduction and tridiagonal QL implicit solution were heavily modified versions of the subroutines available from the book Numerical Recipes (Press, Flannery, Teukolsky & Vetterling, 1992).

To activate the Bayesian normalisation method in the program MITHRIL90, the user enters the normalisation module using option 2 (see Appendix A). This allows the use of all other NORMAL module functions e.g. turning off the error calculations on the E-magnitude calculations. The new MITHRIL90 NORMAL menu has been modified to the form shown in Figure 2.

THE FOLLOWING COMMANDS ARE CURRENTLY AVAILABLE:

NORMAL	<	IK,NB,ISC,MAXD UP,[NOSIG][PHASE]					>
	<	IK=-2/-1/0/1/2 Test/E-input/Wilson/K-curve/Bayes					>
	<	NB=No. points for plot; ISC=1 No parity rescale					>
	<	MAXDUP=Maximum no of duplicates/absences to list					>
	<	[NOSIG] Turns off the calculation of esd of E					>
	<	[PHASE] Reads phase angles with intensity data					>
LIST	<	Print a full E-list	>	SYMM	<	Symmetry operation	>
CELL	<	a,b,c,angles or cos	>	LATTICE	<	A/C P/A/B/C/I/F/R	>
CONTENTS	<	N1,Type,N2,Type etc.	>	SFAC	<	Scattering factor	>
LIMITS	<	Sin max,Sin min,Emax	>	BSCL	<	B,Scale factor(s)	>
NEWE	<	h,k,l new E-value	>	EDIT	<	h,k,l,(F)	>
TRANS	<	Trans matrix	>	GROUP	<	Type,pop,cell	>
ATOM	<	Atom Label,x,y,z	>	NOCHECK	<	Do not check data	>
XRAY	<	X-ray data input	>	NEUT	<	Neutron Data	>
POWD	<	Powder data	>	ELEC	<	Electron data	>
OVER	<	Overlaps:h,k,l,F,sig	>	ENTR	<	Link to MICE prog	>
DATA	<	Format of data	>	MISS	<	Find weak reflns	>

N.B. DATA must be last instruction to Normal
Only the first 4 characters of any dialogue command are significant and a <CR> then terminates current input.

FIGURE 2. The normalisation menu from MITHRIL90.

The output from the Bayesian normalisation method is different from the K-curve or Wilson methods in that no graph of the Wilson plot, or tables of statistical analysis of the graph are produced. All tables of analysis of the E-magnitude distributions are produced, as are errors on the E-magnitudes if these are requested. An example of the output from the normalisation of diamantane is shown in Figure 3.

DIAMANTANE-4-OL(FILE 1)

Normalisation by Bayesian methods
No listing of complete set of E-values
Standard deviations on E-values to be calculated
No index group rescaling
Radiation type: X-rays
This structure is centrosymmetric with lattice type P

The 4 symmetry operations are as follows

X	Y	Z
1/2 - X	1/2 - Y	Z
- Y	1/2 + X	1/2 + Z
1/2 + Y	- X	1/2 + Z

Scale and temperature factor to be calculated

Direct cell is A= 16.704 B= 16.704 C= 7.922 alpha = 90.00 beta = 90.00 gamma = 90.00

Unit cell contents

Atom	Number in cell	Atomic number	Atomic scattering factors f = aa*exp(-a*rho) + bb*exp(-b*rho) +cc)					radii	
C	112	6	2.112	7.827	2.462	31.650	1.412	0.77	1.70
H	160	1	0.388	7.151	0.601	30.180	0.008	0.37	1.20
O	8	8	4.197	6.327	2.218	22.830	1.578	0.74	1.40

Number of atoms in the asymmetric unit = 15.00
Input data with 1 reflections per line - format (3I4,2F8.2,30X,I4)

There are 0 duplicates and 0 systematic absences in the data set containing 1553 reflections
A total of 0 reflections were edited and 0 exceeded the theta limits

Temperature factor = 3.7347 (esd =0.1700) Scale factor = 0.8312 (esd = 0.0570)
Suggested least-squares scale factor = 1.0974
The scale factor - temperature factor correlation coefficient = 0.9287

FIGURE 3. The layout of the Bayesian normalisation report for the MITHRIL90 program

3.0 EXPERIMENTAL AND RESULTS

3.1 Source of Data Sets

During the testing of the crystallographic package SHELXTL (Sheldrick, 1985) Sheldrick compiled a database of structures that are difficult to solve using the techniques of direct methods. Since these are some of the most difficult structures to solve, are in a variety of space groups and have a broad range of atom types, it was felt that a selection of seventeen of these structures would be a good test of our new normalisation method.

A brief summary of the crystallographic data for each test structure is given in Table 2.

Structure	Space Group	Formula	Z
Diamantane-4-ol	$P4_2/n$	$C_{14}H_{20}O$	8
Quinol	$R\bar{3}$	$C_6H_6O_2$	54
HOV1	$C2/m$	$Pr_{14}Ni_6Si_{11}$	4
Selendid	$P2_1$	$C_{22}H_{28}O_2Se$	2
Azet	$Pca2_1$	$C_{21}H_{16}ClNO$	8
TUR10	$P6_322$	$C_{15}H_{24}O_2$	12
Dodecane-Diol	$I\bar{4}2d$	$C_{10}H_{18}O_2$	16
APAPA	$P4_12_12$	$C_{30}H_{37}N_{15}O_{16}P_2 \cdot 6H_2O$	8
MGHEX	$P3_1$	$C_{48}H_{68}N_{12}O_{12}Mg \cdot 2ClO_4 \cdot 4CH_3CN$	3
TOTC	$P6_1$	$C_{33}H_{36}O_6 \cdot O \cdot 2C_{16}H_{33}OH$	6
TPH	$C222_1$	$C_{24}H_{20}N_2$	12
Goldman2	Cc	$C_{28}H_{16}$	8
Munich1	$C2$	$C_{20}H_{16}$	8
MBH2	$P1$	$C_{15}H_{24}O_3$	3
SUOA	$P2_12_12_1$	$C_{28}H_{38}O_{19}$	4
Winter2	$P2_1$	$C_{52}H_{83}N_{11}O_{16} \cdot 3CH_2Cl_2$	2
Loganin	$P2_12_12_1$	$C_{17}H_{26}O_{10}$	4

TABLE 2. The seventeen test structures selected from the Sheldrick database.

All the above data sets normalise well using both the Wilson plot and K-curve fitting technique. Subsets of the observed structure factors were created using a purpose-written program called DATAFILT.

3.2 Random Data

Two subsets of the complete data sets were selected by the program DATAFILT. The first composed of 10% of the total reflections selected at random and the other a data set composed of fifty reflections selected at random. These two tests were performed to compare the capabilities of the new technique and the Wilson plot at dealing with non-systematically sparse data. The results of the seventeen tests are shown in the tables of results starting on page 105.

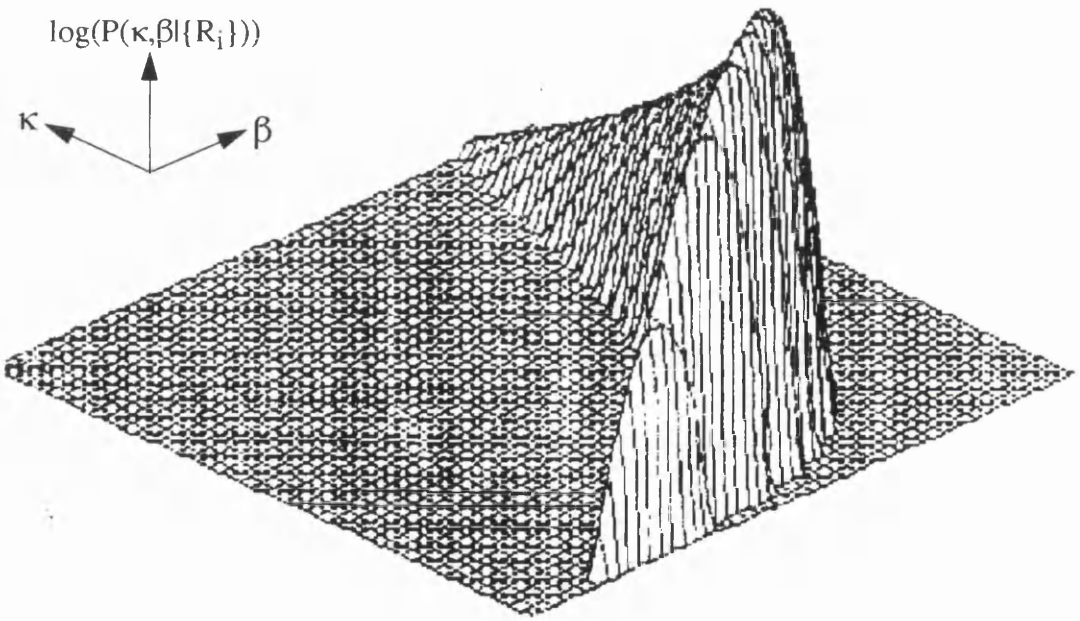
3.3 Low Angle Data

Three subsets of the low angle data were generated using the DATAFILT program. The first was composed of 10% (rounded down) of the data set with the lowest resolution. The second data set was composed of the fifty lowest angle reflections. The last subset contains all reflections with a resolution below 2.0Å. These three subsets of data reflect a realistic problem in crystallography, where data may be sparse and only available at low resolution, i.e. electron diffraction patterns. In these subsets of data frequently the 2.0Å and 10% data sets differ by a total of only 10 reflections, however, both sets were tested for the completeness of the study.

3.4 Examination of Probability Surfaces

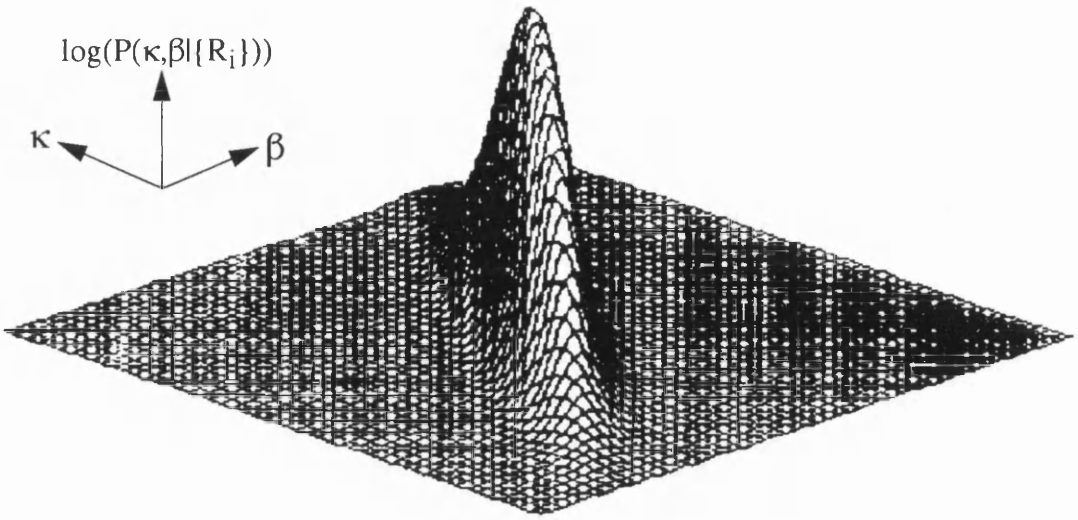
It is possible to calculate the probability for a range of κ, β and plot this as a surface. This was done for two structures, Diamantane and Goldman2, to verify that near the maximum on the probability surface, the cross section of the peak does approximate to an ellipse, and also to study trends in the change of data quality and the subsequent effect on the quality of the probability map. The following diagrams were all produced using SIMPLEPLOT (Butland) routines using probability maps produced by MITHRIL90.

The cut factor refers to a value on the Z-axis below which $\log(P(\kappa,\beta|\{R_i\}))$ is not plotted. This is done to allow clearer maps to be produced and in no way affects the integrity of the data produced in the maps below. However this cut factor on the Z axis is not always constant and care should be taken when comparing maps in a sequence.



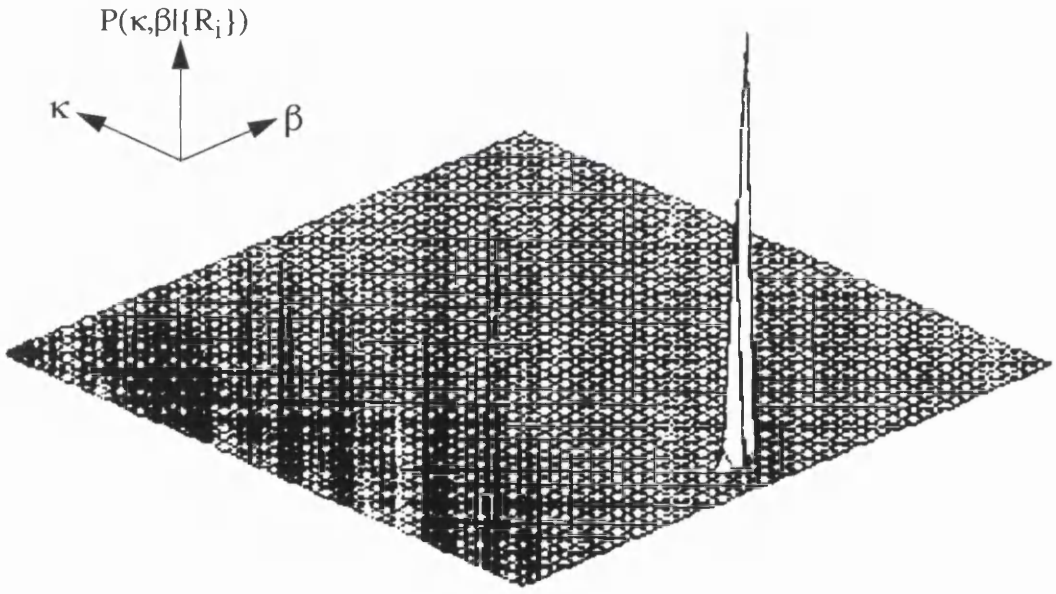
MAP 1A Diamantane produced using the full observed data set.
Range in β = -5.0 - 20.0 Range in κ = 0.1 - 12.0 Cut factor = 100

The above map shows a large range of β,κ with what was found to be a characteristic curvature, which shows a rapid increase in probability with increasing β,κ followed by a slow reduction in probability with increasing κ. There is currently no explanation for this behaviour in the theory.



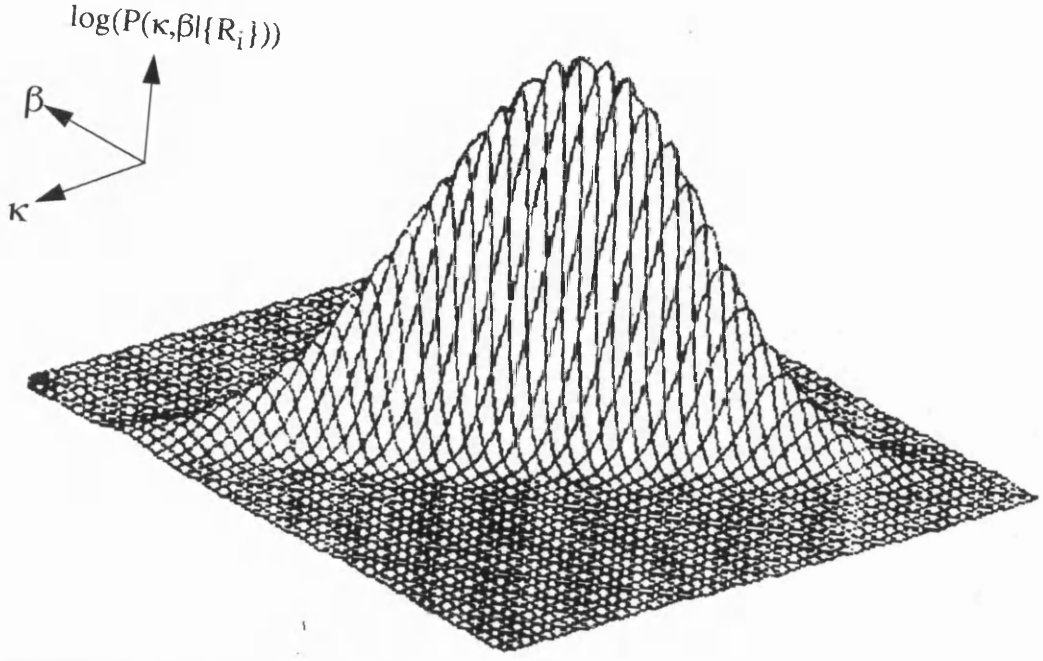
MAP 1B Diamantane produced using the full observed data set.
Range in β = 6.0 - 9.0 Range in κ = 0.95 - 1.45 Cut factor = 290

The above map is a close up of the maximum of MAP 1A. It shows that near the maximum, the probability surface has an elliptical cross section.

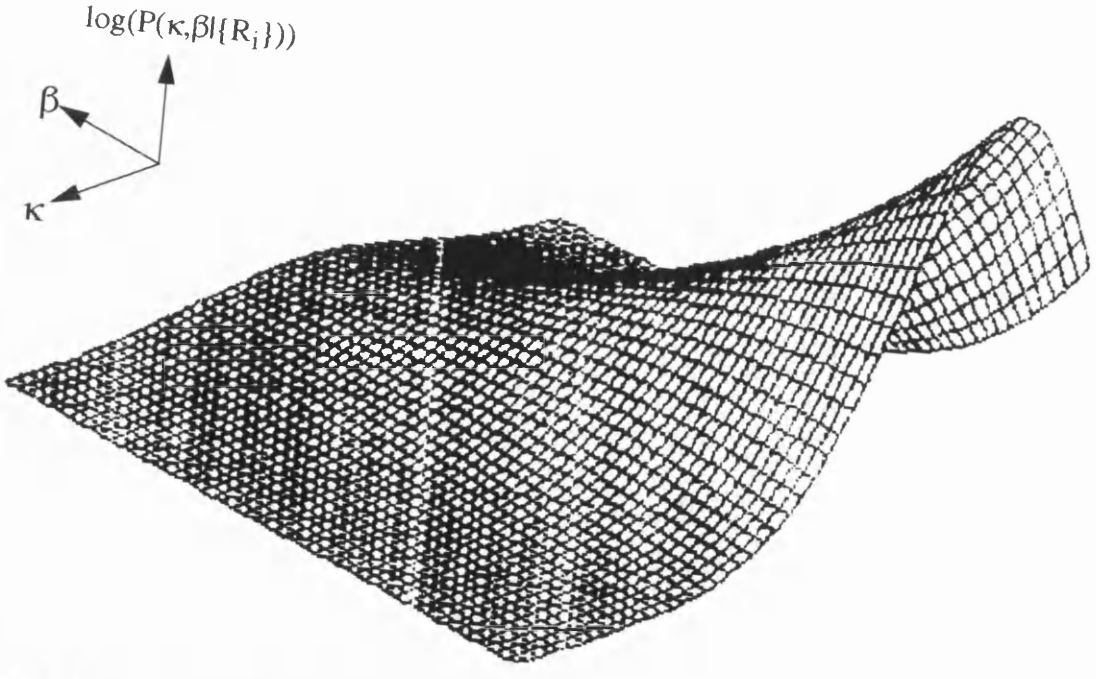


MAP 1C Diamantane produced using the full observed data set.
Range in β = -5.0 - 20.0 Range in κ = 0.1 - 12.0 Cut factor = 290

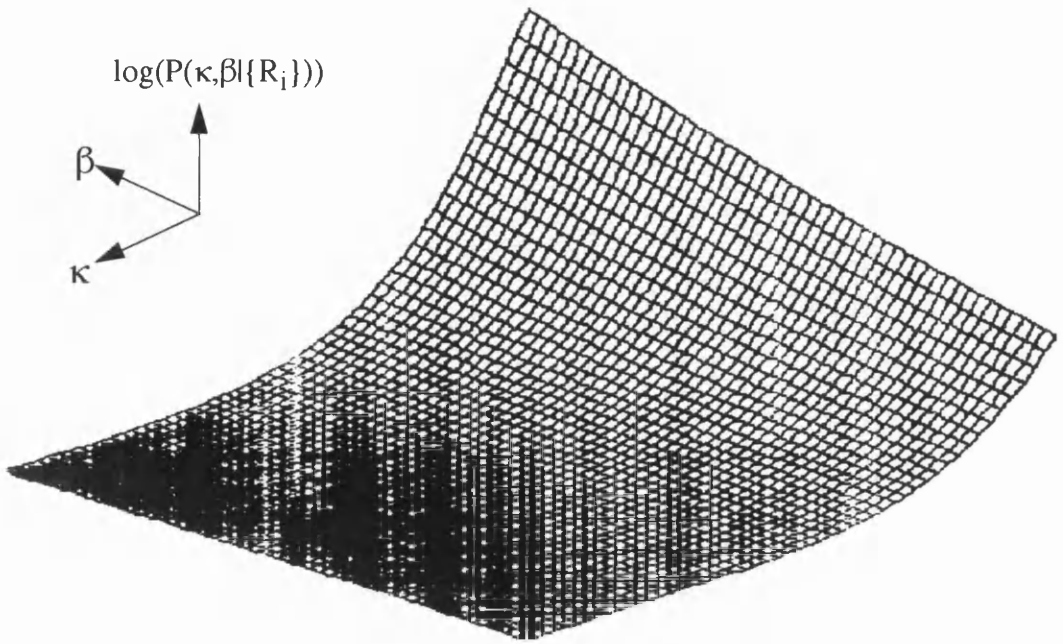
The above map shows not the log of the probability but the probability itself. Notice how sharp the peak is relative to MAP 1A, which is the log probability over the same range of β, κ.



MAP 2A Diamantane produced using the full observed data set.
Range in β = 6.0 - 9.0 Range in κ = 0.95 - 1.45 Cut factor = 290



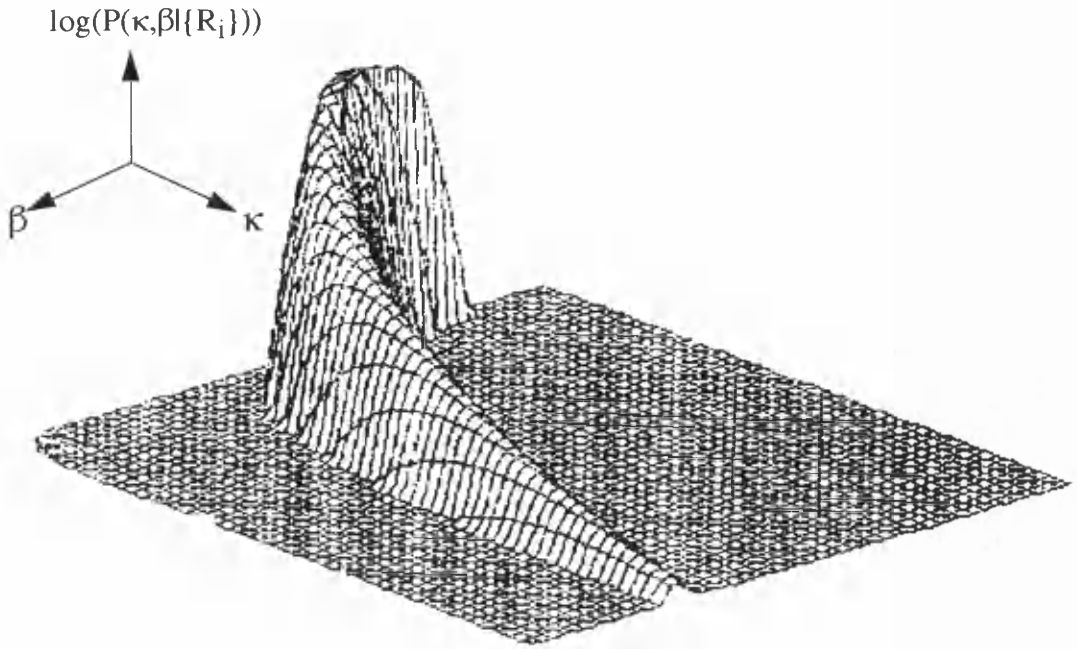
MAP 2B Diamantane produced using the 10% lowest angle reflections data set.
Range in β = 6.0 - 9.0 Range in κ = 0.95 - 1.45 Cut factor = 100



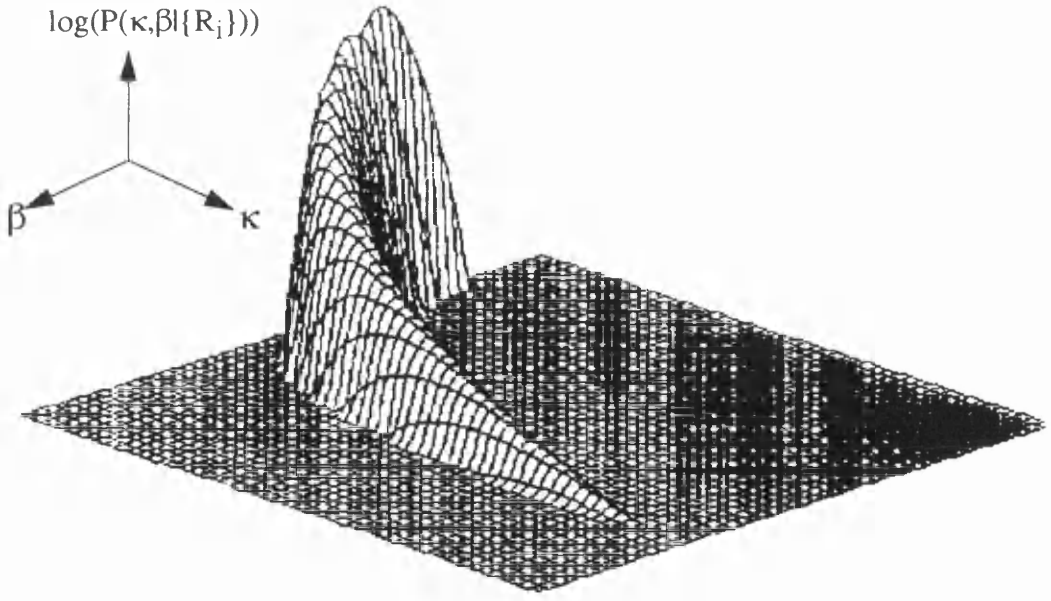
MAP 2C Diamantane produced using the 50 lowest angle reflections data set.
Range in β = 6.0 - 9.0 Range in κ = 0.95 - 1.45 Cut factor = 50

The above sequence of maps MAP2A-2C shows how the maximum probability varies over a fixed range of β and κ as we reduce the number of reflections in the data set, for a specific structure. It shows that the optimum value of β and κ vary with the changing data and that the sharpness of the peak reduces thus increasing the margins of error, for lower quality data.

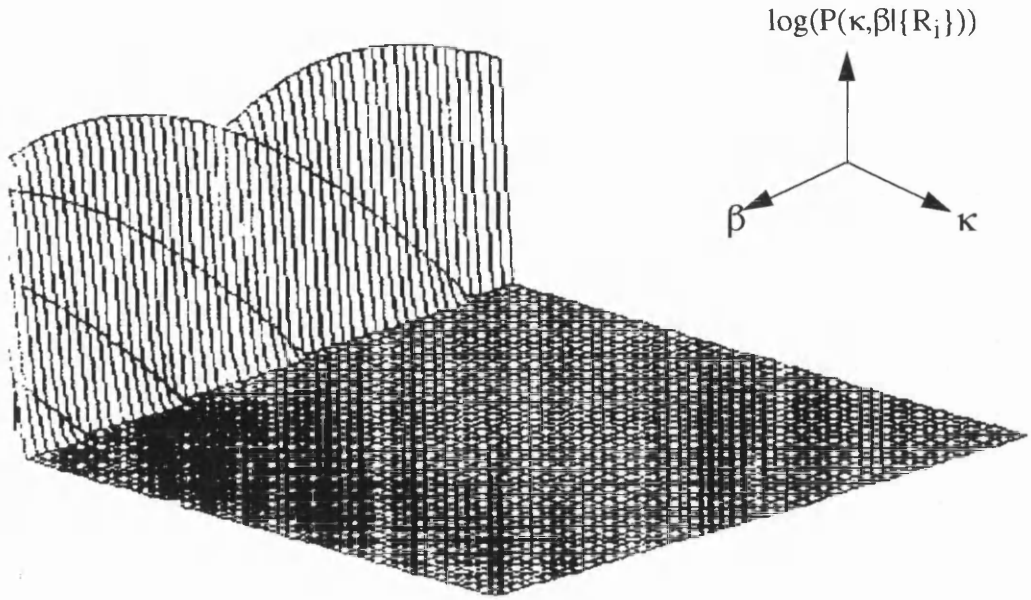
Note that it is necessary to reduce the cut factor to allow for the reduction in the magnitude of the log probability.



MAP 3A Goldman2 produced using a full data set of structure factors calculated from refined atomic positions.
Range in β = -5.0 - 25.0 Range in κ = 0.001 - 35.0 Cut factor = 100



MAP 3B Goldman2 produced using a full data set of observed structure factors.
Range in β = -5.0 - 25.0 Range in κ = 0.001 - 35.0 Cut factor = 100



MAP 3C Goldman2 produced using a data set of observed structure factors with a resolution below 2Å.
Range in β = -5.0 - 25.0 Range in κ = 0.001 - 35.0 Cut factor = 100

The above sequence of maps MAP3A-3C shows the change in the probability surface over a fixed range of β,κ and fixed cut factor with decreasing quality of data. There is minimal degradation in the quality of the probability map in moving from calculated structure factors to observed structure factors for this compound as the quality of the observed structure factors is high. There is a severe degradation in the map in going from all observed structure factors to those with a resolution of 2.0Å or less.

3.5 The Calculated Scale and Temperature Factors

In all tables of scale and temperature factor results, a figure in brackets beside the result indicates the estimated standard deviation (esd). All figures in the tables are given to two decimal places for the sake of clarity.

Any test entry, "no results obtained" indicates that the calculation became unstable and it was impossible to obtain numeric answers. This is caused by the eigenvalues going to zero, at which point the search for a maximum is switched to a grid search of the probability surface, however, if after 100 cycles of searching the grid a solution has not been found then the search is terminated and no values for β, κ are determined.

In all cases where no reference is given for the structure or it is listed as unpublished the reader is referred to the Sheldrick difficult structures database.

		Diamantane-4-ol (Rogers & Kennard, unpublished work)			
	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	1553	3.73 (.72)	3.73 (.17)	0.88 (.13)	0.83 (.06)
10% refs. random	155	3.29 (.50)	3.06 (.49)	1.10 (.11)	1.12 (.21)
50 refs. random	50	3.85 (1.40)	4.63 (.97)	0.72 (.22)	0.55 (.23)
10% refs. low res.	155	-8.71 (6.02)	-2.63 (2.63)	3.01 (.80)	1.55 (.28)
50 refs. low res.	50	36.01 (22.91)	15.87 (7.45)	0.45 (.21)	0.83 (.26)
2Å data	142	-8.30 (7.12)	no results obtained	2.78 (.83)	no results obtained

Quinol (Wallwork & Powell, 1980)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	2844	2.91 (.42)	2.95 (.11)	1.02 (.10)	0.98 (.05)
10% refs. random	284	2.13 (.28)	2.31 (.31)	1.56 (.10)	1.44 (.20)
50 refs. random	50	2.81 (.54)	2.79 (.71)	0.93 (.12)	0.96 (.33)
10% refs. low res.	284	4.65 (2.65)	3.58 (1.42)	0.94 (.12)	1.02 (.15)
50 refs. low res.	50	-52.94 (24.62)	no results obtained	5.22 (2.01)	no results obtained
2Å data	208	6.66 (2.26)	6.96 (2.30)	0.83 (.08)	0.82 (.15)

HOV1 (Hovestreydt, Klepp & Parthé, 1983)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	3502	1.24 (.26)	1.16 (.08)	0.47 (.03)	0.51 (.02)
10% refs. random	350	1.83 (.86)	1.82 (.28)	0.48 (.11)	0.45 (.06)
50 refs. random	50	-0.19 (.39)	-0.07 (.69)	0.72 (.08)	0.67 (.24)
10% refs. low res.	350	-3.19 (8.68)	-2.99 (1.45)	1.24 (.58)	1.04 (.16)
50 refs. low res.	50	no results obtained	no results obtained	no results obtained	no results obtained
2Å data	234	-14.66 (12.71)	-14.08 (2.67)	2.68 (1.41)	2.16 (.45)

Selendid (Clegg et al., 1980)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	2626	4.77 (.18)	4.82 (.12)	0.72 (.03)	0.71 (.04)
10% refs. random	262	4.73 (.21)	4.65 (.35)	0.67 (.03)	0.70 (.10)
50 refs. random	50	4.79 (.42)	4.83 (.76)	0.63 (.05)	0.65 (.20)
10% refs. low res.	262	1.01 (2.41)	-0.07 (2.45)	1.07 (.11)	1.10 (.23)
50 refs. low res.	50	12.13 (.91)	-0.11 (18.43)	0.64 (.01)	1.01 (.61)
2Å data	144	0.62 (2.66)	-0.91 (2.54)	1.10 (.12)	1.16 (.24)

Azet (Colens et al., 1974)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	1910	3.54 (.25)	3.55 (.17)	1.23 (.05)	1.23 (.07)
10% refs. random	191	2.87 (.74)	2.75 (.56)	1.48 (.18)	1.57 (.28)
50 refs. random	50	3.10 (.92)	1.69 (1.21)	1.14 (.17)	1.88 (.71)
10% refs. low res.	191	9.50 (4.02)	7.05 (2.82)	0.75 (.10)	0.91 (.17)
50 refs. low res.	50	-33.72 (11.96)	no results obtained	2.15 (.34)	no results obtained
2Å data	262	8.05 (1.69)	6.09 (1.87)	0.80 (.06)	0.95 (.15)

TUR10 (Braekman et al., 1981)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	1310	4.31 (.42)	4.34 (.17)	1.38 (.12)	1.30 (.09)
10% refs. random	131	4.74 (.77)	5.23 (.56)	0.92 (.15)	0.82 (.19)
50 refs. random	50	4.38 (.75)	5.24 (.73)	1.53 (.26)	1.17 (.36)
10% refs. low res.	131	-2.78 (8.93)	4.57 (2.49)	2.47 (.83)	1.23 (.23)
50 refs. low res.	50	43.14 (10.81)	34.37 (9.86)	0.42 (.81)	0.53 (.18)
2Å data	150	1.59 (5.96)	6.87 (2.26)	1.90 (.47)	1.10 (.20)

Dodecane-Diol

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	786	2.69 (.39)	2.63 (.16)	1.60 (.14)	1.60 (.12)
10% refs. random	78	3.22 (.84)	2.87 (.49)	1.28 (.25)	1.54 (.34)
50 refs. random	50	2.08 (.79)	2.13 (.67)	2.27 (.41)	2.11 (.64)
10% refs. low res.	78	2.59 (2.66)	3.99 (3.95)	1.39 (.16)	1.28 (.45)
50 refs. low res.	50	2.92 (3.33)	3.08 (6.97)	1.30 (.15)	1.38 (.64)
2Å data	75	1.41 (3.06)	3.55 (4.14)	1.52 (.20)	1.32 (.46)

APAPA (Suck, Manor & Saenger, 1976)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	3242	2.96 (.41)	3.07 (.12)	0.96 (.07)	0.92 (.04)
10% refs. random	324	3.56 (.90)	3.66 (.38)	0.76 (.13)	0.71 (.10)
50 refs. random	50	2.11 (2.18)	2.06 (1.11)	0.80 (.34)	0.84 (.34)
10% refs. low res.	324	5.68 (3.27)	5.08 (1.98)	0.85 (.10)	0.83 (.11)
50 refs. low res.	50	-29.55(34.03)	-60.67(22.60)	2.10 (.66)	3.67 (1.42)
2Å data	385	4.02 (2.34)	3.19 (1.54)	0.91 (.09)	0.91 (.11)

MGHEX (Karle & Karle, 1981)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	4595	3.66 (.25)	3.75 (.10)	1.52 (.07)	1.46 (.05)
10% refs. random	459	3.71 (.43)	3.80 (.31)	1.50 (.12)	1.44 (.17)
50 refs. random	50	6.21 (.31)	6.16 (.89)	0.59 (.03)	0.63 (.21)
10% refs. low res.	459	9.19 (4.08)	9.16 (1.42)	0.88 (.14)	0.86 (.10)
50 refs. low res.	50	-45.72 (42.38)	-25.98 (18.61)	2.77 (1.04)	1.82 (.61)
2Å data	447	9.11 (3.81)	9.13 (1.39)	0.88 (.14)	0.86 (.10)

TOTC (Williams & Lawton, 1975)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	3328	5.30 (.36)	5.38 (.09)	1.15 (.11)	1.10 (.05)
10% refs. random	332	4.86 (.43)	5.00 (.28)	1.33 (.14)	1.27 (.18)
50 refs. random	50	5.05 (.75)	4.83 (.65)	1.15 (.21)	1.29 (.41)
10% refs. low res.	332	7.38 (2.45)	8.09 (1.24)	0.90 (.11)	0.82 (.11)
50 refs. low res.	50	12.02 (7.55)	13.97 (12.34)	0.71 (.08)	0.62 (.22)
2Å data	233	9.37 (3.54)	10.17 (1.86)	0.80 (.12)	0.74 (.11)

TPH (Hoekstra, Vos, Braun & Hornstra, 1975)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	4758	2.17 (.21)	2.23 (.05)	1.47 (.11)	1.38 (.05)
10% refs. random	475	1.96 (.31)	2.01 (.15)	1.62 (.18)	1.55 (.16)
50 refs. random	50	2.00 (.37)	2.41 (.48)	1.71 (.24)	1.23 (.44)
10% refs. low res.	475	5.97 (1.65)	6.49 (.81)	0.86 (.10)	0.81 (.09)
50 refs. low res.	50	-37.23 (41.45)	no results obtained	2.53 (1.46)	no results obtained
2Å data	226	9.22 (3.35)	8.17 (1.86)	0.72 (.10)	0.77 (.12)

Goldman2 (Irngartinger Reibel & Sheldrick, 1981)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	3891	2.70 (.32)	2.77 (.07)	1.35 (.12)	1.28 (.05)
10% refs. random	389	2.48 (.34)	2.46 (.23)	1.66 (.15)	1.65 (.21)
50 refs. random	50	3.75 (1.33)	4.00 (.74)	0.52 (.19)	0.47 (.20)
10% refs. low res.	389	2.72 (3.63)	3.59 (1.04)	1.22 (.26)	1.09 (.13)
50 refs. low res.	50	-25.34 (13.03)	-16.10 (13.43)	1.90 (.38)	1.48 (.60)
2Å data	228	9.22 (6.17)	7.83 (1.91)	0.79 (.20)	0.85 (.14)

Munich1 (Szeimies-Seebach et al., 1978)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	2240	2.93 (.57)	2.99 (.13)	1.74 (.20)	1.71 (.09)
10% refs. random	224	2.82 (.90)	3.11 (.39)	1.82 (.34)	1.62 (.26)
50 refs. random	50	4.46 (.72)	4.49 (.87)	0.78 (.11)	0.83 (.28)
10% refs. low res.	224	8.22 (6.19)	8.22 (1.78)	1.13 (.29)	1.19 (.18)
50 refs. low res.	50	-4.30 (24.10)	-7.40 (15.29)	1.12 (.38)	1.33 (.56)
2Å data	220	8.56 (6.09)	8.35 (1.81)	1.10 (.28)	1.18 (.18)

MBH2 (Poyser et al., 1986)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	3764	4.31 (.40)	4.33 (.10)	0.91 (.08)	0.89 (.04)
10% refs. random	376	4.04 (.75)	4.07 (.32)	0.98 (.16)	0.95 (.13)
50 refs. random	50	3.99 (1.07)	3.85 (.82)	0.89 (.24)	0.94 (.38)
10% refs. low res.	376	1.71 (1.86)	2.74 (1.32)	1.29 (.12)	1.15 (.15)
50 refs. low res.	50	39.79 (11.37)	28.62 (15.07)	0.46 (.07)	0.60 (.23)
2Å data	294	0.79 (2.00)	1.39 (1.68)	1.33 (.11)	1.26 (.18)

SUOA (Oliver & Strickland, 1984)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	3142	2.70 (.35)	2.75 (.10)	0.92 (.08)	0.87 (.04)
10% refs. random	314	2.40 (.78)	2.67 (.34)	1.03 (.19)	0.91 (.14)
50 refs. random	50	3.14 (2.02)	3.42 (.85)	0.82 (.40)	0.68 (.27)
10% refs. low res.	314	1.81 (3.67)	1.76 (1.39)	1.03 (.18)	0.99 (.13)
50 refs. low res.	50	-46.61 (12.76)	no results obtained	2.62 (.45)	no results obtained
2Å data	253	4.49 (4.66)	3.68 (1.76)	0.88 (.17)	0.89 (.13)

Winter2 (Butters et al., 1981)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	6980	5.69 (.36)	5.68 (.16)	0.83 (.07)	0.82 (.04)
10% refs. random	698	6.38 (.54)	6.45 (.22)	1.28 (.16)	1.24 (.12)
50 refs. random	50	6.14 (1.97)	6.56 (.88)	1.34 (.64)	1.27 (.52)
10% refs. low res.	698	8.44 (.80)	8.46 (.92)	1.09 (.05)	1.10 (.10)
50 refs. low res.	50	-41.79 (34.05)	-51.11 (19.04)	3.04 (.90)	3.47 (1.15)
2Å data	491	7.88 (2.40)	7.90 (1.36)	1.15 (.11)	1.15 (.13)

Loganin (Jones, Sheldrick, Glüsenkamp & Tietze, 1980)

	Number of Reflections	Wilson Plot Temp. Factor	Bayesian Temp. Factor	Wilson Plot Scale Factor	Bayesian Scale Factor
All data	3498	2.74 (.28)	2.82 (.18)	0.96 (.08)	0.90 (.07)
10% refs. random	349	3.34 (.41)	3.22 (.24)	0.68 (.08)	0.73 (.10)
50 refs. random	50	3.57 (.74)	3.91 (.77)	0.62 (.13)	0.50 (.21)
10% refs. low res.	349	0.47 (2.21)	2.19 (1.40)	1.22 (.15)	0.99 (.15)
50 refs. low res.	50	-1.96 (7.41)	no results obtained	0.90 (.11)	no results obtained
2Å data	224	0.75 (5.27)	2.64 (2.23)	1.19 (.26)	0.96 (.17)

The tables of results on the previous pages show several trends:

(i) The two normalisation techniques both go wrong "in the same direction" i.e they both become negative, or unrealistically large. This is not surprising as the two techniques are mathematically equivalent.

(ii) The standard deviations are generally smaller for the full data sets for the Bayesian technique, and this lowering of $\sigma(K)$ and $\sigma(B)$ leads in turn to a reduction in $\sigma|E|$ and $\sigma|U|$. The values of esds calculated using the Bayesian method are more realistic than those calculated using the methods of Hall and Subramanian (Hall & Subramanian, 1982b).

(iii) The magnitude of the esd on the temperature and scale factors increases with decreasing numbers of reflections and also with decreasing resolution of reflections. This is to be expected from the mathematics and from the plots of the log probability surfaces shown in Section 3.4.

(iv) For the large majority of data sets used in the test the result obtained by the Bayesian technique was within three Bayesian standard deviations of the result obtained by the Wilson plot method, when comparing the results for the same data set.

3.6 Summary

We have shown that the Bayesian normalisation method is perfectly adequate for the normalisation of X-ray diffraction data. The technique is not as robust as the commonly used Wilson plot with very sparse data, but it works well for full data sets. It is no worse than the Wilson plot for full structure solution and can provide a useful alternative to the Wilson plot. As an alternative to the Wilson plot it may be used to give a new phasing path, through which a solution to the phase problem may be discovered. This idea of applying many different techniques to difficult structures was expressed best by George Sheldrick "If you try to solve a crystal structure many different ways, one of them will probably work. Why it works and the others fail may not be obvious". This use of many different techniques in structure elucidation is one of the most important tools available to the crystallographer and this new technique provides us with another method of normalisation with which to approach a data set.

The computer code to perform the calculation is no slower than the calculation of a Wilson plot and would be easily implemented into existing normalisation packages, just as it was implemented into MITHRIL90.

The new error calculation provides smaller and more reliable esd values on the E-magnitudes than the method devised by Hall and Subramanian, which may be advantageous in the apriori phasing of a structure using maximum entropy techniques, which relies on the esd estimates of the U-magnitudes.

4.0 FUTURE WORK

As was stated in Section 2.5 the theory used to derive the formula 2.5.6 is very easily extended by substitution of a new prior set of knowledge. At the moment we state that we know nothing at all about the scale and temperature factors while in reality we do have some prior constraints, i.e. a negative temperature factor has no physical meaning and so should be forbidden by our prior, similarly a temperature factor over 20 must be considered highly suspicious and should also be forbidden. A function that gives a graph as shown in Figure 4 for values of B could be used as a prior. Prior information is not easily used with conventional methods of normalisation.

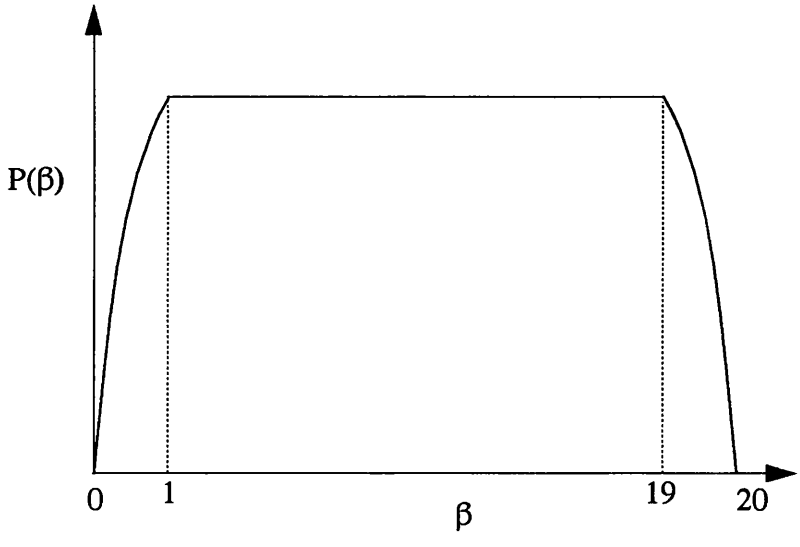


FIGURE 4. A plot of the probability of β varying with the value of β for use as a prior.

If a function as shown above was used as our prior to multiply the Wilson statistics that we currently use, then all the results that give a negative temperature factor or a temperature in excess of 20 would be forbidden, while all those with a value within the range 1-19 would remain unaltered. Those in the ranges 0-1 and 19-20 would be deemed "unlikely" and it would take a very strong indication from the data to produce a temperature factor in one of these ranges, although it would be possible.

The theory can be altered to produce a single overall anisotropic temperature factor and an anisotropic scale factor. The theory for this has been developed (R.K. Henderson, private communication) but has not yet been programmed. The difficulty in producing figures from the data then becomes one of keeping the search on the probability surface stable and new algorithms for this would have to be studied.

This new technique could be used for the normalisation of protein structures by using a multichannel formalism to produce an independent temperature factor for both the solvent and the protein parts of the data. The multichannel technique would be easily coded by using a multi dimensional search of the probability map, this would be impossible if using the Wilson plot technique alone. Using two temperature factors would be of great help in the phase refinement using the maximum entropy method for proteins (Xiang, Carter, Bricogne & Gilmore, 1993).

References

Bayes, T. (1763). *Phil. Trans. Roy. Soc.* **53**, 370-418

Braekman, J.C., Daloze, D., Dupont, A., Tursch, B., Declercq, J.P., Germain, G. & Van Meerssche, M. (1981). *Tetrahedron* **37**, 179-186

Butland, J. *SIMPLEPLOT User's Handbook*, Report No. 253, University of Bradford

Butters, T., Hütter, P., Jung, G., Pauls, N., Schmitt, H., Sheldrick, G.M. & Winter, W. (1981). *Angew. Chem. Int. Ed. Engl.* **20**, 889-890

Clegg, W., Harms, K., Sheldrick, G.M., von Kiedrowski, G. & Tietze, L.F. (1980). *Acta. Cryst.* **B36**, 3159-3162

Colens, A., Declercq, J.P., Germain, G., Putzeys, J.P. & Van Meerssche, M. (1974). *Cryst. Struct. Comm.* **3**, 119-122

French, S. & Wilson, K. (1978). *Acta. Cryst.* **A34**, 517-525

Germain, G., Main, P. & Woolfson, M.M. (1970). *Acta. Cryst.* **B26**, 274-285

Gilmore, C.J. (1984). *J. Appl. Cryst.* **17**, 42-46

Gilmore, C.J. & Brown, S.R. (1988). *J. Appl. Cryst.* **22**, 571-572

Hall, S.R. & Subramanian, V. (1982a). *Acta. Cryst.* **A38**, 590-598

Hall, S.R. & Subramanian, V. (1982b). *Acta. Cryst.* **A38**, 598-608

Hauptman, H. & Karle, J. (1953). *Acta. Cryst.* **6**, 136-141

Hoekstra, A., Vos, A., Braun, P.B. & Hornstra, J. (1975). *Acta. Cryst.* **B31**, 1708-1715

Hovestreydt, E., Klepp, K. & Parthé, E. (1983). *Acta. Cryst.* **C39**, 422-425

Irngartinger, H., Reibel, W.R.K. & Sheldrick, G.M. (1981). *Acta. Cryst.* **B37**, 1768-1771

Jones, P.G., Sheldrick, G.M., Glüsenkamp, K.H. & Tietze, L.F. (1980). *Acta. Cryst.* **B36**, 481-483

Karle, I.L. & Karle, J. (1981). *Proc. Natl. Acad. Sci. USA* **78**, 681-685

Karle, J. & Hauptman, H. (1953). *Acta. Cryst.* **6**, 473-476

Nielsen, K. (1975). *Acta. Cryst.* **A31**, 762-763

Oliver, J.D. & Strickland, L.C. (1984). *Acta. Cryst.* **C40**, 820-824

Poyser, J.P., Edwards, P.L., Anderson, J.R., Hursthouse, M.B., Walker, N.P.C., Sheldrick, G.M. & Walley, A.J.S. (1986). *J. Antibiotics* **39**, 167-169

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1992). *Numerical Recipes* pp 462-475, Cambridge University Press, Cambridge

Rogers, D. (1965). *Computing Methods in Crystallography* Edited by Rollett, J.S. pp 133-148, Pergamon Press, London

Shmueli, U., Rabinovich, S. & Weiss, G.H. (1990). *Acta. Cryst.* **A46**, 241-246

Sheldrick, G.M. (1985). *Crystallographic Computing 3*, Edited by Sheldrick, G.M., Kruger, C. & Goddard, R. pp 184-189, Clarendon Press, Oxford

Subramanian, V. & Hall, S.R. (1982). *Acta. Cryst.* **A38**, 577-590

Suck, D., Manor, P.C. & Saenger, W. (1976). *Acta. Cryst.* **B32**, 1727-1737

Szeimies-Seebach, U., Harnisch, J., Szeimies, G., Van Meerssche, M., Germain, G. & Declercq, J.P. (1978). *Angew. Chem. Int. Ed. Engl.* **17**, 848-850

Wallwork, S.C. & Powell, H.M. (1980). *J. Chem. Soc. Perkin* **2**, 641-646

Williams, D.J. & Lawton, D. (1975). *Tetrahedron Letters* **No. 2**, 111-114

Wilson, A.J.C. (1942). *Nature* **150**, 151-152

Xiang, S., Carter, C.W., Bricogne, G. & Gilmore, C.J. (1993). *Acta. Cryst.* **D49**, 193-212

CHAPTER 4

A LIKELIHOOD FIGURE OF MERIT FOR CONVENTIONAL DIRECT METHODS

1.0 INTRODUCTION

1.1 The Importance of Figures of Merit

Before the structure solution of a large or complex molecule can begin using the multi-solution technique, we need to know the phases of a large number of reflections. This is achieved by use of a large number of permuted reflections in the starting set; generally using magic integer sequences (Main, 1977) to reduce the number of phase sets produced (see Chapter 1 Section 10.2). The multi-solution technique thus produces a large number of phase sets from which the correct solution must be determined. It would be very time consuming to compute and examine the electron density maps for each phase set to determine which one contains structurally relevant information.

Figures of Merit (FOMs) are functions based on quantities which can be expected to have extreme values for a correct phase set, or to determine internal consistency of some aspect of the direct methods procedure. These FOMs are then used as part of a screening procedure to eliminate un-promising phase sets before the calculation of the E-maps.

In this chapter a new figure of merit based on triplet consistency will be developed, tested and examined for correlations with other conventional FOMs.

2.0 THEORY

2.1 The Sayre Equation

The Sayre equation (Sayre, 1952) is given in Chapter 1 EQ 2.0.1. The Sayre equation can be rewritten in terms of the normalised structure factors as shown in EQ 2.1.1.

$$E_{\underline{h}} = \frac{f_{\underline{h}}}{\gamma_{\underline{h}} V} \sum_{\underline{k}} E_{\underline{k}} E_{\underline{h}-\underline{k}} \qquad \text{(EQ 2.1.1)}$$

Where

$f_{\underline{h}}$ is the scattering factor for each atom

$\gamma_{\underline{h}}$ is the Fourier transform of the squared electron density

V is the volume of the unit cell

The above equation allows the mathematical relationship of structure factors, and will allow the determination of $E_{\underline{h}}$ providing the E-magnitude and phase are known for reflections $\underline{k}$ and $\underline{h}$-$\underline{k}$. This is normally used in the determination of the phase of reflection $\underline{h}$ from the other two reflections that form a triplet.

The above equation has been used to solve simple centrosymmetric structures (Karle & Hauptman, 1953), and also for phase extension in the solution of proteins.

2.2 Common Figures of Merit

There are a great many FOMs in use in the various crystallographic packages that are available. These FOMs can be broken down into four main types, all of which are measurements of invariants:

1. Those that measure triplet consistency

2. Those that measure quartet consistency

3. Those that are based on special triplets

4. Those based on one or two-phase structure invariants

All types of FOM have properties to commend them, yet none is good enough to be used in isolation of the other FOMs. The first type of FOM is that which measure the triplet consistency and are good at indicating correct atomic arrangement. Examples of this type of FOM include:

ABSFOM and R_α (both of which are discussed in Chapter 1 sections 12.1 and 12.2 respectively).

R_{Karle} (Karle & Karle, 1966) is a residual of observed and calculated E-magnitudes for phased reflections.

$$R_{Karle} = \frac{\sum_{\underline{h}} \left| |E_{\underline{h}}^{obs}| - |E_{\underline{h}}^{calc}| \right|}{\sum_{\underline{h}} |E_{\underline{h}}^{obs}|} \qquad \text{(EQ 2.2.1)}$$

The second type are those that determine quartet consistency; examples of these include:

NQEST, which is discussed in Chapter 1 Section 12.4.

NQUAL (Sheldrick, 1990) which is a measure of triplet and negative quartet consistency

$$NQUAL = \frac{\sum_{\underline{h}} |\alpha \cdot \eta|}{|\alpha||\eta|} \qquad \text{(EQ 2.2.2)}$$

Where

$$\alpha = \frac{2|E_{\underline{h}}|E_{\underline{k}}E_{\underline{h}-\underline{k}}}{\sqrt{N}}$$

$$\eta = \frac{g|E_{\underline{h}}|E_{\underline{k}}E_{\underline{l}}E_{-\underline{h}-\underline{k}-\underline{l}}}{N}$$

g is a positive constant based on the values of the negative quartet cross terms

NQUAL has proved to be a better discriminator of phase sets for larger structures than NQEST. Other FOMs have been developed based on special quartets, and include:

HKC Harker-Kasper Criterion and P1C (Schenk, 1973)

$$P1C = \sum_{\underline{h}}\sum_{\underline{k}} \frac{(|U_{\underline{h}}| - |U_{\underline{k}}|)^2}{(1-|U_{\underline{h}+\underline{k}}|)\,(1-|U_{\underline{h}-\underline{k}}|)} |\pi - (\varphi_{\underline{h}+\underline{k}} + \varphi_{\underline{h}-\underline{k}} - 2\varphi_{\underline{h}})| \qquad \text{(EQ 2.2.3)}$$

Where

$U_{\underline{h}}$ is the unitary structure factor

$0 \le \varphi_{\underline{h}+\underline{k}} + \varphi_{\underline{h}-\underline{k}} - 2\varphi_{\underline{h}} \le 2\pi$

P1C is the non-centrosymmetric version of HKC and the two are equivalent in centrosymmetric space groups. Both HKC and P1C are particularly discriminative for symmorphic space groups (no glide planes or screw axes).

The Negative Quartet Criterion (NQC) (Schenk, 1974) is highly correlated to ψ_o. This is due to ψ_o using small $|E|$ reflections in the Sayre equation EQ 2.1.1 and that the negative quartets are constructed from two triplets that are composed of one weak reflection and two strong reflections. NQC is defined as

$$NQC = \sum_{\underline{h}}\sum_{\underline{k}}\sum_{\underline{l}} \frac{2}{N} j\,(W_j - m_j)\,|E_{\underline{h}}E_{\underline{k}}E_{\underline{l}}E_{-\underline{h}-\underline{k}-\underline{l}}|\,|\pi - (\varphi_{\underline{h}} + \varphi_{\underline{k}} + \varphi_{\underline{k}} + \varphi_{-\underline{h}-\underline{k}-\underline{l}})|$$

(EQ 2.2.4)

Where

j is the number of quartet cross terms present in the observed data set

W_j is a weight, dependent on the number of quartet cross terms present and must lie in the range $0<W_j<1$, and are usually set to the values $W_1 = 0$, $W_2 = 0.5$, $W_3 = 0.9$

$$m_1 = |E_{\underline{h}+\underline{k}}|$$

$$m_2 = (|E_{\underline{h}+\underline{k}}| + |E_{\underline{h}+\underline{l}}|)/2$$

$$m_3 = (|E_{\underline{h}+\underline{k}}| + |E_{\underline{h}+\underline{l}}| + |E_{\underline{k}+\underline{l}}|)/3$$

$$0 \le \varphi_{\underline{h}} + \varphi_{\underline{k}} + \varphi_{\underline{k}} + \varphi_{-\underline{h}-\underline{k}-\underline{l}} \le 2\pi$$

The equation must also obey the constraint $(W_j - m_j) \ge 0$ to ensure only positive values of NQC are produced.

It must be noted that negative quartets are very sensitive to correctly measured weak reflections. If a small number of strong reflections are incorrectly measured as being weak then this will be very detrimental to the quality of phase information from negative quartets and to the reliability of NQC.

The third type of FOMs are those that involve special triplets and which are good for determination of structural placement within the unit cell. The best example of this is the ψ_o FOM, discussed in Chapter 1 Section 12.3. This FOM is highly correlated with the NQC figure of merit.

The fourth type of FOM is those that measure the consistency of structure seminvariants. Examples of these include consistency of Σ_1 relationships and two phase structure invariants.

2.3 Calculation of E-magnitudes

As stated earlier we intend to calculate values for $|E_{\underline{h}}|$ for unphased reflections using two reflections that have been phased and refined by some process, usually tangent refinement (Karle & Hauptman, 1956). Sayres equation EQ 2.1.1 can be used to determine the phase of a reflection from the phases of two known reflections that are related by a three phase invariant. The Sayre equation may be rewritten in its equivalent form as EQ 2.3.1 (Hughes, 1953).

$$E_{\underline{h}} = \frac{1}{\sqrt{N}} \langle E_{\underline{k}} E_{\underline{h}-\underline{k}} \rangle_{\underline{k}} \qquad \text{(EQ 2.3.1)}$$

Where

N is the number of atoms in the unit cell, assumed equal

$\langle \ldots \rangle_{\underline{k}}$ means the average over all values of $\underline{k}$

For unequal atom structures we must substitute $\frac{\sigma_2^{3/2}}{\sigma_3}$ for $\sqrt{N}$, where $\sigma_n = \sum_{j=1}^{N} Z_j^n$

We may now examine the real and imaginary parts of EQ 2.3.1

$$|E_{\underline{h}}| \cos\varphi_{\underline{h}} \approx \frac{\sigma_2^{3/2}}{\sigma_3} \sum_{\underline{k}} |E_{\underline{k}} E_{\underline{h}-\underline{k}}| \cos(\varphi_{\underline{k}} + \varphi_{\underline{h}-\underline{k}} + b) \qquad \text{(EQ 2.3.2)}$$

$$|E_{\underline{h}}| \sin\varphi_{\underline{h}} \approx \frac{\sigma_2^{3/2}}{\sigma_3} \sum_{\underline{k}} |E_{\underline{k}} E_{\underline{h}-\underline{k}}| \sin(\varphi_{\underline{k}} + \varphi_{\underline{h}-\underline{k}} + b) \qquad \text{(EQ 2.3.3)}$$

Where b is the phase shift that occurrs in moving origins

By squaring and adding EQ 2.3.2 and EQ 2.3.3 we can obtain EQ 2.3.4

$$|E_{\underline{h}}|^2 (sin^2\varphi_{\underline{h}} + cos^2\varphi_{\underline{h}}) \approx \left(\frac{\sigma_2^{3/2}}{\sigma_3}\right)^2 \left(\sum_{\underline{k}} |E_{\underline{k}} E_{\underline{h}-\underline{k}}| \sin(\varphi_{\underline{k}} + \varphi_{\underline{h}-\underline{k}} + b)\right)^2 +$$

$$\left(\frac{\sigma_2^{3/2}}{\sigma_3}\right)^2 \left(\sum_{\underline{k}} |E_{\underline{k}} E_{\underline{h}-\underline{k}}| \cos(\varphi_{\underline{k}} + \varphi_{\underline{h}-\underline{k}} + b)\right)^2 \qquad \text{(EQ 2.3.4)}$$

By making the simplification that $sin^2\varphi_{\underline{h}} + cos^2\varphi_{\underline{h}} = 1$ and then taking the square root of EQ 2.3.4 we obtain

$$|E_{\underline{h}}^{calc}| \approx \frac{\sigma_2^{3/2}}{\sigma_3} \sqrt{\left(\sum_{\underline{k}} |E_{\underline{k}} E_{\underline{h}-\underline{k}}| \cos(\varphi_{\underline{k}} + \varphi_{\underline{h}-\underline{k}} + b)\right)^2 + \left(\sum_{\underline{k}} |E_{\underline{k}} E_{\underline{h}-\underline{k}}| \sin(\varphi_{\underline{k}} + \varphi_{\underline{h}-\underline{k}} + b)\right)^2}$$

(EQ 2.3.5)

To increase the probability of EQ 2.3.5 being true we only use the largest phased E-magnitude reflections for $\underline{k}$ and $\underline{h}$-$\underline{k}$. The above equation allows us to make an estimate for $|E_{\underline{h}}|^{calc}$. This can then be compared with $|E_{\underline{h}}|^{obs}$ to produce a figure of merit, this is already done in the calculation of R_{Karle} but we will be making comparisons only with reflections that have not been a part of the phasing process.

Before any comparison is performed a scale factor, κ, required to place both the observed and calculated $|E_{\underline{h}}|$ must be determined. This is done using equation 2.3.6

$$\kappa = \frac{\sum_{\underline{h}} |E_{\underline{h}}^{obs}|}{\sum_{\underline{h}} |E_{\underline{h}}^{calc}|} \qquad \text{(EQ 2.3.6)}$$

2.4 Derivation of LOGLIK

As was discussed in Chapter 1 Section 15.8 likelihood is the best possible discriminator that can be used to make decisions. The log likelihood gain (LLG) is defined as:

$$LLG = \log\left(\sum_{\underline{h}} \Lambda^{centric}\right) + \log\left(\sum_{\underline{h}} \Lambda^{acentric}\right) + \log\left(\sum_{\underline{h}} \Lambda^{centric}_{null}\right) + \log\left(\sum_{\underline{h}} \Lambda^{acentric}_{null}\right)$$

(EQ 2.4.1)

Where

$\Lambda^{centric}$ is the likelihood of a centric reflection contributing to the FOM

$\Lambda^{acentric}$ is the likelihood of an acentric reflection contributing to the FOM

$\Lambda^{centric}_{null}$ is the centric null hypothesis for a single reflection

$\Lambda^{acentric}_{null}$ is the acentric null hypothesis for a single reflection

For LLG to indicate a correct phase set the value should always be a large as possible.

For the diagonal approximation the following equations apply (Bricogne & Gilmore, 1990) for acentric and centric reflections respectively.

$$\Lambda^{acentric} = \frac{2N}{\varepsilon_{\underline{h}}}\left|U^{obs}_{\underline{h}}\right| e^{\frac{-N}{\varepsilon_{\underline{h}}}\left(\left|U^{obs}_{\underline{h}}\right|^2 + \left|U^{calc}_{\underline{h}}\right|^2\right)} I_o\left[\frac{2N}{\varepsilon_{\underline{h}}}\left|U^{obs}_{\underline{h}}\right|\left|U^{calc}_{\underline{h}}\right|\right] \qquad \text{(EQ 2.4.2)}$$

$$\Lambda^{centric} = \sqrt{\frac{2N}{\pi\varepsilon_{\underline{h}}}} e^{\frac{-N}{2\varepsilon_{\underline{h}}}\left(\left|U^{obs}_{\underline{h}}\right|^2 + \left|U^{calc}_{\underline{h}}\right|^2\right)} \cosh\left[\frac{N}{\varepsilon_{\underline{h}}}\left|U^{obs}_{\underline{h}}\right|\left|U^{calc}_{\underline{h}}\right|\right] \qquad \text{(EQ 2.4.3)}$$

By substituting EQ 2.4.4 into EQs 2.4.2, 2.4.3 and rearranging we obtain equations, 2.4.5 and 2.4.6 respectively.

$$\left|U_{\underline{h}}\right| \approx \frac{\left|E_{\underline{h}}\right|}{\sqrt{N}} \qquad \text{(EQ 2.4.4)}$$

$$\Lambda^{acentric} \approx \frac{2\sqrt{N}}{\varepsilon_{\underline{h}}}\left|E^{obs}_{\underline{h}}\right| e^{\frac{-1}{\varepsilon_{\underline{h}}}\left(\left|E^{obs}_{\underline{h}}\right|^2 + \left|E^{calc}_{\underline{h}}\right|^2\right)} I_o\left[\frac{2}{\varepsilon_{\underline{h}}}\left|E^{obs}_{\underline{h}}\right|\left|E^{calc}_{\underline{h}}\right|\right] \qquad \text{(EQ 2.4.5)}$$

$$\Lambda^{centric} \approx \sqrt{\frac{2N}{\pi\varepsilon_{\underline{h}}}} e^{\frac{-1}{2\varepsilon_{\underline{h}}}\left(\left|E^{obs}_{\underline{h}}\right|^2 + \left|E^{calc}_{\underline{h}}\right|^2\right)} \cosh\left[\frac{1}{\varepsilon_{\underline{h}}}\left|E^{obs}_{\underline{h}}\right|\left|E^{calc}_{\underline{h}}\right|\right] \qquad \text{(EQ 2.4.6)}$$

Traditionally the null hypothesis assumes that our calculated values are zero i.e. $\left|E^{calc}_{\underline{h}}\right| = 0$, and this is substituted into EQs 2.4.5 and 2.4.6 to yield $\Lambda^{acentric}_{null}$ and $\Lambda^{centric}_{null}$ respectively.

$$\Lambda_{null}^{acentric} \approx \frac{2\sqrt{N}}{\varepsilon_{\underline{h}}} \left| E_{\underline{h}}^{obs} \right| e^{\frac{-\left| E_{\underline{h}}^{obs} \right|^2}{\varepsilon_{\underline{h}}}} \qquad \text{(EQ 2.4.7)}$$

$$\Lambda_{null}^{centric} \approx \sqrt{\frac{2N}{\pi \varepsilon_{\underline{h}}}} e^{\frac{-\left| E_{\underline{h}}^{obs} \right|^2}{2\varepsilon_{\underline{h}}}} \qquad \text{(EQ 2.4.8)}$$

This now allows us to calculate the LLG which has been renamed LOGLIK, and use it as a FOM in the MITHRIL90 program.

2.5 FOM Correlations

In order to determine if our new figure of merit is indeed different from all others it was decide to perform an investigation into the linear correlations of the new FOM with those currently used in the MITHRIL program.

Linear Correlation Coefficient (Pearson r) is given by the following equation (Press, Flannery, Teukolsky & Vetterling, 1992).

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}} \qquad \text{(EQ 2.5.1)}$$

Where

N is the number of pairs of points (x_i, y_i)

$\bar{x}_i$ is the mean of the x_i's and $\bar{y}_i$ is the mean of the y_i's

r lies in the range -1 < r <1 and indicates that there is full correlation as r approaches 1 or -1 and that the two variables being tested are completely un-correlated when r=0. The linear correlation coefficient is not a good indicator of whether a correlation is significant or not as there is no way to introduce a null hypothesis into the formula. It is used as the conventional measure of the strength of a known correlation.

2.6 Implementation into MITHRIL90

The source code for the new figure of merit was written in standard FORTRAN77, to be transportable across systems. The theory lends itself to be written and implemented in two parts:

1) A subroutine to generate the additional triplets for the reflections that will remain unphased, in addition to the invariants used for phasing and other FOM calculations. A temporary file of all triplets involving two reflections that would be phased and one that would not is written to disk. The code that generates these triplets is not re-entrant unless renormalisation has taken place, to prevent any unnecessary processing. This subroutine is called at the time of invariant generation.

2) A subroutine to read in the FOM triplets from the temporary file and produce $\left|E_{\underline{h}}^{calc}\right|$. The values for $\left|E_{\underline{h}}^{calc}\right|$ can then be compared with $\left|E_{\underline{h}}^{obs}\right|$ to produce our figure of merit. This subroutine is called following tangent refinement.

To request the calculation of the LOGLIK FOM the user must enter the request at the time of invariant generation. This has required an alteration to the menu for invariant generation in MITHRIL90 and the new menu is shown in Figure 1.

THE FOLLOWING COMMANDS ARE CURRENTLY AVAILABLE:

TRIPlet	<	Parameters	(1)	No. of reflections for triplets		>
	<		(2)	0/1 Default/all refs for psi-zero		>
	<		(3)	Cut-off for (Sin(T)/L)**2		>
	<	[FOm] Turns on the calculation of the new F.O.M.				>
WEIGht	<	Weight E's according to (Sin(T)/L)**2 values				>
	<	1/0 = Print/Don't print table of weighted E's				>
L.E.	<	Use Linear Equations to estimate the triplets				>
	<	Parameters	(1)	Max. No. of missing terms	(0)	>
	<		(2)	Minimum diagonal term	(1)	>
	<		(3)	Max. terms for Eqns. 1-3	(0.5)	>
	<		(4)	Minimum No. of equations	(5)	>
LIST	<	Print triplets - Please use this only if essential				>
MDKS	<	Invokes the MDKS formula to estimate triplets				>
	<	Parameter is min. no. of contributors to <D>				>
	<	*** Both MDKs and L.E. options are very slow ***				>

Only the first 4 characters of any dialogue command are significant and a <CR> then terminates current input.

FIGURE 1. The triplet menu from MITHRIL90.

The output from the tangent refinement is different in that LOGLIK now appears on the reports alongside all other FOMs. An example of the output produced for LOGANIN (Jones, Sheldrick, Glüsenkamp & Tietze, 1980) is shown in Figure 2. Only the first 10 phase sets and not all permuted reflections are shown.

Tangent formula phase determination

LOGANIN

The early figures of merit are not to be applied

	Figures of Merit					Undet.	Starting set phases generated by programme				
Set	Abs.	Psi-Zero	Resid.	Nqest	Loglik	Phases	2	5	6	9	21
1	0.7478	2.054	26.38	0.00	-131.13	0(18)	90	360	50	45	360
2	0.8960	1.992	24.23	0.00	-126.61	0(18)	90	360	150	45	360
3	0.7250	2.225	28.38	0.00	-146.88	0(14)	90	360	250	45	360
4	0.7124	1.708	25.71	0.00	-137.92	0(16)	90	360	350	45	360
5	0.5444	2.943	33.70	0.00	-136.00	0(15)	90	360	90	45	360
6	0.9352	1.503	18.26	0.00	-119.11	0(13)	90	360	190	45	360
7	0.5815	2.762	34.20	0.00	-135.96	0(14)	90	360	290	45	360
8	0.7520	2.717	27.16	0.00	-139.49	0(15)	90	360	30	45	360
9	0.8943	1.688	21.12	0.00	-126.69	0(22)	90	360	130	45	360
10	0.6227	2.161	30.12	0.00	-143.01	0(14)	90	360	230	45	360

The figures in parenthesis in the above list refer to the number of cycles of tangent refinement that were required

FIGURE 2. The layout of the tangent phase refinement report showing the LOGLIK FOM

Additional code was added to the REVIEW module which is used to rank phase sets according to the magnitude of individual FOMs. This module now has an option for ranking phase sets by LOGLIK. It was decided not to use LOGLIK in the calculation of CFOM (see Chapter 1 Section 12.5) until further refinements of the theory had been implemented.

3.0 EXPERIMENTAL AND RESULTS

3.1 Source and Processing of Data Sets

The test structures were selected from the Sheldrick difficult structures database, and also from a group of readily solved structures used at Glasgow for the testing of the MITHRIL program. Only the full observed data sets were used. Table 1 shows the relevant crystallographic data for each of the structures used in the tests.

Structure	Space Group	Formula	Z
Diamantane-4-ol	$P4_2/n$	$C_{14}H_{20}O$	8
Quinol	$R\bar{3}$	$C_6H_6O_2$	54
HOV1	$C2/m$	$Pr_{14}Ni_6Si_{11}$	4
Selendid	$P2_1$	$C_{22}H_{28}O_2Se$	2
Azet	$Pca2_1$	$C_{21}H_{16}ClNO$	8
TUR10	$P6_322$	$C_{15}H_{24}O_2$	12
Loganin	$P2_12_12_1$	$C_{17}H_{26}O_{10}$	4
Dodecane-Diol	$I\bar{4}2d$	$C_{10}H_{18}O_2$	16
APAPA	$P4_12_12$	$C_{30}H_{37}N_{15}O_{16}P_2 \cdot 6H_2O$	8
MGHEX	$P3_1$	$C_{48}H_{68}N_{12}O_{12}Mg \cdot 2ClO_4 \cdot 4CH_3CN$	3
TOTC	$P6_1$	$C_{33}H_{36}O_6 \cdot O \cdot 2C_{16}H_{33}OH$	6
TPH	$C222_1$	$C_{24}H_{20}N_2$	12
Goldman2	Cc	$C_{28}H_{16}$	8
Munich1	$C2$	$C_{20}H_{16}$	8
MBH2	$P1$	$C_{15}H_{24}O_3$	3
SUOA	$P2_12_12_1$	$C_{28}H_{38}O_{19}$	4
Winter2	$P2_1$	$C_{52}H_{83}N_{11}O_{16} \cdot 3CH_2Cl_2$	2
Synthanecine	$P2_1/n$	$C_{11}H_{15}NO_4$	4
Ibuprofen	$P2_1/c$	$C_{13}H_{18}O_2$	4
Platynecine	$P2_12_12_1$	$C_8H_{15}NO_2$	4
Cytisine	$P2_12_12_1$	$C_{11}H_{14}N_2O$	8
Heliotridine	$P2_12_12_1$	$C_{15}H_{21}NO_4$	4

TABLE 1. Crystallographic information for the test structures used

The data set for each structure was processed through a test version of MITHRIL90 that contained the code for the calculations of LOGLIK. The structure was then solved using a variety of options that always included a request for LOGLIK to be calculated. The structure solution then continued until tangent refinement was completed. Table 2 shows the options used in the trial structure solution runs and also results of these test runs.

Structure	Quartets (Y/N)	Tangent weighting scheme	No. of Phase sets generated	Solution found (Y/N)
Diamantane-4-ol	N	None	16	Y
Quinol	Y	None	16	Y
HOV1	Y	None	32	N
Selendid	N	None	24	Y
Azet	N	Hull-Irwin	36	N
TUR10	N	Hull-Irwin	60	Y
Loganin	N	Hull-Irwin	56	Y
Dodecane-Diol	N	Hull-Irwin	54	N
APAPA	N	Hull-Irwin	64	N
MGHEX	N	Hull-Irwin	54	N
TOTC	N	None	48	N
TPH	N	Hull-Irwin	28	N
Goldman2	N	Hull-Irwin	40	N
Munich1	Y	Hull-Irwin	16	N
MBH2	N	Hull-Irwin	65	N
SUOA	N	Hull-Irwin	48	N
Winter2	N	None	60	N
Synthanecine	N	None	16	Y
Ibuprofen	N	Hull-Irwin	24	Y
Platynecine	N	Hull-Irwin	56	Y
Cytisine	N	None	24	Y
Heliotridine	N	Hull-Irwin	15	Y

TABLE 2. Options used during testing of structures

The data used for the correlation experiments were the figures of merit output from the final tangent refinement.

3.2 The Figure of Merit Results

All results given here are comparisons with the conventional FOMs output by the MITHRIL90 program.

The first table shows the FOMs for sixteen of the default number of phase sets (shown in Table 2), normally this is the first sixteen sets but may be the sixteen phase sets of most interest. Phase set numbers that are marked with an asterisk are correct phase sets i.e. show the whole or a partial molecular structure. In this table if the column of figures for NQEST is given as zero then no quartets were calculated. In all cases where no reference is given for the structure or it is listed as unpublished the reader is referred to the Sheldrick difficult structures database for further information.

The second table shows the correlations of the conventional FOMs with LOGLIK and 1/LOGLIK, for all the phase sets that were generated, not only those that are shown in the first table. After each set of results there is a short discussion describing the behaviour of LOGLIK with this data.

Diamantane-4-ol (Rodgers & Kennard, unpublished work)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1*	0	1.1899	0.702	18.30	0.00	7.72
2	0	0.8880	1.455	23.90	0.00	-8.58
3	0	0.8849	1.442	23.83	0.00	-8.58
4*	0	1.1899	0.702	18.30	0.00	7.72
5*	0	1.1899	0.702	18.30	0.00	7.72
6*	0	1.1380	0.796	18.42	0.00	3.52
7	0	0.4243	2.780	44.77	0.00	-36.40
8*	0	1.1380	0.796	18.42	0.00	3.52
9*	0	1.1380	0.804	18.42	0.00	3.52
10	0	0.9082	1.246	23.33	0.00	-5.37
11	0	0.9068	1.276	23.28	0.00	-5.37
12*	0	1.1380	0.804	18.42	0.00	3.52
13*	0	1.1381	0.804	18.43	0.00	3.52
14*	0	1.1899	0.678	18.30	0.00	7.72
15	0	0.4547	2.977	43.51	0.00	-30.10
16*	0	1.1899	0.678	18.30	0.00	7.72

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.9946	-0.9772	-0.9873	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	0.6111	-0.4843	-0.5402	0.0000

For Diamantane-4-ol LOGLIK behaves in an identical manner to the other FOMs indicating the best phase set and giving a reduced value for the second best set.

Quinol (Wallwork & Powell, 1980)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1*	0	0.8385	1.460	19.16	-0.478	-17.19
2	0	0.5061	2.806	40.34	-0.067	-53.95
3	0	0.9682	2.989	21.48	-0.749	18.57
4	0	0.6058	2.557	34.36	-0.112	-54.91
5*	0	0.8388	1.460	19.15	0.444	-17.19
6	0	0.5097	2.905	40.26	0.086	-54.52
7	0	0.9684	2.989	21.52	0.585	18.43
8	0	0.6047	2.581	34.67	0.028	-50.10
9	0	1.0491	5.299	27.62	-0.502	86.83
10	0	0.4522	2.744	43.45	0.035	-96.17
11	0	0.4099	2.674	46.97	-0.749	-80.57
12	0	0.5559	2.645	36.70	-0.169	-43.53
13*	0	0.8412	1.451	19.14	0.288	-17.19
14	0	0.4556	2.709	43.10	0.086	-89.58
15	0	0.9689	2.935	21.60	0.543	22.41
16	0	0.5468	2.897	37.61	0.035	-37.62

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.9319	-0.7394	0.4584	0.5654
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	0.3917	-0.1006	0.5925	0.8669

This is the first example of the LOGLIK FOM preferentially selecting sets that have a low residual but with a high value for ψ_o. On examination of the electron density maps for such phase sets, no meaningful structure or fragments can be seen. The correlation between LOGLIK and ABSFOM is pronounced.

HOV1 (Hovestreydt, Klepp & Parthé, 1983)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	1.0601	3.180	2.84	-1.00	114.06
2	0	1.0609	3.316	2.99	1.00	113.06
3	0	0.8413	3.640	16.51	-0.68	89.08
4	0	0.8367	3.311	16.30	0.68	74.56
5	0	1.0591	3.224	2.87	-1.00	115.64
6	0	1.0599	3.361	3.00	1.00	117.32
7	0	0.8413	3.589	16.48	-0.68	86.92
8	0	0.8408	3.209	16.59	0.68	69.69
9	0	1.0601	3.180	2.84	-1.00	114.06
10	0	1.0609	3.316	2.99	1.00	113.82
11	0	0.8413	3.640	16.51	-0.68	89.08
12	0	0.8408	3.128	16.55	0.68	71.67
13	0	1.0591	3.224	2.87	-1.00	115.64
14	0	1.0599	3.361	3.00	1.00	117.32
15	0	0.8413	3.589	16.48	-0.68	86.92
16	0	0.8408	3.076	16.52	0.68	67.95

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.8978	-0.8963	0.0981	0.1593
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.8220	0.8199	-0.2409	-0.1810

HOV1 is a heavy atom inorganic structure with a single heavy atom on a special position. The above results do show that LOGLIK is closely related to RESID, which is unrealistically small in the above tables and that this appears to be making LOGLIK appear to be unrealistically large. Note that the correlation to ABSFOM is still present.

Selendid (Clegg et al., 1980)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1*	0	1.2718	2.394	9.45	0.00	-97.42
2	0	0.8448	2.589	22.75	0.00	-127.83
3*	0	1.2719	2.394	9.45	0.00	-97.42
4	0	0.8438	2.584	22.68	0.00	-127.88
5	0	1.0668	2.108	18.58	0.00	-114.82
6*	0	1.2718	2.395	9.45	0.00	-97.42
7	0	1.0673	2.107	18.62	0.00	-114.83
8	0	1.0670	2.108	18.61	0.00	-114.84
9	0	1.0668	2.108	18.61	0.00	-114.85
10*	0	1.2718	2.394	9.45	0.00	-97.42
11	0	1.0672	2.108	18.61	0.00	-114.85
12*	0	1.2720	2.394	9.46	0.00	-97.42
13*	0	1.2722	2.394	9.46	0.00	-97.42
14	0	1.0666	2.109	18.50	0.00	-114.78
15	0	1.0669	2.108	18.55	0.00	-114.80
16	0	0.9987	2.863	20.95	0.00	-114.14

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.9878	-0.9838	0.1948	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.9774	0.9919	-0.2432	0.0000

The LOGLIK figure of merit indicates the correct solution strongly, which the ψ_o FOM does not.

Azet (Colens et al., 1974)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	1.1031	1.271	29.44	0.00	-94.42
2	0	0.9675	1.327	38.71	0.00	-96.53
3	0	1.7210	2.250	43.66	0.00	-93.68
4	0	1.1302	1.096	31.66	0.00	-93.99
5	0	1.7458	2.294	46.65	0.00	-94.59
6	0	0.8835	2.086	31.52	0.00	-93.00
7	0	1.1564	1.263	32.93	0.00	-92.63
8	0	1.0439	1.397	40.64	0.00	-94.87
9	0	1.3954	1.319	40.61	0.00	-93.52
10	0	0.7883	1.811	31.59	0.00	-96.74
11	0	0.9997	1.266	42.45	0.00	-95.93
12	0	1.0889	1.393	35.12	0.00	-95.18
13	0	1.6097	1.969	39.21	0.00	-93.85
14	0	1.6684	2.126	43.74	0.00	-93.89
15	0	0.8080	1.502	47.63	0.00	-95.83
17	0	0.5168	4.417	36.79	0.00	-100.53

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.5189	0.1631	-0.4400	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.5152	-0.1589	0.4284	0.0000

In the above tables of results there are no correct solutions. The values of LOGLIK are very similar with no clear indication of a true minimum in phase space. This structure is a good example of LOGLIK not giving a false indication of correctness where none is warranted.For this structure there is no correlation between LOGLIK and the conventional FOMs.

TUR10 (Braekman et al., 1981)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.6009	4.316	40.64	0.00	-41.85
2	0	0.4885	3.509	43.39	0.00	-66.39
3	0	0.6741	3.303	36.82	0.00	-44.01
4	0	0.5397	3.656	39.91	0.00	-55.17
5	0	0.6333	2.616	38.69	0.00	-44.80
6	0	0.5515	2.441	36.53	0.00	-72.79
7	0	0.6069	2.324	33.45	0.00	-66.48
8	0	0.5812	2.498	35.21	0.00	-68.51
9	0	0.5903	2.153	34.69	0.00	-69.52
10	0	0.5264	3.449	39.89	0.00	-57.59
11	0	0.5136	2.994	38.57	0.00	-67.24
12	0	0.5219	2.912	38.43	0.00	-66.38
13	0	0.5375	2.620	37.05	0.00	-60.67
14	0	0.5215	2.425	37.88	0.00	-71.31
16	0	0.7228	3.360	32.92	0.00	-38.24
47*	0	0.8313	1.406	21.47	0.00	-33.25

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.4760	-0.3661	0.1118	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.5938	0.4877	-0.0209	0.0000

In the above phase sets only one correct solution was produced. The highest value given for LOGLIK corresponds to the correct set and has a value similar to a phase set with a low residual - high ψ_o value, as was seen in many test sets. The spread of values of LOGLIK is large within the incorrect phase sets.

Loganin (Jones, Sheldrick, Glüsenkamp & Tietze, 1980)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	115	0.7590	1.911	22.89	0.00	-43.12
2	115	0.6615	1.530	25.56	0.00	-45.58
3	115	0.7527	1.463	23.99	0.00	-38.87
4	115	0.7564	2.016	23.92	0.00	-43.48
5	115	0.7059	1.844	27.96	0.00	-44.31
6	115	0.6601	1.887	28.54	0.00	-48.05
7	115	0.7034	1.969	24.73	0.00	-48.86
8	115	0.7746	1.499	22.88	0.00	-42.42
9	115	0.7480	1.883	24.85	0.00	-43.84
10	115	0.5714	1.919	30.77	0.00	-51.53
11	115	0.7055	1.750	23.51	0.00	-48.32
12	115	0.6322	2.216	28.43	0.00	-41.23
13	115	0.7264	2.066	24.68	0.00	-42.33
14	115	0.6584	1.937	29.72	0.00	-44.81
15	115	0.6239	1.988	28.59	0.00	-44.43
43*	115	0.9735	0.996	15.01	0.00	-34.73

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.6164	-0.5403	-0.5779	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.6520	0.5858	0.6257	0.0000

As for TUR10 only one solution was determined from the phase sets generated. We see a broad spread of values for LOGLIK for the incorrect phase sets, however the correct phase set does possess the highest value, and is thus identified.

Dodecane-Diol

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.2194	4.771	52.98	0.00	-7.57
2	0	0.3992	3.120	41.83	0.00	-10.17
3	0	0.4307	2.462	39.60	0.00	-10.44
4	0	0.2648	5.382	50.73	0.00	-10.94
5	0	0.2079	6.272	53.42	0.00	-8.81
6	0	0.3419	4.677	47.14	0.00	-8.07
7	0	0.5234	5.394	36.24	0.00	-8.17
8	0	0.4294	2.790	39.75	0.00	-9.32
9	0	0.3527	1.955	44.71	0.00	-9.95
10	0	0.5688	3.543	32.98	0.00	-9.07
11	0	0.4836	2.764	36.92	0.00	-8.74
12	0	0.2520	4.131	51.03	0.00	-9.39
13	0	0.4332	2.965	39.85	0.00	-9.38
14	0	0.1239	8.781	58.93	0.00	-9.33
15	0	0.5325	3.695	35.35	0.00	-9.76
16	0	0.5020	2.282	35.09	0.00	-8.47

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.1625	-0.1495	-0.0694	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.1363	0.1241	0.0454	0.0000

No solutions were determined from tangent refinement using 54 sets. There is a very small spread in the values calculated for LOGLIK. The similar values for LOGLIK for all phase sets indicates that there is no false identification of a correct phase set.

APAPA (Suck, Manor & Saenger, 1976)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.7292	1.635	27.90	0.00	-71.25
2	0	0.5331	1.698	32.42	0.00	-82.64
3	0	0.5980	1.752	33.05	0.00	-73.79
4	0	0.6714	1.300	26.35	0.00	-58.57
5	0	0.4675	2.212	35.22	0.00	-78.18
6	0	0.6200	1.642	29.34	0.00	-70.51
7	0	0.5501	2.048	33.44	0.00	-72.96
8	0	0.4686	2.032	34.26	0.00	-65.05
9	0	0.4031	2.409	37.06	0.00	-66.14
10	0	0.5802	2.048	30.53	0.00	-64.74
11	0	0.4390	2.164	35.59	0.00	-64.78
12	0	0.2705	3.756	42.93	0.00	-68.51
13	0	0.4231	2.122	36.07	0.00	-68.36
14	0	0.4815	2.007	33.64	0.00	-64.87
15	0	0.2747	3.465	43.78	0.00	-70.09
16	0	0.2876	3.022	42.34	0.00	-71.89

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	-0.1254	0.0590	-0.0107	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	0.1044	-0.0314	0.0352	0.0000

No solution was found in the 54 phase sets generated, but the LOGLIK FOM has the highest figure for the phase set with the best conventional figures of merit. All other values of LOGLIK are similar with a relatively small spread. This structure, with all incorrect solutions displays no correlation between LOGLIK and the other conventional FOMs.

MGHEX (Karle & Karle, 1981)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.2731	0.000	50.37	0.00	-111.66
2	0	0.1796	0.000	56.77	0.00	-119.03
3	0	0.3708	0.000	43.87	0.00	-113.70
4	0	0.3842	0.000	42.85	0.00	-115.56
5	0	0.2073	0.000	54.67	0.00	-116.84
6	0	0.1872	0.000	56.05	0.00	-115.20
7	0	0.1994	0.000	55.46	0.00	-114.54
8	0	0.3375	0.000	45.99	0.00	-112.58
9	0	0.4878	0.000	36.39	0.00	-115.33
10	0	0.3640	0.000	44.35	0.00	-116.50
11	0	0.3269	0.000	46.67	0.00	-118.57
12	0	0.2652	0.000	50.89	0.00	-117.20
13	0	0.2207	0.000	54.02	0.00	-118.93
14	0	0.3326	0.000	46.43	0.00	-119.10
15	0	0.2718	0.000	50.47	0.00	-113.52
16	0	0.2990	0.000	48.53	0.00	-113.91

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.0017	-0.0050	0.0000	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.0048	0.0081	0.0000	0.0000

No figures are given for ψ_o as there were too few contributors to this FOM for it to be calculated by the program. Again there is no solution and LOGLIK has very similar values for all phase sets and no correlations with the conventional FOMs.

TOTC (Williams & Lawton, 1975)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.5262	1.792	31.87	0.00	-142.32
2	0	0.5016	1.815	33.23	0.00	-139.88
3	0	0.4644	1.887	34.92	0.00	-152.93
4	0	0.4521	2.036	35.93	0.00	-145.75
5	0	0.4542	2.014	35.82	0.00	-139.65
6	0	0.5041	2.229	33.25	0.00	-140.36
7	0	0.4976	1.931	32.96	0.00	-142.24
8	0	0.4927	1.949	33.36	0.00	-139.21
9	0	0.5027	2.014	32.85	0.00	-139.57
10	0	0.5150	1.709	32.23	0.00	-137.47
11	0	0.4967	1.840	34.10	0.00	-146.80
12	0	0.4874	1.908	33.89	0.00	-151.49
13	0	0.4938	1.779	33.45	0.00	-151.10
14	0	0.4755	2.116	34.85	0.00	-146.30
15	0	0.4804	1.940	34.31	0.00	-139.33
16	0	0.4855	1.779	34.10	0.00	-144.51

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.2065	-0.2491	-0.0078	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.2072	0.2501	0.0107	0.0000

No correct solution was found in the 48 phase sets refined by tangent refinement. As has been seen before, LOGLIK selects no individual phase set although a broader spread of LOGLIK is observed than normal for equally bad phase sets.

TPH (Hoekstra, Vos, Braun & Hornstra, 1975)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.8522	2.025	27.56	0.00	-138.55
2	0	0.7859	1.983	26.15	0.00	-151.14
3	0	0.7165	1.842	25.46	0.00	-150.69
4	0	0.7754	2.257	24.11	0.00	-146.94
5	0	0.7514	2.242	28.00	0.00	-145.62
6	0	0.8290	1.755	21.04	0.00	-154.60
7	0	0.8379	1.810	25.15	0.00	-142.99
8	0	0.8800	1.727	22.75	0.00	-140.44
9	0	0.7181	1.823	24.28	0.00	-153.37
10	0	0.7634	2.273	26.20	0.00	-152.17
11	0	0.7501	2.313	25.72	0.00	-144.91
12	0	0.7431	2.373	25.97	0.00	-140.90
13	0	0.7725	2.094	26.15	0.00	-152.98
14	0	0.8372	2.013	27.31	0.00	-139.00
15	0	0.8022	1.716	20.51	0.00	-162.27
18	0	0.8411	1.635	19.87	0.00	-150.21

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.6612	0.0185	0.3593	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.6609	-0.0432	-0.3650	0.0000

No correct phase sets were generated for this structure. The spread of LOGLIK is large and the phase set with the best conventional FOMs is not selected as preferred by LOGLIK, indeed the largest values of LOGLIK are associated with phase sets that have high values of ψ_o and low values of RESID.

Goldman2 (Irngartinger Reibel & Sheldrick, 1981)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.8059	2.027	27.71	0.00	-149.54
2	0	0.8911	1.673	27.91	0.00	-149.65
3	0	0.9237	1.754	24.83	0.00	-150.86
4	0	0.8672	1.675	24.51	0.00	-152.37
5	0	0.9617	1.526	24.81	0.00	-149.30
6	0	0.9804	1.519	25.56	0.00	-150.71
7	0	0.9469	1.614	23.92	0.00	-148.73
8	0	0.9372	1.610	25.20	0.00	-147.99
9	0	0.8719	1.940	26.72	0.00	-151.14
10	0	0.9034	1.798	24.74	0.00	-151.79
11	0	0.9561	1.777	21.57	0.00	-152.83
12	0	0.8967	1.756	26.07	0.00	-151.65
13	0	0.9964	1.636	23.87	0.00	-152.91
14	0	0.8735	1.697	25.68	0.00	-149.44
15	0	0.9459	1.570	26.58	0.00	-150.92
16	0	0.9729	1.670	26.34	0.00	-152.39

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.3766	0.2814	-0.2554	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.3710	-0.2835	0.2578	0.0000

Once again no correct set was determined by tangent refinement. We see little correlation between LOGLIK and the conventional FOMs and the spread in LOGLIK is small.

Munich1 (Szeimies-Seebach et al., 1978)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.7744	2.357	33.19	0.07	-51.23
2	0	1.0121	2.457	30.85	0.19	-10.12
3	0	1.1168	2.510	29.63	0.58	2.40
4	0	1.3311	2.543	34.16	0.75	13.43
5	0	1.0513	2.239	31.92	0.31	-25.69
6	0	1.1460	2.246	34.82	0.25	-6.73
7	0	1.2369	2.179	36.52	0.52	-4.90
8	0	1.2035	2.399	36.45	0.48	1.61
9	0	0.9471	2.527	32.21	0.31	-21.83
10	0	0.7417	2.380	33.41	0.16	-49.42
11	0	1.1034	2.428	31.66	0.39	1.14
12	0	1.1534	2.503	34.34	0.49	3.01
13	0	1.6796	3.491	46.98	1.00	50.83
14	0	1.1192	2.434	32.36	0.46	-0.18
15	0	1.2154	2.327	34.05	0.58	2.96
16	0	1.3958	2.552	36.41	0.64	15.24

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.9592	0.6181	0.6659	0.9113
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	0.0408	0.0959	0.0381	0.0249

This structure is LOGLIKs' greatest failure. It falsely indicates correct solutions and selects as the most probable set the one with the worst conventional figures of merit. For this phase set NQEST has a value of 1.0, indicating that all negative quartets are incorrectly fitted, this may be a false minimum in phase space that is being indicated by LOGLIK. The maps were examined that were indicated by LOGLIK but they contained no relevant structural information.

MBH2 (Poyser et al., 1986)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	1.3089	0.000	22.50	0.00	-125.11
2	0	1.3331	0.000	18.88	0.00	-124.21
3	0	0.9468	0.000	18.97	0.00	-131.72
4	0	1.3362	0.000	19.27	0.00	-124.22
5	0	1.3420	0.000	20.37	0.00	-124.19
6	0	1.3685	0.000	18.26	0.00	-124.21
7	0	1.3928	0.000	22.08	0.00	-124.08
8	0	1.3468	0.000	19.78	0.00	-125.01
9	0	0.9008	0.000	18.84	0.00	-137.94
10	0	1.3721	0.000	19.20	0.00	-124.56
11	0	1.3938	0.000	20.82	0.00	-124.24
12	0	1.2477	0.000	18.31	0.00	-124.28
13	0	1.0187	0.000	19.21	0.00	-125.80
14	0	0.8547	0.000	18.61	0.00	-145.33
15	0	1.4039	0.000	21.82	0.00	-124.10
16	0	1.3613	0.000	22.62	0.00	-124.20

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.8307	0.3247	0.0000	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.8446	-0.3333	0.0000	0.0000

Despite good values for RESID no solutions for this structure were found in the phase sets. There is a very narrow spread of values for LOGLIK showing no discrimination between sets. There is a correlation between LOGLIK and ABSFOM, which is a feature normally seen in collections of phase sets that contain correct solutions.

SUOA (Oliver & Strickland, 1984)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.6963	2.002	27.31	0.00	-140.95
2	0	0.5787	2.368	32.98	0.00	-130.11
3	0	0.7274	1.794	26.78	0.00	-132.70
4	0	0.7788	1.866	27.44	0.00	-134.38
5	0	0.5628	2.044	31.50	0.00	-148.06
6	0	0.7443	1.798	31.46	0.00	-130.52
7	0	0.7305	1.759	27.36	0.00	-137.98
8	0	0.6656	2.097	28.59	0.00	-144.68
9	0	0.7175	1.998	28.27	0.00	-133.76
10	0	0.7040	1.949	27.37	0.00	-134.17
11	0	0.5827	2.211	32.49	0.00	-130.45
12	0	0.6258	1.737	27.71	0.00	-144.50
13	0	0.6798	1.942	26.08	0.00	-130.26
14	0	0.6732	1.925	27.30	0.00	-134.43
15	0	0.6999	1.942	27.43	0.00	-131.01
16	0	0.6240	1.791	28.17	0.00	-137.09

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.5732	0.2204	0.1217	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.5826	-0.2335	-0.1189	0.0000

No correct phase sets were produced for SUOA but LOGLIK does exhibit a broad range of values giving an indication of correctness to phase sets with higher values of ψ_o. This behaviour is seen in other structures where no solution was found.

Winter2 (Butters et al., 1981)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	1.0775	2.786	28.03	0.00	-105.85
2	0	1.0272	2.839	26.50	0.00	-111.69
3	0	1.1541	2.775	24.76	0.00	-104.43
4	0	1.2194	3.034	26.97	0.00	-105.50
5	0	1.0771	2.777	27.33	0.00	-105.65
6	0	1.1492	2.752	25.20	0.00	-107.28
7	0	0.9041	3.066	23.47	0.00	-108.49
8	0	1.0795	2.773	27.87	0.00	-105.84
9	0	1.0244	2.778	27.02	0.00	-110.96
10	0	1.1489	2.727	25.11	0.00	-106.88
11	0	1.0263	2.850	26.65	0.00	-111.66
12	0	1.0668	2.748	26.48	0.00	-104.42
13	0	1.0251	2.854	26.86	0.00	-112.00
14	0	1.1423	2.724	25.30	0.00	-107.75
15	0	1.0817	2.765	27.50	0.00	-105.38
16	0	1.2196	3.016	26.64	0.00	-105.94

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.7312	-0.1088	-0.0379	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.7302	0.1100	0.0373	0.0000

No solution was found for Winter2 among the phase sets generated. LOGLIK shows a small range of values with little discrimination between phase sets.

Synthanecine (Barbour, Freer, Robins, 1987)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1*	0	1.1759	0.752	17.39	0.00	43.94
2*	0	1.1759	0.752	17.39	0.00	43.94
3	0	0.7903	1.841	25.87	0.00	-5.22
4	0	0.6471	1.824	31.50	0.00	-27.07
5	0	0.9077	1.559	24.85	0.00	-15.13
6	0	0.9097	1.489	24.64	0.00	-14.16
7	0	0.6777	1.448	30.76	0.00	-17.68
8	0	0.6950	1.429	29.37	0.00	-10.23
9*	0	1.1759	0.752	17.39	0.00	43.94
10*	0	1.1759	0.752	17.39	0.00	43.94
11	0	0.7860	1.690	25.09	0.00	-7.72
12	0	0.7132	1.669	27.76	0.00	-12.28
13	0	0.8871	1.593	25.66	0.00	-9.91
14	0	0.9163	1.551	24.78	0.00	-18.63
15	0	0.6589	1.990	29.97	0.00	-37.67
16	0	0.6975	2.753	30.98	0.00	-34.27

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.9152	-0.9374	-0.8929	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	0.5836	-0.5053	-0.5023	0.0000

LOGLIK strongly identifies the correct phase set. The other incorrect phase sets have LOGLIK values in a broad range but the correct value is so large that the others could be discounted.

Ibuprofen (McConnell, 1974)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.7944	2.318	24.41	0.00	-148.82
2	0	0.8791	2.021	20.23	0.00	-138.74
3	0	0.7920	2.304	21.35	0.00	-145.76
4	0	0.9349	1.729	19.45	0.00	-137.60
5	0	0.9114	2.031	16.69	0.00	-134.47
6	0	0.7954	2.035	19.43	0.00	-147.67
7	0	0.9216	1.984	18.66	0.00	-133.07
8	0	0.8304	2.253	21.70	0.00	-140.08
9*	0	0.9778	1.125	10.99	0.00	-133.52
10	0	0.8372	2.174	18.21	0.00	-137.59
11	0	0.9075	1.511	14.33	0.00	-137.14
12	0	0.8748	1.650	16.27	0.00	-138.38
13*	0	0.9858	1.080	12.12	0.00	-134.25
14	0	0.8666	1.579	16.41	0.00	-140.89
15*	0	0.9775	1.251	12.85	0.00	-135.94
16	0	0.8020	1.743	19.56	0.00	-149.16

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.8714	-0.6410	-0.5002	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.8757	0.6459	0.5073	0.0000

In this structure LOGLIK identifies the correct solutions but also identifies many other incorrect solutions as being correct. RESID has a very similar behaviour for this structure. The correlations with other figures of merit are not as large as other structures where a correct solution was found.

Platynecine (Freer, Kelly & Robins, 1987)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1*	0	0.9705	0.993	11.13	0.00	-17.81
2*	0	0.9659	0.994	11.32	0.00	-17.81
3	0	0.6312	3.320	36.80	0.00	-22.31
4	0	0.7230	2.646	34.97	0.00	-20.43
5	0	0.6295	3.280	37.25	0.00	-21.03
6*	0	0.9356	1.015	13.14	0.00	-18.01
7	0	0.7444	2.272	33.70	0.00	-22.33
8	0	0.6845	2.816	34.29	0.00	-22.32
9	0	0.6217	3.315	34.19	0.00	-22.62
10	0	0.5834	3.332	36.46	0.00	-23.15
11*	0	0.9788	0.969	11.07	0.00	-17.82
12	0	0.7293	2.664	35.22	0.00	-20.22
13	0	0.7266	2.666	35.59	0.00	-20.29
14	0	0.6968	3.398	32.25	0.00	-19.87
15	0	0.6513	3.027	33.85	0.00	-21.07
16	0	0.6313	3.233	37.23	0.00	-21.13

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.7997	-0.7604	-0.7602	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.8206	0.7867	0.7763	0.0000

LOGLIK successfully identifies and ranks the correct solutions that were found. The identification of the correct sets is not strong but they are the sets with the highest value of LOGLIK. The range of LOGLIK is very small, a feature normally associated with a collection of incorrect sets.

Cytisine (Freer, Robins & Sheldrake, 1987)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.6318	2.510	30.53	0.00	-124.64
2	0	0.6777	2.271	31.45	0.00	-121.91
3	0	0.7810	2.034	28.51	0.00	-111.73
4*	0	0.9390	1.125	16.12	0.00	-96.20
5	0	0.5481	1.887	33.63	0.00	-115.96
6	0	0.7043	2.003	29.68	0.00	-115.27
7	0	0.6595	2.327	30.14	0.00	-115.76
8	0	0.6196	2.396	32.36	0.00	-120.97
9	0	0.6763	2.336	31.28	0.00	-119.47
10	0	0.7221	2.049	27.62	0.00	-120.51
11	0	0.5829	2.350	32.46	0.00	-120.05
12	0	0.6211	2.437	32.83	0.00	-117.87
13	0	0.6772	2.704	32.16	0.00	-117.60
14	0	0.6178	2.505	32.16	0.00	-121.39
15	0	0.6885	2.405	29.74	0.00	-119.69
16	0	0.6643	2.705	31.16	0.00	-117.93

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.7941	-0.8754	-0.9127	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.7754	0.8945	0.9332	0.0000

LOGLIK successfully identifies the only correct solution produced by tangent refinement. The spread of values of LOGLIK for the incorrect phase sets is very small and the correct solution is strongly indicated. This is one of the few structures for which LOGLIK is strongly correlated with ψ_o.

(+) Heliotridine (Freer, Hagan & Robins, 1988)

Set Number	Undet. Phases	ABSFOM	ψ_o	Residual	NQEST	LOGLIK
1	0	0.8568	3.426	30.04	0.00	-86.09
2	0	0.8823	2.999	27.53	0.00	-81.90
3	0	0.8614	3.192	29.77	0.00	-84.98
4	0	0.8358	2.852	24.68	0.00	-82.70
5*	0	1.0795	1.293	14.48	0.00	-70.66
6	0	0.8965	2.969	28.33	0.00	-84.03
7	0	0.8823	2.999	27.53	0.00	-81.90
8*	0	1.0800	1.292	14.40	0.00	-70.67
9	0	0.8738	3.128	30.15	0.00	-84.68
10	0	0.8385	3.253	26.61	0.00	-83.74
11	0	0.8209	2.776	26.89	0.00	-89.26
12	0	0.3542	3.077	44.08	0.00	-121.68
13	0	0.6129	3.082	33.97	0.00	-88.93
14	0	0.8698	3.056	26.00	0.00	-82.24
15*	0	1.0800	1.287	14.31	0.00	-70.58

	ABSFOM vs. LOGLIK	Residual vs. LOGLIK	ψ_o vs. LOGLIK	NQEST vs. LOGLIK
Correlation Coefficient	0.8579	-0.9085	-0.9159	0.0000
	ABSFOM vs. 1/LOGLIK	Residual vs. 1/LOGLIK	ψ_o vs. 1/LOGLIK	NQEST vs. 1/LOGLIK
Correlation Coefficient	-0.8617	0.9138	0.9186	0.0000

The correct phase set is successfully identified by LOGLIK. While the incorrect solutions have a small range of values the correct solution is strongly indicated. For this structure there is strong correlation between LOGLIK and all other conventional FOMs.

The results on the previous pages show several trends

(i) Where a solution is to be found then it will be indicated as the preferred phase set by LOGLIK.

(ii) LOGLIK indicates correct solutions for phase sets with low residual and high ψ_o values. This can be the cause of LOGLIK indicating incorrect results over correct phase sets. Despite this trend in results there is little correlation between LOGLIK and ψ_o when compared directly.

(iii) It is very difficult to define a range of LOGLIK that indicates a correct solution. We can only say that the highest value of LOGLIK for any given structure is the preferred phase set. In most structures the indication of correctness is given by all incorrect phase sets having similar values of LOGLIK while correct sets have much increased values.

(iv) Correlations to other FOMs are not consistent across structures. In those that are easily solved there appears to be close correlation, as one would expect, however in those where all phase sets are incorrect there is little or no correlation. This lack of correlation for incorrect sets is indicative of new information being contained in the LOGLIK FOM. The FOM with the highest number of close correlations is ABSFOM, which would be expected as both FOMs are a measure of internal triplet consistency.

(v) The values for many of the phase sets is large and negative. This would be expected only for very incorrect sets, yet appears for correct sets also.

3.3 Summary

It has been shown that LOGLIK is a good figure of merit that contains information not normally available to the conventional figures of merit i.e. the triplet consistency of the Es in the mid-range that have not been used in the phasing process. Like all conventional FOMs it should not be used in isolation, as false indications of correctness can be given. Although LOGLIK contains new information it is not superior to the conventional FOMs and as such aids little in the identification of correct phase sets. It should remain in the MITHRIL package in its current form as it is optional and does not contribute to CFOM.

The source of incorrect indications is the sensitivity of LOGLIK to the magnitudes of E^{calc}. We produce E^{calc} from the tangent formula which is flawed in its assumption of the non-correlation of invariants. This leads to the over consistency of phase sets, seen so often in tangent refinement, and reduces the ability to calculate high quality, low error E-magnitudes. Use of another technique to produce the extrapolated E^{calc} may remove these problems.

The problem of the lack of a fixed range of LOGLIK that indicates a correct solution can be overcome by using the FOM only to rank phase sets. This ranking can then be used to examine a reduced number of the other FOMs to indicate correctness, in a similar manner to the CFOM.

For a small increase in computer time it is possible to obtain a new figure of merit which can be used to corroborate indications given by the other FOMs. On a modern work station this increase in computer time is negligible.

4.0 FUTURE WORK

The large negative values of LOGLIK that indicate a correct solution is not to be expected. It does indicate a very poor fit between $|E^{calc}|$ and $|E^{obs}|$, which is not expected for a correct phase set. This could be due to a scaling problem of $|E^{calc}|$ and a new scaling technique should be used based on resolution or groups of intensities.

When deciding on which Es should be incorporated into LOGLIK those with high estimated standard deviations of the observed E-magnitude should be excluded or down-weighted in their contributions to the FOM. This would reduce the standard deviation of LOGLIK itself and so hopefully reduce the number of instances where the FOM only marginally indicates a solution.

The largest problem with this FOM appears to be the quality of the $|E^{calc}|$ data, and a better way of extrapolating the E-magnitudes should be found. LOGLIK works very well as part of the maximum entropy method and a mix of the two techniques should be very powerful. The phased reflections produced and refined from the tangent formula would be used to produce an E-map. This E-map would then pass through cycles of entropy maximisation until a maximum map had been determined. This ME-map would then be Fourier transformed back to phased reflections which could used in the $|E^{obs}|$, $|E^{calc}|$ comparison. This solution would be expensive in computer time but should yield a very powerful FOM.

The reflections being used for comparison to produce a value of LOGLIK should be split into two sets, those that contain only low resolution reflections i.e. >2.5Å and those of atomic resolution i.e. <2.0Å. This would have the effect of producing two figures of merit, the first used to determine molecular placement within the unit cell and the second to determine the correctness of the atomic arrangement.

References

Barbour, R.H., Freer, A.A. & Robins, D.J. (1987). *J. Chem. Soc. Perkin Trans.* **I**, 2069-2072

Braekman, J.C., Daloze, D., Dupont, A., Tursch, B., Declercq, J.P., Germain, G. & Van Meerssche, M. (1981). *Tetrahedron* **37**, 179-186

Bricogne, G. & Gilmore, C.J. (1990). *Acta. Cryst.* **A46**, 284-297

Butters, T., Hütter, P., Jung, G., Pauls, N., Schmitt, H., Sheldrick, G.M. & Winter, W. (1981). *Angew. Chem. Int. Ed. Engl.* **20**, 889-890

Clegg, W., Harms, K., Sheldrick, G.M., von Kiedrowski, G. & Tietze, L.F. (1980). *Acta. Cryst.* **B36**, 3159-3162

Colens, A., Declercq, J.P., Germain, G., Putzeys, J.P. & Van Meerssche, M. (1974). *Cryst. Struct. Comm.* **3**, 119-122

Freer, A.A., Hagan, D.B. & Robins, D.J. (1988). *Acta. Cryst.* **C44**, 666-668

Freer, A.A., Kelly, H.A. & Robins, D.J. (1987). *Acta. Cryst.* **C43**, 2020-2022

Freer, A.A., Robins, D.J. & Sheldrake, G.N. (1987). *Acta. Cryst.* **C43**, 1119-1122

Hoekstra, A., Vos, A., Braun, P.B. & Hornstra, J. (1975). *Acta. Cryst.* **B31**, 1708-1715

Hovestreydt, E., Klepp, K. & Parthé, E. (1983). *Acta. Cryst.* **C39**, 422-425

Hughes, E.W. (1953). *Acta. Cryst.* **6**, 871

Irngartinger, H., Reibel, W.R.K. & Sheldrick, G.M. (1981). *Acta. Cryst.* **B37**, 1768-1771

Jones, P.G., Sheldrick, G.M., Glüsenkamp, K.H. & Tietze, L.F. (1980). *Acta. Cryst.* **B36**, 481-483

Karle, I.L. & Karle, J. (1981). *Proc. Natl. Acad. Sci. USA* **78**, 681-685

Karle, J. & Karle, I.L. (1966). *Acta. Cryst.* **21**, 849-859

Karle, J. & Hauptman, H. (1953). *Acta. Cryst.* **6**, 473-476

Karle, J. & Hauptman, H. (1956). *Acta. Cryst.* **9**, 635-651

Main, P. (1977). *Acta. Cryst.* **A33**, 750-757

McConnell, J.F. (1974). *Cryst. Struct. Comm.* **3**, 73-75

Oliver, J.D. & Strickland, L.C. (1984). *Acta. Cryst.* **C40**, 820-824

Poyser, J.P., Edwards, P.L., Anderson, J.R., Hursthouse, M.B., Walker, N.P.C., Sheldrick, G.M. & Walley, A.J.S. (1986). *J. Antibiotics* **39**, 167-169

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1992). *Numerical Recipes* pp630-633, Cambridge University Press, Cambridge

Sayre, D. (1952). *Acta. Cryst.* **5**, 60-65

Schenk, H. (1973). *Acta. Cryst.* **A29**, 480-481

Schenk, H. (1974). *Acta. Cryst.* **A30**, 477-481

Sheldrick, G.M. (1990). *Acta. Cryst.* **A46**, 467-473

Suck, D., Manor, P.C. & Saenger, W. (1976). *Acta. Cryst.* **B32**, 1727-1737

Szeimies-Seebach, U., Harnisch, J., Szeimies, G., Van Meerssche, M., Germain, G. & Declercq, J.P. (1978). *Angew. Chem. Int. Ed. Engl.* **17**, 848-850

Wallwork, S.C. & Powell, H.M. (1980). *J. Chem. Soc. Perkin* **2**, 641-646

Williams, D.J. & Lawton, D. (1975). *Tetrahedron Letters* **No. 2**, 111-114

APPENDIX A

INTRODUCTION TO APPENDIX A

This appendix contains the manual produced for the program MITHRIL90 in October 1990. It contains a full explanation of all required commands and optional commands that can be used within the MITHRIL90 package. It shows the use of the two new options described in chapters three and four of this thesis. It also shows as new features added by Stephen Brown, Karine Lesley and Chris Gilmore.

MITHRIL90

A COMPUTER PROGRAM FOR THE AUTOMATIC SOLUTION OF CRYSTAL STRUCTURES FROM X-RAY DATA

VERSION 2.0 dated OCTOBER 1990

by

A. N. HENDERSON
C. J. GILMORE
S. R. BROWN

Department of Chemistry
University of Glasgow
Glasgow G12 8QQ
Scotland.

On every thought I have the countless shadows fall
Of other thoughts as valid that I cannot have;
Cross-lights of errors, too, impossible to me,
Yet somehow truer than all these thoughts, being with more power aglow.

May I never lose these shadowy glimpses of unknown thoughts
That modify and minify my own, and never fail
To keep some shining sense of the way all thoughts at last
Before life's dawning meaning like the stars at sunrise pale

Hugh MacDiarmid

Light and Shadow

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

1.0 AN INTRODUCTION

It doesn't.....so much matter where you begin the examination of a subject, so long as you keep on until you get round again to your starting point. As it were, you start on a sphere or a cube; you must keep on until you have seen it from all sides.

Ezra Pound - A B C of Reading

The 1970's were an exciting time for direct methods in X-ray crystallography. Groups in Europe and the U.S.A. were involved in producing a whole new set of ideas and concepts which were greatly extending the power of the method. Although there was undoubtedly a spirit of friendly competition between the groups involved, most workers carved out an area in which to work that did not overlap much with other groups. One consequence of this was that the practical implementation of new methods by other crystallographers began to lag behind the theory because no single computer program existed in which these ideas were implemented.

Professor Michael Woolfson organised a meeting of researchers in direct methods in January, 1978 at York University. This meeting was made possible by a NATO Research Grant. It was attended by T. Debaerdemaeker (Ulm), G. Germain (Louvain), C. Giacovazzo (Bari), C. Gilmore (Glasgow), H. Hauptman (Buffalo), S. Hull (York), P. Main (York), P. Rentzeperis (Salonika), H. Schenk (Amsterdam), G. Tsoucaris (Paris), D. Viterbo (Turin) and M. Woolfson (York). It was agreed to pool the efforts of the groups represented by these people, and attempt to produce a single computer program incorporating all the relevant new developments in direct methods. NATO continued to support this work by providing travel funds for groups to meet and discuss implementation problems, and Chris Gilmore took on the job of coordinating all the material into a computer program. The job was assisted by a grant from the S.E.R.C. and the University of Glasgow.

In 1988 it was decided that a new release of MITHRIL was required to bring the programme up to date with the theory developed during the 1980's. This work was done by two Ph.D. students, Stephen R. Brown and Allan N. Henderson both of whom worked with Chris Gilmore, and MITHRIL90 is the result of their efforts.

The manual is organised as follows. The next section (Section 2.0) gives a brief listing of the changes between MITHRIL and MITHRIL90. Section 3.0 gives a general view of the program, and a brief theoretical outline of the processes involved. These descriptions are necessarily sparse, and those unfamiliar with the topics under discussion will need to refer to the literature. Sections 5.0-17.0 are concerned with detailed descriptions of program input, and the remaining sections (18.0-24.0) deal with examples and suggestions on the methods used to solve difficult structures.

2.0 PRINCIPAL CHANGES IN MITHRIL90

In addition to bug-fixes, there have been several improvements made in MITHRIL since 1983, some were obvious omissions in the original program, e.g. a SYMBOLIC ADDITION module; some were made following advances in theory e.g. the calculation of the scale and temperature factors by a Bayesian technique, while others were made to improve the user friendliness of the program e.g. the REVIEW module.

There follows a list of the modules with a brief summary of the changes made. You are referred to the relevant part of section3.0, and the commands sections (5.0-17.0) for more detailed information on each change.

NORMAL:
Bayesian method of normalisation is added.
Powder diffraction data including overlaps are now processed.
The ability to normalise Electron and Neutron diffraction data is added.
Missing reflections can be inserted.
Calculation of errors on scale and temperature factors is added.

TRIPLETS & QUARTETS:
Both have cut-offs and weighting schemes based on sin θ/λ.

QUINTETS:
Removed with all subsquent references.

SYMB:
There is a new module to solve centro-symmetric structures by symbolic addition.

TANGENT:
Phase refinement is now performed 50% faster.
New figure of merit, LOGLIK.

REVIEW:
A new module to sort and output phase sets after tangent refinement is provided.

GENERAL:
Output and menus are largely in lower case.
All input is now case independent.
The ability to deal with larger structures with more reflections has been added.

3.0 MITHRIL90 - AN OVERVIEW

MITHRIL90 is a crystallographic direct methods program based on MULTAN80 (Main, Fiske, Germain, Hull, Declercq, Lessinger & Woolfson, 1980) with the incorporation of MAGEX (Hull, Viterbo, Woolfson & Shao-Hui, 1981), YZARC (Baggio, Woolfson, Declercq & Germain, 1978) and RANTAN (Yao Jia-Xing, 1981), quartet invariants, and symbolic addition. It contains numerous additional features all of which are outlined in this section.

The name of the program comes, like so much software, from Tolkien. Readers of "The Hobbit", "The Lord of the Rings" or "The Silmarillion" will know all about Mithril. It was the light, malleable and infinitely workable metal beloved of Dwarves and Elves alike. It was also unbreakable, and so can be considered a rather pretentious name for a direct methods program. However, the name MITHRIL is also an acronym for Multan with Interactive facilities, Triplet checking, Higher invariants, Random phasing, Intelligent control of flow and options and Linear equations phasing.

A flowchart of the program modules is shown in Figure 1. Sections 3.2 - 3.15 give a detailed description of the package. It is assumed here that the reader has a working knowledge of direct-methods, and in particular the MULTAN package. There is a useful summary by Main (1980).

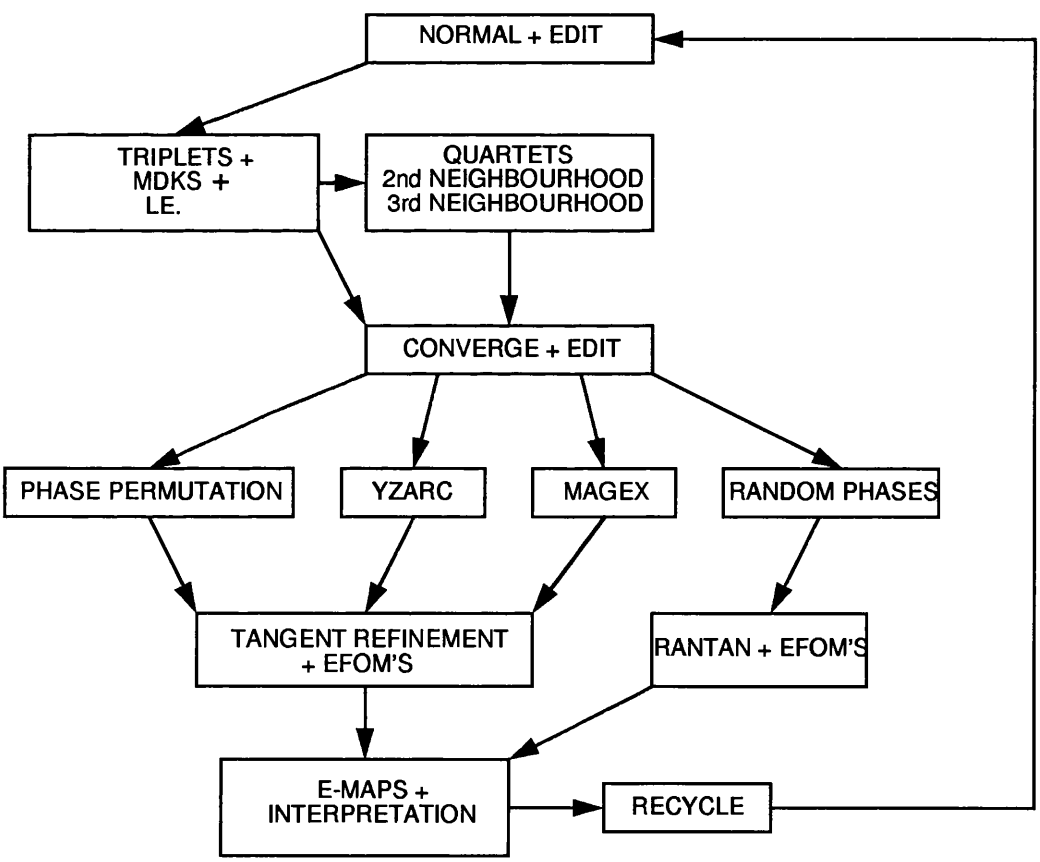


FIGURE 1. A flowchart of the MITHRIL90 program

3.1 The User Interface

Four levels of user interaction are provided, ranging from the batch mode to a full dialogue requiring user decisions at key points. This level is set dynamically and can be changed at any time, so that some modules may be run interactively and others under complete program control. The interactive modes are menu-driven with a separate menu for each module outlining available commands. Each instruction consists of a 4-character key word and a free-format list of input fields. Zero rather than blank fields generate default options. Each command has a sensible default value so that the user need only enter non-default options. Furthermore, there are four levels of default complexity, designed to cover all situations between the simplest and most difficult problems. These default levels are also dynamic and may be altered from module to module. Finally, the user need only specify those modules for which non-default options are to be used, the program will automatically run any modules that are needed but that have not been explicitly called by the user.

3.2 Normalisation

For detailed discussions of the normalisation process see Rodgers (1965, 1980) and Giacovazzo (1980a). The code used here is an extensive modification of the MULTAN80 program. The normalisation procedure is often neglected by crystallographers since it is so automatic, but decisions made at this point have drastic implications for subsequent steps in the analysis, and full control over the process may be needed. Accordingly, the package offers the following features:

(1) Equivalent reflections and systematic absences are removed. The former distort the phasing procedure, while the latter can render relationships which rely on small E-magnitudes incorrect. In this latter category are the Q triplets and the negative quartets. Missing reflections may be added to complete a data set using Wilson statistics to estimate the magnitudes of the missing E values.

(2) Allowance is made for a lack of knowledge of the detailed contents of the unit cell.

(3) Should the traditional K-curve or Wilson plot techniques prove inadequate, a new Bayesian method of normalisation has been introduced. In addition direct input of E-magnitudes and/or phases is permitted and the normalisation procedure is bypassed. This allows interfacing to other normalisation programs.

(4) Editing facilities are provided to remove or modify structure factors before normalisation, or to remove or modify the subsequent E-magnitudes. The former facility alters the normalisation process, whereas the latter modifies only some of the normalised structure factors. Allied to these provisions is a set of

optional theta limits and a maximum permitted E-magnitude. Experience has shown that E-magnitudes greater than approximately 3.5 can prevent the weighting schemes in a weighted multi-solution phasing environment from working with optimum efficiency as they tend to drive most weights to unity very rapidly. The relatively crude device of setting an upper limit can often remove this problem. The θ limits can be useful to exclude low-angle data which are subject to large systematic error, or the high angle reflections which can be very sensitive to small changes in overall temperature factor.

(5) Full control is provided over scaling and the temperature factor.

(6) The ability to input groups of known stereochemistry whose position and orientation in the unit cell can be either fixed or random (Main, 1976).

(7) The ability to normalise X-ray, electron, and neutron diffraction data.

(8) Normalise powders including overlapped reflections. The latter are used for normalisation but are not included in the list of reflections to be phased.

(9) An interface to the Maximum entropy program, MICE (Gilmore & Bricogne, 1990).

(10) Standard deviations are calculated for the scale and temperature factors; the method used depends on the normalisation method: Wilson or K-Curve techniques use the method of Hall & Subramanian (1982). Whilst in the Bayesian method the errors come from the underlying theory. Standard deviations are also calculated for each individual E-magnitude. It may be advisable to edit invariants involving more than one reflection with unusually large sigma(|E|).

3.3 Triplets

The Cochran distribution (Cochran & Woolfson, 1955) is used with the addition of two formulae for independently checking the triplet cosine. These are the MDKS formula (Hauptman, 1972) and a related technique (LE) in which a quintet extension of a triplet is used to derive a joint conditional probability distribution involving six E-magnitudes (Gilmore & Hauptman, 1985). The distribution is manipulated to give a system of 10 simultaneous linear equations in which the triple-phase invariant is one of the undetermined variables, and can be calculated in two ways. One is from the 10 simultaneous equations. Another method is via linear least-squares since three of the variables in the least-squares system are in fact the E-magnitudes involved in the triplet itself, and are thus known. The module requires that both estimates should agree within reasonable limits, otherwise both are ignored. Both MDKS and LE methods give only approximate estimates for the cosines, but they can be used to alter the relative weights of the triplets and for indicating which relationships may be troublesome.

The triplets are searched for those which give indications of the phases of one-phase seminvariants. If MDKS or LE has been used, an analysis of these triplets in terms of the estimated cosine is also given. This can be a useful adjunct in the decisions concerning the reliabilities of the Sigma-1 phases.

A cut-off, input as a fraction of the largest value of sin^2/λ^2 can be used to remove triplets with two or more reflections having a resolution greater than this cut-off. A weighting scheme based on sin^2/λ^2 may be used to down weight unreliable triplets that involve any high resolution reflections. (See Gilmore & Brown, 1988). This can be very useful for very high resolution data sets.

3.4 Quartets

Users unfamiliar with quartet theory are recommended to read a review by Hauptman (1980). Quartet invariants of the form:

phi{h} + phi{k} + phi{l} + phi{-h-k-l} = PHI{4}

are a very important component of MITHRIL90. Three types of quartet must be distinguished - that for which cos PHI{4} is estimated to be zero (a positive quartet), one in which cos PHI{4} is estimated to be 180 (a negative quartet), and the enantiomorph sensitive relationships between these two extremes. The negative quartets in particular are very useful. They can be used both as a figure of merit NQEST (DeTitta, Edmonds, Langs & Hauptman, 1975; Gilmore, 1977), and in an active mode to generate new phases (Freer & Gilmore, 1980). Two formulae are provided - the 7-magnitude, 2nd neighbourhood formula, and the more powerful 3rd neighbourhood, 13-magnitude formula (Hauptman, 1977a, 1977b). Missing members of these neighbourhoods are permitted, and the missing magnitudes are assigned values of unity. The use of the 2nd. neighbourhood formula is now widespread, but the 3rd. neighbourhood formula is still somewhat neglected.

Let us define the quartet which we wish to estimate as follows:

phi = phi{h} + phi{k} + phi{l} + phi{m} (1)

such that $\underline{h} + \underline{k} + \underline{l} + \underline{m} = 0$

The first neighbourhood consists of E{h},E{k},E{l} and E{m}, whilst the second neighbourhood adds the three cross terms E{h+k},E{k+l} and E{l+h}. The addition of the third neighbourhood is accomplished by introducing an arbitrary vector p and its associated vector q such that:

$$\underline{h} + \underline{k} + \underline{p} + \underline{q} = 0$$

It is necessary that E{p} and E{q} are 'large'. We now have a second quartet invariant:

phi{pq} = phi{h} + phi{k} + phi{p} + phi{q} (2)

We have also indirectly defined a third quartet invariant:

phi{lm} = phi{l} + phi{m} + phi{-p} + phi{-q} (3)

Quartet (2) has a second neighbourhood comprising:

E{h}, E{k}, E{p}, E{q}, E{h+k}, E{k+p} and E{p+h}

Quartet (3) has a second neighbourhood comprising:
E{l}, E{m}, E{p}, E{q}, E{l+m}, E{m-p} and E{-p-q}
However, an identity exists between (1), (2) and (3), such that:

phi + phi{pq} + phi{lm} = 0 (4)

so that phi can be estimated not only by its own 7-magnitudes comprising its second neighbourhood, but also by the two invariants (2) and (3). A total of 21 magnitudes are now involved in the estimation of phi, of which only 13 are unique. These 13-magnitudes define the third neighbourhood. The three quartets (1), (2) and (3) define a trio.

It is possible to construct a joint conditional probability distribution of the pair of structure invariants phi and phi{pq} given these 13 unique magnitudes. This is the P{2/13} distribution of Hauptman (1977a). It is possible to extract from this the P{1/13} distribution which gives phi as a function of 13 E- magnitudes.

Clearly, the third neighbourhood is not unique. There is a multiplicity of third neighbourhoods, each giving rise to an estimate for f, and each estimate having its own variance. Direct methods will only allow us to use one of these. The best way to handle this situation is to use that estimate which has the lowest associated variance, at the same time discarding any invariants for which there is disagreement between the third neighbourhood calculations. It must also be stressed that the individual phi estimates are not independent because they will have many terms in common. The 3rd neighbourhood formula also goes some way towards alleviating the 1/N dependence of quartet reliability (N is the number of atoms -assumed equal- in the unit cell). It can also be very useful with weak or limited data sets (Gilmore, Hardy, MacNicol & Wilson, 1977).

In practical terms, it is necessary to restrict the vectors p and q to span the top 50-100 E-magnitudes, otherwise the calculations consume considerable amounts of computer time. There is very little loss of accuracy in doing this.

Positive quartets may also be generated if requested. They can sometimes be useful in situations where there is a deficiency in the number of triplets available, e.g. powder diffraction data. Positive quartets are correlated with triplets (Giacovazzo, 1980b) and this correlation is dependent on the E's

involved. It is handled in the way described by Freer and Gilmore (1980). Higher invariants are put on the same scale as triplets using the concept of an equivalent k which is obtained from the variance of the appropriate probability distribution. (Freer & Gilmore, 1980).

The negative quartet module is always called in situations where the space group is symmorphic. They are also generated in cases where difficulty in solving the structure is indicated by the user. As in triplets, there is an optional sin^2/λ^2 cut-off which eliminates all quartets that involve two or more reflections above a user specified limit, and a weighting scheme to down weight quartets that involve reflections with any reflections having a resolution above a specified limit.

3.5 Convergence Mapping

The convergence mapping module follows invariant generation. It performs two quite distinct functions:

(1) The collection of invariants. All the invariants generated by previous modules are loaded for active use in the phasing procedures which follow, although it is possible to exclude higher invariants. Triplets are optionally weighted via their MDKS or LE cosine estimate, if they are available. A rather simpler method than that used by Bussetta and Comberton (1974) is employed. The invariants phi{3} are split into 4 classes:

(i) cos phi{3} >= 0.7
(ii) 0.7 > cos phi{3} >= 0.0
(iii) 0.0 > cos phi{3} >= -0.7
(iv) cos phi{3} < -0.7

Each class (i)-(iv) is assigned a weight greater than or equal to zero by the user, and this is used to multiply the kappa value for the relationship, where:

kappa = 2 E{$\underline{h}$} E{$\underline{k}$} E{$\underline{h}$-$\underline{k}$} / root(N)

(E{$\underline{h}$}, E{$\underline{k}$}, E{$\underline{h}$-$\underline{k}$} are the E's involved in the triplet).

By giving relationships in class (i) weights greater than unity, they can be upweighted and play a larger role in the phasing; in a similar way, those in class (iv) can be down-weighted or removed completely. The MDKS and LE formulae are unreliable in the two remaining classes and unit weights are usually used here. It is worth emphasising that this weighting scheme radically alters the phasing path as decided by the convergence method, and even if the MDKS estimates are unreliable, the resulting convergence map may be sufficiently different from the original that previous problems may disappear. In particular, unreliable triplets which appear early in the convergence map, even if not detected by the MDKS/LE tests, may now appear in less critical phasing areas.

Specific relationships may be deleted from or added to the list. Any 2-, 3- or 4-phase invariant or seminvariant may be added. This can be particularly useful as an adjunct to symbolic addition (Karle & Karle, 1966) where indications of relationships between symbolic phases or of potentially unreliable invariants can emerge. This information can now be supplied to the multisolution process in a simple way. The user must estimate the reliability of any input relationship by supplying a kappa value.

(2) Convergence mapping itself. Once the relationships have been suitably ordered, the convergence map chooses suitable reflections which define the origin, the enantiomorph (if relevant), the known phases and the permuted phases. The user is provided with the usual options to define or partially define origin and enantiomorph, the permuted reflections, the maximum and minimum number of phase sets and the values of any known phases. Three criteria are applied in the acceptance of the Sigma-1 results:

(i) A minimum probability.
(ii) A minimum number of indications (= (N+) + (N-),
where N+ is the number of zero indications and N- is the number of 180 degree indications)
(iii) A consistency ratio (c) defined as:

$$c = \max(N+,N-) / ((N+) + (N-))$$

Caution in accepting a Sigma-1 determined phase cannot be overemphasised.

Care is also needed with the active employment of quartets. It is obvious that with a triplet, two known phases can derive a third, whereas for a quartet, three are required to give a fourth. If the phase angles are very approximate, as they are in the early stages of direct methods, quartets will tend to propagate errors more than triplets. Accordingly, the number of higher invariants should be a maximum of ca. 20% of the total number of invariants used, although this rule may be broken for powder data sets.

At this point the negative invariants are assembled for use in NQEST. It is important to use reliable invariants, and thus only relationships with an equivalent k greater than 1.0 are accepted. A minimum of 25 invariants is needed for NQEST. The ψ_o figure of merit is also set up at this point, and again there is a minimum number of relationships needed for this figure of merit to be invoked.

The reflections which are eliminated last in the convergence procedure are those which will be phased first in the tangent refinement module. In other words, the map represents an inverted phasing path which is stored, and used by all subsequent modules. In difficult circumstances, it is necessary to examine this map carefully, and it should always warrant more than just a cursory glance.

Reading the convergence map output is straightforward. A typical line of output looks like this:

```
27 3 0 5 12.3        12 -4 0 12 6.5 17 -6 8 0 4.8
                     -4 18 1 0 1.7 1 4 88 12 1.8
```

The reflection to be phased at this point is number 27 with indices 3 0 5. It has an alpha value of 12.3 at its point of elimination. Four relationships are shown as contributors. The first is a triplet:

Phi{12} - Phi{4} + (12/24) = Phi{27}

with a kappa-value of 6.5, and a phase shift of (12/24) x 360 = 180deg. The second is a quartet involving reflections 17, 6, 8 and 27; the phase shift is zero and the kappa-value is 4.8. The next relationship is also a quartet, involving reflections 1, 4, 18 and 27. This is followed by another triplet.

The phases to be permuted are listed at the end of the convergence output, along with origin and enantiomorph definitions. Permutation of general phases is carried out using magic integers, where each general phase (one with no phase restriction) is defined by:
Phi = M{T}X

M{T} is the magic integer, and X is a variable which takes on different values at equal intervals in the range 0-360 deg. The interval of X is chosen to make the r.m.s. difference from one set to the next equal to the r.m.s. error in the 'best' phase set.

The user may specify the number of general reflections (Ngen), and the number of special reflections (Ns) the starting set. General reflections have no phase restrictions; special reflections have only two possible values. The number of starting sets so generated can be calculated by:

No. of sets = Nsets * 2 Ns * Ef

Nsets is the number of phase sets produced from the general reflections via magic integer phase permutation (Main, 1978). Ef is an enantiomorph factor. If the magic integer variable is set at 2 (the default - see Section 9.0), Nsets can be derived from the number of general reflections in the starting set (Ngen) from the following table:

Ngen	0	1	2	3	4	5	6	7	8	9	10
Nsets	1	4	14	18	30	48	80	130	214	348	568
Rms Error	0	26	28	39	42	45	47	48	49	50	50

The value of Ef depends on how the enantiomorph is fixed:

= 1/2 if an unknown phase defines the enantiomorph - the range of permutation is restricted to 1/2 the phase circle.

= 1 if the enantiomorph is defined by restricting the origin defining phases to fixed values or the enantiomorph is defined by the space group.

= 2 if the enantiomorph is defined by restricting the origin defining phases to two sets of values-one phase must take on two values.

= 4 if the enantiomorph is defined by restricting the origin defining phases to four sets of values-two phases each take on two values.

3.6 Symbolic Addition

MITHRIL90 has a symbolic addition module, called SYMB derived from LSAM (Germain & Woolfson, 1968). Many modifications have been made including the use of quartets and new more powerful figures of merit.

SYMB accepts all triplet and quartet phase relationships (up to a maximum of 2000) which have probabilities greater than a minimum set by the user. These probabilities are calculated using the Cochran & Woolfson(1955) formula

$$P = 1/2 + 1/2 \tan(\text{kappa}/2)$$

where kappa was defined in section 3.5
Using program assigned symbols and origin reflections, signs are developed by symbolic addition (Karle & Karle 1966). Starting set signs are assumed true (i.e. Prob. = 1.0). Sign indications are then computed for all other reflections and the new reflection that is indicated with the highest probability is accepted. This is then used in another round of sign indication. In this iterative fashion all signs with a probability greater than a set minimum are found. When sign determination is done SYMB checks the consistency of the signs by rejecting all those with probabilities less than a specified cut-off and re-determining the signs that remain. This continues until there is no change in sign from one re-determination to the next. Unlike many symbolic addition programs, SYMB allows for the failure of individual triplets.

Note that SYMB can only be used for centro-symmetric structures, and is most useful in case where tangent refinement causes phases to become over consistent, particularly heavy atom structures.

3.7 Yzarc

Usually one proceeds directly to tangent refinement from convergence mapping, but MITHRIL90 offers two additional modules which can be run between CONVERGE and TANGENT. The first of these is YZARC. It uses sets of random phases as a starting point, and refines them via least-squares or

steepest-descents. Users unfamiliar with these concepts should first read two papers - Baggio, Woolfson, Declercq and Germain (1978) and Declercq, Germain and Woolfson (1979). Neither of these papers discuss the steepest descents method. It is sufficient to state here that standard steepest-descents algebra is used with appropriate weighting schemes. Steepest descents and least squares will, in general, produce different phase sets even when all other conditions are the same.

Usually only a subset of reflections is phased, and these phases are passed to tangent refinement. Normally some 50-100 different sets are processed in this way. Some of the facilities offered in MITHRIL90 differ from the standard YZARC procedure:

(1) All the relationships collected during convergence are used. The inclusion of quartets alters the refinement; sometimes it improves the radius of convergence, but sometimes it does not.

(2) The problem of when to stop refinement has always been difficult in YZARC. An alternative method based on NQEST is offered as an option. After an initial round of n cycles, NQEST is calculated. If at this point the NQEST figure of merit is greater than a specified cut-off (e) refinement is terminated, otherwise it continues. It checks this figure of merit after each cycle, and refinement continues whilst it continues to fall until it hits a minimum at which point it stops. Experience dictates that suitable values for n and e above are 7 and 0.0 respectively, but they are user options. This method has the effect of reducing the number of refinement cycles - usually by a factor of six - with a commensurate fall in the computer time required. As with the traditional techniques of stopping YZARC refinement, it is not always successful.

(3) YZARC normally phases the bottom 100 or so reflections from the convergence map. An option is provided whereby the convergence map is bypassed except for origin definition, and the top reflections ordered on E-magnitude alone are phased. This usually creates a singular matrix if least-squares is employed so that steepest descents is normally used. This can be useful in circumstances where the convergence map selects a subset of highly linked reflections in which certain parity groups are not represented. Bypassing the map can phase a more representative set of reflections.

(4) The final figures of merit are augmented by the inclusion of NQEST.

3.8 Magex

MAGEX is another module run between convergence mapping and tangent refinement. Users unfamiliar with the concepts employed here should first read two papers - White and Woolfson (1975) and Declercq, Germain & Woolfson (1975). These deal with a three-dimensional form of MAGEX - the MAGIC procedure. A paper by Hull, Viterbo, Woolfson and Shao-Hui(1981) shows how the method is adapted to the one-dimensional case employed here.

MAGEX has been inserted in the MITHRIL90 package without any major modifications. Higher invariants are not employed here except that an NQEST value is assigned to each possible solution just as in YZARC.

3.9 Tangent Refinement

The tangent formula carries out expansion and refinement of phase angles. The following options are provided:

(1) All relationships accepted by CONVERGE are used actively to generate new phase angles.

(2) Three early figures of merit - the two MULTAN80 options of ψ_o combined with R{Karle} and ψ_o alone to which has been added NQEST. Any combination of these three can be specified. Early figures of merit are useful when it is proposed to generate large numbers of phase sets. There is a link here with SHEXTL which relies very heavily on NQEST to promote efficiency. These figures of merit can also be used to filter out unacceptable phase sets from YZARC and MAGEX after only a few cycles of tangent refinement and expansion.

(3) Two weighting schemes as in MULTAN80 - the traditional MULTAN scheme and Hull - Irwin statistical weights (Hull & Irwin, 1978). The latter is useful in cases of pseudo-symmetry, over consistent phase sets (as demonstrated by very low R{Karle} values, and very high ABSFOM's), symmorphic space groups, and heavy-atom cases.

(4) In its interactive mode, the program will attempt to identify any solution with exceptional figures of merit. It is possible to stop at this point, compute an E-map then continue with tangent refinement if the map seems unpromising.

(5) Four figures of merit: ABSFOM, ψ_o, R{Karle}, and NQEST, combined together with user controlled weights give a single figure of merit and this is used to rank the solutions. The R{Karle} is based on alpha values.

(6) A new figure of merit, LOGLIK, based on the agreement between observed and calculated magnitudes of unphased reflections has been implemented. It is based on the prediction of an E-magnitude from triplets, and the FOM keyword in the triplet call will allow calculation of this figure of merit.

(7) Both tangent and Hull-Irwin weighted tangent refinement is now performed without disc i/o to give a 50% reduction in the time taken for the refinement of one phase set, at the cost of increased memory requirements.

3.10 RANTAN

The RANTAN module offers all the facilities provided by the tangent refinement module discussed above. It differs from the latter in that it does not use phase permutation, but assigns all unknown phases random values. There is thus no phase expansion, but only refinement. The phases are refined to convergence in the usual way. The theory of this technique is discussed by Yao Jia-Xing (1981). As in section 3.9 the LOGLIK figure of merit is available and run times are decreased by 50%.

3.11 Review

The new interactive module, Review, has been added to MITHRIL as a direct consequence of the experience gained in solving difficult structures. For some structures it has been found necessary to generate many phase sets (up to 2000 have been produced). To facilitate in the selection of a correct phase set MITHRIL90 will now sort the sets and rank them according to one of six figures of merit.

3.12 E-Maps

The final step of a direct methods analysis is the calculation of one or more E-maps. It is E-map interpretation that the high level of user interaction which can be provided becomes important, particularly when there are many maps to search. It is possible to generate and search all the appropriate maps quickly, and reject unsuitable solutions on the basis of peak heights, peak positions or fragmentation patterns before attempting a full interpretation. Simple graphics facilities are also provided for those users with graphics terminals.

3.13 Recycling

Four methods of recycling are provided:

(1) Weighted Fourier syntheses (Sim, 1959, 1960).
(2) Karle recycling (Karle, 1968).
(3) Karle recycling with random phases for the unphased reflections (Yao Jia-Xing, 1983).
(4) Groups of correct orientation but known or unknown position (Main, 1976).

3.14 Patterson and Vector Maps

Provision is made for the calculation of an E^2 -1 vector map. User control of the sharpening procedure is possible. Examination of such maps can be useful in difficult cases (Nixon, 1978).

3.15 Graphics

A program, PLOTQ, is also provided. This stand-alone program reads in the maps produced by MITHRIL90 after the MAPS module has been run (providing they have been saved as peranent files). It allows the user considerable scope for viewing the electron density map, and may be invaluable for defining molecular boundaries and placing fragments.

4.0 DATA COLLECTION

It may seem strange to include a discussion of data collection, but all direct methods calculations depend critically on the quality and resolution of the intensity data. These factors become more and more important as structural complexity increases or in cases of pseudo-symmetry, structural regularity (such as fused six-membered rings) or high thermal motion in the crystal. It is therefore important to plan your data collection with care especially if any of these features are expected. In particular:

(1) Get your intensities as accurately as possible with at least some equivalent reflections which are subsequently merged to give a unique data set. Be suspicious of any discrepancies.

(2) Measure weak reflections with a similar accuracy to the strong ones. Weak reflections generate small E-magnitudes and these play a critical part in the quartet and Q relationships which have a vital role in the MITHRIL90 package. Some diffractometer software places great reliance on pre-scans of peaks before measuring the intensity accurately. Weak peaks are then ignored. This increases efficiency but if the small E-magnitudes are only measured in a pre-scan, then they will have very large standard deviations and the relationships which use them will be unreliable. So make sure you input data collection parameters which will override the pre-scan options.

(3) The higher the resolution the better. The Cu sphere should be considered the minimum although for some crystals this is just not possible because of poor diffraction quality. The program does its best in these circumstances, but the chances of success are reduced. Sheldrick's rule quantifies this as follows "If less than 50% of the reflections in the resolution range 1.1-1.2Å are observed (i.e. $|F|>2.0\sigma(|F|)$), then the structure will be difficult to solve by conventional direct methods". Two exceptions to this rule are very small structures and heavy atom structures. Be careful with very high, observed, weak reflections which become very large E-magnitudes. This is especially likely with high intensity tubes. This problem can be dealt with by the use of triplet and quartet weighting and cut-off techniques. These are discussed in more detail in the relevant sections. Another problem of unusually large E's can be caused by solvent, which may produce very large E's at very low resolution. It is found that it is best to remove these reflections from a data set.

5.0 MITHRIL90 COMMANDS - A GENERAL INTRODUCTION

In the remainder of this manual and in the program itself the following nomenclature is used:

(1) Any text or part of a command text in lower case is optional e.g. NORMal means that only NORM is required but that NORMA or NORMAL is accepted.

(2) All commands appear in **LARGE, BOLD, CAPITALS**

(3) An essential blank is written as an underscore

MITHRIL90 commands consist of four or more characters which begin in the first column of each line, they may be in upper or lower case. Only the first four characters are significant, the rest are ignored. The parameters, if any, then follow on the same line in free format. Only columns 1-72 are scanned. The parameters are separated by blanks or commas. To get a default insert a parameter of zero - not a blank. This is important when entering parameters for those commands which allow more than one.

For example, in the module NORMAL the command LIMIt can be used. It has three parameters - sin theta maximum, sin theta minimum, and maximum permitted E-magnitude. If it is desired to use the first two parameters with their default values, but to specify a maximum E of 3.5 the command:

LIMIt 0,0,3.5

must be issued. The theta limits are given their default values and E-max is set to 3.5. Note that the command:

LIMIt, 3.5 or LIMIt,3.5 is not acceptable.

If the fields fill the line, continuations are possible by using '='. The rest of the line is then ignored and another line is read. This should be blank in columns 1-4. As many continuations as needed may be used, but there is a limit of 200 parameters in total for any command. Example:

```
PHASe 1 3 4 5 6 7 8 9 11 23 34 45 46 56 102 =
      104 119
```

Some commands require keywords instead of, or in addition to the numerical parameters. These keywords can be placed anywhere on the line. Only the first two characters are significant. E.g.

SIGMa ALL where 'ALL' is a keyword.

Comments can be inserted on a command by using '!'. Everything that follows this symbol is ignored.

There are three different types of command:

(1) The general commands which alter the mode in which the program is run - TITLE, END, MENU, LEVEL NOPRINT, PRINT, DEFAULT, HARD, VERY_HARD, MODEL SHOW, X.

These can be entered as often and whenever desired.

(2) Those which call modules -

NORMAL, TRIPLETS, QUARTETS, CONVERGE, YZARC, MAGEX, SYMB, TANGENT, RANTAN, REVIEW, MAPS, RECYCLE, PATTERSON

These commands cause the relevant module to be entered.

(3) Those which are particular commands for a given module only.

5.1 The General Commands

TITLe Enters a title which is printed at the top of each page. The default is a blank line. Continuations ('=' sign) are not allowed with this command.

END Tells the program that no further input will follow. The current module will be run to completion, and the package will then stop.

MENU In the interactive modes (LEVEL = 1,2,3) the screen is cleared and the current menu is displayed. If LEVEL = 0 the command is ignored.

NOPRint Switches off output to the print file (but not the secondary file used under LEVEL 0). This is useful if you are re-running a job for which the output has already been printed. The PRINt command turns the printing on again. You can use these commands as often as necessary.

LEVEl N The parameter N is an integer which can have the values 0, 1, 2 or 3. This signifies the degree of user interaction required. The levels operate as follows: N=0 This is used for batch jobs. No real-time user interaction is expected. No menus are displayed. A secondary print file is created which contains a list of the input commands as they are executed, the error and warning messages, and a summary of progress through the program.

N=1 This is the lowest level of real-time operation. The user is expected to be sitting at a terminal. Menus are displayed on the screen, modules are run in real-time as selected by the user. A limited summary of progress and results appears on the screen. There is no secondary print file.

N=2 This gives the same as N=1 but the screen output is more detailed and interrogation of the user may occur in some modules.

N=3 This expects a high degree of user-machine interaction. For example, the screen is never cleared without user permission, tangent refinement stops after each solution with a request to continue, the MAPS module becomes more interactive. All the facilities of N=2 are also included.

The default will depend on your installation (at Glasgow we use N=2). Remember that the LEVEL parameter is quite flexible and can be changed as you proceed through a task.

DEFAult This removes further control of the package from the user, and runs the program to completion from the point at which the command is issued. Default parameters are used in all the subsequent modules, and the HARD and VERY_HARD options are still relevant. The default flow through the package is as follows: (1) For structures with translational symmetry (non-symmorphic):

NORMAL - TRIPLETS - CONVERGE - TANGENT - MAPS

(2) For structures without translational symmetry (symmorphic) or for which the commands HARD or VERY HARD have been issued:

NORMAL - TRIPLETS - QUARTETS - CONVERGE - TANGENT - MAPS
(3) For weighted Fourier recycling:

NORMAL - MAPS

(4) For Karle recycling of the non-symmorphic structures:

NORMAL - TRIPLETS - TANGENT - MAPS

(5) For Karle recycling of the structures of type (2) above:

NORMAL - TRIPLETS - QUARTETS - TANGENT - MAPS

(6) For situations where MAGEX or YZARC is being run:

NORMAL - TRIPLETS - QUARTETS - CONVERGE - MAGEX (or YZARC) - TANGENT - MAPS

or, in non-symmorphic situations with standard defaults:

NORMAL - TRIPLETS - CONVERGE - MAGEX (or YZARC) - TANGENT - MAPS

(7) RANTAN behaves just like TANGENT for default paths.

So, for example, if you have a symmorphic space group, and during or after running the triplets module you issue the command DEFAULT the modules QUARTET, CONVERGE, TANGENT and MAPS will be run under appropriate defaults, and the program will then halt. The modules MAGEX and YZARC are never called by DEFAULT, they are always user called.

HARD This gives a new level of defaults for structures which are proving difficult. It has the effect of generating more triplets; it automatically calls the quartets module with the third neighbourhood option; no reflections are rejected by converge; both MAGEX and YZARC phase larger starting sets; there is no automatic stopping during tangent refinement; more than one E-map is produced. It is possible to return to the standard level of defaults by entering the command HARD a second time.

VERY hard This extends the options of the HARD command further by including the generation and use of positive quartets. Both commands make considerable demands on computer time and need to be used with care. It is possible in times of great need to enter both commands. This results in the cumulative effect of both options. It is possible to remove the VERY HARD option in the same way as for HARD.

MODEl At Glasgow, each structure has associated with it a 'MODEL' file in which the unit cell parameters, symmetry, lattice type, cell contents etc. are stored as MITHRIL90 compatible commands. The command MODEL is usually issued in NORMAL and it causes the MODEL file to be scanned, and any relevant information extracted. This saves entering the information directly from the keyboard. It is also possible to store MITHRIL90 commands on this file. Under these circumstances, the instruction MODEL will cause the commands to be executed before returning control to the user.

SHOW When operating interactively, it is possible to forget which options have been selected via PRINT and NOPRINT, or whether HARD and/or VERY HARD options are currently in operation. The command SHOW displays these parameters on the screen for an interactive job, and on the secondary print file for batch jobs (LEVEL 0).

X___ Gives an immediate abort from the package.

5.2 The Commands Which Call Modules

The following commands call modules:

NORMal, TRIPlets, QUARtets, CONVerge, SYMB, MAGEx, YZARc, TANGent, RANTan, REVIew, MAPS, RECYcle.

In the interactive mode, calling a module will cause the terminal screen to be cleared, and a menu appears outlining the available commands which may be entered in any order. If you make a mistake with any command just re-enter it. The modules do, however, need to be run in a fixed order, and this is outlined in the table below. (The diagram shown in figure 1 may be useful here).

MODULE	MODULES WHICH MUST BE RUN BEFORE THIS MODULE (An * signifies an optional module call)
NORMAL	NONE (This is the package entry point)
TRIPLETS	NORMAL
QUARTETS	NORMAL, TRIPLETS
CONVERGE	NORMAL, TRIPLETS, QUARTETS*,
MAGEX	NORMAL, TRIPLETS, QUARTETS*, CONVERGE
YZARC	NORMAL, TRIPLETS, QUARTETS*, CONVERGE
SYMB	NORMAL, TRIPLETS, QUARTETS*, CONVERGE*
TANGENT	NORMAL, TRIPLETS, QUARTETS*, CONVERGE
RANTAN	NORMAL, TRIPLETS, QUARTETS*, CONVERGE
REVIEW	NORMAL, TRIPLETS, QUARTETS*, CONVERGE, TANGENT
MAPS	NORMAL, TRIPLETS, QUARTETS*, CONVERGE, TANGENT

The consequence of this requirement, is that modules are called automatically with suitable defaults whenever the user enters a sequence of commands in which one or more of the necessary modules are missing. For example, the sequence:

NORMAL - TANGENT - MAPS

has TRIPLETS and CONVERGE missing. (QUARTETS is also missing if the structure is symmorphic, hard or very hard), so these modules are run with the appropriate default options in between NORMAL and TANGENT.

The sequence:

NORMAL - QUARTET - MAGEX - DEFAULT

will run NORMAL under user control, the TRIPLETS under default, CONVERGE under default, and MAGEX under user control. The command DEFAULT then invokes the flow already outlined on page 15, and the modules TANGENT and MAPS are then run under default. Note that although all these examples enter the package via NORMAL, the program can be entered at any point provided that the relevant modules have been run on a previous occasion and the necessary files have been kept (See Section 21.0)

It is possible to go backwards as well as forwards in the program. When this is done the program goes directly to the module which has been called, it does not go through the remaining modules first. All the modules can be re-entered except TRIPLETS in which the MDKS option is invoked. E.g.

NORMAL - MAGEX - TRIPLETS - MAPS - YZARC - DEFAULT

will run the modules NORMAL, TRIPLETS, QUARTETS (if appropriate), CONVERGE, TANGENT, MAPS, YZARC, TANGENT, MAPS.

It is possible to override the default flow, if you really know what you are doing. Any module, except NORMAL, can be called with a "-1" as the first parameter. The module is entered, the program is informed of this, but the module is not actually run. One of the main uses of this is with the QUARTET module where it is possible to prevent quartets being generated in symmorphic space groups by using the command:

QUARTET -1

After all the commands for a module have been entered, one of the following operations will cause the module to be run:

(1) A call to another module. The current module is run and the new module entered. Calling the same module as the current one has two possible effects. In the interactive modes (LEVEL 1-3) the module is abandoned without running it, and re-entered. All the options must be re-typed. In the batch mode (LEVEL set to 0) the module is run, and then re-entered.

(2) The commands END or DEFAult (See pages185 and 186).

(3) If the LEVEL parameter is 1, 2 or 3, a carriage return or blank line. (If LEVEL is zero, blank lines are ignored on input)

6.0 NORMAL

The normalisation module provides the entry point into the MITHRIL90 package. It must always be the first module run in an analysis. The following commands are available; they can appear in any order except the DATA instruction which must come last.

NORMal IK, NB, ISC, MAXDUP [PHase] [NOsigma]

IK = -2 read E's from the Sheldrick data base
= -1 for do not normalise input structure factors. This is used when inserting one's own set of E's.
= 0 for use Wilson plot.
= 1 for use K-curve.
= 2 for use Bayesian method.

NB is the number of points to use in the Wilson plot. The default is

8*log [0.05 * Max(No. of reflections, 100)]

ISC = 0 for one scale factor per parity group. = 1 for use one overall scale factor.

MAXDUP. The NORMAL module checks for duplicates and systematic absences. Any that are found are output on the secondary print file (batch job) or the terminal (interactive job). Only the first MAXDUP found are listed. The default is 50.

[PHase] A keyword to indicate that phases are to input along with h, k, l, F, sigma(F).

[NOsigma] A keyword to prevent the calculation of errors on the temperature and scale factors, and to signal that sigma(F) is omitted

LIST

Causes a full list of E's (and F's on an absolute scale) to be printed. Only use this if you really need it.

CELL A, B, C, alpha, beta, gamma or A, B, C, $\cos\alpha$, $\cos\beta$, $\cos\gamma$

Enters the unit cell parameters. Any missing or zero angles are assumed to be 90 deg. So for the orthorhombic system, for example, it is only necessary to enter the cell edges.

SYMM

A symmetry operation as written in International Tables Vol. I as far as possible,

although "-" is needed instead of a bar. One command is needed for each operation. The identity operation is not required, and is ignored if input. Operations arising from a centre of symmetry at the origin, or from lattice centerings must not be entered. Example of a valid symmetry command:

SYMM -X, 1/2+Y, Z

In the interactive mode, it is possible that the user detects an error in the symmetry operations whilst entering commands. In this case the instruction:
SYMM
without any associated fields will cause all the previous symmetry operations to be removed, and the symmetry information can then be re-entered.

LATTice A / C P / A / B / C / I / F / R

The first field defines whether the space group is centrosymmetric (C) or not (A). The default is C. The lattice type then follows. The default is P, so if you have a primitive, centrosymmetric space group then this command can be omitted.

SFAC ATOM TYPE, a 1, a 2, b 1, b 2, c, COVALENT RADIUS, VAN DER WAALS RADIUS.

Inputs a scattering factor curve of the form:

f = a{1} * exp(-a{2}ro^2) + b{1} * exp(-b{2} ro^2) + c

where ro = (sin theta)/lambda. (Moore, 1963).

Note the use of the five-parameter form here as opposed to the more usual nine-parameter form. The program has stored scattering curves for the following elements on a file MITHRIL.DAT, an ASCII file which maybe edited if desired, and is read in by NORMAL.

H, Li, Be, B, C, N, O, F, Na, Mg, Al, Si, P, S, Cl, K, Ca, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Se, Br, Rb, Sr, Zr, Mo, Ru, Rh, Pd, Ag, Cd, Sn, Sb, I, Cs, Ba, W, Os, Pt, Au, Hg, Pb, Bi.

You only need to enter an SFAC command if you have an atom type not in the above list. Quite often you can approximate to the nearest atom in atomic weight without any serious lack of accuracy as far as direct methods is concerned. The covalent and van der Waals radii are used by the MAPS module, and they can be omitted if desired. They are input in Angstroms.

CONTents ATOM TYPE(S), NUMBER IN THE UNIT CELL

This enters the unit cell contents (not the asymmetric unit). A list of symbols and the number of this type of atom in the unit cell must be included. It does not

matter if the symbols come first or the types, the module associates the first symbol with the first numeric field, the second symbol with the second numeric field etc., so the following are all permitted:
CONTENTS 120 44 180 C O H
CONTENTS C 120 O H 44 180 etc.

If the CONTENTS command is missing, a cell in which only CH{2} groups are present, and for which the crystal density is 1.25 g/ml is assumed. Certain default options will be slightly modified in a pessimistic way if this is done e.g. the MAPS module will give you an extra 10 peaks. If an atom type is named which does not have a corresponding atomic scattering curve stored in the MITHRIL.DAT file (see SFAC) then an SFAC command will be needed for this atom.

LIMIts $(\sin\theta^{max})/\lambda$, $(\sin\theta^{min})/\lambda$ MAXIMUM PERMITTED E.

Any reflection with Bragg angles outside these limits is rejected. The defaults for the first two parameters are 0.0 and 1.0. The default for the third is 8.2, which represents a programming induced limit. Very large E- magnitudes can prevent a successful direct methods attempt, and this parameter merely sets any E greater than the maximum to the maximum itself. Values around 3.5 are usually successful.

NEWE H, K, L, NEW E-MAGNITUDE

This command enables the user to alter the magnitude of the specified normalised structure factor. If the new magnitude is given as zero, then the E is removed. Note that this takes place after the normalisation procedure, so the structure factor is used in the usual way in the calculation of the scaling parameters. Note also that the h, k, l indices must refer to the final E as output by NORMAL. This may have been symmetry transformed from its input form. A maximum of 100 NEWE commands are accepted.

MISSing

Find missing reflections, and, using Wilson statistics, calculate an expected F-magnitude, insert this reflection into the data, and re-normalise.

BSCL TEMPERATURE FACTOR, SCALE FACTOR(S)

This command allows the user to input his own temperature factor or scale factors. If only the temperature factor appears, then NORMAL calculates its own scale factor or factors depending in the ISC parameter on the NORMAL call. If the scale factor(s) are included on this command, then the temperature factor must be supplied as well. If it is not then a default of 4.0 is used. It is not possible to specify scale factor(s) alone and ask NORMAL to compute the temperature factor. If individual scale factors have been requested for each

parity group and a single scale factor is input, this scale is applied to all the parity groups.
EDIT H, K, L, NEW F

This works like the NEWE command, except that it operates on the structure factors as input before normalisation, so the indices must refer to the input indices. A zero structure factor magnitude will cause the input reflection to be deleted. A maximum of 100 EDIT commands are accepted.

TRANs TRANSFORMATION MATRICES

This command allows the input reflections to have their indices transformed from their old values of h, k, l to new ones h', k' and l' as follows:
h' = x.h + x.k + x.l + x 1 2 3 4
k' = x.h + x.k + x.l + x 5 6 7 8
l' = x.h + x.k + x.l + x 9 10 11 12

The elements x{1} to x{12} are input in that order. If less than 12 elements are input, the remainder will be assumed zero.

If the TRANS command is being used in conjunction with the EDIT command, note that the reflections are checked via EDIT before the transformation matrices are applied.

ATOM ATOM LABEL X, Y, Z

This defines the coordinates of the group atoms. The label consists of an atom type as written on SFAC or CONTENTS commands with an optional label inside parentheses e.g. C(12) or C(12') or Li(2A) or Li. Labels such as C12 or C2A are invalid. If the atom type and label is missing or unrecognised, then it is assumed to be carbon. A maximum of 200 atoms is permitted when totalled from all the groups with a limit of 100 atoms for all the type 3 or 4 groups.

GROUp TYPE(2/3/4/5/6), NO. IN CELL, CELL PARAMETERS

This command inputs the groups of known stereochemistry into the package. Each group (up to a limit of 10 groups) is followed by a set of ATOM commands (see below). The group type and the default cell associated with the group works as follows:

TYPE GROUP POSITION ORIENTATION CELL DEFAULT 2 Random Random Orthogonal A 3 Random Correct Orthogonal A -3 Random Correct Crystal Cell 4 Correct Correct Orthogonal A -4 Correct Correct Crystal Cell 5 Karle Recycling Crystal Cell 6 Weighted Fourier Recycling Crystal Cell

Groups of type 5 and 6 imply a recycling procedure is required, and this situation is discussed fully in section 18.0. The cell parameters must refer to edges and angles not their cosines.

A missing or zero TYPE parameter is assumed to be 2. If the number of groups in the unit cell is missing or zero then it is assumed to be the number of equivalent positions including centres of symmetry and lattice centerings.

NOCHeck

This switches off the checks for duplicate reflections and systematic absences in the intensity data. This can be useful if you already use a program to remove these reflections at data reduction time, since it prevents duplicate calculations and reduces the running time of NORMAL. However, only use it if you are certain that the duplicates and absences are missing otherwise you will cause havoc.

ELECtron

Data input will be electron diffraction data.

POWDer

Data input will be powder diffraction data. In this case only a single scale factor is assigned (i.e. there is no parity group rescaling).

NEUTron

Data input will be neutron diffraction data.

XRAY

Data input will be X-ray diffraction data.

OVERlap H,K,L,H,K,L,....H, K, L, $\sum|F|^2$, $\sigma(\sum|F|^2)$ 1 1 1 2 2 2 n n n

Each entry defines the reflections under a given overlap. There are n sets of h,k,l indices defining the reflections plus an entry for the total F and its e.s.d.. If there are n reflections there must be 3n+2 entries on this line. This command automatically sets the powder option.

ENTRopy

Will make the binary, unformatted file that normal outputs readable by the MICE maximum entropy program. This option should used with caution during a normal structure solution, since the E-magnitudes produced in this are quasi-normalised E's.

DATA NUMBER OF REFLECTIONS PER LINE, FORMAT OF DATA

Usually intensity data will be input from a standard file that has been interfaced

to the package by the programmer who set it up at your installation (See Section 21.0). Under these circumstances no DATA command should be issued; once all the commands have been input to NORMAL the intensity data will be read automatically. However, if you are using formatted data sets from another source, this command can be used to tell the system how many reflections are on each line and the FORMAT of the data. The FORMAT must obey all the FORTRAN rules and must be enclosed by parentheses e.g.

DATA 2 (3I3, 2F10.5, 10X, 3I3, 2F10.5) The module expects the reflection parameters to be in the order h, k, l, F and sigma(F). If the PHASE keyword is involved in the NORMAL call line then the phase (in degrees) of the reflection is expected as an integer. It may take any value. The end of the data can be specified by:

(1) An end of file marker. (2) A reflection with h = k = l = 0 (3) A reflection with |F| < 0.0

The DATA command must be the last input to NORMAL.

Reflections having |F| < 2sigma(|F|) are not phased even if they have large E-magnitudes.

The commands TITLE, END, MENU, LEVEL, NOPRINT, PRINT, DEFAULT, HARD, VERY HARD, MODEL, SHOW and X are always available.

7.0 TRIPLETS

This module must always be run, and will be executed with defaults if not called explicitly. The TRIPLETS module initialises the invariants file so that all the invariants (if any) previously stored on it are lost. (See Section 21.0). The available commands, which can be entered in any order, are as follows:

TRIPlets NO. OF REFLECTIONS, CUT-OFF, [FOm]

The first parameter is the number of reflections for which triplets will be generated. The default is:

No. of reflections = 4 * No. of atoms in asymmetric unit + 100 + 10 * Isymp + 50 * Idif + 150 * Ivdif
where:
Isymp = 0/1 for non symmorphic / symmorphic space group
Idif = 0/1 for a standard / hard structure
Ivdif = 0/1 for a standard / very hard structure

This is subject to a minimum of 250, or the number of E's greater than 1.0, whichever is the smaller. The maximum permitted is 800.

The second parameter is a cut-off, input as a fraction of the maximum sin^2/λ^2 of the data. This will prevent the use of triplets that contain two or more reflections with a resolution larger than the cut-off limit. The default is to use all reflections. [FOm] is a keyword designed to cause the calculation of an extended triplet list, involving E's not used for phasing. This is used later in tangent refinement for the calculation of the LOGLIK figure of merit.

LIST

This causes a full list of each triplet to be output to the printer. In the interactive modes (LEVEL = 1, 2 or 3), this information is also output to the terminal. By using NOPRINT in conjunction with LIST it is possible to switch off the printing, but the user will still get the screen output. The print option uses one line per triplet, so use this option with care.

MDKS NUMBER OF CONTRIBUTORS REQUIRED FOR D AVERAGE

This invokes the MDKS calculations. The only parameter here specifies the maximum number of contributors required for the D average in the MDKS formula. The smaller this number is, the quicker the calculations will run, but the less reliable they will be. The default value is given by:

No. = max (5 * no. of atoms in asymmetric unit, 200)

Sometimes triplets appear where this number of contributors cannot be found.

The program will accept such calculations provided there is at least half the required number. Any triplets with less than this are flagged as having no available MDKS estimates.

L.E.

This invokes a similar calculation to that of the MDKS formula, although it requires no scaling calculations. There are no parameters. Note the two periods in this command.

** Both the MDKS and L.E. options are very slow and demanding; for this reason they are never called by default. Use them with caution. **

WEIGht N

Use a weighting scheme based on sin^2/λ^2 to down-weight triplets that contain one or more reflections with a resolution greater than 0.95 times the resolution maximum. (See Gilmore & Brown, 1988)

N = 1 Print a list of the weighted E's
N = 0 Don't print a list of the weighted E's

As usual the commands TITLE, END, MENU, LEVEL, NOPRINT, PRINT, DEFAULT, HARD, VERY HARD, MODEL, SHOW and X are always available.

8.0 QUARTETS

This module generates the quartet invariants. It is always run in situations where the space group is symmorphic, or where the HARD or VERY_HARD command has been issued. The quartets are put on to the invariants file after the triplets. Any existing quartets will be overwritten. (See Section 21.0) The available commands are as follows:

QUARtets NO. OF REFLECTIONS, CUT-OFF

The first parameter is the number of reflections for which quartets are to be generated. The default is:

No. of reflections = 100 + 15 * Isymp + 25 * Idif + 25 * Ivdif

where: Isymp = 0 / 1 for non-symmorphic / symmorphic space group Idif = 0 / 1 for standard / difficult structure Ivdif = 0 / 1 for standard / very difficult structure

The maximum number of reflections is 256; it may not exceed the number of E's greater than 1.0. The default is often unsuccessful and may result in either far too many or too few invariants. You may need to experiment with it.

The second parameter is a cut-off, input as a fraction of the current maximum sin^2/λ^2. It will prevent the use of any quartet that has two or more reflections with a resolution greater than that of the cut-off. The default uses all reflections.

POSItive

Usually the program generates only the negative quartets. This command will cause the positive invariants to be generated as well. It is a default option when the VERY_HARD instruction has been issued. It increases the computer time needed by a factor of at least two, and will generate a very large number of relationships - so use it with care.

LIST

This causes a listing of all the quartets to appear on the printer, and the terminal as well in the interactive mode. There is a considerable amount of output per invariant which will be even more extensive if the third neighbourhood has been invoked, so use this with care. It is possible just to get the screen output by using the NOPRINT and LIST commands together.

NEIGhbour SIZE OF 3rd NEIGHBOURHOOD, MAX NO. OF 3rd NEIGHBOURS TO EXPLORE

This command invokes the third neighbourhood calculations. There are two parameters:

The first dictates the range of the floating vector p, which defines the neighbourhood. A value, for example, of 100 means that p is restricted to the top 100 E-magnitudes. The smaller this parameter is, the faster the calculation will run. The default is 100 which usually works well.

The second parameter specifies how many third neighbourhoods are to be found for a given quartet before moving on to the next. A full calculation in which all the neighbourhoods are explored is very time consuming, and may not be any better than a situation in which the search stops after 10-20 neighbours. The default is 10 unless the HARD or VERY_HARD command has been issued, in which case it is increased to 20.

MISSIng MAX. NO. OF MISSING 2nd NEIGHBOURS, MAX. NO. OF MISSING 3rd NEIGHBOURS.

This command specifies how many missing reflections are allowed in the 2nd and 3rd neighbourhoods. The defaults are both zero unless the VERY_HARD command has been issued in which case they become 1 and 2 respectively. This command can be useful in situations where the data set is limited in some way, or in large structures. It increases the computer time required.

RESTart REFLECTION NUMBER

This is usually used in the batch mode to restart a quartet generation which ran out of time. It works as follows:

In batch mode, the secondary output file will print a record of the quartet generation as it proceeds. The message takes the form:

ALL QUARTETS GENERATED FOR VECTOR 1: 205 GENERATED 200.....
ALL QUARTETS GENERATED FOR VECTOR 2: 320 GENERATED 290.....
ALL QUARTETS GENERATED FOR VECTOR 3: 450 GENERATED 327.....
etc.

This means that all the quartets have been generated for the first E-magnitude, then the second etc. To restart examine this output, go back a few lines from the last line to allow for buffering effects, and hence select a vector number from which to restart and enter it on the RESTART command. It does not matter if you go back too far; the duplicate quartets will get overwritten.

WEIGht N

Use a weighting scheme based on sin^2/λ^2 to down-weight quartets that contain one or more reflections with a resolution greater than 0.95 times the maximum resolution of the intensity data. (Gilmore & Brown, 1988)

N = 1 Print a list of the weighted E's
N = 0 Don't print a list of the weighted E's

As usual the commands TITLE, END, MENU, LEVEL, NOPRINT, PRINT, DEFAULT, HARD, VERY_HARD, MODEL, SHOW and X are always available.

9.0 CONVERGE

The CONVERGE module carries out the dual functions of collecting together the invariants (and seminvariants), then selecting the reflections which comprise the starting set. Any complete direct methods analysis must include this module, and it will be run under defaults if not called explicitly. The available instructions, which may appear in any order, are as follows:

CONVerge NDET, IQ4,

NDET is the number of reflections which are to be phased by tangent refinement. The default works as follows: The maximum possible value is defined by the number of reflections for which invariants were generated. This value is usually reduced by 10% to remove those reflections which are mostly weakly linked into the phasing procedure and this is the default unless:

(1) There are less than 101 reflections to be phased or
(2) The command HARD or VERY_HARD has been issued.

In both these cases the 10% reduction is not applied. If this procedure is unsatisfactory then enter your own value of NDET.

IQ4 = 0 for negative quartets to be used actively and passively; no positive quartets are to be used. This is the default. = 1 for use negative quartets as figures of merit only; no active use of positive or negative quartets. = 2 for use positive and negative quartets actively, and negative quartets as figures of merit as well. This is the default for VERY_HARD structures.

EXCLude H, K, L or REFLECTION NUMBER

All the invariants involving this reflection are removed from the input list so that the reflection is not phased. The reflection can either be specified by its serial number or by its indices.

REMOve RELATIONSHIP

This removes any specified phase relationship from the list of invariants. Any 1-, 2-, 3-, or 4-phase relationship can be deleted in this way. The relationship is specified by arranging the moduli of the serial numbers involved in descending order. The associated kappa-value, and the phase shift are ignored. The serial numbers are then included as parameters of the REMOVE command. A limit of 100 REMOVE commands is imposed.

SIGMa [ALl] /[NOne] or MIN. PROBABILITY, MIN. CONSISTENCY RATIO.

This defines the criteria by which one-phase seminvariants (the Sigma-1 reflections) are selected for the starting set. This command either uses the

keywords "ALL" / "NONE" or utilises two numeric fields which define the acceptance parameters.

NONE requests that no Sigma-1 type reflections are accepted. This is the default for symmorphic space groups.

ALL requests that all the Sigma-1 reflections be put into the starting set suitably weighted.

MINIMUM PROBABILITY specifies the minimum acceptable probability. the default is 0.95 unless the space group is symmorphic, in which case the default is set so that no reflections are accepted.

MINIMUM CONSISTENCY RATIO (See Section 3.5). The default is 0.67. Note that a minimum of three contributors is also required for a Sigma-1 reflection to be accepted.

RELAtionship SERIAL NUMBER(S), PHASE SHIFT, k-VALUE

Inputs a 1-, 2-, 3-,or 4-phase invariant or seminvariant into the list of relationships. The relationship is arranged in descending order of the moduli of the serial numbers involved; the phase shift is expressed in 24ths (e.g. a phase shift of 180deg is 12); the kappa-value is usually a guess at this point, but it must be included. The larger the value, the more reliable the relationship is deemed to be. If you want it to have a very large weight, then use a value of 10.0 - 20.0. The relationship is then input via the RELATIONSHIP command in which the parameters are the serial number(s) involved (complete with signs), the phase shift in 24ths, and the kappa-value.

Suppose that symbolic addition strongly suggests that:

phi{12} = phi{23} + 180

rearranged it becomes:

phi{23} - phi{12} + 180 = 0

and it can now be input to CONVERGE via:

RELATIONSHIP 23 -12 12 10.0

KMIN MINIMUM ACCEPTABLE kappa-VALUE

Specifies a minimum acceptable kappa-value. The default is 0.6. Lower this value for difficult situations, or occasions where there is a paucity of phase relationships.

MDKS W1, W2, W3, W4

If the MDKS or L.E. command was issued when the TRIPLETS module was run, then a record of the estimate of the cosine of each invariant is stored on the invariants file. This information can be used to change the kappa-values of the triplets. The estimated cosines are split into four classes (i) - (iv) and each class is assigned a weight which multiplies the kappa-value of the triplet. As already described in section 3.5, the classes are:

(i)	cos phi{3} >= 0.7	Weight is W1, default = 1.2
(ii)	0.7 > cos phi{3} >= 0.0	Weight is W2, default = 1.0
(iii)	0.0 > cos phi{3} >= 0.7	Weight is W3, default = 1.0
(iv)	cos phi{3} < -0.7	Weight is W4, default = 0.0

A weight of 0.0 or less means that the relationship is deleted. Unless the MDKS command is issued, the cosine estimates from the TRIPLETS module are not used.

MAX_ MAXIMUM NUMBER OF PHASE RELATIONSHIPS

Usually the module uses all the available relationships up a limit of 28350, which is the program limit. If you want less than this then use this command. There is a minimum of 100.

LIST [ALl] / [NOne] / [PArtial]

This command uses keywords to specify how much of the convergence map is to be printed.

ALL prints the entire map. NONE suspends all convergence map printing. PARTIAL prints only the last 60 phasing steps of the map. This is the default.

MAGIc MAGIC INTEGER SEED

Specifies the number to be used as the seed for the series which acts as the magic integer sequence generator. The default is 2. Note that this command can decrease the number of phase sets generated, at the expense of reduced accuracy of the starting set.

ORIGin REFLECTION NUMBER

Usually, CONVERGE defines its own origin, but this command can be used to specify a reflection that is to be used for origin definition. The module will reject any reflection that is invalid. Note that it is not necessary to fully define the origin via a sequence of several ORIGIN commands. The module recognises partially defined origins, and will complete the definition during convergence.

ENANtiomorph REFLECTION NUMBER

Usually, CONVERGE defines its own enantiomorph, but this command can be used for a user defined enantiomorph. The module will reject an invalid reflection.

PERMute REFLECTION NUMBER

Specifies a reflection which is to be included in the starting set. Its phase is unknown, but it is given a sequence of values as generated by the magic integer routines in tangent refinement, unless it is a special reflection in which case it is given its two possible phase values in turn.

KNOWn REFLECTION NUMBER, PHASE (in DEGREES), WEIGHT

Allows a reflection of known phase to be included in the starting set. Its phase (in degrees) must be specified, and an associated weight in the range 0.0 - 1.0. The default value is 1.0.

SPECial NUMBER OF SPECIAL REFLECTIONS IN THE STARTING SET

This command defines the total number of special reflections to be included in the starting set, and whose phases are to be permuted. This includes any relevant reflections entered via the PERMUTE command.

GENEral NUMBER OF GENERAL REFLECTIONS IN THE STARTING SET

Specifies the total number of general reflections to be included in the starting set. The phases of these reflections are permuted in the usual way.

ANY_ NUMBER OF ANY SORT OF REFLECTIONS IN THE STARTING SET

Specifies the number of any sort of reflection, special or general, which are to be included in the starting set, and whose phases are to be permuted.

SETS MAXIMUM NO. OF PHASE SETS, MINIMUM NO. OF PHASE SETS

This command specifies the maximum and minimum number of phase sets to be generated. The maximum number of phase sets is only used if the user does not specify the starting set in any way via the GENERAL, SPECIAL or ANY commands. The default is 65 with a maximum of 2250, and a minimum of zero.

The minimum number of phase sets can also be specified. Again it is only used if the starting set has not been specified in any way. The default is calculated as follows:

Min. no. of sets = min(2*Nasu/3, Max. no. of sets/2 - 1) + 64*Ivdif + 32*Idif
where:

Nasu = No. of atoms in asymmetric unit.
Ivdif = 0 / 1 for standard / very hard structure.
Idif = 0 / 1 for standard / hard structure.

This is subject to a minimum of 12. If the minimum number of phase sets is specified, but the maximum number is left as a default, then the latter is given the value:

Max. no. of sets = Min. no. of sets * 2 + 1

The general commands TITLE, END, MENU, LEVEL, NOPRINT, PRINT, DEFAULT, HARD, VERY_HARD, MODEL, SHOW and X can be issued at any time.

10.0 MAGEX

MAGEX is run between CONVERGE and TANGENT. Is is always user called, and never entered by default. The available commands, which can appear in any order, are as follows:

MAGEx NO. OF PHASE SETS TO GENERATE

This specifies how many phase sets are to be passed to tangent refinement. The default is:

No. of sets = 80 + 50 * Idif + 50 * Ivdif

where:
Idif = 0 / 1 for Standard / Hard structure.
Ivdif = 0 / 1 for Standard / Very hard structure.

LIST

This causes the primary and secondary reflections and the triplets linking them to be listed on the printer.

PRIMary NUMBER OF PRIMARY REFLECTIONS TO USE

This command defines how many primary reflections are to be used to initialise the MAGEX procedure. The default is:

No. of primaries = 10 + 6 * Max (Idif, Ivdif)

There is a maximum of 22 primaries.

FUNCtion FUNCTION TO USE IN PARAMETER SHIFT REFINEMENT (N)

There are four options here:

(1) N = 1 uses

$$\sum k \frac{I_1(k)}{I_o(k)} \cos\varphi_3$$

Special reflections are allowed to take any value, and then reset to the closest permitted value at the end of refinement. This is equation (16) in the paper by Hull, Viterbo, Woolfson and Shao-Hui (1981).

(2) N = -1. Uses option (1) above, but special reflections are only allowed to take one of their two possible values throughout.

(3) N = 2 uses Equation 17 in the paper of Hull et al (1981). Special reflections are treated as in option (1).

(4) N = -2 uses the same function as option (3) but constrains the special reflections in the same way as option (2).

The default value is option (3) i.e. N = 2. For centrosymmetric cases only options (1) and (2) i.e. N = 1 or -1 are permitted.

MAGIc N

Defines the Magic Integer sequence. If N is a small positive integer then the integers are based on the sequence:

$F_n = F_{n-1} + F_{n-k}$

where k = N as defined above. A value of N = 1 gives a power of two series, whilst N = 2 gives the Fibonacci series which is the default. A value greater than 2 gives the hyper - Fibonacci series. If N is negative the sequence:

$F_n = 2 * F_{n-1} + F_{n-k}$

where k = N, is used instead.

ALIMit MINIMUM ALPHA VALUE FOR ACCEPTING A SECONDARY

This command defines the lowest estimated alpha for accepting a secondary reflection. The default is 1.7, but for centrosymmetric cases this should be increased to around 4.0. If the option chosen results in too many relationships, it will be automatically reduced by MAGEX.

KALImit LOWEST kappa-VALUE FOR SECONDARY DEFINITION

This defines the minimum kappa-value a triplet may have if it is to be used in secondary definition. The default is 1.0.

EXPAnd NO. OF CYCLES

This repeats the MAGEX process for the specified number of cycles. Use this option with care. It is not recommended that more than two cycles are performed, but more are possible if you wish. The default is zero.

WSPEc WEIGHT

Each potential primary reflection has an omega function associated with it (See equation (5) in Hull, Viterbo, Woolfson and Shao-Hui, 1981), and the reflections with the largest omega values are selected as working primaries. The weight defined on this command multiplies each omega value when a special

reflection is involved. Setting the weight to a value less than unity can be used to avoid having too many special reflections as primaries. The default is 1.0.

SELEct IFUNC

This selects the function to use for ranking the solutions derived from the small Q-maps. IFUNC can have only 2 possible values:

(1) IFUNC = 1 uses function (1) defined in the FUNC command.
(2) IFUNC = 0 uses function (2) defined in the FUNC command.

The default is function (2) i.e. IFUNC = 0.

As always the commands TITLE,END,MENU, LEVEL, NOPRINT, PRINT, DEFAULT, HARD, VERY_HARD, MODEL, SHOW and X are available.

11.0 YZARC

This is another module that is run between CONVERGE and TANGENT. It is always user called and never entered automatically. The following commands may be entered in any order:

YZARc NO. OF PHASE SETS, MAX. NO. OF CYCLES, MAX. MEAN SHIFT

This is the YZARC calling command. It has three parameters:

(1) The number of phase sets to pass to TANGENT. The default is calculated as:
No. of sets = 60 + 50 * Idif + 50 * Ivdif

where:
Idif = 0 / 1 for Standard / Hard structure
Ivdif = 0 / 1 for Standard / Very hard structure.

(2) The maximum number of cycles of refinement for any phase set. The default is 70.

(3) The maximum value of the mean phase shift in degrees. If the mean phase shift falls below this value in any YZARC refinement cycle, then refinement ceases on this phase set and moves on to the next. The default is 4 deg.

Be careful about tampering with options (2) and (3). In general, they are set at their most useful values.

NREF NO. OF REFLECTIONS TO BE PHASED

This specifies the number of reflections to be phased by the YZARC procedure. The default is calculated via:

No. of reflections to phase = 100 + 50 * Idif + 50 * Ivdif

where Idif and Ivdif have already been defined.

L.S.

The module normally uses steepest-descents to refine the phase sets. This command invokes the standard least-squares procedure with its associated weighting scheme. Note the periods in the L.S. command.

STARt IX, IY

These are two odd integers which act as the seed for the random number generator which is used to give the reflections random phases. The default

values are 190907 and 568835.

NQESt CYCLE NUMBER, MAXIMUM NQEST

This command causes NQEST to be used as the criterion for stopping refinement of a given phase set. The cycle number specifies the cycle at which the NQEST test is first applied. All previous cycles proceed untested. The default is 7. The second parameter defines the maximum allowed value of NQEST after these cycles. The default is zero. If a solution has a value greater than this, then refinement terminates, otherwise it continues as long as NQEST is falling. If there is an increase in NQEST, refinement stops. Refinement is still subject to any constraints input on the YZARC calling command.

TOP_
Usually YZARC phases the bottom reflections in the convergence map; this command causes the map to be bypassed, apart from origin and enantiomorph definition, and the top reflections, based on E-magnitude alone, are phased. The number to be phased is that defined by the NREF command. This instruction should usually be used only in conjunction with steepest descents, since it is likely to give a singular matrix in a least-squares environment unless all the reflections are linked to each other by the available triplets. If least-squares has been specified and a singular matrix is found, the module will automatically change to steepest descents.

RANDom IX, IY

This instruction defines two odd integers which act as a seed for the random number generator. Unlike the START command, however, only a single phase set is produced, the module then expects another RANDOM command for another solution to be produce, and so on. In batch mode these commands are stacked together in sequence and terminated by a RANDOM instruction with IX=IY=0 or by a call to another module. In the interactive mode both these options are available, but the user can merely press a carriage return to terminate input of random numbers. This command is useful when re-running refinements in a situation where only certain solutions are to be investigated.

As usual the commands TITLE, END, MENU, LEVEL, NOPRINT, PRINT, DEFAULT, HARD, VERY_HARD, MODEL, SHOW and X are available as required.

12.0 SYMBOLIC ADDITION

The SYMB module may only be entered if a structure is centrosymmetric, and is only ever called by the user. The following commands may be entered in any order:

SYMB

Calls and initialises the SYMBOL module.

SYMboL NO. OF SYMBOLS

Allows the user to choose how many symbols, up to a maximum of 11, are to be used in the symbolic addition procedure. The default is four. Note that this command is SYML and not SYMB as may be expected.

SIGN MINIMUM PROBABILITY

This sets the minimum probability acceptable for a sign determination. The default is 0.95.

NUMBer NUMBER OF EQUATIONS

This determines the number of symbol equations to be considered in the solution for the symbols. It may be set at a number higher than the number of relationships in order to ensure that all relationships present are considered. The default is 10.

LIMIt PROB

All signs predicted with a probability greater than PROB will be included in the final output. The default is 0.8.

ORIGin Determine the origin by the SYMB module rather than accepting the origin determined by the CONVERGE module.

LIST

Output a table of signs developed by symbolic addition, and relationships between the symbols to be displayed in the output file.

As usual, the commands TITLE, END, MENU, LEVEL, NOPRINT, PRINT, DEFAULT, HARD, VERY_HARD, MODEL, SHOW and X are available as appropriate.

13.0 TANGENT

The TANGENT module performs the dual tasks of phase expansion and refinement. It must always be run except when Fourier recycling is being used, and will be entered and run with defaults if there is no explicit user call. The following commands can be entered in any order:

TANGent

Calls and initialises the TANGENT module.

SWTR [NO]

This command causes the Hull - Irwin weighting scheme to be used rather than the traditional MULTAN80 weights. This is a default when Karle recycling is used, otherwise the standard procedure is used. If for some reason the Hull - Irwin scheme is to be used and you wish to revert to the standard weighting method, then use the command:

SWTR NO where "NO" is a keyword.

SKIP N
This command is usually used for restarts. It causes the first N phase sets to be skipped before starting tangent refinement. These N phase sets must already exist on the file on channel 11 (See section 21.0). When re-running a job which ran out of time, N should be the number of the last set output on the printer, but make allowances for buffering by reducing this. The exact extent of this reduction depends on your installation.

SKIP 0 has a special function. When either the YZARC or MAGEX modules have been run in a job, the TANGENT module automatically takes its input phases from those produced by them. Phase permutation is not carried out. The SKIP 0 command causes these YZARC / MAGEX calculations to be ignored, and routine tangent refinement to be carried out instead.

SETS N1, N2, N3, N4..... etc.

With this option only the phase sets with numbers N1, N2 etc. are investigated by the TANGENT module. This is useful for re-runs. Unlike the SKIP command, the other phase sets do not need to be on file 11.

EFOM [ALl]/[NOne] or EFOM NO., CUT; EFOM NO., CUT; EFOM NO., CUT

This invokes the early figures of merit. There are three EFOM's numbered as follows:

(1) ψ_o + Resid. (R{Karle}) (2) NQEST. (3) ψ_o alone (applied later in refinement than (1) above).

These are the numbers used on the EFOM command. The cut values listed on the command are the maximum values that the early figure of merit may have when tested; any solution with values greater than these cut-offs are rejected. The defaults are: (1) 1.8 (2) 0.0 (3) 1.6

Usually the EFOM's are not used. The command:

EFOM NONE

also has this effect, and can be used to switch off EFOM's already invoked. If you wish to use all three early figures of merit then the command:

EFOM ALL CUT(1) CUT(2) CUT(3)

can be used, where CUT(1), CUT(2) etc. refer to the cut-off parameters discussed above. If all three cut-off's are absent then the default values are used for all these parameters. If only one appears, it is presumed to apply to the first EFOM; two parameters are assumed to apply to the first two etc. A zero parameter gives defaults. E.g.

EFOM ALL 1.2 0.1 2.0

applies cut-offs of 1.2, 0.1 and 2.0 respectively; whereas:

EFOM ALL 0 0 2.0

will give defaults (1.8 and 0.0) for the first two EFOM's and a cut-off of 2.0 for the third.

If only certain EFOM's are wanted, then do not use the "ALL" keyword. Instead enter the EFOM number followed by the required cut-off. In these circumstances, only the specified early figures of merit are invoked. E.g.

EFOM 2 0.1 3 2.0 or EFOM 3 2.0 2 0.1

invokes the second EFOM with a cut-off of 0.1 and the third with a cut-off of 2.0; the first EFOM is not used. Note that all the EFOM requirements must appear on a single EFOM command.

It is quite difficult to devise suitable defaults applicable under all situations, and some parameter tuning may be necessary under certain circumstances. EFOM's are not available with the Hull-Irwin weighting scheme.

WTFOm W1, W2, W3, W4

This command defines the relative weights of the four figures of merit used by the TANGENT module when calculating a combined figure of merit (CFOM).

The defaults are as follows:

Figure of Merit	Weight	Default	Default in symmorphic cases
ABSFOM	W1	1.0	0.6
PSI-ZERO	W2	1.0	1.2
RESID	W3	1.0	0.6
NQEST	W4	1.0	1.3

If there are no appropriate relationships for a particular figure of merit then a weight of zero is assigned. The relative weights are normalised such that the maximum CFOM value is equal to the total number of figures of merit contributing to it. The LOGLIK figure of merit, if calculated, is not used in the combined figure of merit.

NOSTop

If the TANGENT module finds a solution with figures of merit that satisfy the conditions:

(1) Resid (R{Karle}) less than 20.0.
(2) ψ_o less than 1.25
(3) NQEST less than -0.15.
(4) The figures of merit above are within 5% of the best so far

Then the module assumes that this is the correct solution and exits. Users in the interactive mode will be questioned first if they wish to accept this solution, but batch users will not. The command NOSTOP switches off these tests. So do the HARD and VERY_HARD options. NOSTOP is often worthwhile as the correct solution can often be missed under the early stop algorithm.

SERIal SERIAL NUMBERS N1, N2, N3, N4..... etc.

The commands SERIAL, MARK, WEIGHTS and PHASES are used together to input a set of known phases into tangent refinement. The SERIAL command gives the serial numbers (in any order) of the reflections whose phases are to be input.

MARK M1, M2, M3, M4..... etc.

These are the markers associated with the reflection numbers entered on the SERIAL command. A value of 1 means that the input phase can be refined immediately, whereas a value of -1 signifies that the phase is to be held constant until the last two cycles. The default values are -1, so the MARK command is only needed if this is unsatisfactory.

WEIGhts W1, W2, W3, W4..... etc.

The weights associated with each input phase. The default values are 0.9, so

this command is only needed in cases where this is inappropriate.

PHASes PHASE 1, PHASE 2, PHASE 3, PHASE 4..... etc.

Inputs the known phase angles in degrees. Each angle must appear in the same order as the serial numbers on the SERIAL command. You can input as many sets as you require; each set requires its own PHASES command and is refined immediately after it has been input. The first input phase set via the PHASES command terminates current input to the TANGENT module. In the interactive mode the user will be prompted for more phase sets as required-a simple carriage return or a call to another module is used to signify the end of input phase sets. In the batch mode only another module call, or an END, X or DEFAULT command can be used to do this.

As usual, the commands TITLE, END, MENU, LEVEL, NOPRINT, PRINT, DEFAULT, HARD, VERY_HARD, MODEL, SHOW and X are available as appropriate.

14.0 RANTAN

In one respect the RANTAN module is similar to TANGENT in that it refines phase angles. The difference is that RANTAN uses random phases for all the unknown phases and refines them using the tangent formula, rather than using the phase permutation techniques of TANGENT. RANTAN is never called automatically. The following commands, apart from the first, may be entered in any order:

RANTan NO. OF PHASE SETS TO GENERATE.

Calls and initialises the RANTAN module. The number of phase sets to generate is entered. The default is:

No. of phase sets = 100 + 50 * Idif + 100 * Ivdif
where:
Idif = 0 / 1 for Standard / Hard structure.
Ivdif = 0 / 1 for Standard / Very hard structure.

SWTR [NO]

This command causes the Hull - Irwin weighting scheme to be used rather than the traditional MULTAN80 weights. This is a default when Karle recycling is used, otherwise the standard procedure is used. If for some reason the Hull - Irwin scheme is to be used and you wish to revert to the standard weighting method, then use the command:

SWTR NO where "NO" is a keyword.

SKIP N

This command is usually used for restarts. It causes the first N phase sets to be skipped before starting tangent refinement. These N phase sets must already exist on the file on channel 11 (See section 21.0) When re-running a job which ran out of time, N should be the set number of the last set output on the printer, but make allowances for buffering by reducing this by 3 or 4 depending on your installation, and operating system.

WTMIn WEIGHTS OF UNKNOWN REFLECTIONS.

The random phases have weights of 0.25 assigned to them before refinement begins, but if a different value is wanted, then this command can be used. Some experimentation with these weights can be useful in difficult cases.

EFOM [ALl]/[NOne] or EFOM NO., CUT; EFOM NO., CUT; EFOM NO., CUT

This invokes the early figures of merit. There are three EFOMs numbered as follows:

(1) ψ_o + Resid (R{Karle})
(2) NQEST.
(3) ψ_o alone (applied later in refinement than (1) above).

These are the numbers used on the EFOM command. The cut values listed on the command are the maximum values that the early figure of merit may have when tested; any solution with values greater than these cut-offs are rejected. The defaults are:
(1) 1.8 (2) 0.0 (3) 1.6

Usually the EFOMs are not used. The command:

EFOM NONE

also has this effect, and can be used to switch off EFOMs already invoked. If you wish to use all three early figures of merit then the command:

EFOM ALL CUT(1) CUT(2) CUT(3)

can be used, where CUT(1), CUT(2) etc. refer to the cut-off parameters discussed above. If all three cut-offs are absent then the default values are used for all these parameters. If only one appears, it is presumed to apply to the first EFOM; two parameters are assumed to apply to the first two etc. A zero parameter gives defaults. E.g.

EFOM ALL 1.2 0.1 2.0

applies cut-offs of 1.2, 0.1 and 2.0 respectively; whereas:

EFOM ALL 0 0 2.0

will give defaults (1.8 and 0.0) for the first two EFOM's and a cut-off of 2.0 for the third. If only certain EFOM's are wanted, then do not use the "ALL" keyword. Instead enter the EFOM number followed by the required cut-off. In these circumstances, only the specified early figures of merit are invoked. E.g.

EFOM 2 0.1 3 2.0 or **EFOM** 3 2.0 2 0.1

invokes the second EFOM with a cut-off of 0.1 and the third with a cut-off of 2.0; the first EFOM is not used. Note that all the EFOM requirements must appear on a single EFOM command.

WTFOm W1, W2, W3, W4

This command defines the relative weights of the five figures of merit used by the RANTAN module when calculating a combined figure of merit (CFOM). The defaults are as follows:

Figure of Merit		Weight Default	Default in symmorphic cases
ABS FOM	W1	1.0	0.6
PSI-ZERO	W2	1.0	1.2
RESID	W3	1.0	0.6
NQEST	W4	1.0	1.3

If there are no appropriate relationships for a particular figure of merit then a weight of zero is assigned. The relative weights are normalised such that the maximum CFOM value is equal to the total number of figures of merit contributing to it. The LOGLIK figure of merit, if calculated, is not used in the combined figure of merit.

NOSTop

If the RANTAN module finds a solution with figures of merit that satisfy the following conditions:

(1) Resid (R{Karle}) less than 20.0. (2) ψ_o less than 1.25 (3) NQEST less than -0.15. (4) The figures of merit above are within 5% of the best so far.

(assuming that these figures of merit are available), then the module assumes that this is the correct solution and exits. Users in the interactive modes will be questioned first if they wish to accept this solution, but batch users will not. The command NOSTOP switches off these tests. So do the HARD and VERY_HARD options.

SETS N1, N2, N3, N4..... etc.

With this option only the phase sets with numbers N1, N2 etc. are investigated via the RANTAN module. This is useful for re-runs. Unlike the SKIP command, the other phase sets do not need to be on file 11.

STARt IX, IY

Two odd integers used to seed the random number generator. The default values are 1 and 1 (cf. YZARC).

As usual, the commands TITLE, END, MENU, LEVEL, NOPRINT, PRINT, DEFAULT, HARD, VERY_HARD, MODEL, SHOW and X are available as appropriate.

15.0 MAPS

The MAPS module calculates Fourier maps, picks the peaks and attempts to identify chemically reasonable fragments (Main & Hull, 1978). Usually an E-map is computed, but where Fourier recycling has been requested a Sim-weighted electron density map is calculated and interpreted. Following the MAPS command, the following commands may appear in any order:

MAPS N1 [, N2]... etc. or -N or [ALl]

Calls and initialises the MAPS module. There are three possibilities for parameters associated with this command:

(1) One or more positive integers. Each of the specified solutions from tangent refinement will be used to compute an E-map. E.g.

MAPS 2, 4, 1, 19

will compute the maps for solutions 2, 4, 1 and 19.

(2) A single negative integer (-N) will compute maps for the +N best solutions as ranked on combined figure of merit.

(3) The keyword "ALL" causes all the E-maps to be calculated. This is usually used in the interactive mode with LEVEL = 3 as a way of examining maps in a quick and efficient way. Under LEVEL 3, it is possible to exit from the MAPS module once a suitable map has been found. It is not necessary to actually investigate all the maps.

You may only choose one of these methods on a MAPS command. The default is to calculate the single best E-map. In the situation where Fourier recycling is being carried out, all parameters on the MAPS command are ignored.

GRID RESOLUTION, X LIMIT, Y LIMIT, Z LIMIT

This command defines the resolution and extent of the Fourier computation. RESOLUTION is the resolution required in Angstroms. The default is 0.333Å.

X LIMIT is the limit of the Fourier calculation along the x-axis. The calculation always starts at x = 0.0. The default is 1.0.

Y LIMIT operates like the X limit above but along y.
Z LIMIT operates like the X limit above but along z.

PEAKs MAX NO. OF PEAKS, NO. OF HEAVY ATOMS, MIN. NO. OF PEAKS PRESENT FOR PLOTTING

This command specifies how many peaks are to be picked from the Fourier map, and used in the interpretation routines. The default is calculated as follows:

Max. no. of peaks = (11*Nasu+13)/9 + 10*Idif + 10*Ivdif + 10*Iapx

where:
Nasu = No. of atoms in asymmetric unit.
Idif = 0 / 1 for Standard / Hard structure.
Ivdif = 0 / 1 for Standard / Very Hard structure.
Iapx = 0 / 1 for Yes / No CONTENTS command issued in the NORMAL module.

This default is increased when two clusters are within 2.8Å of each other.

The number of heavy atoms present in the asymmetric unit (N) is used as follows: when carrying out an interpretation of a cluster, the MAPS module will assume that the N highest peaks correspond to these heavy atoms. They will not be included in the interpretation, and all peaks within the maximum bond length will be marked as spurious. This can be useful in heavy-atom cases to prevent diffraction ripples around the heavy atom(s) being treated as real atoms, but it does not always work well, so be wary of E-map interpretations that incorporate this option. The default is zero.

The final parameter defines the minimum size of cluster, defined by the number of atoms it contains, which will be plotted on the printer (and the terminal if graphics options are used). The default is four.

NOJOin

Stops all connectivity and peak interpretation calculations from being performed.

PROJect NO. OF PROJECTIONS, MAX. NO. OF INTERPRETATIONS

The first parameter here defines the number of orthogonal projections of each cluster to be plotted. The default is always to plot the least-squares projection first, and then to plot the projection orthogonal to the least- and most-squares planes if the cluster is spherical or cylindrical in shape. The maximum number of projections is three. The third projection is on to the most-squares plane.

The second parameter specifies the maximum number of possible peak interpretations to be output for a cluster. The default is three.

DOUT MAXIMUM BOND LENGTH TO OUTPUT

All interpeak distances less than this parameter are tabulated. The default is 2.4Å.

DMIN MINIMUM ACCEPTABLE BOND LENGTH

This is the minimum allowed bond length for peak interpretation. The default is 1.1Å.

DMAX MAXIMUM ACCEPTABLE BOND LENGTH

This is the maximum allowed bond length for peak interpretation. The default is 1.95Å.

AMIN MINIMUM ACCEPTABLE BOND ANGLE

This is the minimum allowed bond angle for peak interpretation. The default is 85 deg. You may need to lower this when three-membered rings are involved, although it can result in a rather more messy collection of peaks.

AMAX MAXIMUM ACCEPTABLE BOND ANGLE

This is the maximum allowed bond angle for peak interpretation. The default is 145 deg. This, too,will not always be a suitable figure.

VDU_
This command requests that all peak interpretations be output to the screen of a graphics terminal as well as the printer. It is only accepted if:

(1) The graphics routines have been implemented at your installation.
(2) The LEVEL parameter is set at 2 or 3.

The user will be asked to confirm that this is so.

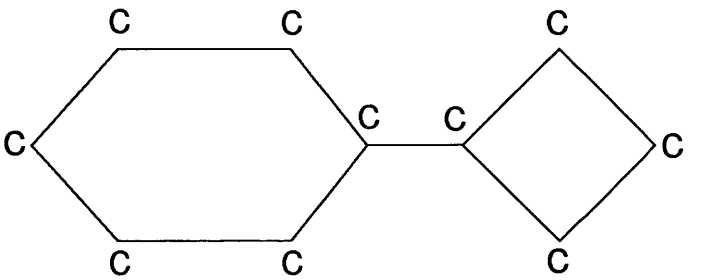
LIST N

This command causes the full Fourier map to be output on the printer. All the peaks greater than or equal to N are underlined with asterisks.

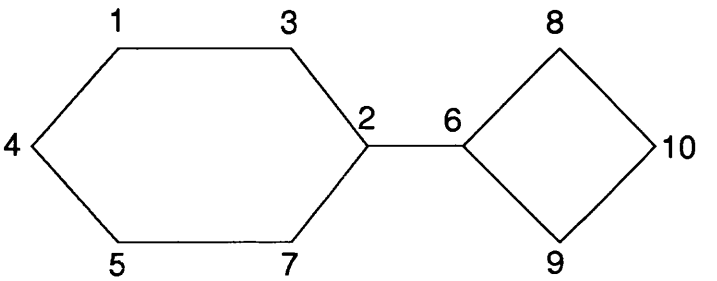
MOLEcule CONNECTIVITY

This instruction inputs a set of molecular connectivies which are used for comparison with the molecular fragments found in the map itself. They are not used by the MAPS module in its search for a fragment, and are therefore optional. The connectivities are input as follows:

(1) Draw the molecule or fragment whose chemical structure and connectivity is known. The stereochemistry is not important. E.g.



(2) Number the atoms in any order. E.g.



(3) Specify the connectivities as follows:

MOLE 1 3 4 / 2 3 6 7 / 3 1 2 / 4 1 5 / 5 4 7 / 6 2 8/ =
7 2 5 / 8 6 10 / 9 6 10 / 10 8 9 ///

where atom 1 is joined to atoms 3 and 4, atom 2 to atoms 3, 6 and 7 etc.

Redundant information may be removed. For example the above sequence can be simplified to:

MOLE 1 3 4 / 2 3 6 7 / 4 5 / 5 7 / 6 8 9 / 8 10 / 9 10 /// or
MOLE 1 3 4 / 2 3 6 7 / 5 4 7 / 8 6 10 / 9 6 10 ///

In both these cases each bond is specified only once.

(4) The double slash (//) signifies the end of a fragment. Another can follow, although it must be part of the original MOLE command (only one such command is permitted). The maximum number of atoms which can be entered is 200. Up to five fragments or molecules can be entered each separated by the double slash (//). The end of the command is specified by a triple slash (///). This must appear.

The commands TITLE, END, MENU, LEVEL, NOPRINT, PRINT, DEFAULT, HARD, VERY_HARD, MODEL, SHOW and X are available as required.

16.0 PATTERSON

This module computes an E^2 -1 vector map followed by a peak search, but without peak interpretation. The following commands are available:

PATTerson

Calls and initialises the Patterson map calculations.

LIST N

This command causes the map to be output on the printer. All the peaks greater than or equal to N are underlined.

GRID RESOLUTION, X LIMIT, Y LIMIT, Z LIMIT

This command defines the resolution and extent of the Patterson computation.

RESOLUTION is the resolution required in Angstoms. The default is 0.333Å.

X LIMIT is the limit of the calculation along the x-axis. The calculation always starts at x = 0.0. The default is 1.0.

Y LIMIT operates like the X limit above but along y.
Z LIMIT operates like the X limit above but along z.

PEAKs MAX. NO. OF PEAKS TO LIST

This specifies the maximum number of peaks to output. The default is 30.

As usual, the commands TITLE, END, MENU, LEVEL, NOPRINT, PRINT, DEFAULT, HARD, VERY_HARD, MODEL, SHOW and X are available as appropriate.

17.0 REVIEW

This module sorts and outputs the phase sets for each of the individual figures of merit as well as CFOM. The output will also give an indication as to whether a map has been investigated earlier or has an equivalent set. The following commands are available:

Absfom	Sort the sets on the Absolute figure of merit
Psi-zero	Sort the sets on the ψ_o figure of merit
Resid	Sort the sets on the R{Karle} figure of merit
Nqest	Sort the sets on the Nqest figure of merit
Loglik	Sort the sets on the Loglik figure of merit
Cfom	Sort the sets based on the Combined figure discussed in section13.0
Exit	Exits from the module and returns to the master MITHRIL90 menu.

Note: The above commands require only the first letter each command -any additional letters are ignored. This is the only module in which the general commands are not available.

18.0 RECYCLING

If a direct methods analysis provides only a partial fragment or fragments, some form of recycling is required to complete the structure. It is usually assumed initially that the fragment(s) are correctly oriented and positioned in the unit cell. Under these circumstances, four types of recycling are provided:

(1) Weighted Fourier calculations using the RECYCLE module. The fragment is recycled using a Sim weighted Fourier.
(2) Karle recycling using the RECYCLE module, the tangent formula is used to extend and refine phases.
(3) Karle recycling where the unknown phases are given random values using the RECYCLE module.
(4) Using type 4 groups in NORMAL.

We will deal with each of these in turn, and then the situation where the fragment(s) are in fact misplaced in the cell.

18.1 Weighted Fourier Recycling

This is the recommended recycling procedure and the best way to proceed is as follows:

(a) Set up an ASCII file containing the relevant CELL, LATTICE, SYMM, CONTENTS, SFAC, LIMITS, EDIT, TRANS and NOCHECK commands as required. These are all commands to NORMAL (See section 6.0). Add to this set the GROUP command without any parameters followed immediately by a set of atomic coordinates for the fragment which is to be recycled using the ATOM command. Store this as a model file.

(b) Now run RECYCLE with this file treated as a MODEL file (See section 5.1). The RECYCLE command must contain the keyword "FOURIER", all other parameters are ignored. RECYCLE calls module NORMAL.

(c) Either enter the command DEFAULT or call the MAPS module explicitly. You will get a weighted Fourier map. The whole sequence looks like this:

```
RECYCLE
FOURIER          ! Requests Fourier recycling.
MODEL            ! Scans MODEL file-else insert CELL etc.
DEFAULT          ! Runs MAPS under default.
```

Note that for the MODEL command to work the MODEL file must be attached to the job at run time (See Section 21.0).

18.2 Karle Recycling without Random Phases

The best way to proceed is as follows:

(a) Set up a model file containing the relevant CELL, LATTICE, SYMM, CONTENTS, SFAC, LIMITS, EDIT, TRANS and NOCHECK commands as required. These are all commands to NORMAL (See section 6.0) Add to this set the GROUP command without any parameters followed immediately by a set of atomic coordinates using the ATOM command. Store this as a model file.

(b) Now run RECYCLE with this file treated as a MODEL file (See section 5.1). The RECYCLE command must contain the keyword "KARLE", but unlike Fourier recycling, the parameters IK, NB, ISC and MAXDUP are available as on the NORMAL instruction.

(c) The program will now need to run the TRIPLETS, QUARTETS (if relevant), TANGENT and MAPS modules. Either call these explicitly or use the DEFAULT instruction in the usual way. You may control the refinement procedure in the same way as an a-priori analysis. You will finally obtain an E-map. E.g. for the simplest possible recycling enter the commands:

```
RECYCLE
KARLE
MODEL          ! Scans MODEL file
DEFAULT        ! Runs necessary modules under default.
```

18.3 Karle Recycling with Random Phases

This is run in exactly the same way as Karle recycling, but an explicit call to the RANTAN module is required in the correct place. For example, to run the recycling job outlined in section 18.2 above but with random phases for the unknowns, the following sequence of commands could be used:

```
NORMAL
KARLE
MODEL          ! Scans MODEL file
RANTAN 100     ! Asks for 100 phase sets to be generated
DEFAULT        ! MAPS will follow under default
```

18.4 Using Type 4 Groups in Normal

The type 4 groups in NORMAL are those which are correctly oriented and placed in the cell. To recycle in this way it is sufficient to run a typical a-priori phasing calculation but to include a:

GROUP 4

instruction followed by a set of ATOM commands in the input to NORMAL. Then proceed in the usual way. The use of the MODEL file as an auxiliary input is much recommended since it makes the input of the atomic coordinates less prone to error.

18.5 Incorrectly Placed Groups

If none of these recycling procedures are effective, then it is reasonable to assume that the fragment(s) are incorrectly placed. If they make chemical sense, however, they are probably correctly oriented in the unit cell. Under these circumstances recycling as in section 18.4, but using GROUPS of type 3 may well prove successful. If this does not work either, then relaxation to GROUPS of type 2 should be tried.

19.0 WHAT TO DO WHEN MITHRIL FAILS

This section outlines some of the options available to you if MITHRIL90 has failed to solve a structure. The possibilities are not quoted in any order of preference, except perhaps for the first set involving normalisation.

19.1 NORMAL

(1) Check the data. If there are many duplicates or systematic absences list them all by setting the MAXDUP parameter on the NORMAL command to 100000, or any large integer. Then investigate this list with suspicion-are you sure about the space group? Have you collected at least the unique data? How good is the resolution of the data? If there are missing reflections in your data reinsert them using the MISSING command.

(2) Read the output from NORMAL very carefully. Are the statistics sensible? Do the large E-magnitudes form a readily identifiable subset where some parity groups are missing? If some of the E-magnitudes are very large, it is often useful to use a LIMIT or EDIT command to reduce or remove them. The variation of E^2-1.0 as a function of Bragg angle is a good guide to the applicability of the calculated temperature factor. If there is a fall- off in this average as the angle increases, use the BSCL command to input a larger value of B than that calculated by the normalisation module. If some or all of the stereochemistry is known, then input this as a type 2 group. This is especially important with molecules containing planar fused rings.

(3) If you have generated the best set of E-magnitudes that you can, and the structure still will not solve, then a systematic distortion of the E's can be successful. Changes in the E- magnitudes cause changes in the invariants, and these in turn give rise to drastic modifications of the convergence map, and the subsequent phasing path. There are several ways of distorting the E's:

(a) Modify the unit cell contents using the CONTENTS command. Doubling the contents is the best starting point.

(b) Use artificially raised or lowered temperature factors via the BSCL instruction. Often only a small change in B gives rise to drastic changes in the E-magnitudes. This is especially recommended for situations where resolution is less than the Cu sphere.

(c) Insert a molecular fragment via the GROUP command that does not correspond to any group expected in the molecule.

(4) Try using the Bayesian normalisation technique by setting IK=2. If this does not work try another normalisation program that has different facilities. For

example the NORMAL module in the X-ray system permits the use of an overall anisotropic temperature factor. These E's can be input using IK=-1 on the NORMAL instruction. Nixon (1978) for example, has described an alternative approach to normalisation via the Patterson function.

19.2 Invariants - Triplets and Quartets

There are two problems which can arise here. One concerns a paucity of suitable relationships which can be common in situations of low symmetry. The other concerns the accuracy of the invariants themselves.

(1) If there is a paucity of triplets:

(a) Use quartets as well. Try just the negative ones first, then add the positives if this is unsuccessful.
(b) Increase the number of reflections on the TRIPLETS command.
(c) Reduce the minimum kappa-value from its default of 0.6 by using the KMIN command in CONVERGE. This will, however, introduce a number of very unreliable relationships.

(2) If there is a paucity of quartets:

(a) Increase the number of reflections on the QUARTETS command. This is usually the best way.
(b) Invoke the third neighbourhood via the NEIGH instruction.
(c) Ask for positive quartets as well. This can also be useful when triplets are scarce.
(d) Allow more missing second (and third) neighbours.

(3) If the problem is thought to be the reliability of the invariants then use the cut-off based on sin^2/λ^2, and/or the weighting scheme also based on sin^2/λ^2. These are less time consuming and have proven more useful than MDKS and L.E. when dealing with unreliable invariants.

(4) Even if the space group is not symmorphic, quartets often have a very beneficial effect on a direct methods analysis, and can be recommended as an option to try early in the list of weapons in the armoury. Only the negative quartets should be tried first, since they are independent of the triplets. Even a few four-phase invariants can drastically alter the phasing path.

(5) The MDKS and L.E. options coupled with convergence map weighting also has a drastic effect. The two options are different so if one is unsuccessful, it is worth trying the other. However, they are both very time consuming.

(6) Invariant generation is obviously of critical importance. MITHRIL90 allows a good deal of user control over the process, so use the options provided.

19.3 Convergence Mapping

Convergence mapping lies at the very heart of the multisolution approach to direct methods employed by MITHRIL90. It is essential to examine the convergence map carefully in cases of difficulty.

(1) Make sure that the starting set is a good one with all the starting set reflections used early in the phase determination. This can be checked by examining the bottom of the convergence map. If a starting set reflection is not used at all early on, then a better starting point can often be obtained by including at least one other reflection whose phase depends on that of the late starter. Introducing quartets may also have a similar effect.

(2) If there are gaps near the bottom of the convergence map (i.e. reflections with a zero estimated alpha and no invariants contributing), or the map is very 'thin' with many phases determined by only one or two relationships, then the phasing often fails. This can be remedied by increasing the size of the starting set or introducing higher invariants, particularly quartets.

(3) Be wary of the Sigma-1 determined phases. If they play a major role in the early stages of phasing, it is often worthwhile excluding them via the SIGMA NONE command. If MDKS or L.E. has been run, examine the s(1) triplet analysis. The triplets should have estimated cosines close to unity.

(4) If it still proves impossible to obtain a suitable convergence map without a massive amount of computer time, then several options are possible:

(a) Run MAGEX. Apart from origin and enantiomorph definition, it largely ignores the convergence map.

(b) Run YZARC. Try both least-squares and steepest descents - they give different results. The TOP_ instruction by-passes the convergence map.

(c) Run RANTAN instead of regular tangent refinement.

(5) Check to see if all the reflections at the bottom of the map have something in common e.g. they all have h even or k+l divisible by 3. If so, then make sure that the average value of E^2-1.0 is unity for such reflections. Renormalisation may be necessary. It may be possible to use the editing facilities of NORMAL to juggle these magnitudes. Try introducing new reflections into the starting set which do not belong to these groups.

(6) Altering the origin and enantiomorph is often unsuccessful, particularly if only small changes are made. The same relationships are still used in the early stages of phasing, but in a different form. For example, the triplet:

Psi{1} - Psi{2} + Psi{3}

may appear in one map generating Psi{1} from Psi{2} and Psi{3}. If the origin is partially re-defined by the user, this triplet may well appear again in a critical place but this time generating Psi{2} from Psi{1} and Psi{3}. If this triplet is erroneous it will be erroneous however it is used. This said, juggling with the starting set can be successful on some occasions, and is worth a try.

(7) Hand applied symbolic addition, even in a limited form, can give rise to possible relationships between phases, and these can be introduced into the convergence map by the RELATIONSHIP command. The relationships linking two phases (the pair relationships) are the most valuable. The inclusion of only one or two with high associated kappa-value will drastically alter a convergence map. There is the added bonus that symbolic addition can give valuable insights into the causes of phasing difficulties (Karle & Karle 1966). Do not use the convergence map for symbolic addition; get a list of triplets and work with this. The convergence procedure has too many weak relationships early in the phasing path.

19.4 Phase Expansion and Refinement

(1) The only way to monitor phase expansion and refinement is by inspecting the final figures of merit, so it is important to examine these closely in difficult cases. In particular do not just inspect the final CFOM's, but look also at the individual contributors:

(a) ABSFOM is the least reliable. If the ABSFOM values all tend to be large then the refined phases are over-consistent. Using Hull-Irwin weights will often give better results.

(b) NQEST and ψ_o are the most reliable figures of merit provided that the weak reflections have been accurately measured. (See section 4.0.) Do not expect NQEST to be very negative, particularly if quartets are being used actively in phase refinement. In these circumstances values around -0.1 are often satisfactory.

(c) Heavy atom cases often give extreme figures of merit. The correct solution may well be present even if the figures of merit seem unrealistic.

(d) Calculate the LOGLIK figure of merit. You do this by using the FOM keyword in the call to the TRIPLET module. It must never be examined on its own but always in conjunction with ψ_o.

(e) If all the phase sets have similar figures of merit too few invariants may be present.

(2) In case of pseudo-symmetry, the presence of heavy atoms or substantial planar moieties in the structure, use the Hull-Irwin scheme.

(3) Do not forget that there are two weighting schemes - if one does not work,

the other may, even using the same starting set and convergence map.

(4) Be careful of the early stop option. It is often found that the set selected by the TANGENT module during refinement as being an obvious solution is not the correct one. The NOSTOP command will prevent this happening.

(5) If you are using early figures of merit, and most sets are getting rejected, it is possible that the correct set is also being discarded. It is an easy matter to turn off these figures of merit. Do not use the first EFOM for planar ring structures, and set the cut-off for the third EFOM to at least 1.7.

(6) Do not confine your attentions to the one or two phase sets with the highest CFOM's. It may be necessary to examine maps with quite low associated CFOM's.

(7) The CFOM's are dependent on the relative weights of the individual figures of merit. Adjusting the weights to reflect your own intuition concerning the contributing figures of merit will often result in a drastic re-ranking of the solutions.

(8) In a case with over-consistent phases and a centrosymmetric structure it is strongly advised to use the SYMB module in preference to the TANGENT module.

19.5 E-Maps

Look at the resulting E-maps carefully. Remember that the interpretation assumes that you have well-resolved peaks, and this may not be the case. The routines which perform the chemical interpretation are quite sophisticated, but they are never as good as a trained crystallographer. Do not, therefore, accept the given interpretations as the only possibilities. Some other points to note are:

(a) If the map contains one or two large peaks and no heavy atoms are expected, the phases are probably incorrect - but not always. Sometimes something can be salvaged. If heavy atoms are present direct methods will probably only produce these atoms.

(b) If the E-maps show pseudosymmetry switch to Hull-Irwin weights in TANGENT.

(c) One or two missing peaks coupled with one or more spurious ones, can quickly make a map uninterpretable. Increasing the number of peaks can make parts of the map more readily interpretable at the expense of producing more noise peaks.

In very difficult cases be tenacious. If just a small portion of the expected structure is found, stick with it through all the possible recycling schemes. It may be correctly oriented but misplaced, so the use of type 3 groups in NORMAL will be useful, but this procedure is not infallible. It sometimes does not work even with correct information.

20.0 EXAMPLES

(1) A first time run, batch mode:

```
LEVEL 0          ! This may be your installation default
NORMAL           ! Calls the NORMAL module.
CELL             ! Unit cell.
SYMM             ! As many commands as needed.
LATTICE          ! Lattice type.
CONTENTS         ! Unit cell contents.
DEFAULT          ! TRIPLETS .... MAPS under default.
```

Notes.
(a) The commands CELL through CONTENTS could be stored in a model file and accessed via the command MODEL .
(b) For interactive mode use LEVEL 1, 2 or 3.

(2) A situation as in (1) but RANTAN is required instead of the standard TANGENT module; 100 phase sets are required with Hull-Irwin weights. The Fourier module is to follow, run under defaults.

```
LEVEL 0          ! Set this as early as possible.
NORMAL           ! Calls the NORMAL module.
CELL             ! Unit cell.
SYMM             ! As many commands as needed.
LATTICE          ! Lattice type.
CONTENTS         ! Unit cell contents.
RANTAN 100       ! TRIPLETS-CONVERGE run first under default
SWTR             ! Hull-Irwin scheme.
DEFAULT          ! MAPS under default.
```

(3) As in (1) but MAGEX is to be run with default options between CONVERGE and TANGENT.

```
LEVEL 0          ! Set this as early as possible.
NORMAL           ! Calls the NORMAL module.
CELL             ! Unit cell.
SYMM             ! As many commands as needed.
LATTICE          ! Lattice type.
CONTENTS         ! Unit cell contents.
MAGEX            ! TRIPLETS,(QUARTETS),CONVERGE run first
DEFAULT          ! TANGENT and MAPS under default.
```

(4) A full interactive run with user control on all modules:

```
LEVEL 2          ! Set this as early as possible.
NORMAL           ! Calls the NORMAL module.
CELL             ! Unit cell.
SYMM             ! As many commands as needed.
LATTICE          ! Lattice type.
CONTENTS         ! Unit cell contents.
TRIPLETS 400     ! Triplets for top 400 E's.
L.E.             ! Use linear equations option.
QUARTET 100      ! Quartets for top 100 E's.
NEIGH 0,20       ! Invokes 3rd neighbourhood.
CONVE            ! Convergence map.
SIGMA NONE       ! No S 's- note keyword.
SETS 0,32        ! A minimum of 32 phase sets.
MDKS 1.2,1,1     ! Weight triplets from L.E. calculation.
YZARC 60         ! YZARC called-60 phase sets.
RANDOM 3 101     ! Seed random numbers.
TANGENT          ! Call tangent refinement.
NOSTOP           ! Don't stop for good solutions.
MAPS ALL         ! All the maps to be examined.
LEVEL 3          ! Highly interactive.
                 ! Blank line initiates MAPS
CONVERGE         ! No success-go back to converge.
SETS 128         ! 128 phase sets.
TANGENT          ! Tangent refinement again.
SKIP 0           ! Ignore previous YZARC phases.
LEVEL 2          ! Lower interactive level.
DEFAULT          ! TANGENT and MAPS under default.
```

Note the ability to go backwards as well as forwards.

(5) It is not necessary to start with NORMAL each time, as long as the required files have been made permanent. See section 21.0. In this example the files from NORMAL and TRIPLETS are both permanent files, so that we can enter CONVERGE without re-running these modules. This is an essential feature of jobs where invariant generation is time consuming.

```
LEVEL 2          ! Interactive job.
CONV             ! Calls CONVERGE.
LIST ALL         ! Full list.
NOPRINT          ! No printer output.
MAPS -4          ! TANGENT then 4 best maps.
MAGEX            ! Back to MAGEX.
PRINT            ! Printer output on again.
YZARC            ! MAGEX no good-try YZARC.
TOP              ! Ignore Convergence map.
MAPS -5          ! Tangent then 5 best maps.
RANTAN 200       ! Back to RANTAN-200 sets.
END              ! Run RANTAN then exit.
```

(6) Karle recycling with full user control.

```
RECYCLE KARLE
MODEL                ! All atoms etc on MODEL file.
DATA 1 (3I3,2F10.2)  ! Data in card image form.
TRIPLET 450          ! Triplets for top 450 E's.
QUARTET 125          ! Quartets for top 125 E's
TANGENT              ! Tangent refinement.
MAPS                 ! E-map.
END                  ! Run MAPS then stop.
```

(7) Sharpened, origin-removed Patterson:

```
NORMAL           ! Normalise F's first.
MODEL            ! CELL etc. on MODEL file.
PATTERSON        ! Run Patterson.
END
```

(8) Normalisation using the Bayesian technique, with all CELL, DATA etc. commands stored on a model file. Run in batch mode.

```
LEVEL 0    ! Set this as early as possible
NORM 2     ! The 2 calls the Bayesian module
MODEL      ! All commands read from model file
DEFAULT    ! TRIPLETS.....MAPS on defaults
```

(9) An interactive run using the Bayesian normalisation technique, the LOGLIK figure of merit, and reviewing the phases sets after tangent.

```
LEVEL 2          ! Run in interactive mode
NORMAL 2         ! Perform a Bayesian normalisation
MODEL            ! All NORMAL commands on model file
TRIPLET FOM      ! Run the triplets in preparation for the F.O.M.
                 ! Blank line to start triplet module
TANG             ! Run (QUARTETS), CONVERGE under defaults before
                   tangent
NOSTOP           ! Don't stop for a good solution
                 ! Blank line to start tangent refinement
REVIEW           ! Call review module
ABSFOM           ! Sort and list sets on Absolute figure of merit
P                ! Sort and list sets on Q figure of merit
C                ! Sort and list sets on Combined figure of merit
E                ! Exit REVIEW return to main menu
MAPS 145 83 27   ! Calculate E-map for each of the three sets
END              ! Once the maps are calculated end
```

(10) Run SYMB from E-magnitude and invariant files already present (See example 5)

```
CONV     ! Call converge module
SYMB     ! Run converge module under defaults and call SYMB
ORIGIN   ! Use SYMB origin in preference to Converge origin
REVIEW   ! Call review
LOGLIK   ! Sort sets on the LOGLIK figure of merit
C        ! Sort sets on the combined figure of merit
E        ! Exit REVIEW and return to main menu
MAPS     ! Calculate E-map for set with the highest CFOM
END      ! End run after map is calculated
```

Note: The full command need not be entered :- only the first four letters for the majority of commands, the first two letters of keywords and only the first letter of all commands used in the REVIEW module. Upper and lower case are both accepted.

21.0 COMPUTER FILES

This section describes the files used by MITHRIL90. It is of importance to both programmer and user, particularly for users who are likely to be working with difficult structures, and hence will need to store certain information (such as invariants) on permanent rather than scratch files. The files that MITHRIL90 uses are as follows : (the words in parenthesis are the FORTRAN variable names)

FILE 1 (NOUT) : A null file used by the NOPRINT option. Execution of the NOPRINT command causes the printer output to be set to this channel.

FILE 3 (NSPEC) : The secondary output file used in the batch mode for output of warnings, and a summary of input commands. This should be a print file.

FILE 4 (NSPEC) : The user terminal in the interactive mode. Setting the LEVEL parameter to 1, 2 or 3 causes the secondary output file defined above for channel 3 to be changed to this channel.

FILE 5 (NIN) : The input channel. This is a terminal for interactive use, or a disc file in batchmode. Note that changing from (say) interactive (LEVEL 1 or 2 or 3) to batch (LEVEL 0) does not change the input device.

FILE 6 (NOUT) : The standard printer output file.

FILE 8 (NTAPEA) : Contains the results of NORMAL for input to TRIPLETS, QUARTETS, CONVERGE, PATTERSON and for MAPS when Fourier recycling is invoked. It contains the cell dimensions and contents, the symmetry, the group atoms, a full set of E's and the subset greater than 1.0. This is usually a scratch file, but your system should permit it to be a permanent user file as well. It is a binary, unformatted file. It is also an input file for the MICE maximum entropy program.

FILE 9 (NTAPEB) : Contains invariants. The triplets always appear first, then the quartets if they have been generated. This file is necessary for input to CONVERGE - all runs of CONVERGE require it. It is a binary, unformatted file. Allowance should be made for making it a permanent, user file.

FILE 10 (NTAPEC) : Contains the results of CONVERGE. All modules which follow CONVERGE read this file, except MAPS and PATTERSON. Apart from the convergence map, it contains symmetry information, the cell dimensions and a collated set of invariants. It is sufficient input for TANGENT, RANTAN, MAGEX and YZARC, which do not need to read Files 8 and 9. It is a binary, unformatted file.

FILE 11 (NTAPED) : Contains the phase sets from TANGENT or RANTAN for input to MAPS. The latter module needs only this file as input. It is a binary,

unformatted file.

FILE 12 (NTAPEE) : This is a binary file used by most modules. The MAPS module writes an electron density map onto this file after a FOURIER calculation. If you intend to examine this map using the PLOTQ program it must be a permanent file. Note Only the last map to be calculated is stored.

FILE 13 (NTAPEF) : This is a binary scratch file used by most modules. There is no need for it to be permanent.

FILE 14 (NTAPEG) : This file contains the user intensity data. The DATAIN routine of NORMAL will be modified for your installation to read this file if no DATA command is encountered whilst running NORMAL. The same file is also used when a DATA command is encountered, the difference being that formatted data is expected on NTAPEG, which means that there is one extra command to remember to use each time NORMAL is run. This is usually a permanent file.

FILE 16 (NTAPEH) : This is the MODEL file, which will be a permanent, formatted user file capable of being edited via your system editor. It can contain CELL, SYMM, LATT, GROUP, ATOM, CONTENTS, and LATTICE commands. It is only accessed if the MODEL command is issued. The use of MODEL files is strongly recommended

FILE 17 (NTAPEI) : This is a formatted file written by MAPS. It contains a list of peaks in the form of ATOM commands which can be subsequently edited, and placed in a MODEL file and used in recycling procedures. It removes some of the tedium, and errors inherent in entering a large set of atomic coordinates manually.

FILE 21 : Contains the results of the TRIPLET module, when it is run using the keyword FOM, on the call line. These results are used in TANGENT for the calculation of the LOGLIK figure of merit. It is a binary scratch file but due to the time involved in its generation it is recommended that this file is made permanent.

22.0 FUTURE ENHANCEMENTS

The following enhancements to the MITHRIL90 package are either planned or underway :

(1) Divergence mapping to follow the convergence map as a method of strengthening the convergence and phasing processes.

(2) Tests for phase oscillation during tangent refinement.

(3) The use of covalent and van der Waals' radii in the interpretation routines of the MAPS module to assist the derivation of chemically sensible fragments, and the identification of incorrectly positioned fragments.

(4)The incorporation of the SAYTAN (SAYre TANgent refinement) program.

23.0 PLOTQ

PLOTQ is a stand-alone program designed to display MITHRIL90 electron density maps on a graphics terminal. The program is fully interactive, menu driven and requires no external documentation. Those wishing further information on PLOTQ are referred to Henderson, Bannister & Gilmore (1990). Densities may be displayed as :

(i) Single sections viewed along any crystallographic axis.

(ii) Multiple sections viewed along any crystallographic axis.

(iii) A three dimensional surface plot viewed from any direction

For small molecules its principal uses are :

(i) To examine situations of disorder, either molecular or solvent.

(ii) As a final check at the end of Fourier synthesis to detect unaccounted for density

(iii) To define molecular envelopes for larger structures where peak searching is inappropriate, or smaller structures where an uninterpretable peak list has been produced.

The MITHRIL90 maps file (output on channel 12, NTAPEE) must be a, permanent file for this program to run. It is possible to superimpose atom positions onto a map by reading the relevent model file whilst running PLOTQ.

24.0 PROGRAMMING CONSIDERATIONS

This section is only intended for those users who are setting up MITHRIL90 on their own computer, or who wish to modify it. You should read Section 21.0 on files first.

To run MITHRIL90, your computer should provide the following facilities:

(1) 32 bit integers and floating point numbers with facilities for double precision.

(2) The ability to address at least 3 Mbyte of memory either directly or via memory management. If the program is not to use virtual memory, then 3 Mbyte of physical memory will be needed. Section 24.2 has a full discussion of memory requirements.

(3) A FORTRAN 77 compiler.

(4) At least 5 Mbytes of disc space for files. See section 21.0.

(5) In addition, to run in the interactive mode, a terminal is required. For the graphics options then the addition of graphics boards is obviously necessary.

As supplied, the program is written for a Concurrent computer running under the UNIX operating system. There are a few machine specific lines of code which may need to be altered for your own machine. Each section of machine specific code is marked by a comment statement which states "Warning machine specific". They are as follows:

(1) You may need a PROGRAM statement on the first line of the main routine.

(2) The BLOCK DATA subroutine has a name (INITAL). Some compilers may object to this (in which case remove the name), or they may have their own naming requirements (adjust accordingly).

(3) Subroutine CLSCN(N) is machine specific. Its purpose is to clear the screen of a terminal after an N second delay (to allow the user time to scan the last output). Two versions are supplied, one for a VAX running under VMS, and one for UNIX. If you are not using one of these operating systems, you will need to make your own version, or you can just leave a dummy routine. This latter option is not recommended since it means that the screen output will not be so neat. However, if you never intend to use the interactive mode of MITHRIL90, then it will be sufficient.

(4) Subroutine DATAIN in NORMAL reads the raw intensity data. You will need to tailor it to your own installation. There are some comments on this in the subroutine itself.

(5) Several subroutines use DATA statements to initialise non-character variables to characters. This is legal FORTRAN IV, but your compiler may object. In this case set the offending variables to CHARACTER types, and the problem should disappear.

(6) There is extensive use of END= and ERR= in READ statements. Some compilers may object. Dummy reads are used, and so are partial reads of a logical record. You may need to pad out these READ statements.

24.1 Graphics Routines

The graphics facilities use the University of Bradford "SIMPLEPLOT" routines. These are not supplied with MITHRIL90, and can be obtained from:

D. Butland
Bradford University Research Limited,
University of Bradford,
Richmond Road,
Bradford BD7 1DP
U. K.

However, you may well have your own graphics routines which are quite suitable. The routines called by MITHRIL90 are as follows:

(1) CALL NEWPLT(XLEFT,XRIGHT,XCMS,YBTM,YTOP,YCMS)

XLEFT - Horizontal scale limit for left edge.
XRIGHT - Horizontal scale limit for right edge.
XCMS - Length of horizontal scale in cms.
YBTM - Vertical scale limit for bottom edge.
YTOP - Vertical scale limit for top edge.
YCMS - Length of vertical scale in cms.

All the parameters are real. This is the basic setting-up routine. It defines the user's Cartesian units for all subsequent plots, and clears the screen.

(2) CALL NUMBPT (X,Y,I,N)

X - x-coordinate of a point in units of the graph scales.
Y - y-coordinate of a point in units of the graph scales.
I - Integer defining type of symbol placed at point (x,y).
N - Integer number to label point.

This draws a symbol centred at the point (x,y). These points are real numbers.

(3) CALL JOINPT (X,Y)

This draws a straight line from the current pen position ,if it was reached by a previous JOINPT, to the specified point (x,y). Each first call of JOINPT moves the pen to the point (x,y) without drawing. BREAK is used if a gap is required between groups of joined points.

(4) CALL BREAK

The first call of JOINPT after calling subroutine BREAK moves the pen to the point defined in the call to JOINPT without drawing a line.

(5) CALL VDUMOD (I)

I=1 is used to take the terminal into graphics (4010) from alphanumeric mode. I=-1 is used to return the terminal from graphical (4010) mode to standard alphameric mode.

(6) CALL TITLE(IV,IH,TITLE,N)

IV,IH integer values for the vertical and horizontal positions of the title respectively.
TITLE the title string.
N the number of characters in the TITLE string.

This is used to output text to the screen.

(7) CALL CLOSE(1)

Closes all graphical work.

24.2 Memory Requirements

The requirement of 3 Mbyte of memory in which to run MITHRIL90 may not be a problem if you have an operating system which gives you virtual memory capabilities. There is, however, one problem which can arise with virual memory. The COMMON block labelled /BLK1/ requires 116K words (464K bytes). The TRIPLETS, and QUARTETS modules use this block to store a full hemisphere of E-magnitudes packed in a singly dimensioned array IESN(100000). This array is accessed in a wholly random way, and some virtual memory systems may spend an excessive amount of time paging because of this. Some operating systems allow you to specify data areas which may not be paged (i.e. which are memory resident); if this is possible in your case, then use it to keep IESN in memory. Alternatively, even a virtual machine environment may permit overlays or segmentation to reduce memory requirements, and hence the need for pagination.

25.0 REFERENCES

BAGGIO, R., WOOLFSON, M.M., DECLERCQ, J.P. & GERMAIN, G. (1978). *Acta. Cryst.* **A34** , 883-892.

BUSETTA, B. & COMBERTON, G. (1974). *Acta Cryst.* **A30** , 564-568.

BUTLAND, J., SIMPLEPLOT *User's Handbook , Report No. 253*, University of Bradford.

COCHRAN, W. & WOOLFSON, M.M. (1955). *Acta. Cryst.* **8**, 473-478

DECLERCQ, J.P., GERMAIN, G. & WOOLFSON, M.M. (1975). *Acta Cryst.* **A31**, 367-372.

DECLERCQ, J.P., GERMAIN,G. & WOOLFSON M.M. (1979). *Acta. Cryst.* **A35**, 622-626.

DeTITTA, G.T., EDMONDS, J.W., LANGS, D.A. & HAUPTMAN, H. (1975). *Acta. Cryst.* **A31**, 474-479.

FREER, A.A. & GILMORE C.J. (1980). *Acta Cryst.* **A36**, 470-475.

GERMAIN, G. & WOOLFSON, M.M. (1968). *Acta Cryst.* **B24**, 91.

GIACOVAZZO, C. (1980a). *Direct Methods in Crystallography*, Academic Press.

GIACOVAZZO, C. (1980b). *Acta Cryst.* **A36**, 74-82.

GILMORE, C.J. (1977). *Acta Cryst.* **A33**, 712-716.

GILMORE, C.J., BRICOGNE, G. & BANNISTER, C. (1990) *Acta Cryst.* **A46**, 297-308

GILMORE, C.J. & BROWN, S.R. (1988). *Acta Cryst.* **A44** , 1018-1021.

GILMORE, C.J., HARDY, A.D.U., MacNICOL, D.D. & WILSON, D.R. (1977). *J.C.S. Perkin* **II**, 1427-1434.

GILMORE, C.J., & HAUPTMAN, H. (1985) . *Acta Cryst.* **A41**, 457-462.

HALL, S.R. & SUBRAMANIAN, V. (1982). *Acta Cryst.* **A38**, 598-608

HAUPTMAN, H. (1972). *Crystal Structure Analysis: The Role of the Cosine Seminvariants*, New York : Plenum Press.

HAUPTMAN, H. (1977a). *Acta Cryst.* **A33**, 556-564.

HAUPTMAN, H. (1977b). *Acta Cryst.* **A33**, 565-568.

HAUPTMAN, H. (1980). In *Theory and Practice of Direct Methods in Crystallography*, pp 151-197. Ed. M.F.C.Ladd & R.A.Palmer, Plenum Press.

HENDERSON, R.K., BANNISTER, C. & GILMORE, C.J. (1990). *J. Appl. Cryst.* **23**, 143-144

HULL, S.E. & IRWIN, M.J. (1978). *Acta Cryst.* **A34**, 863-870.

HULL, S.E., VITERBO, D., WOOLFSON, M.M. & SHAO-HUI, Z. (1981). *Acta Cryst.* **A37**, 566-572.

KARLE, J. (1967). *Acta Cryst.* **B24**, 182-186.

KARLE, J. & KARLE, I. (1966). *Acta Cryst.* **21**, 849-859.

MAIN, P. (1976). In *Crystallographic Computing Techniques* pp 97-105, Copenhagen: Munksgaard.

MAIN, P. (1978). *Acta Cryst.* **A34** , 31-38.

MAIN, P. (1980). In *Computing in Crystallography* pp 8.01-8.13, Indian Academy of Sciences.

MAIN, P & HULL, S.E. (1978). *Acta Cryst.* **A34**, 353-361.

MAIN, P., FISKE, S.J., GERMAIN, G., HULL, S.E., DECLERCQ, J.P., LESSINGER, L. & WOOLFSON, M.M. (1980). *Multan-80 A System of Computer Programs for the Automatic Solution of Crystal Structures from X-ray Diffraction Data*, Univ. of York.

MOORE, F.H. (1963). *Acta Cryst.* **16**, 1169-1175.

NIXON, P.E. (1978). *Acta Cryst.* **A34**, 450-453.

RODGERS, D. (1965). In *Computing Methods in Crystallography* pp 117-148. Ed. J.S.Rollett, Pergamon Press.

RODGERS D. (1980). In *Theory and Practice of Direct Methods in Crystallography* pp. 82-92, Ed. M.C.F. Ladd & R.A. Palmer, Plenum Press.

SIM, G.A. (1959). *Acta Cryst.* **12** , 813-815.

SIM, G.A. (1960). *Acta Cryst.* **13** , 511-512.

WHITE, P.S. & WOOLFSON, M.M. (1975). *Acta Cryst.* **A31**, 53-56.

YAO JIA-XING (1981). *Acta Cryst.* **A37**, 642-644.

YAO JIA-XING (1983). *Acta Cryst.* **A39**, 35-37.