Acta Cryst. (1993). D**49**, 37–60

Direct Phase Determination by Entropy Maximization and Likelihood Ranking: Status Report and Perspectives

BY G. BRICOGNE

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England, and LURE, Bâtiment 209D, Université Paris-Sud, 91405 *Orsay, France*

(*Received* 29 *June* 1992; *accepted* 13 *October* 1992)

Abstract

A new multisolution phasing method based on entropy maximization and likelihood ranking, proposed for the specific purpose of extending probabilistic direct methods to the field of macromolecules, has been implemented in two different computer programs and applied to a wide variety of problems. The latter comprise the determination of small crystal structures from X-ray diffraction data obtained from single crystals or from powders, and from electron diffraction data partially phased by image processing of electron micrographs; the *ab initio* generation and ranking of phase sets for small proteins; and the improvement of poor quality phases for a larger protein at medium resolution under constraint of solvent flatness. These applications show that the primary goal of this new method - namely increasing the accuracy and sensitivity of probabilistic phase indications compared with conventional direct methods - has been achieved. The main components of the method are (1) a tree-directed search through a space of trial phase sets; (2) the saddle-point method for calculating joint probabilities of structure factors, using entropy maximization; (3) likelihood-based scores to rank trial phase sets and prune the search tree; (4) efficient schemes, based on error-correcting codes, for sampling trial phase sets; (5) a statistical analysis of the scores for automatically selecting reliable phase indications. They have been implemented to varying degrees of completeness in a computer program (*BUSTER*) and tested on two small structures as well as on the small protein crambin. The main obstructions to successful *ab initio* phasing in the latter case seem to reside in the accumulation of phase sampling errors and in the lack of a properly defined molecular envelope, both of which can be remedied within the methods proposed. A review of the Bayesian statistical theory encompassing all phasing procedures, proposed earlier as an extension of the initial theory, shows that the techniques now available in *BUSTER* bring closer a number of major enhancements of standard macromolecular phasing techniques, namely isomorphous replacement, molecular replacement, solvent flattening and non-crystallographic symmetry averaging. The gradual implementation of the successive stages of this 'Bayesian programme' should lead to an increasingly integrated, effective and dependable phasing procedure for macromolecular structure determination.

0. Introduction

The *ab initio* determination of macromolecular structures by a purely computational solution to the phase problem used to be the secret dream of most protein crystallographers beginning their careers. A whole decade of biotechnological frenzy may however have distracted the attention of many away from this lofty challenge, towards applications whose feasibility does not require such a radical advance, and it is therefore a particular pleasure to see a meeting unashamedly devoted to precisely this high-risk endeavour. I hope it will result in a rekindling of interest for the subject among the younger generation of crystallographers.

In 1984 I proposed a new multisolution phasing method based on entropy maximization and likelihood ranking for the specific purpose of extending probabilistic direct methods to the field of macromolecules (Bricogne, 1984*a*, hereafter called I). The need for such a new approach followed from a critical analysis of direct methods as practised at the time - a snapshot of which was captured, for instance, in the papers devoted to this topic at the Ottawa computing school (Sayre, 1982). This analysis brought to light (I, §2) a number of shortcomings in the implementation of probabilistic ideas within direct methods which made them unfit to tackle macromolecules, namely:

(1) the ubiquitous assumption of a uniform prior distribution of the atoms in the crystal in the derivation of probabilistic formulae;

(2) the intrinsic limitations of the Edgeworth series as an approximation to joint-probability distributions involving many structure factors with large moduli;

(3) the reinterpretation, for practical purposes, of sharply peaked joint-probability distributions as yielding 'estimates' of triplet and quartet phase invariants, thus giving rise to as many 'equations' which were to be 'solved for the phases'.

These objections could be - and were - regarded as rather academic in the field of small structures: (1) is only a mild concern because small-molecule crystals are usually close-packed; as for (2) and (3), the enormous overdetermination available in a copper sphere of data (typically 50 to 100 observations per non-hydrogen atom) can be relied upon to compensate for any but the most severe methodological deficiencies. In the macromolecular field, how-

0907-4449/93/010037-24$06.00

ever, the situation is much less forgiving: the conjunction of (1) and (2) results in the relations normally considered in (3) being not only very weak (because they involve only a few phases at a time, and the number of atoms is large) but also systematically in error (because the presence of solvent regions creates extreme non-uniformity in the distribution of atoms belonging to the macromolecule); and the degree of overdetermination in the data is, except in rare cases, much reduced or even non-existent, hence unable to remedy these deficiencies.

As a consequence of this analysis, the main thrust of the work published in (I) was to urge a return to the fundamental problem of calculating joint-probability distributions of structure factors and to find methods better suited to the macromolecular field which would *increase the accuracy and the sensitivity of probabilistic phase indications*. For this purpose I proposed the saddlepoint method as a powerful enhancement of the analytical techniques previously used, showed it to be closely related to entropy maximization, and demonstrated both its power and its practical applicability by a test calculation on a macromolecule. I also pointed out the necessity of adopting a tree-directed search strategy and of using likelihood as a look-ahead criterion to guide the search according to the principles of Bayesian inference. Readers wishing to gain some familiarity with the basic notions and the terminology used throughout this work (*e.g.* Edgeworth series, maximum entropy, saddlepoint method, likelihood) are advised to consult the present writer's contribution (Bricogne, 1991*a*) to a collection of expository essays on the maximum-entropy method. A cursory glance through this paper should at least fulfil the rôle of a glossary.

An implementation of these ideas has been taking place in collaboration with Chris Gilmore and coworkers in the form of a computer program called *MICE* (Bricogne & Gilmore, 1990). The basic validity of the approach was demonstrated by the solution of some test small-molecule structures (Gilmore, Bricogne & Bannister, 1990) and by the successful identification of the best phase sets produced by the *SAYTAN* procedure of Woolfson & Yao (1990) for a small protein (Gilmore, A. N. Henderson & Bricogne, 1991). An adaptation of the likelihood criterion to intensity data corrupted by overlap (Bricogne, 1991*b*) also enabled *MICE* to successfully tackle the *ab initio* determination of small inorganic structures from powder diffraction data (Gilmore, K. Henderson & Bricogne, 1991; Shankland, Gilmore, Bricogne & Hashizume, 1992). Another related application has been in electron crystallography (Dong, Baird, Fryer, Gilmore, MacNicol, Bricogne, Smith, O'Keefe & Hovmoller, 1992) where *MICE* was able to extend initial phases to 3 Å resolution for a projection of perchlorocoronene, obtained from electron micrographs by image-processing techniques, to the full set of 1 Å resolution amplitudes measured by electron diffraction. In macromolecular crystallography Charles Carter and co-workers have shown at this meeting (Xiang, Carter, Bricogne & Gilmore, 1993) that *MICE* can be used to carry out phase extension very effectively by entropy maximization to maximum likelihood, a technique which yields results far superior to conventional solvent flattening when some initial phase information is available together with a sufficiently good molecular envelope. In all these applications it was noted that the combination of the saddlepoint approximation and of likelihood as a decision criterion did provide a degree of sensitivity hitherto unachievable by other methods.

The theoretical work initiated in (I) was subsequently extended (Bricogne, 1988*a*, hereafter called II) by incorporating into the calculation of probabilities and likelihoods all sources of phase information normally used in macromolecular crystallography, thus defining an extensive research and programming project. Here I will describe my own computer program called *BUSTER* (Bricogne, 1991*d*,*e*), the purpose of which is precisely to serve as an experimental medium for the systematic implementation of this comprehensive Bayesian scheme. The program has been written entirely from scratch rather than as an extension of an existing direct-methods package or tradition. It incorporates many hitherto unpublished items of what might be called 'mathematical technology', of which the saddlepoint method may be viewed as the first. These have been developed to provide new and sometimes radical solutions to several classical problems of direct methods such as the processes of normalization, of origin definition, of enantiomorph discrimination, of phase permutation and of extraction of reliable phase indications. These are described here at the level of pre-publication outlines, pending fuller accounts in separate papers. The applications have been directed specifically towards macromolecular structures, although smaller ones have also been used for testing.

The present meeting is the first opportunity to present this body of work in the very context in which it was conceived. I hope this will result in a clearer perception of its internal logic and of the mathematical technology it mobilizes, both by macromolecular crystallographers and by those practitioners of conventional direct methods who feel that the accumulation of practical experience had perhaps obscured the mathematical essence of the subject.

The programming project defined in (I) and (II) is still only in the early stages of its realization. However the applications carried out so far demonstrate beyond reasonable doubt that the expected gains in accuracy and sensitivity have indeed materialized, and suggest strongly that the *ab initio* phasing of macromolecular structures is now within reach.

1. Overview of theory and implementation

1.1. *Theoretical summary*

The rationale of the theory on which this work is based has been described in many publications (Bricogne, 1982, 1984*a*,*b*, 1988*a*,*b*, 1991*a*,*b*,*c*) so it will only be outlined here. It views an unknown crystal structure as made up of

atoms with known chemical identity but unknown positions, and considers the latter as random, with an initially uniform distribution in the asymmetric unit of the crystal. Structure determination consists in the gradual removal of that randomness. For this purpose, limit theorems of probability theory are invoked to estimate the joint-probability distribution of suitably chosen structure factors; formal substitution of the observed amplitudes of the latter into these distributions is then expected to yield conditional joint distributions for the phases, indicating that certain combinations of phase values are more probable than others once the amplitudes are known.

Traditionally the *Edgeworth series* was used to approximate joint distributions, but this turns out to be unsuitable for large structure-factor amplitudes. In addition, this series was previously derived under the assumption that the distribution of the random atomic positions is always uniform - a severe handicap in the macromolecular field because of the existence of solvent regions. These limitations can be overcome simultaneously by using the *saddlepoint method* instead of the Edgeworth series; this always yields optimal estimates of joint probabilities involving large amplitudes, and is equivalent to requiring that the distribution of random atomic positions should be updated whenever phase assumptions are made so as to retain *maximum entropy* under the constraints embodied in these assumptions. As discussed elsewhere (Bricogne, 1991*a*) this approach does not amount to a wholesale adoption of the so-called 'maximum-entropy principle' but instead to a consolidation of the analytical apparatus of direct methods.

This revised statistical analysis of the phase problem leads naturally to a general multisolution strategy of structure determination (I, §2.4, §8.1) in which the space of hypothetical phase sets is to be explored in a hierarchical fashion by building a search tree, similar to those used in game-playing computer programs. Each trial phase set is ranked according to a certain statistical criterion (the *log-likelihood gain*, or the Bayesian score) which acts as a heuristic function in guiding the subsequent growth of the tree.

1.2. *Intuitive content of the method*

In spite of the relative mathematical complexity of its detailed formulation, this method has a simple intuitive content which may be easily grasped by reference to the use of 'structure-factor graphs' (Bragg & Lipson, 1936; Lipson & Cochran, 1968, pp. 77-82) in the structure determination of penicillin (Crowfoot, Bunn, Rogers-Low & Turner-Jones, 1949) or of outstandingly strong reflexions (Lipson & Cochran, 1968, pp. 134-136) in solving hexamethylbenzene (Lonsdale, 1929) and coronene (Robertson & White, 1945). In these procedures a strong Bragg reflexion is interpreted as indicating a bias, away from uniformity, in the distribution of atomic positions. This leads to placing any strong scatterer (in the first procedure), or planar molecular fragment (in the second procedure), near the maxima of the structure-factor graph for that reflexion, in accordance with its known or assumed phase. For very small structures a strong scatterer or fragment tentatively placed in this way can be treated as a 'heavy atom' and used to try and complete the structure. The chances of success of such an attempt will depend crucially on whether the tentative placement of the strong scatterer or fragment can account not only for the strong reflexions used in inferring that placement, but also for the most salient features in the pattern of intensities in the rest of the data: if these features are correctly predicted, there is a good chance of being able to complete the structure by subsequently placing the remaining atoms.

The present method consists in a statistical generalization of this simple idea, in which assumed phases for a 'basis set' of strong reflexions give rise not to a tentative position for some dominant scatterer but to a tentative redistribution of the random positions of all scatterers. By virtue of the maximum-entropy criterion used in its construction, this redistribution is the most non-committal with respect to the rest of the data: it is therefore a much safer decision than the tentative placement of a 'heavy atom', so that it can use reflexions which are merely strong rather than outstanding, and scatterers which need not be dominant. In spite of this minimal commitment, the maximum-entropy redistribution of the scatterers cannot help but predict (*via* maximum-entropy extrapolation) certain biases in the intensity distribution outside the basis set. The log-likelihood gain affords a measure of the degree of corroboration of that prediction (and hence of the basis-set phase assumptions on which it is based) by the observed intensities of the non-basis reflexions, and hence the increase in the chances of being able to 'complete the structure' by introducing more reflexions into the basis set. In this sense it is a 'look-ahead' criterion.

To summarize, the present multisolution strategy may be thought of as a statistical, quantitative but 'model-free' version of the intuitive and simple methods mentioned above.

1.3. *Principles of implementation*

The purpose of this section is to establish the terminology which will be used throughout in describing the mathematical methods, in analysing the detailed logic of the program and in discussing the results.

The first step is always the determination of global scale and temperature factors to put the observed amplitudes on absolute scale, allowing the calculation of unitary and normalized structure-factor amplitudes $|U_{\mathbf{h}}|$ and $|E_{\mathbf{h}}|$. This normalization is carried out by a maximum likelihood method, allowing the use of an arbitrary non-uniform distribution $m(\mathbf{x})$ for the random atomic positions and the incorporation of known partial structures and/or non-crystallographic symmetries. The resulting absolute scale is not considered as having a permanent value, but is instead recalculated whenever any new assumptions are

introduced so as to maintain likelihood at a maximum (see §2.1.3). Usually the initial normalization corresponds to a uniform distribution $m(\mathbf{x}) = 1/\mathrm{vol}(V)$ where V denotes the unit cell and $\mathrm{vol}(V)$ its volume.

At any stage of the calculation, the symmetry-unique non-origin reflexions are divided into two sets: the *basis set* $H = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{|H|}\}$ comprising those $|H|$ reflexions for which explicit phase assumptions have been made, and the complementary set K of *non-basis* reflexions for which amplitude measurements are available but which are unphased. A *node* ν of the phasing tree consists of a set $\Phi(\nu) = \{\varphi_\nu(\mathbf{h}_1), \varphi_\nu(\mathbf{h}_2), ..., \varphi_\nu(\mathbf{h}_{|H|})\}$ of trial phase values for the reflexions in H. The basis set is *extended* in the course of the calculation by incorporation of new reflexions, defining successive *levels* of the phasing tree. Each node at level l has a unique *parent* node at level $l - 1$ and may have a *progeny* at level $l + 1$. The process of generating the progeny of a node (which may be of arbitrary size) is called the *expansion* of that node and consists in appending trial phase values for the reflexions incorporated as the latest level of the basis set to the phase values belonging to the original (now parent) node.

Given a node ν with corresponding basis set H, the unique distribution q_ν^{ME} having maximum entropy compatible with the data attached to ν is constructed by maximizing the relative entropy

$$S_m(q) = -\int_V q(\mathbf{x}) \log[q(\mathbf{x})/m(\mathbf{x})]\, \mathrm{d}^3\mathbf{x} \qquad (1.1)$$

under the constraints

$$\int_V q(\mathbf{x}) \exp(2\pi i \mathbf{h}.\mathbf{x})\, \mathrm{d}^3\mathbf{x} = |U_\mathbf{h}|^{\mathrm{obs}} \exp[i\varphi_\nu(\mathbf{h})] \quad \text{for all } \mathbf{h} \in H \qquad (1.2)$$

Besides reproducing the amplitudes and phases attached to ν for reflexions in H, q_ν^{ME} has Fourier coefficients $U_{\nu,\mathbf{k}}^{\mathrm{ME}}$ with non-negligible amplitude for many non-basis reflexions $\mathbf{k} \in K$. This is especially the case when $\mathbf{k}$ is in the *second neighbourhood* $\mathcal{N}_2(H)$ of H:

$$\mathcal{N}_2(H) = \{\mathrm{UNIQUE}(\mathbf{h}_1 \pm \mathbf{R}_g^T \mathbf{h}_2) \mid \mathbf{h}_1 \in H,\ \mathbf{h}_2 \in H,\ g \in G\} \qquad (1.3)$$

where G is the space group of the crystal, $\mathbf{R}_g^T$ denotes the transpose of the integer matrix associated to $g \in G$, and UNIQUE($\mathbf{h}$) denotes the unique representative of $\mathbf{h}$ under symmetry and Friedel equivalence. This *maximum-entropy extrapolation* from H into $\mathcal{N}_2(H)$ causes the conditional distribution of $|U_{\nu,\mathbf{k}}|$ to deviate systematically, in a manner which depends on the phases $\Phi(\nu)$ attached to node ν, from the Gaussian or Rayleigh distributions of Wilson statistics (corresponding to $|U_{\nu,\mathbf{k}}^{\mathrm{ME}}| = 0$) and to become a Rice distribution instead. We may formalize this situation by denoting $(\mathcal{H}_0)$ the null hypothesis that the atoms are uniformly distributed, and denoting by $(\mathcal{H}_1)$ the alternative hypothesis that they are distributed according to q_ν^{ME}. These two hypotheses can be tested against each other by calculating the log-likelihood gain:

$$\mathrm{LLG}(\nu) = \log \frac{\mathcal{P}\{|U_\mathbf{k}| = |U_\mathbf{k}|^{\mathrm{obs}} \text{ for } \mathbf{k} \in K \mid U_\mathbf{h} = |U_\mathbf{h}|^{\mathrm{obs}} \exp[i\varphi_\nu(\mathbf{h})] \text{ for } \mathbf{h} \in H\}}{\mathcal{P}(|U_\mathbf{k}| = |U_\mathbf{k}|^{\mathrm{obs}} \text{ for } \mathbf{k} \in K \mid U_\mathbf{h} = 0 \text{ for } \mathbf{h} \in H)} \qquad (1.4)$$

where $\mathcal{P}(B|A)$ denotes the conditional probability of event B if event A is assumed to have occurred. $\mathrm{LLG}(\nu)$ will be largest when the phase assumptions attached to ν lead one to expect deviations from Wilson statistics for the unphased amplitudes $|U_\mathbf{k}|$, $\mathbf{k} \in K$, that most closely match those present in the distribution of the actual measurements $|U_\mathbf{k}|^{\mathrm{obs}}$: it is therefore a quantitative measure of the degree of corroboration by the unphased data of the phase assumptions attached to ν.

The tree-directed search through all possible combinations of phases is carried out by successive incorporations of reflexions initially in K to form the successive levels of H. The first major problem is to decide which reflexions to pick at each of these steps. A selected subset of the nodes at the previous level is then subjected to node expansion by giving 'permuted' values to the phases of the newly incorporated reflexions, and the second major problem is to design optimal methods for sampling the vast number of possible phase sets. Each progeny node ν generated in this way is evaluated by numerically constructing the corresponding q_ν^{ME} and calculating the log-likelihood gain $\mathrm{LLG}(\nu)$ resulting from the maximum-entropy extrapolation into $\mathcal{N}_2(H)$. The LLG may then be used to prune the search tree by retaining only those nodes with the highest LLG values, or used in some form of post-processing to detect and recycle phase indications. The third problem is then to make such decisions automatic and reliable. This procedure is essentially Bayesian, the consideration of the LLG alone often being justifiable by the fact that it consults much more experimental data than does the entropy loss $S(\nu) = S_m(q_\nu^{\mathrm{ME}})$. The full 'Bayesian score' $B(\nu) = NS(\nu) + \mathrm{LLG}(\nu)$, where N is the number of non-hydrogen atoms in the asymmetric unit, has also been successfully used in the applications described below; it measures the increase in the logarithm of the *a posteriori* probability of the measured amplitudes, compared with that given by Wilson statistics.

The implementation of this scheme within *BUSTER* is fully automatic, space-group general, and accommodates any mixture of centric and acentric reflexions.

2. Mathematical technology

2.0. *Introduction*

Most phasing methods consist in designing a *score* whose maximization with respect to the phases is supposed to encourage the buildup of some desirable property of the distribution of atomic positions. The scores used in

conventional direct methods and in the present approach, although very similar in principle, differ greatly from a practical point of view.

In conventional direct methods the primary concept is that of phase relation; the score is a secondary concept and is rather loosely defined as a consistency criterion between large numbers of phase relations. This has the advantage that once such a criterion has been chosen, it comes automatically with an *explicit and global functional representation* in terms of the phases, and thus lends itself to symbolic processes such as, for instance, elimination or least-squares solution. The drawback is that this functional representation is not accurate (I, §2.1) so that by the time a large subset of phase values has been obtained or assumed on a trial basis, their relations *with* the other phases and the relations *between* the other phases are no longer well represented (if only because the normalization should be redone). Fortunately, the overdetermination of the relations helps overcome their low individual accuracy so that the score remains useful.

In this approach the score is defined from first principles and as a result enjoys various properties of optimality: that of the saddlepoint approximation is well known (I, §5.3; Bricogne, 1991*a*, §8.3.6), and the optimality of likelihood as a hypothesis-testing criterion is asserted by a fundamental theorem of Neyman & Pearson (1933). Unfortunately the mode of construction of this score only gives a method for its *accurate pointwise evaluation* and does not provide a global functional representation. Indeed the possibility of the latter is ruled out by the very analysis (the problem of large deviations, see I, §2.1) on which the method is based and which dictates the use of a phasing tree.

To overcome this handicap and have the best of both worlds, two strategies may be used: either (1) calculate a large number of scores on a grid of trial phase values, then carry out a Fourier analysis or apply some interpolation scheme to recover an accurate and global functional representation of the score, at least with respect to that subset of phases; here the grid must be chosen to economize the number of (expensive) evaluations (see §2.2.2) and thus allow calculations on the greatest number of phases simultaneously; or (2) use Taylor expansions around fewer grid points to build local approximations to the desired functional dependence; then use pointwise evaluations in regions of interest to improve accuracy [see §2.2.3 (6)]. In both cases efficiency will depend critically on the choice of the grid and on the careful selection of sets of reflexions which are most strongly coupled with each other.

The mathematical techniques around which *BUSTER* has been written fall into three main categories. First, the techniques of calculation of probabilities and likelihoods, based on structure-factor algebra, the solution of the maximum-entropy equations, and the evaluation and maximization of likelihoods with respect to a great variety of parameters. Second, the multidimensional search techniques used in the optimal scheduling of basis-set growth, in designing efficient phase-sampling schemes and in surveying multidimensional Fourier series approximations to the score function. Third, the numerical and statistical techniques used in analysing the likelihood-based node scores and making reliable and automatic decisions about the pruning of the phasing tree and its subsequent growth.

2.1. *Calculation of probabilities and likelihoods*

2.1.0. *Structure-factor algebra and statistics.* The calculation of moments of trigonometric structure-factor expressions under assumptions of *non-uniform* random distribution of atomic positions rests on a fundamental algebraic identity, the Bertaut linearization formula, the importance of which could not be overstated. The primary references for this material are Bertaut (1955*a*,*b*,*c*, 1956*a*,*b*, 1959*a*,*b*), Bertaut & Dulac (1956) and Bertaut & Waser (1957). Bertaut used this formula to identify non-vanishing moments when the atoms are uniformly distributed. The extension of his theory to non-uniform distributions of atoms and to more general situations was first given in §*A*1 and §*A*2 of (II).

The contribution $\Xi(\mathbf{h},\mathbf{x})$ of a point atom of unit scattering factor placed at $\mathbf{x}$ to the structure factor at $\mathbf{h}$ may be written:

$$\Xi(\mathbf{h},\mathbf{x}) = (1/|G_{\mathbf{x}}|) \sum_{g\in G} \exp[2\pi i \mathbf{h}.S_g(\mathbf{x})] \qquad (2.1)$$

where $G_{\mathbf{x}}$ is the isotropy subgroup of $\mathbf{x}$ and $|G_{\mathbf{x}}|$ denotes the number of its elements. If we consider this quantity as a function of $\mathbf{x}$ indexed by $\mathbf{h}$, the family of functions defined by $\Xi_{\mathbf{h}}(\mathbf{x}) = \Xi(\mathbf{h},\mathbf{x})$ constitutes an *algebra* in the sense that products of such functions may be rewritten as linear combinations of other functions in that family, generalizing well known trigonometric identities such as $\cos a \cos b = (1/2)[\cos(a + b) + \cos(a - b)]$. This relation is given by *Bertaut's linearization formula*:

$$\Xi(\mathbf{h},\mathbf{x}) \times \Xi(\mathbf{k},\mathbf{x}) = (1/|G_{\mathbf{x}}|) \sum_{g\in G} \exp(2\pi i \mathbf{k}.\mathbf{t}_g)\Xi(\mathbf{h} + \mathbf{R}_g^T\mathbf{k},\mathbf{x}) \qquad (2.2)$$

(see §*A*2 of II for a proof).

Let the position $\mathbf{x}$ of a generic atom now be considered as a random vector distributed in the asymmetric unit D of the crystal with probability density $q(\mathbf{x})$. The trigonometric structure-factor expressions $\Xi_{\mathbf{h}}(\mathbf{x})$ then become random variables with complex or real values, and the calculation of their moments is a fundamental operation in the statistical approach to the phase problem. For our present purposes, only moments of order 1 and 2 will be needed.

Let $M_{\mathbf{h}}$ [or $M(\mathbf{h})$] denote the Fourier coefficient with indices $\mathbf{h}$ of the function obtained by symmetry-expanding $q(\mathbf{x})/|G_{\mathbf{x}}|$ to the whole unit cell and renormalizing it. Then the first-order moment of $\Xi_{\mathbf{h}}$ is

$$\langle \Xi_{\mathbf{h}} \rangle = |G| M_{\mathbf{h}} \qquad (2.3)$$

while the second-order moments can be obtained, by linearizing according to (2.2), as

$$\langle \Xi_{\mathbf{h}} \Xi_{\pm\mathbf{k}} \rangle = |G| \sum_{g \in G} \exp(\pm 2\pi i \mathbf{k}.\mathbf{t}_g) M(\mathbf{h} \pm \mathbf{R}_g^T \mathbf{k}). \quad (2.4)$$

At this point it becomes desirable to take into account the centric character of the reflexions by defining:

$$\alpha_{\mathbf{h}} = \mathrm{Re}\Xi_{\mathbf{h}}, \beta_{\mathbf{h}} = \mathrm{Im}\Xi_{\mathbf{h}} \quad \text{for } \mathbf{h} \text{ acentric} \quad (2.5a)$$

$$\gamma_{\mathbf{h}} = \mathrm{Re}[\exp(-i\theta_{\mathbf{h}})\Xi_{\mathbf{h}}] \quad \text{for } \mathbf{h} \text{ centric} \quad (2.5b)$$

where, for $\mathbf{h}$ centric, $\theta_{\mathbf{h}} = \pi \mathbf{h}.\mathbf{t}_g$ with g any element of the group G such that $\mathbf{R}_g^T \mathbf{h} = -\mathbf{h}$. Elementary calculations according to a general procedure described in §*A*1 and §*A*2 of (II) then yield the following expressions for the second-order moments:

$$\langle \alpha_{\mathbf{h}} \alpha_{\mathbf{k}} \rangle = (1/2)(\mathrm{Re}\langle \Xi_{\mathbf{h}} \Xi_{+\mathbf{k}} \rangle + \mathrm{Re}\langle \Xi_{\mathbf{h}} \Xi_{-\mathbf{k}} \rangle) \quad (2.6a)$$

$$\langle \alpha_{\mathbf{h}} \beta_{\mathbf{k}} \rangle = (1/2)(\mathrm{Im}\langle \Xi_{\mathbf{h}} \Xi_{+\mathbf{k}} \rangle - \mathrm{Im}\langle \Xi_{\mathbf{h}} \Xi_{-\mathbf{k}} \rangle) \quad (2.6b)$$

$$\langle \beta_{\mathbf{h}} \alpha_{\mathbf{k}} \rangle = (1/2)(\mathrm{Im}\langle \Xi_{\mathbf{h}} \Xi_{+\mathbf{k}} \rangle + \mathrm{Im}\langle \Xi_{\mathbf{h}} \Xi_{-\mathbf{k}} \rangle) \quad (2.6c)$$

$$\langle \beta_{\mathbf{h}} \beta_{\mathbf{k}} \rangle = (1/2)(\mathrm{Re}\langle \Xi_{\mathbf{h}} \Xi_{-\mathbf{k}} \rangle - \mathrm{Re}\langle \Xi_{\mathbf{h}} \Xi_{+\mathbf{k}} \rangle) \quad (2.6d)$$

$$\langle \alpha_{\mathbf{h}} \gamma_{\mathbf{k}} \rangle = (1/2)\mathrm{Re}[\exp(-i\theta_{\mathbf{k}})\langle \Xi_{\mathbf{h}} \Xi_{+\mathbf{k}} \rangle + \exp(+i\theta_{\mathbf{k}})\langle \Xi_{\mathbf{h}} \Xi_{-\mathbf{k}} \rangle] \quad (2.7a)$$

$$\langle \beta_{\mathbf{h}} \gamma_{\mathbf{k}} \rangle = (1/2)\mathrm{Im}[\exp(-i\theta_{\mathbf{k}})\langle \Xi_{\mathbf{h}} \Xi_{+\mathbf{k}} \rangle + \exp(+i\theta_{\mathbf{k}})\langle \Xi_{\mathbf{h}} \Xi_{-\mathbf{k}} \rangle] \quad (2.7b)$$

$$\langle \gamma_{\mathbf{h}} \gamma_{\mathbf{k}} \rangle = (1/2)\mathrm{Re}\{\exp[-i(\theta_{\mathbf{h}} + \theta_{\mathbf{k}})]\langle \Xi_{\mathbf{h}} \Xi_{+\mathbf{k}} \rangle + \exp[-i(\theta_{\mathbf{h}} - \theta_{\mathbf{k}})]\langle \Xi_{\mathbf{h}} \Xi_{-\mathbf{k}} \rangle\}. \quad (2.8)$$

These formulae, considered together with (2.3) and (2.4), completely specify the elements of the vector of first moments and of the covariance matrix of the trigonometric structure-factor expressions under the assumption of an arbitrary distribution $q(\mathbf{x})$ of random atoms. They are implemented in two distinct ways in *BUSTER*:

(1) with $\mathbf{h} = \mathbf{k}$ to generate the individual trigonometric covariance matrices $\mathbf{V}_{\mathbf{hh}}$ for all reflexions $\mathbf{h}$ to be included in a likelihood calculation using the block-diagonal approximation, *viz.*

$$\mathbf{V}_{\mathbf{hh}} = \begin{pmatrix} \langle \alpha_{\mathbf{h}} \alpha_{\mathbf{h}} \rangle - \langle \alpha_{\mathbf{h}} \rangle \langle \alpha_{\mathbf{h}} \rangle & \langle \alpha_{\mathbf{h}} \beta_{\mathbf{h}} \rangle - \langle \alpha_{\mathbf{h}} \rangle \langle \beta_{\mathbf{h}} \rangle \\ \langle \beta_{\mathbf{h}} \alpha_{\mathbf{h}} \rangle - \langle \beta_{\mathbf{h}} \rangle \langle \alpha_{\mathbf{h}} \rangle & \langle \beta_{\mathbf{h}} \beta_{\mathbf{h}} \rangle - \langle \beta_{\mathbf{h}} \rangle \langle \beta_{\mathbf{h}} \rangle \end{pmatrix} \quad \text{for } \mathbf{h} \text{ acentric,} \quad (2.9a)$$

$$\mathbf{V}_{\mathbf{hh}} = (\langle \gamma_{\mathbf{h}} \gamma_{\mathbf{h}} \rangle - \langle \gamma_{\mathbf{h}} \rangle \langle \gamma_{\mathbf{h}} \rangle) \quad \text{for } \mathbf{h} \text{ centric;} \quad (2.9b)$$

(2) with $\mathbf{h}, \mathbf{k} \in H$ to generate the Hessian matrix of $\log \mathcal{Z}$ which is later used to solve the maximum-entropy equations (§2.1.1) and compute the saddlepoint approximation (§2.1.2).

2.1.1. *Solving the maximum-entropy equations.* The central numerical task in the process described in §1.3 and §3.5 consists in satisfying the saddlepoint condition – or, in an equivalent terminology, solving the maximum-entropy equations (see §5.4 of I and §0.5 of II for the details of this equivalence). In the notation used in (II) the general form of these equations is

$$\nabla_{\lambda}(\log \mathcal{Z}) = \mathbf{F}^* \quad (2.10)$$

where $\mathbf{F}^*$ is the vector of trial structure-factor values for the basis-set reflexions and $\log \mathcal{Z}$ is the cumulant-generating function of the distribution of $\mathbf{F}$. Its Hessian (the covariance matrix of $\mathbf{F}$) is positive definite, which guarantees that there is a one-to-one correspondence between λ and the gradient vector of $\log \mathcal{Z}$ at λ wherever it is defined, and hence that the solution to the maximum-entropy equations, if it exists at all, is unique (I, §5.2, §6.1.2). The global convexity of $\log \mathcal{Z}$ as a function of λ makes Newton's method very effective at solving (2.10) for small values of the total dimension n. The elements of the Hessian matrix can be evaluated at each iteration by means of expressions (2.6*a*,*b*,*c*,*d*), (2.7*a*,*b*) and (2.8) from the Fourier coefficients of the current approximation to the maximum-entropy distribution, as pointed out in §7.1.1 of (I).

It is worth noting that the maximum-entropy equations are *not* solved by maximizing the entropy $\mathcal{S}$ with respect to λ: $\mathcal{S}$ is globally concave as a function of the constraint values $\mathbf{F}^*$, but *not as a function of the Lagrange multipliers* λ. The transformation of the constrained maximization of $\mathcal{S}$ into the problem of solving equations (2.10) is an instance of the use of *duality* and rests in this particular case on the Legendre duality which relates $\mathcal{S}$ and $\log \mathcal{Z}$. This duality may be visualized geometrically (see Fig. 1 for a one-dimensional illustration) as the equivalence between

(1) finding the unique λ^* where $\log \mathcal{Z}$ has a gradient (here, slope) equal to $\mathbf{F}^*$;

(2) finding the unique λ^* where the *Lagrangian* $\mathcal{L}(\lambda) = \log \mathcal{Z}(\lambda) - \lambda.\mathbf{F}^*$ is a minimum, causing the graphs corresponding to the two summands to be tangent at the desired λ^*.

Analytically, this equivalence amounts to the trivial rewriting of (2.10) as

$$\nabla_{\lambda}(\log \mathcal{Z} - \lambda.\mathbf{F}^*) = 0, \quad (2.11)$$

the positive-definiteness of the Hessian guaranteeing that any stationary point of $\mathcal{L}$ is indeed a minimum. Setting up Newton's method to solve (2.10) or (2.11) leads to identical numerical computations, but the latter form makes more conspicuous a very useful property of the Lagrangian $\mathcal{L}$: for λ other than λ^*, the value of the Lagrangian $\mathcal{L}(\lambda)$ {which may be read off as the intercept with the vertical axis of the line through the point $[\lambda, \log \mathcal{Z}(\lambda)]$ with slope $\mathbf{F}^*$} gives an upper bound on the final entropy $\mathcal{S}(\lambda^*)$ (similarly marked as an intercept in Fig. 1), while at the solution $\mathcal{L}(\lambda^*) = \mathcal{S}(\lambda^*)$. From the practical point of view this bounding property of the intermediate values of the Lagrangian can be exploited to endow the duality method with extraordinary robustness. An ordinary Newton method would occasionally overshoot past λ^* (especially in multidimensional situations), giving rise to per-

ilously unstable behaviour. By contrast, the systematic use of a line search along the direction indicated by Newton's method so as to make $\mathcal{L}(\lambda)$ a minimum along that direction will guard very effectively against such overshoots and keep the algorithm stable even under very demanding conditions.

The use of duality for solving general maximum-entropy equations was first suggested by Alhassid, Agmon & Levine (1978) and its implementation was described by Agmon, Alhassid & Levine (1979). It is this method which is used in *BUSTER*: the Hessian matrix is calculated in full by (2.6*a*,*b*,*c*,*d*), (2.7*a*,*b*) and (2.8) at each cycle, and the line search along the Newton direction is carried out by the method of Davidon (see Scales, 1985, pp. 39–40). This algorithm is extremely robust, and most of the time is consumed in calculating fast Fourier transforms which, since they are to be exponentiated with minimal aliasing, must be very finely sampled. When the number of reflexions in the basis set H exceeds a few hundred, a matrix method is no longer practical and the exponential modelling technique with line or plane search, described by Bricogne & Gilmore (1990), may be used.

Recently Prince and coworkers (Prince, 1989; Collins & Prince, 1991; Sjölin, Prince, Svensson & Gilliland, 1991) have separately re-examined the problem of solving the maximum-entropy equations, and it may be useful to give a brief review of similarities and differences between the various approaches. The Fourier coefficients of their exponential model are the Lagrange multipliers λ (I, §7.1.2). The use of duality to turn the constrained maximization of $\mathcal{S}$ into the unconstrained minimization of $\mathcal{L}$ is the same as that proposed by Agmon *et al.* (1979). The possibility

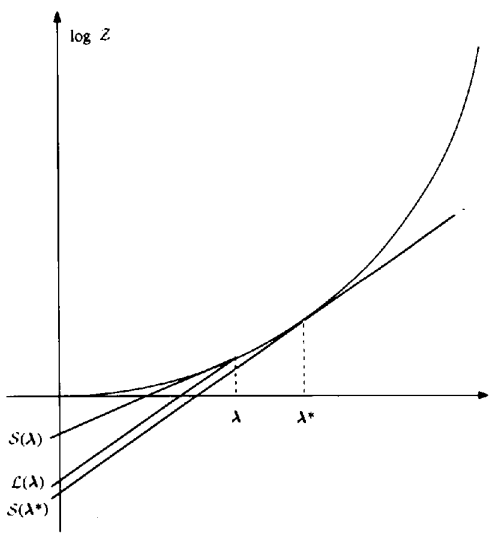

Fig. 1. Schematic graphical representation of the solution of the maximum-entropy equations for $n = 1$. The solution λ^* is the value of λ where the graph of $\log\mathcal{Z}$ as a function of λ has a slope equal to the constraint value $\mathbf{F}^*$; the solution is unique because of the convexity of that graph. For λ other than λ^*, the value of the Lagrangian $\mathcal{L}(\lambda)$ gives an upper boundary on the final entropy $\mathcal{S}(\lambda^*)$, while at the solution $\mathcal{L}(\lambda^*) = \mathcal{S}(\lambda^*)$.

of expressing the elements of the Hessian matrix in terms of the Fourier coefficients of the current estimate of the ME density was described in (I). The linearization of the product of two cosines given in Collins & Prince (1991) and in Sjölin *et al.* (1991) is the *P*1 version of structure-factor algebra, and hence does not make full use of symmetry; whereas the general formula, given by equation (7.7), §7.11 of (I) and implemented here, which uses the full structure-factor algebra for the space group at hand, is both general and optimal. Prince's implementation of the duality method avoids the computation of the full Hessian by using a quasi-Newton method, the BFGS procedure [see Press, Flannery, Teukolsky & Vetterling (1986, pp. 308–309), or Scales (1985, pp. 89–90)], but this is of no avail here since the Hessian is required to compute the saddlepoint approximation to $\mathcal{P}(\mathbf{F}^*)$, as shown by equations (2.12) to (2.14) below.

2.1.2. *The saddlepoint approximation.* Once the maximum-entropy equations have been solved with data $\mathbf{F}^*$ the saddlepoint approximation $\mathcal{P}^{\rm SP}(\mathbf{F}^*)$ to the joint probability of the components of $\mathbf{F}^*$ is given by (I, §5.3; II, §0.4):

$$\mathcal{P}^{\rm SP}(\mathbf{F}^*) = \exp(\mathcal{S})/[\det(2\pi\mathcal{Q})]^{1/2} \qquad (2.12)$$

where

$$\mathcal{S} = \log\mathcal{Z}(\lambda) - \lambda.\mathbf{F}^* \qquad (2.13)$$

and where

$$\mathcal{Q} = \nabla\nabla^T(\log\mathcal{Z}) \qquad (2.14)$$

is the Hessian matrix used in (and hence obtained as a by-product of) solving the maximum-entropy equations.

A remark is in order at this point so as to avoid a possible misunderstanding – even an apparent paradox – about the fact that $\det\mathcal{Q}$ is in the denominator and thus would be encouraged to become small, whereas the conventional wisdom (Tsoucaris, 1970) has it that such a determinant should be maximal. The crucial observation is that $\mathcal{Q}$ is a Goedkoop matrix associated with $q^{\rm ME}$, and therefore is not an ordinary Goedkoop matrix (Goedkoop, 1950) until *all* reflexions are in the basis set. It is normally rather sparse, as its elements are made up *precisely* of the maximum-entropy extrapolates attached to the second neighbourhood of the basis set [see equations (1.3) and (2.4)–(2.9)]. It is therefore the need to keep the LLG as large as possible, *i.e.* the requirement that these extrapolates should agree well with the unphased data, which prevents this matrix from becoming singular. Since the largest off-diagonal elements correspond to triplets within the basis set, this observation sheds new light on the subtle way in which the LLG will prevent intra-basis-set relationships from becoming over-consistent. It also shows that if and when the LLG does endorse a strong maximum-entropy extrapolation pattern, the rank of $\mathcal{Q}$ will level off, causing $\det\mathcal{Q}$ to drop rapidly; the loss of entropy will slow down (since the future constraints have already been largely anticipated), and therefore $\mathcal{P}^{\rm SP}$ will start to soar, signalling that the atomicity

of the structure has now taken over as the main agent of phase extrapolation. This should happen when the number of reflexions approaches the number of atoms, or more generally of 'effective scatterers' at low resolution. This behaviour has been observed with small molecules as the related phenomenon of a sudden decrease of the Σ parameter in the *MICE* program (Gilmore *et al.*, 1990; Shankland *et al.*, 1992) and has been investigated in another context by J. Navaza (personal communication). The role of the determinant - which, it may be noted, does not appear in the pure ME theory - is therefore an important one in signalling, albeit in a rather unexpected way, the successful completion of a structure determination.

2.1.3. *Likelihood evaluation and maximization.* All calculations related to the evaluation of log-likelihood for a given hypothesis ($\mathcal{H}$) and its optimization with respect to any set of parameters associated with that hypothesis are carried out in a single routine, *MLNORM*.

Typically a hypothesis is specified by the following entities (which are not all independent):

(1) the list $\{U_{\mathbf{h}}^{\rm ME} | \mathbf{h}$ observed$\}$ of Fourier coefficients of $q^{\rm ME}$ for the observed reflexions;

(2) the diagonal blocks $\mathbf{V_{hh}}$ of the trigonometric covariance matrix calculated as described by (2.9*a*,*b*) for all observed reflexions $\mathbf{h}$;

(3) the sums $\sigma_1(\mathbf{h})$ and $\sigma_2(\mathbf{h})$, calculated over the unit cell, of first and second powers of the scattering factors of the random atoms which are distributed according to $q^{\rm ME}$;

(4) the list $\{F_{\mathbf{h}}^{\rm par} | \mathbf{h}$ observed$\}$ of structure factors for a partial structure, if any is specified, together with the partial temperature factor $B^{\rm par}$ such that the structure factor at $\mathbf{h}$ for the complete structure be distributed with mean value

$$\langle F_{\mathbf{h}} \rangle = \sigma_1(\mathbf{h}) U_{\mathbf{h}}^{\rm ME} + \exp[-(1/4)B^{\rm par}(d_{\mathbf{h}}^*)^2] F_{\mathbf{h}}^{\rm par} \qquad (2.15)$$

and covariance matrix

$$\mathbf{Q_{hh}} = \sigma_2(\mathbf{h})\mathbf{V_{hh}} \qquad (2.16)$$

(5) the overall scale factor k and temperature factor B such that the structure-factor amplitudes $k\exp[-(1/4)B(d_{\mathbf{h}}^*)^2]|F_{\mathbf{h}}|^{\rm obs}$ be on absolute scale.

As stated earlier (§1.3), k and B are defined implicitly in terms of the other parameters by the condition that the log-likelihood L be a maximum with respect to them:

$$L(\mathcal{H}) = L(p, q, ...) = \log \max_{k,B} \Lambda(k, B, p, q, ...) \qquad (2.17)$$

where p, q, ... stand for the collection of parameters in ($\mathcal{H}$) other than k and B, and the likelihood function Λ is the conditional probability of the observations under ($\mathcal{H}$).

The current program uses an approximation to conditional probabilities, called the *block-diagonal approximation*, which consists in using as a covariance matrix $\mathcal{Q}$ a block-diagonal matrix consisting of the blocks $\mathbf{Q}$ defined above. This decouples the various reflexions so that the resulting log-likelihood L may be written as a sum of contributions $L_{\mathbf{h}}$ from individual reflexions

$$\mathrm{LLG}(\nu) = \sum_{\mathbf{k}\in K} \log \frac{\mathcal{P}\{|U_{\mathbf{k}}| = |U_{\mathbf{k}}|^{\rm obs} \mid U_{\mathbf{h}} = |U_{\mathbf{h}}|^{\rm obs}\exp[i\varphi_\nu(\mathbf{h})] \text{ for } \mathbf{h}\in H\}}{\mathcal{P}(|U_{\mathbf{k}}| = |U_{\mathbf{k}}|^{\rm obs} \mid U_{\mathbf{h}} = 0 \text{ for } \mathbf{h}\in H)} \qquad (2.18)$$

An exact expression has been obtained for the log-likelihood, using a general symmetric rather than block-diagonal covariance matrix for the Gaussian conditional probability out of which phases and signs are to be integrated. It will be published separately in a sequel to (II).

Two types of likelihood functions are used: (i) phased likelihoods for basis-set reflexions, and (ii) unphased likelihoods for non-basis reflexions. The rationale for using a phased likelihood for basis-set reflexions is to reflect faithfully the increased probability of the phased values and let it offset the loss of entropy caused by that fitting. It is easily checked that, to the quadratic approximation, the phased LLG exactly compensates the effect of the entropy loss (for equal atoms), so that the Bayesian score remains stationary to begin with.

For each likelihood type, one must distinguish the centric case from the acentric case.

(i) *Phased likelihood.* In the acentric case, writing a complex F as a 2-vector $\mathbf{F}$:

$$\mathcal{P}(\mathbf{F_h}) = [1/2\pi(\det\mathbf{Q_{hh}})^{1/2}]\exp[-(1/2)(\mathbf{F_h} - \langle\mathbf{F_h}\rangle)^T\mathbf{Q_{hh}}^{-1} \times (\mathbf{F_h} - \langle\mathbf{F_h}\rangle)] \qquad (2.19)$$

while in the centric case ($Q_{\mathbf{hh}}$ is then a scalar):

$$\mathcal{P}(F_{\mathbf{h}}) = [1/(2\pi Q_{\mathbf{hh}})^{1/2}]\exp[-(1/2Q_{\mathbf{hh}})|F_{\mathbf{h}} - \langle F_{\mathbf{h}}\rangle|^2]. \qquad (2.20)$$

Then

$$L_{\mathbf{h}}(\mathcal{H}) = \log\mathcal{P}\{k\exp[-(1/4)B(d_{\mathbf{h}}^*)^2]|F_{\mathbf{h}}|^{\rm obs}\exp(i\varphi_{\mathbf{h}})\}. \qquad (2.21)$$

If $q^{\rm ME}$ is constructed by solving the maximum-entropy equations exactly, the exponential is equal to 1 and only the normalization factor contributes. For the null hypothesis, $\langle\mathbf{F_h}\rangle = 0$ and the variances are given by Wilson statistics, *viz.* $\mathbf{V_{hh}} = \mathrm{diag}[(1/2)\varepsilon_{\mathbf{h}}, (1/2)\varepsilon_{\mathbf{h}}]$ for $\mathbf{h}$ acentric and $V_{\mathbf{hh}} = \varepsilon_{\mathbf{h}}$ for $\mathbf{h}$ centric, where $\varepsilon_{\mathbf{h}}$ is the statistical weight (Stewart & Karle, 1976; Iwasaki & Ito, 1977; Stewart, Karle, Iwasaki & Ito, 1977).

Strictly speaking, the basis-set constraints should not be fitted exactly, but only as far as will result in an increase in the Bayesian score (*i.e.* in the *a posteriori* probability of these phases). This remark will be implemented in the near future as a phase refinement mechanism.

(ii) *Unphased likelihood.* In the acentric case, the current program uses an approximation to the exact likelihood, obtained by replacing the 2×2 matrix $\mathbf{Q_{hh}}$ by $\mathrm{diag}(\Sigma_{\mathbf{h}}^{\mathbf{g}}, \Sigma_{\mathbf{h}}^{\mathbf{g}})$ with $\Sigma_{\mathbf{h}}^{\mathbf{g}} = (\det\mathbf{Q_{hh}})^{1/2}$. This leads to a Rice distribution (Rice, 1944, 1945) for the amplitude $|F_{\mathbf{h}}|$ and

hence to:

$$L_{\mathbf{h}}(\mathcal{H}) = \log\mathcal{R}\{\langle\mathbf{F}_{\mathbf{h}}\rangle, k\exp[-(1/4)B(d^*_{\mathbf{h}})^2]|F_{\mathbf{h}}|^{\rm obs}, \Sigma^{\rm f}_{\mathbf{h}}\} \tag{2.22}$$

where

$$\mathcal{R}(r, R, \Sigma) = (R/\Sigma)\exp[-(r^2 + R^2)/2\Sigma]I_0(rR/\Sigma). \tag{2.23}$$

In the centric case, putting $\Sigma^{\rm f}_{\mathbf{h}} = Q_{\mathbf{hh}}$, one has a centric Rice distribution and hence:

$$L_{\mathbf{h}}(\mathcal{H}) = \log C\{\langle\mathbf{F}_{\mathbf{h}}\rangle, k\exp[-(1/4)B(d^*_{\mathbf{h}})^2]|F_{\mathbf{h}}|^{\rm obs}, \Sigma^{\rm f}_{\mathbf{h}}\} \tag{2.24}$$

where

$$C(r, R, \Sigma) = (2/\pi\Sigma)^{1/2}\exp[-(r^2 + R^2)/2\Sigma]\cosh(rR/\Sigma). \tag{2.25}$$

The null-hypothesis values are obtained by the same substitutions as were described for phased likelihoods.

The refinement of k and B is the simplest instance of optimization of the global LLG with respect to certain parameters attached to $(\mathcal{H})$. Other instances would be its optimization with respect to the rotational and translational parameters defining the placement of the fragment(s) which constitute the partial structure and with respect to the associated B factor(s), and with respect to the components of the $U^{\rm ME}_{\mathbf{h}}$ for the purpose of phase refinement.

In anticipation of these developments, and of others relating to the MIRAS and MAD methods, and to the use of non-crystallographic symmetry, a general procedure has been programmed for systematically applying the chain rule to the calculation of first- and second-order derivatives of nested functional expressions. For the first-order derivatives this is the well known multiplication of gradients by Jacobian matrices; more involved expressions are needed to propagate second-order derivatives. The availability of the latter gives access to the use of the Newton–Raphson method (if desired) in all refinements, but more importantly it allows the detection of bifurcations along directions with negative curvature in the course of phase refinement (see *e.g.* I, §7.2).

2.2. *Multidimensional search techniques*

2.2.1. *Optimal scheduling of basis-set growth.* The large number of atoms in macromolecules results in the general weakness of individual phase interactions, although those mediated by the presence of solvent regions and/or by non-crystallographic symmetry are notable exceptions to this rule. This is reflected in the relative weakness of the maximum-entropy extrapolation pattern outside the basis set. If the criteria constructed in §2.1 are to be of any use in discriminating good trial phase sets from bad ones, and thus avoid a bush-like growth of the phasing tree, their sensitivity to these phases must be maximized. This can be achieved by carefully planning the progressive enlargement of the basis set so as to concentrate the maximum number of the available interactions at the early stages of the tree growth. This is a well known concern in standard direct methods. It has been my experience that choosing basis-set reflexions as a function of their strength alone, as described by Sjölin *et al.* (1991), is very unlikely to produce the necessary concentration of phase interactions.

The following procedure, implemented in routine *COBWEB*, has been designed to maximize the sensitivity of the LLG or of the Bayesian score with respect to the basis-set phases at any stage of its growth.

(1) Identify the 'strong' reflexions (as defined for the current shell of data - see §3.1) eligible as basis-set candidates, and add them to the existing basis set (if any) to form a *mock basis set.*

(2) Compile a list of the effects of maximum-entropy extrapolation from these candidates on the offsets and the variances of the conditional distributions of the reflexions in the second neighbourhood of the mock basis set, calculating their strength from the partial derivatives of the entropy, of the determinant (§2.1.2) and of the log-likelihood gain.

(3) Eliminate the newly introduced candidates from the mock basis set one by one, by taking out at each stage that candidate which is most weakly coupled to the reflexions still in the mock basis set, and disabling all interactions in which it is involved. The reverse order of elimination is then the optimal order of 'priority' in which reflexions should be incorporated into the basis set so as to maximize the sensitivity of the Bayesian score to the associated phase values, in the sense of maximizing the range of variation of the score around its mean value. A record is kept of the interactions which are lost when each reflexion is eliminated, *i.e.* gained when it is introduced. This record can later be accessed to generate a list of multidimensional Fourier coefficients (see §2.2.2 below) giving an approximate functional representation of the score as a Fourier series in the phases associated to any segment of consecutive reflexions in this priority list.

There is a general air of similarity between this procedure and the *CONVERGENCE* mapping, but there are also substantial differences. The first resides in evaluating coupling terms between phases from the derivatives of the score function chosen: although the *functional form* remains that of a dependence on triplet and quartet phase invariants, the *values* of the coefficients will incorporate the effect of higher-order interactions generated numerically by the process of ME extrapolation. The second resides in the strict book-keeping of the dependence of the score on the various phases as the mock basis set is being thinned out by elimination. For instance, if a given basis-set candidate $\mathbf{h}$ forms a triplet with candidates $\mathbf{h}_1$ and $\mathbf{h}_2$ there will be an interaction term (from the phased LLG for $\mathbf{h}$) involving that triplet invariant; if however $\mathbf{h}$ comes up for elimination from the mock basis set while $\mathbf{h}_1$

and $\mathbf{h}_2$ remain in it, the phase of $\mathbf{h}$ will be integrated out to produce an unphased LLG, resulting in quartet terms between $\mathbf{h}_1$, $\mathbf{h}_2$ and any other pair of uneliminated reflexions also involved in a triplet with $\mathbf{h}$. Thus the list of interactions never simultaneously contains triplets and quartets which are mutually redundant. This desirable feature follows naturally from using a globally defined and theoretically sound score function, rather than an *ad hoc* score obtained by piecing together indications for separate invariants, with problems of non-independence and questionable weighting (see I, §2.2).

The *COBWEB* procedure will be described in greater detail in a separate publication.

2.2.2. *Optimal sampling of trial sign and phase combinations.* (i) *The need for efficient sampling.* The process of node expansion is fundamental to the tree-directed multisolution strategy. It bears a clear resemblance to the phase permutation methods used since the early days of *MULTAN*, but its incorporation into the hierarchical structure of a search tree is novel and is related to the necessity of partitioning structure-factor space into small regions and of constructing distinct approximations to the joint-probability distribution within each region (I, §2.4; Bricogne, 1988*b*, §2). The problem of economizing on the number of nodes, or equivalently of finding the most efficient partitions of structure-factor space, arises from the conflicting requirements of (1) introducing the largest possible number of new reflexions in each extension of the basis set (to maximize sensitivity) and using the finest possible sampling of phase values (to maximize accuracy), and (2) keeping under control a combinatorial explosion of crippling proportions.

The converse problem occurs if we want to use the Fourier coefficients of the score function with respect to a set of phases to calculate a multidimensional Fourier 'map' giving an approximate value of the score at all points of a grid of trial phase combinations: such a map will be more useful when more phases can be made to vary simultaneously, but this number is severely restricted by the fact that even relatively small numbers of phases (*e.g.* 12) give rise to unwieldy maps.

(ii) *Previous methods: substantialization and magic integers.* The very idea of a computerized multisolution approach to the phase problem, which may be traced to papers by Cochran & Douglas (1953, 1955, 1957) and by Vand & Pepinsky (1956), was formulated in response to this combinatorial explosion. Its practical implementation was limited by the scarcity of computing and graphical display resources, which created as urgent a need as that which exists today for methods of economizing on the number of trial solutions which had to be calculated and inspected. Woolfson (1954), working with Huggins masks rather than a computer, gave an example of how such economy could be achieved by exhibiting a set of 16 combinations of seven signs with the property that any of the $2^7 = 128$ possible combinations of these signs differed from one of the given 16 in at most one place. Woolfson gave no proof of his statement other than suggesting its brute force verification. In the same year Good (1954) coined the term 'substantialization' for this process and showed how to construct a set of 2048 combinations of 15 signs having the same property with respect to the full set of $2^{15} = 32768$ possible combinations. Good did prove his statement but unfortunately he provided no indication as to whether these two examples were isolated recreational curiosities, or were instead instances of some underlying mathematical theory and hence were capable of generalization. I have found that such a theory does indeed exist, namely the theory of error-correcting codes (MacWilliams & Sloane, 1977) and of the packing and covering properties of high-dimensional lattices associated with them (Conway & Sloane, 1988) which will be described below in (iii). The two examples mentioned above are in fact the two simplest Hamming single-error correcting codes, Woolfson's being the [7,4,3] code [described in Shannon & Weaver (1949, pp. 80)] and Good's the [15,11,3] code, where the notation $[n,k,d]$ is explained in (iii) below.

Later, in the course of developing the *MULTAN* program, White & Woolfson (1975) showed how the 'phase correlation function' of Riche (1970, 1973), similar to criteria defined for centric cases by Cochran & Douglas, could be surveyed by means of a Fourier summation in much fewer dimensions than the number of independent phases or symbols. This economy was achieved by defining several phases φ_1, φ_2, ..., φ_n in terms of a single symbol x by means of congruential relations of the form

$$\varphi_i = 2\pi m_i x \pmod{2\pi}. \qquad (2.26)$$

Certain 'magic' integer sequences m_1, m_2, ..., m_n were found to produce smaller systematic errors in this representation than others of comparable size. Main (1977, 1978) gave a quantitative analysis of this phenomenon in terms of the shape of the Voronoi region associated to each of the n-dimensional lattice points representing the n-tuple of phases defined by regularly spaced sample values of x. This analysis will be seen within the context of coding theory [see (iii) below] to have come quite close to a reformulation of the optimal sampling problem as a search for 'magic' high-dimensional lattices with optimal packing, covering or quantizing properties. It may have been hampered by its adherence to the rigid mode of representation (2.26) in which each phase depends on *one* symbol only. As will be shown below, the optimal solutions may be viewed as *matrix* analogues of the magic-integer representation, with several generator symbols all contributing to defining all phases.

(iii) *A new method: magic lattices from coding theory.* Let the latest extension of the basis set H consist of m reflexions, m_a of which are acentric and m_c of them centric; denote by $n = 2m_a + m_c$ the total number of real components of the associated structure factors (the 'number of degrees of freedom'), and by Φ the vector made up of the m phases. Assume for the sake of definiteness that the

score function of interest for the trial values of Φ is the Bayesian score $B(\Phi)$ defined at the end of §1.3. By virtue of its periodicity, B can be expressed as an m-dimensional Fourier series

$$B(\Phi) = \sum_{\mathbf{K}} G_{\mathbf{K}} \exp(i\mathbf{K}.\Phi). \qquad (2.27)$$

Here each $\mathbf{K}$ denotes an m-tuple of indices $(k_1, k_2, ..., k_m)$ with k_j an unrestricted integer if the corresponding phase is acentric, and $k_j = 0$ or 1 if it is centric. The coefficients $G_{\mathbf{K}}$ have Hermitian symmetry (*i.e.* $G_{-\mathbf{K}} = \overline{G_{\mathbf{K}}}$) because B is real-valued. Those multi-indices $\mathbf{K}$ likely to contribute the most to B (which constitute the *spectrum* of B, hereafter denoted Spec B) are a subset of those for which $|\mathbf{K}| = \sum_{j=1}^{m} |k_j|$ is small, and a list of them can be compiled by the scheduling routine *COBWEB* (§2.2.1). Typically $|\mathbf{K}|$ is at most 3 or 4 but the effect of higher-order terms will be present in the *values* of the coefficients $G_{\mathbf{K}}$ if the previous basis set was already fairly large.

The optimal sampling problem may then be rephrased as follows: what is the most economical 'grid' of Φ values on which to either (1) carry out accurate pointwise evaluations of $B(\Phi)$ so that the dominant $G_{\mathbf{K}}$ can be extracted from the score values by m-dimensional Fourier analysis (see §2.2.4); or (2) calculate a Fourier synthesis giving approximate values of $B(\Phi)$ from the partial list of coefficients $G_{\mathbf{K}}$ compiled by *COBWEB* so as to avoid or minimize aliasing, and thus allow good interpolation to locate the peaks of B. The naive answer to this question is that the Shannon criterion should be satisfied along each phase independently. In the typical case where no $|k_j|$ exceeds 1 this would require at least $2^{m_c} 4^{m_a} = 2^n$ grid points for a full set of sign and quadrant phase combinations. Observe that this grid is the 'unit cell' of an m-dimensional crystal made up by repeating the motif $B(\Phi)$ for these grid points by the translations of a rectangular periodizing lattice having two division points along each centric phase and four along each acentric phase. The naive Shannon criterion then ensures the absence of overlap along each phase axis between the projections of translates of Spec B.

At this point '*the miraculous enters*' [to quote Conway & Sloane (1988)]: some unexpected properties of high-dimensional lattices can be exploited to achieve spectacular savings. Recall that the essence of Shannon's criterion is to prevent overlap of lattice translates of Spec B so that sampling at the points of the reciprocal lattice should lose no information. This criterion, however, needs only be met in m-dimensional space rather than in each one-dimensional principal projection. This remark prompts one to use m-dimensional periodizing lattices which give a much *denser* packing of translates of Spec B so that the sampling lattice, which is reciprocal to the periodizing lattice, will be much *looser* and yet will retain the same amount of information about Spec B as the conventional rectangular lattice.

In dimension 3 this method would consist (Bricogne, 1992*a*, §1.3.2.2.2.1) of periodizing a Bragg sphere of structure factors by the translations *e.g.* of a face-centered lattice (with optimal *packing* properties) rather than a primitive lattice, and computing electron densities on a body-centered grid (with optimal *covering* properties). Besides saving about a third on computation, this method would provide much better interpolation in real space than the ordinary rectangular sampling since each grid point would have 12 nearest neighbours rather than six.

In dimension $m \gg 3$ the possible gains increase dramatically. From the point of view of m-dimensional crystallography this phenomenon corresponds to the existence of centered lattice modes with high multiplicities and large holohedries, giving rise to very dense, highly symmetrical packings; the reciprocal (sampling) lattices then have optimal covering properties, *i.e.* have Voronoi regions with the smallest outer diameter, and allow high-quality interpolation because of the large number of nearest neighbours (called the 'kissing number') around each point. Certain values of m are especially favourable. For $m = 7$ and $m = 8$ the 'root lattices' E_7 and E_8 have multiplicities of 8 and 16 respectively; their duals E_7^* and E_8^* = E_8 have kissing numbers 126 and 240 instead of 14 and 16 for their rectangular counterparts. For $m = 24$ the Leech lattice achieves the extraordinary gain in packing efficiency of 2^{24} compared with the rectangular lattice, and gives a sampling grid with kissing number 196560 rather than 48.

Coding theorists have long taken advantage of the existence of such lattices to construct error-correcting codes. In the simplest case this consists of choosing a subset of 2^k combinations of n binary digits (0 or 1) called 'words' among the 2^n possible ones in such a way that any two distinct words differ in at least d places. Many of these codes are *linear*, *i.e.* consist of the linear span of k n-dimensional binary vectors under modulo 2 arithmetic. Such a code, denoted $[n,k,d]$, may be viewed as a centered lattice mode in dimension n, the $n - k$ linear relations satisfied by its points modulo 2 being *parity checks* expressing the 'reflexion conditions' of that lattice. There are numerous other types of codes (MacWilliams & Sloane, 1977). Some of them are non-linear, and the notation (n,M,d) is used to denote a non-linear code comprising M codewords of length n with minimum distance d.

These connections between lattices and codes, and their use in optimizing the sampling of trial phase combinations, will be described in full detail in a separate publication. They will be shown to lead to the best possible sampling schemes (as gauged by the shape of their Voronoi region), surpassing the performance of the best magic-integer sequences.

2.2.3. *Methods of node expansion.* Several options are available in *BUSTER* for generating the progeny of a given node, which should be self-explanatory in the light of the previous section and will in any case be described in more detail elsewhere. They are listed below as keywords followed by the relevant parameters.

(1) *Full permutation* (φ_0, $\Delta\varphi$, $n_{\max}$): creates one node per point of a standard rectangular grid, each centric

phase being given its two possible values while acentric phases run (in °) from φ_0 in steps of $\Delta\varphi$; this generates $2^{m_c}(360/\Delta\varphi)^{m_a}$ progeny nodes, with $n = 2m_a + m_c \leq n_{max}$, the latter being the maximum number of degrees of freedom to be activated in the expansion.

(2) *Hadamard code* (n_{max}): creates $2n_{max}$ progeny nodes from the rows of an $n_{max} \times n_{max}$ Hadamard matrix and their complements, where $n = n_{max} - 1$ or $n = n_{max}$; each row is made up of + and - signs, a centric reflexion using up one sign to define one of its restricted phases, and an acentric reflexion using up two signs to define a quadrant phase. The values of n_{max} currently supported are 4, 8, 12, 16, 24, 32. When the last bit is not used the corresponding code is said to be 'punctured', *i.e.* obtained by deleting a given coordinate from each original codeword.

(3) *Reed–Muller code* (r, m): creates a progeny node for each of the 2^k words of the rth-order binary Reed-Muller code of length $n_{max} = 2^m$, where k is given by the sum of binomial coefficients $k = 1 + \binom{m}{1} + \binom{m}{2} + \ldots + \binom{m}{r}$. The binary digits of these codewords are translated into phase combinations in the same way, interpreting 0 as + and 1 as -. For $r = 1$ these codes are Hadamard codes of length 2^m. The RM(1,3) and RM(2,4) codes are respectively [8,4,4] and [16,11,4] codes whose punctured versions are the [7,4,3] and [15,11,3] Hamming single-error-correcting codes; these have a covering radius of 1, which gives rise to the 'substantialization' property mentioned in §2.2.2 (i).

(4) *Golay code*: this generates 4096 codewords of length 24 forming a [24,12,8] code called the extended Golay code; the Golay code itself is the [23,12,7] code obtained by puncturing the former by removal of any one of the coordinates. These codes have covering radii of 4 and 3 respectively, giving rise to a staggering efficiency of substantialization: any of the $2^{24} = 16777216$ ($2^{23} = 8388608$) combinations of 24 (23) signs differ in at most 4 (3) places from a suitable word of these codes. They are translated into phase combinations as above.

(5) *Nordstrom–Robinson code*: this first generates the extended Golay code, then shortens it by 8 bits in a special way to produce 256 words of length 16 forming a (16,256,6) non-linear code called the Nordstrom-Robinson code; this in turn can be punctured into a (15,256,5) code. They have covering radii of 4 and 3 respectively so that their substantialization properties, although less spectacular than those of the Golay codes, are still quite attractive.

(6) *Survey node list* (φ_0, $\Delta\varphi$, n_{max}, l_{max}): this option activates the conversion of the relevant phase interactions compiled by *COBWEB* for a segment of m reflexions corresponding to at most n_{max} degrees of freedom into a list of m-dimensional Fourier coefficients, as explained in §2.2.2 (ii). An m-dimensional Fourier synthesis is then calculated on a rectangular grid defined by φ_0 and $\Delta\varphi$ as in (1) to produce a map of a 'synthetic' score approximating the true score; these scores are then tag-sorted and the trial phase sets associated with the grid points with the top l_{max} (typically 32) scores are used to create the progeny of the parent node. This procedure is being upgraded by the introduction of magic-lattice sampling of the synthetic score map (thus allowing higher values of n_{max} for a given memory size) and by locating the peaks of this map rather than simply selecting the top values. This option can be recognized as a modern version of the ψ map of White & Woolfson (1975), with some advantages: the map will give a better synthetic score because the coefficients are evaluated from derivatives of the true score at the position of the parent node, rather than at the origin of structure-factor space; the magic-lattice sampling is the most efficient possible; and the mapping from rectangular to magic-lattice coordinates is smooth (because of dimension-preserving) and thus does not distort peaks; in contrast the magic-integer mapping (2.26) is not smooth (because of dimension-collapsing) and induces shearing of the peaks.

(7) An enhancement of the previous method has been developed which first determines the rank of the list of m-dimensional indices attached to the input Fourier coefficients by means of an elimination process. This yields a redefinition of all the phases involved in terms of independent symbols, together with a detailed description of which symbols are likely (1) to have a single optimal value, or (2) to have a discrete multiplicity of equally optimal values, or (3) to have little or no effect on the synthetic score. The generation of progeny nodes can then exploit this information by designing permutation schemes for symbols under categories (2) and (3) which will allow the accurate scores subsequently calculated for these nodes to resolve the degeneracies of the synthetic scores. The algebraic basis and the applications of this method will be described in a separate publication.

2.2.4. *Methods of progeny-score analysis.* Once all the nodes belonging to the progeny of a given parent node have been evaluated, some decisions have to be made as to whether preferred values for certain phases or combinations of phases are being indicated by the distribution of the scores LLG(ν) or $B(\nu)$ attached to these nodes.

At this stage it must be recalled that LLG in the diagonal approximation is defined by (2.18) as a sum of logarithms of probability ratios calculated for a sample of observed values of structure-factor amplitudes in the second neighbourhood of the basis set. As such it is itself a random variable since different samples drawn from a population with a given theoretical distribution (which would combine both statistical dispersion and measurement errors) will in general yield different values of LLG. This intrinsic randomness of LLG results in the possibility that LLG($\mathcal{H}_0$) may be greater than LLG($\mathcal{H}_1$) even if ($\mathcal{H}_1$) is true because of the chance fluctuations in LLG($\mathcal{H}_0$) and LLG($\mathcal{H}_1$). It is therefore of the utmost importance to compare the 'observed' value of LLG with the statistical distribution of its fluctuations so as to gauge the level of significance of any indication of preference for ($\mathcal{H}_1$) over ($\mathcal{H}_0$). This significance level is defined as the probability that the observed LLG be due to statistical fluctuations in

LLG($\mathcal{H}_0$) and LLG($\mathcal{H}_1$) rather than to the change of distribution associated with the alternative hypothesis ($\mathcal{H}_1$).

From a practical point of view, this implies that any rejection of a hypothesis regarding trial phase values (*i.e.* any pruning of the phasing tree) should only be carried out on the basis of a significance test found to be conclusive at a pre-set significance level. To quote Cochran & Cox (1957, p. 5): '*A useful property of a test of significance is that it exerts a sobering influence on the type of experimenter who jumps to conclusions on scanty data, and who might otherwise try to make everyone excited about some sensational [effect] that can well be ascribed to the ordinary variation in his experiment.*' This word of caution needs to be borne in mind at all times, especially when working with known structures - as is necessary when developing and testing a new method - if it is to be ensured that no advantage is being taken of the fact that the answer is known in advance. Even then, 'insider dealing' remains possible in that such knowledge may exert an indirect influence on certain strategic choices which would otherwise be highly uncertain, and thus may lead to illusory successes which cannot be reproduced on unknown structures. It is therefore imperative to make the pruning of the phasing tree an automatic, quantitative and totally objective process.

In the present work no attempt has been made to calculate the theoretical variance of the LLG and to thus ascertain absolute significance levels: these are strongly dependent on the proper estimation of the 'effective' number of atoms which, in the macromolecular case, requires a careful study. I have instead adopted an empirical standpoint in which the set of scores attached to the various phase assumptions are Fourier-analysed with respect to the phases involved, and the largest Fourier coefficients are tested for significance against variances estimated from the sample itself.

The simplest instance of this procedure consists of detecting the effect of the sign of the real component of a single centric reflexion. Let μ^+ and μ^- denote the averages of the scores attached to the nodes where this sign is + and - respectively, and let V^+ and V^- denote the variances of the scores in these two subsets. Then the first Fourier coefficient of the two-point Fourier transform is, up to a factor of 1/2, the *contrast* ($\mu^+ - \mu^-$). Its significance level is the probability that it could arise solely from statistical fluctuations of the score in each subset (as measured by V^+ and V^-) if the two distributions of score had the same theoretical mean μ. It can be calculated by a Student t-test, as described for instance by Press *et al.* (1986, pp. 484-488). The same procedure can be applied to the detection of effects associated with products of several signs, the averages and variances being computed for the two subsets of nodes where that product is + or -. For acentric reflexions a general Fourier transform (*e.g.* on four or six points) can be applied, provided the phase values have been sampled accordingly. It is interesting to note that if the sample phase values of an acentric reflexion are chosen by 'quadrant permutation' ($\varphi = 45° + k \times 90°$, $k = 0, ..., 3$) the resulting origin-shifted four-point transform can be rearranged into a tensor product of two two-point transforms, along the real and imaginary components respectively of the complex phase factor: then the phases of acentric reflexions may be handled as a *pairs of signs* within the same binary formalism as centric reflexions. This approach, first programmed and extensively tested in *BUSTER*, was recently implemented in the *MICE* program and successfully applied to the determination of several powder structures (see *e.g.* Shankland *et al.*, 1992).

In the general case where scores have been evaluated on a rectangular grid or at the points of a lattice defined by a linear code, the Fourier coefficients are calculated by a multidimensional Fourier analysis. If the sample of scores has been evaluated at points defined by a non-linear code, the coefficients are calculated by least squares. The coefficients are then tested for significance.

2.3. *Methods of origin and enantiomorph specification*

2.3.1. *Enantiomorph control.* When dealing with non-centrosymmetric crystals of small molecules, the initial enantiomorph ambiguity is usually broken by fixing the sign of the imaginary part of the structure for *one* enantiomorph-sensitive (or *chiral*) reflexion whose phase is sufficiently strongly coupled to others for its chirality to propagate. When dealing with macromolecular crystals, the weakness of phase interactions implies that this mechanism can no longer be relied upon: a critical amount of chiral phase information must be allowed to accumulate before enantiomorph differentiation becomes locked in.

The phasing tree affords a convenient practical solution to this problem: letting the pool of competing nodes develop until the chirality of each node becomes unambiguous. Chirality is treated as a logical attribute of each node, with the property of being *hereditary* since any descendant of a chiral node is chiral. It is ensured at each level of node expansion that the collection of trial values for the new phases is globally invariant under enantiomorph switch. If a parent node is chiral, its entire progeny is chiral and there can be no duplication among them. If a parent node is non-chiral, its progeny is examined to determine which nodes remain non-chiral and which become chiral; the latter do so in pairs, and are duly marked so that only one member of each pair is subsequently evaluated (its scores are later copied to the other member prior to progeny score analysis) and possibly subjected to further expansion. When a mean absolute phase difference from a reference phase set is to be evaluated for a chiral node, it is calculated for both possible enantiomorphs and the smallest value is taken.

Another method is to expand the root node using the survey node list option [§2.2.3, method (6)]. The synthetic score map will be symmetric under central inversion in the 'chiral subspace' spanned by the m_a imaginary parts of the acentric structure factors. The highest pairs of distinct peaks related by this symmetry will correspond to

chiral phase sets, and selecting one peak in each pair in a consistent manner (*i.e.* with a positive coordinate in a direction along which the pairs are maximally separated) will fix the enantiomorph. Those peaks whose coordinates in the chiral subspace are all zero give rise to non-chiral nodes, so that the same procedure needs to be applied at the next level of expansion.

This scheme of 'deferred' enantiomorph discrimination gives a robust solution to the initial problem, the price to pay being the possible occurrence of a mixture of different enantiomorphs in the pool of nodes, which has to be sorted out in the late stages.

2.3.2. *Deferred origin definition.* A new algebraic method has been developed and implemented to determine whether a given collection of reflexions can determine the origin of a given space group uniquely, and if not, to describe algebraically the remaining degree of ambiguity. This procedure is valid for all space groups, centrosymmetric or not, primitive or not. It will be described in a separate publication.

It is well known that some structures may possess pseudosymmetries, sometimes so strong as to produce superstructure effects. In this case the top of the priority list compiled by *COBWEB* will contain reflexions belonging to a sublattice of the reciprocal lattice, and it will be necessary to use reflexions very far down that list in order to define the origin 'uniquely'. This uniqueness may then be largely illusory since the behaviour of the strongest terms will for some time be almost unaffected by the choice of origin: any decisions (*e.g.* tree pruning) taken during that time may then lead to the loss of origin coherence in the node pool.

This nettle may be grasped in the same way as that relating to enantiomorph choice. The reflexions are considered in strict order of priority, and the new algebraic procedure mentioned above is used to describe the degree of ambiguity which persists in the joint definition of origin and enantiomorph when a given phase set, belonging to a possibly non-primitive collection of reflexions, has been specified. This algebraic descriptor is related to the subgroup structure of the normalizer of the space group of the crystal. It is then treated as a node attribute whose properties under 'inheritance' are slightly more complex that those of chirality but can be exploited in the same way. The phasing tree is therefore developed on the basis of the strongest interactions available at each stage; origin and enantiomorph definition then occur at their own pace, *i.e.* when, and only when, the relevant pseudosymmetry-breaking phase interactions become dominant. The sampling designs have to be symmetrized with respect to the invariance group (*i.e.* the relevant subgroup of the normalizer) attached to the parent node before being applied to the expansion of that node, and redundant evaluations of progeny nodes are avoided by a supervision mechanism which generalizes that described for enantiomorph control. In particular, the residual ambiguity will induce some m-dimensional crystallographic symmetry in the synthetic score map, which may be exploited to save space and computation. This remark shows once again [as in §2.2.2 (ii)] that the design of optimal strategies to solve the phase problem leads naturally to problems in m-dimensional crystallography!

This radical procedure avoids all the well known pitfalls which result from prematurely assuming that origin and enantiomorph have been irreversibly specified.

3. Program description

This is only a brief outline, as the program will be described in more detail elsewhere.

3.1. *Initialization stage*

The following information is read from an external file to define the characteristics of the structure to be solved and provide the main guidelines as to the strategy to be followed.

(1) Cell geometry: the constants defining the direct lattice are read, and are used to calculate the reciprocal-lattice constants and the metric tensors of both lattices.

(2) Space-group symmetry: the name of the space group is read, its first letter being interpreted as a code for the lattice mode; the transformations themselves are read, without replication by centering translations (if any); the multiplication table and inverse table of the group are set up; a list is drawn of all the possible isotropy subgroups, and for each such subgroup a list of unique coset representatives is compiled to allow non-redundant symmetry expansions.

(3) Chemical composition: the number and type of scatterers in the asymmetric unit are read; standard scatterers have their scattering factors defined by tabulated coefficients; it is also possible to have user-defined group scatterers, described by a subroutine returning increments to $\sigma_1(\mathbf{h})$ and $\sigma_2(\mathbf{h})$ for any given $\mathbf{h}$; provision is made for entering fragments defining a partial structure at this stage.

(4) Solvent correction parameters: mean macromolecular electron density, mean solvent electron density, and solvent temperature factor.

(5) Resolution limits for accepting reflexion data from the external file.

(6) Definition of the sequence of data 'shells' which will be used in the successive stages of phase generation. These shells must be *nested*, *i.e.* each must strictly contain the previous ones. A shell is characterized by separate lower and upper resolution limits for data normalization, for basis-set selection and for likelihood evaluation; by E-value thresholds for defining weak, medium and strong reflexions in the charting of phase interactions; by the specification of the method of phase permutation to be used in the expansion of nodes, and of various constants pertaining to that method; and by the specification of a method of analysis of the scores (LLG or B) attached to the progeny of each node for the purpose of inferring phase information, together with a significance-level threshold for accepting these phase indications.

The program then initializes all the pointers involved in the book-keeping of the phasing tree and creates a null node corresponding to a void basis set.

3.2. *Data input and preparation*

This comprises three main stages.

(1) Read **h**, $|F|^{obs}$, $\sigma(|F|^{obs})$ and if available (for comparison purposes only) $|F^{calc}|$ and φ^{calc}. It should be noted that all the calculations to be described below use $|F|^{obs}$, not $|F^{calc}|$! For each **h**, determine its symmetry attributes (isotropy type, centric character and phase restriction) and store them. Apply solvent correction if required.

(2) Calculate initial values of the scale factor k and temperature factor B required to put the solvent-corrected data on absolute scale and compute $|U|$ and $|E|$ values. Rough values of k and B are first obtained by means of a Wilson plot, then refined by likelihood maximization in a routine *MLNORM*, described in §2.1.3, which carries out all likelihood evaluations and optimizations.

(3) Generate the list of associated Miller indices to be used in the frequent re-evaluation of structure-factor variances in the diagonal approximation by means of structure-factor algebra (§2.1).

3.3. *Survey of phase interactions and ranking of basis-set candidates*

This crucial step is carried out in routine *COBWEB*, whose logic was described in §2.2.1. The strongest interactions are those coming from the modulation of variances through (2.9*a*,*b*), leading to Σ_1-type interactions, and from the modulation of first moments through (2.3), leading to triplet-type and quartet-type interactions for phased and unphased LLG's, respectively.

3.4. *Origin definition*

Origin definition is carried out as follows:

(1) Reflexions are considered in order of priority as potential contributors to origin definition (§2.3.2). Deferred origin definition is almost but not totally implemented.

(2) Once origin-defining reflexions (if any are needed) have been chosen, phases are assigned to them, in accordance with their known values if these are available; care is then taken to shift all available known phases so as to give all *acentric* origin definers a zero phase value and thus avoid introducing spurious enantiomorph-sensitive information.

(3) The root node of the phasing tree (level 1) is created from these choices. If no origin-defining reflexions are needed, this remains the null node.

(4) The remaining candidate reflexions are re-ranked by *COBWEB*, preventing the basis-set reflexions from being eliminated.

This completes the initialization stage of the calculation. It results in a *pooled node list at level* 1 containing a single node, the root node. At subsequent levels this list will contain many nodes, those of them deemed to be worthy of further expansion being marked as 'seed' nodes.

3.5. *Basic recursion mechanism*

Let us take as an 'induction hypothesis' that, at the completion of level n, there are at most eight nodes in the *pooled node list at level n* which have been marked as 'seed' nodes. The number eight has no special significance beyond that of having kept most test calculations performed so far within manageable size while not proving damagingly restrictive. In some cases this number can be made self-adjusting (see below, step 3).

For $n = 1$ there is one such node (the root node), which is duly marked as a seed node so recursion can start. It now proceeds through the following steps.

1. Choose a segment of new reflexions and add it to the previous basis set to make up the basis set at level $n + 1$. Prepare all data which depend only on level, such as the list of reflexions in the second neighbourhood, and the list of associated Miller indices to be used in computing (by structure-factor algebra) the covariances between the basis-set structure factors in order to solve the maximum-entropy equations.
2. For each seed node in the pooled node list:
   - 2.1. Create the progeny of that node by permuting the phases of the new reflexions according to the method specified for the current shell.
   - 2.2. For each progeny node ν:
     - 2.2.1. Solve the maximum-entropy equations to construct q_ν^{ME}. This is achieved in routine *MAXIMS* by the duality method (Agmon *et al.*, 1979) as described in §2.1.1.
     - 2.2.2. Calculate the log-likelihood gain LLG(ν) and the Bayesian score $B(\nu)$.
     - 2.2.3. Analyse the dependence of LLG(ν) or $B(\nu)$ on the permuted phases by the method specified for the current shell. Select phase indications at chosen significance level.
     - 2.2.4. Cull the progeny according to the selected phase indications or substitute the phases and re-evaluate as in steps 2.2.1 and 2.2.2.
3. Pool surviving nodes, sort them on LLG(ν) or $B(\nu)$, and mark the top eight (or fewer) as seed nodes. The marking of obvious losers can be avoided by calculating the exponentials of the pooled node scores, normalizing them so they add up to 1 and interpreting them as probabilities: the marking of the top nodes as seed nodes is then stopped when the remaining nodes account for (say) less than 1% of the total probability.

The induction hypothesis is now fulfilled with n replaced by $n + 1$.

3.6. *Higher levels of control*

Higher levels of control can be achieved by:

(1) Recursion on shells: if at step 1 the end of the ranked candidate list is reached, the next shell specified in

Table 1. *Test of ab initio phase determination on platynecin*

Basis-set history

Level	m	n	New reflexions included
1	3	3	3 1 0, 5 0 6, 0 4 9
2	11	16	2 0 8, 2 2 0, 2 6 5, 4 4 5, 4 6 3, 6 4 3, 5 1 8, 5 3 0
3	18	29	7 1 0, 3 3 8, 5 1 9, 7 1 1, 4 2 2, 4 2 10, 4 4 1
4	25	42	4 5 1, 5 2 6, 1 7 5, 2 0 2, 1 7 1, 1 6 7, 1 3 10
5	33	56	9 1 1, 8 2 3, 8 0 8, 6 2 8, 5 5 8, 7 1 10, 8 2 0, 9 2 1
6	41	70	7 5 3, 7 6 1, 6 1 6, 7 0 2, 7 4 2, 5 2 7, 7 5 0, 3 7 9

Tops of pooled node lists at successive levels

Level	Node	Parent	Figure of merit	Phase error	Phase error in $\mathcal{N}_2(H)$	Cumulative probability
2	7*	1	48.80045	59.695	86.753	0.433158
2	9*	1	48.80039	14.812	46.467	0.866291
2	8*	1	46.90649	51.550	81.974	0.931471
2	6*	1	46.90642	6.667	35.024	0.996645
2	31	1	42.62264	59.695	84.436	0.997544
2	30	1	42.62262	14.812	38.407	0.998443
2	11	1	41.76640	67.195	89.526	0.998825
2	12	1	41.76636	22.312	52.372	0.999207
3	129*	8	91.60087	53.825	55.048	0.500036
3	67*	6	91.60072	8.147	24.224	1.000000
3	141	8	73.39848	47.395	55.635	1.000000
3	79	6	73.39845	14.577	33.703	1.000000
3	70	6	72.08923	10.793	32.248	1.000000
3	132	8	72.08908	51.179	57.234	1.000000
3	85	6	70.74586	19.694	36.137	1.000000
3	147	8	70.74561	65.825	66.817	1.000000
4	208*	67	92.73958	10.410	34.982	0.463907
4	239*	129	92.73522	60.019	71.709	0.925798
4	240*	129	89.87247	59.521	71.701	0.952178
4	209*	67	89.87243	10.907	34.903	0.978556
4	221*	129	88.97218	56.794	60.052	0.989278
4	190*	67	88.96849	8.534	24.791	0.999960
4	246	129	82.45014	61.419	71.084	0.999976
4	215	67	82.44959	13.159	37.524	0.999992
5	353*	221	113.33800	49.204	61.735	0.493145
5	260*	190	113.33773	19.362	37.433	0.986158
5	259*	190	109.07172	18.248	39.857	0.993079
5	352	221	109.07076	47.473	60.779	0.999993
5	349	221	101.44053	48.359	58.198	0.999997
5	256	190	101.44035	17.362	38.799	1.000000
5	357	221	95.21426	50.359	65.449	1.000000
5	264	190	95.21358	17.111	41.337	1.000000
6	456*	259	−125.14119	19.555	40.889	1.000000
6	498	353	−234.50043	54.410	69.552	1.000000
6	467	260	−234.50247	19.460	47.645	1.000000
6	457	259	−241.63847	21.966	53.962	1.000000
6	469	260	−246.85487	20.778	52.167	1.000000
6	500	353	−246.89444	52.831	70.949	1.000000
6	504	353	−255.36122	51.252	74.556	1.000000
6	473	260	−255.38322	22.357	50.100	1.000000

Table 2. *Test of ab initio phase determination on sucrose octaacetate*

Basis-set history

Level	m	n	New reflexions included
1	3	3	12 11 0, 9 0 1, 5 9 0
2	7	7	6 0 6, 2 0 2, 0 12 4, 0 6 6
3	9	11	1 5 1, 10 11 2
4	12	15	11 0 3, 8 13 0, 7 6 3
5	16	20	2 17 0, 1 5 0, 1 0 5, 7 8 1
6	19	25	6 2 7, 13 12 0, 5 14 5

Tops of pooled node lists at successive levels

Level	Node	Parent	Figure of merit	Phase error	Phase error in $\mathcal{N}_2(H)$
2	16	1	24.65388	0.000	41.390
2	14	1	24.35492	45.000	69.948
2	8	1	24.04166	45.000	65.013
2	6	1	23.62893	90.000	84.191
2	17	1	22.23748	45.000	57.264
2	12	1	22.08409	45.000	61.120
2	10	1	21.75921	90.000	90.149
2	9	1	21.63463	90.000	75.396
3	208	16	31.87217	8.567	58.675
3	207	16	31.73194	12.467	58.403
3	206	16	31.65525	27.467	68.521
3	222	16	31.63401	25.033	68.893
3	222	16	31.63401	25.033	68.893
3	223	16	31.63214	10.033	56.319
3	209	16	31.61659	23.567	67.850
3	211	16	31.51924	27.467	63.215
3	210	16	31.40901	42.467	70.664
4	284	206	42.38756	43.011	72.059
4	312	207	42.27244	40.367	70.099
4	320	207	42.27152	38.184	66.881
4	288	206	42.26751	44.377	71.105
4	346	208	42.21553	15.611	62.636
4	476	222	42.19426	38.011	72.707
4	504	223	42.19419	35.367	70.680
4	480	222	42.15837	38.243	72.101
5	551	284	60.48391	55.669	80.989
5	559	284	60.36)92	49.239	76.555
5	543	284	60.21621	48.746	75.845
5	991	504	60.18473	47.592	76.897
5	871	476	60.18429	56.346	82.264
5	623	288	60.15242	52.229	76.185
5	671	312	60.14440	46.915	76.735
5	550	284	60.10799	69.515	83.454
6	1508	991	82.90446	47.475	81.648
6	1451	871	82.90424	48.225	83.445
6	1520	991	82.90270	47.661	81.534
6	1455	871	82.90265	48.411	83.405
6	1380	671	82.68206	46.925	81.558
6	1392	671	82.66939	47.366	81.249
6	1373	671	82.40872	58.175	82.336
6	1381	671	82.37637	52.550	85.254

the input is created. The survey of interactions and ranking of new candidates is carried out as in §3.3, inhibiting the elimination of reflexions already in the basis set, and the desired segment is taken from the top of this new list. If no next shell is specified, the calculation terminates.

(2) Supervision of enantiomorph control and (presently) of deferred origin definition.

4. Status report on current capabilities

4.1. *Examples and results*

The test calculations presented here are to be seen more as demonstrations of the use of *BUSTER* than as genuine results. They are summarized in Tables 1–3 by a recapitulation of the history of the basis-set enlargement at successive levels of the phasing tree, m denoting the number of reflexions and n the number of degrees of freedom, and a listing of the top eight nodes of the pooled node list at each level. In Table 1 a subset of these top eight nodes is sometimes chosen by a cumulative probability cutoff, the nodes selected in this way being shown by an asterisk; in the other two cases all eight top nodes are expanded and no asterisk is used.

4.1.1. *Platynecin.* Platynecin is a small acentric structure used as a test by Gilmore *et al.* (1990). It crystallizes in $P2_12_12_1$ (a = 7.81, b = 8.35, c = 12.46 Å) with 11 non-hydrogen atoms in the asymmetric unit. All data to 0.80 Å were used. Node expansion was carried out by the survey node list method with up to 14 degrees of freedom per segment, acentric phases being sampled at 30 + 60 k degrees. The top 32 synthetic scores were used for node expansion, and the progeny nodes were sorted on Bayesian score (defined at the end of §1.3) after their accurate evaluation.

Table 3. *Test of ab initio phase determination for crambin*

Basis-set history

Level	m	n	New reflexions included
1	3	4	1 0 0, 0 0 1, 1 1 0
2	12	19	4 0 0, 2 0 0, 6 1 4, 7 2 4, 6 3 4, 7 4 4, 3 0 0, 9 3 4, 5 3 4
3	21	35	0 0 8, 0 4 4, −1 3 4, 4 0 4, 4 4 0, 5 5 0, 10 2 5, 11 5 1, −4 3 4
4	30	50	9 2 0, 10 0 4, 4 0 6, 7 3 3, 10 2 3, 12 1 1, 16 1 0, 2 0 6, −6 4 1
5	38	66	4 4 4, 8 1 0, 2 5 0, 4 1 8, 9 1 3, 12 3 0, 11 2 1, −7 6 3
6	48	82	2 4 3, 5 6 2, −2 0 5, 4 5 6, 3 3 7, 4 0 5, 5 0 5, 0 6 2, 11 0 3, −13 2 5
7	56	97	3 3 4, 3 6 0, 10 1 6, 11 1 5, −10 4 1, 0 2 6, 5 2 8, −1 0 1
8	65	112	−6 5 4, −6 2 5, 12 4 3, −11 2 2, −11 4 2, −13 1 2, −3 0 4, −4 0 5, −3 0 8

Tops of pooled node lists at successive levels

Level	Node	Parent	Figure of merit	Phase error	Phase error in $\mathcal{N}_2(H)$	Cumulative probability
2	4	1	21.00938	36.406	84.696	0.144474
2	15	1	20.88882	61.423	101.063	0.272539
2	8	1	20.79058	38.170	85.872	0.388622
2	14	1	20.70145	75.594	94.645	0.494805
2	22	1	20.67354	69.423	102.602	0.598066
2	23	1	20.49953	73.466	96.064	0.684835
2	2	1	19.79727	56.406	82.041	0.727826
2	28	1	19.75916	42.577	91.366	0.769209
3	126	8	41.89569	67.287	92.384	0.027248
3	223	23	41.76482	76.322	91.936	0.051154
3	144	14	41.74904	75.482	92.787	0.074686
3	98	8	41.72868	57.287	90.181	0.097744
3	231	23	41.70387	86.322	94.907	0.120236
3	233	23	41.70050	68.322	91.204	0.142652
3	151	14	41.65844	72.322	92.305	0.164146
3	220	23	41.65089	63.243	87.343	0.185478
4	428	220	54.38519	58.389	88.070	0.464195
4	427	220	52.09603	75.503	87.332	0.511242
4	429	220	52.02784	67.394	88.017	0.555188
4	340	126	51.79845	77.263	92.840	0.590126
4	387	151	51.75230	80.433	93.140	0.623488
4	517	233	51.71685	77.766	92.326	0.655688
4	412	220	51.50346	74.769	88.719	0.681700
4	356	144	51.36866	82.540	93.087	0.704433
5	709	428	79.53835	60.201	87.837	0.523084
5	654	427	77.39612	67.369	87.109	0.584490
5	740	429	77.32279	63.540	87.914	0.641554
5	628	412	77.29583	66.803	88.903	0.697101
5	701	428	77.16959	63.055	87.183	0.746060
5	692	428	76.98388	68.189	88.925	0.786721
5	681	427	76.76070	73.265	88.069	0.819248
5	691	428	76.66429	66.700	88.644	0.848786
6	814	654	73.10498	73.792	87.759	0.120484
6	978	709	72.47102	68.057	88.586	0.184399
6	868	681	72.33136	72.533	87.949	0.239983
6	827	654	72.26581	62.507	86.125	0.292041
6	905	692	72.11742	76.670	89.903	0.336919
6	992	709	71.81743	73.572	88.690	0.370166
6	914	692	71.78978	70.506	89.314	0.402506
6	809	654	71.69795	66.592	87.193	0.432009
7	1117	827	110.75179	62.952	85.767	0.470695
7	1076	814	110.35623	72.534	87.411	0.787617
7	1135	868	109.22388	71.465	87.522	0.889753
7	1273	992	108.02380	72.347	88.151	0.920513
7	1078	814	107.74713	71.925	88.801	0.943839
7	1099	827	107.24278	65.669	85.394	0.957925
7	1042	809	106.55161	69.138	86.667	0.964982
7	1209	914	106.40523	69.744	88.848	0.971078
8	1306	1076	92.19655	75.062	87.969	0.802646
8	1401	1117	90.07094	68.508	88.099	0.898450
8	1470	1209	88.90709	69.742	89.234	0.928367
8	1338	1078	88.34831	72.694	88.901	0.945478
8	1423	1117	88.23022	66.414	87.090	0.960682
8	1441	1135	88.18805	75.785	89.539	0.975259
8	1402	1117	87.14124	63.276	86.400	0.980376
8	1308	1076	87.08044	72.707	87.327	0.985191

Phase determination proceeded as shown in Table 1. The discrimination power of the Bayesian score B is excellent. Its sudden decrease between levels 5 and 6 may be attributed to the phenomenon discussed in §2.1.2, namely the near-vanishing of determinants, since at that stage we are handling 41 basis-set reflexions for an 11-atom structure. The negative values of B are caused by the accumulation of phase sampling errors; earlier tests using coarser sampling led to arithmetic overflows in the large negative B values attached to some nodes. In the present case the bottom of the node list at level 6 contained nodes with $B < -122854.0$.

The final result is excellent, the mean absolute phase error being 19.6° for all reflexions with $|E| \geq 1.8$ (the final basis set) and 41° for those in the second neighbourhood, which comprise the near-totality of the data.

4.1.2. *Sucrose octaacetate*. This structure of 47 non-hydrogen atoms in $P2_12_12_1$ (a = 18.35, b = 21.44, c = 8.35 Å) belongs to Sheldrick's data bank of difficult test structures. Sheldrick himself used it to calibrate his phase annealing procedure (Sheldrick, 1990) where it was by no means the easiest of the test cases, giving at best 50 successes in 10 000 tries.

All data to 0.80 Å were used. Node expansion was carried out by the full permutation method with up to five degrees of freedom per segment, acentric phases being sampled at 45 + 90 k degrees. Phase indications were extracted by Fourier analysis of the LLG [defined by equation (2.18)] and were selected at the 2% significance level.

Phase determination proceeded more laboriously (see Table 2), especially as the early incorporation into the basis set of reflexions with large indices makes the FFT calculations rather slow. The LLG is nevertheless able to produce good phase indications even for small basis sets, an encouraging observation in view of the low hit rate of phase annealing on this structure. The large phase errors in the second neighbourhood indicate the urgent need for a phase refinement procedure.

4.1.3. *Crambin*. This structure was used in (I) to assess the usefulness of maximum-entropy extrapolation for a macromolecule. The data set used is the same as in (I): it extends only to 1.5 Å resolution but is otherwise unusually complete and accurate.

Crambin crystallizes in $P2_1$ (a = 40.96, b = 18.65, c = 22.52 Å) with 324 non-hydrogen atoms in the asymmetric unit, excluding the solvent atoms which have high temperature factors. The basis-set reflexions were restricted to spacings greater than 2.5 Å but all data to 1.5 Å were used to compute log-likelihood gains.

Node expansion was performed by the survey node list method with up to 16 degrees of freedom per segment, acentric phases being sampled at 36 + 72 k degrees. The top 32 synthetic scores were used for node expansion, and the progeny nodes were sorted on Bayesian score following their accurate evaluation.

The Bayesian score shows good power of discrimination (see Table 3), especially in view of the very low resolution (by direct-methods standards) at which the computation takes place. The accumulation of phase sampling errors eventually asphyxiates the search process, as shown by the decrease of B between levels 7 and 8. Nevertheless

the final basis-set phases obtained (for the 65 reflexions with $|E| \geq 1.8$ to 2.5 Å resolution) are significantly away from total randomness. Useful results are thus likely to emerge once the coding of the phase-refinement procedure (by maximization of B with respect to the basis-set phases) is completed.

4.2. *Summary of tests and applications*

The results presented above are of a preliminary nature but they should be viewed together with those obtained in collaboration with Chris Gilmore and coworkers and with Charles Carter and coworkers (see *Introduction*). A number of clear conclusions then emerge.

(1) The tree-directed multisolution strategy is a versatile computational procedure capable of dealing in a fully automatic fashion with a wide variety of problems, whether decisions can be taken very early (as with small molecules) or have to be deferred until unpredictably large amounts of phase information have accumulated.

(2) The phase-extension capabilities of constrained entropy maximization are very substantial and keep increasing as phase information accumulates. For macromolecules, the availability of a good approximation to the molecular envelope has a profound influence on the strength and quality of this extrapolation (Xiang *et al.*, 1993) as had been anticipated on theoretical grounds (I, §8.3; II, §2.3; see also §5.2 below).

(3) The log-likelihood gain and the Bayesian score are extremely powerful statistical criteria for the early discrimination of good trial phase sets, even for macromolecules, provided care is taken to maximize their sensitivity to these phases by appropriately scheduling the growth of the basis set.

(4) The statistical analysis of scores and the selection of significant phase indications is a robust and automatic procedure which enhances the usefulness of these scores. In a recent application to a powder structure (Shankland *et al.*, 1992), a large number of nodes with high LLG values were obtained in the first round of phase permutation, between which no clear choice was possible. Applying the t-test removed many of these nodes. Among the survivors the node with the highest LLG (by a comfortable margin) gave an essentially perfect map, and the others all showed useful amounts of correct structure. By contrast all nodes deleted by the t-test, some of which had higher LLG's than some of the survivors, were devoid of correct detail. Furthermore the correct node turned out to have the *lowest* entropy of all, the node with maximum entropy being devoid of any correct detail.

To conclude, the main difficulty in direct macromolecular phasing seems to reside not in any deficiencies of the statistical criteria available in *BUSTER* - since the tests carried out on the *SAYTAN* phase sets for APP (Gilmore, A. N. Henderson & Bricogne, 1991) show that the LLG possesses outstanding powers of discrimination for large basis sets - but rather in conducting sufficiently thorough searches through the vast number of trial phase sets pertaining to such large basis sets and in preventing the accumulation of phase sampling errors through phase refinement. The chances of reliable success are therefore likely to depend critically on the full implementation of the efficient phase sampling methods described in §2.2.2, using magic lattices or non-linear error-correcting codes, and of the phase refinement method described below in §4.4. The implementation of the full likelihood function mentioned in §2.1.3 would undoubtedly afford further help by giving a more discriminating score.

4.3. *Summary of methodological developments*

The main guiding principle in this work has been that the shortest path to the most powerful and efficient solution to the phase problem resides in the correct identification of the underlying mathematical phenomena, followed be a systematic computer implementation structured around these phenomena. This quest for optimality has led to the following developments.

(1) Previous methods of calculation of joint-probability distributions of structure factors have been replaced by the powerful *saddlepoint approximation* (based on maximum-entropy distributions of random atomic positions). This method is optimal in the sense that it is as accurate for arbitrary feasible structure-factor values as the Edgeworth series was for vanishingly small values.

(2) The *a priori* probability of a set of trial phase values has been supplemented by a powerful statistical criterion, the *log-likelihood gain* (LLG), which measures the degree of corroboration of the phase assumption by biases present in the distribution of the unphased amplitudes and therefore consults much more data in the critical early stages of phase determination. This criterion is an optimal figure of merit by the Neyman-Pearson theorem, and can be combined with the entropy loss to provide a *Bayesian score*. The latter also enjoys optimality with respect to minimization of the risk attached to accepting or rejecting various trial phase sets.

(3) The hierarchical exploration of the space of all possible phase sets is represented by the growth of a search tree, whose nodes label the various trial phase sets, the maximum-entropy distributions associated with them, and local expansions of the LLG and the Bayesian score with respect to groups of yet inactive phases. The latter expansions have been used to schedule optimally the inclusion of new reflexions into the basis set so as to maximize the sensitivity of the score to the next set of trial phases and also to provide approximate functional expressions of the scores as multidimensional Fourier series.

(4) The design of efficient sampling schemes capable of varying large numbers of trial phases simultaneously, or of allowing the survey of multidimensional Fourier series approximations to the LLG or Bayesian score, has been based on the theory of error-correcting codes which provides an optimal generalization of previous methods

(substantialization and magic integers). This approach can afford gains of several orders of magnitude in sampling efficiency.

(5) The decision-making process involved in accepting or rejecting trial phase sets in the course of a structure determination has been made entirely automatic and objective by Fourier-analysing the score values and assessing the significance level attached to the various coefficients.

(6) New techniques for handling systems of modular linear equations with integer coefficients have been identified and applied to the problems of origin definition (with possible deferment) and of quantifying the degree of multimodality inherent in systems of phase relations.

(7) Finally a general multidimensional Fourier transform procedure has been implemented which can accommodate any number of dimensions and any number of division points along them. It has been applied to the computation of approximate scores and to the analysis of accurate node scores.

The generality and optimality of these additions to previous approaches endow them with an unusual degree of permanence and stability, so that large-scale programming efforts can confidently be undertaken on their basis.

4.4. *Forthcoming developments*

In the immediate future the programming efforts in *BUSTER* will concentrate on consolidating the existing features and seeking improved performance in the phasing of difficult small structures. These comprise:

(1) phase refinement by maximization of the log-likelihood gain or of the Bayesian score with respect to the basis-set phases; the necessary derivatives can all be calculated by structure-factor algebra (as will be shown in the sequel to II) and most of the computer code is already in place; this should greatly reduce the rate of accumulation of phase error with increasing basis-set size;

(2) calculation of a full likelihood criterion, without diagonal approximation, for which an analytical expression has been obtained;

(3) no longer fitting the constraints of the entropy maximization problem exactly but only to the point where the Bayesian score reaches a maximum; this will already provide a first round of refinement of the basis-set phases;

(4) improving the quality of the local approximation to the Bayesian score by a multidimensional Fourier series, particularly in conjunction with (2);

(5) exploiting the multidimensional crystallographic symmetries induced into the synthetic score map by the actions of the space-group normalizer and of its subgroups (§2.3.2) so as to avoid redundant computations at this stage;

(6) optimizing the use of symmetry in the calculation of ordinary crystallographic Fourier transforms, which are heavily used in solving the maximum-entropy equations; numerous new and efficient algorithms are described in the new volume B of *International Tables for Crystallography* (Bricogne, 1992*a*) which are awaiting implementation.

As far as macromolecular structures are concerned, there are strong theoretical reasons to expect that an absolute prerequisite to successful *ab initio* phasing without extraordinary data resolution will hinge on the proper characterization and exploitation of solvent regions, which in turn will require the implementation of the multichannel formalism described in §5.1 below.

5. Perspectives: the Bayesian programme

The conclusions reached above in §4.2 apply only to the hardest form of the phase problem, namely *ab initio* phase determination. With many macromolecular structures, however, it is often the case that a wide variety of sources of phase information may be available which, although too weak individually to enable the structure to be solved, could be used to prime or guide a statistics-driven phasing process.

From the converse point of view it can be argued that the standard phasing methods of macromolecular crystallography (isomorphous replacement, molecular replacement, solvent flattening and non-crystallographic symmetry averaging) tend to be somewhat deficient in the various forms of statistical analysis leading to the phase probability distributions through which they pool and communicate phase information. In most cases the sources of phase uncertainty can be traced back to an underlying statistical model involving randomly positioned scatterers, whose distribution often needs to possess varying degrees of non-uniformity which the usual treatments (typically based on Wilson statistics) are unable to accommodate. The mathematical apparatus described in §2.1 is an obvious candidate to provide the statistical techniques which had hitherto been missing in each of the standard methods.

It is precisely this rationale which motivated the synthesis, within a Bayesian theory of structure determination, between direct methods and conventional macromolecular methods presented in (II). The work described above has brought closer the actual implementation of this 'Bayesian programme' and has provided an infrastructure for its rapid development. This section will give an overview of the various components of this interface with standard methods and of the bonuses which can be expected.

5.1. *Multichannel formalism*

The present implementation of structure-factor statistics in *BUSTER* already accommodates (§2.1.3) any heterogeneous chemical composition - provided all atoms have the *same* non-uniform distribution of random positions - as well as the possibility of specifying a partial structure.

The main new ingredient described in (II) is a 'multichannel formalism' which can accommodate not only different types of atoms (specified by their scattering factors), but also different non-uniform spatial distributions for each atom type. This is the minimum prerequisite for a proper statistical treatment of solvent effects in macromolecular crystals. It can also accommodate the possibil-

ity of varying the scattering factors of certain subsets of atoms between different sets of intensity data collected from the same unknown crystal structure, as specified by a *matrix* of scattering factors. The treatment then encompasses the multiple isomorphous replacement (MIR) and multiwavelength anomalous-dispersion (MAD) methods which vary the scattering factors of localized substituents, as well as the contrast variation method which modifies that of the solvent. Finally, the possible existence of non-crystallographic symmetry and/or of multiple crystal forms can be accommodated by generalizing Bertaut's structure-factor algebra so as to include the action of the geometric operations relating the crystallographically independent copies of the unknown macromolecule in the crystal(s).

From a programming point of view the implementation of the multichannel formalism requires only benign modifications of the existing code, and leaves intact the general structure of *BUSTER* as far as solving the maximum-entropy equations is concerned.

5.2. *Solvent flatness and contrast variation*

Devising a correct statistical treatment capable of allowing for the existence of solvent regions in macromolecular crystals is a particularly intractable problem for conventional direct methods. The latter would use the same uniform distribution for solvent atoms (with high temperature factors) and macromolecule atoms (with low temperature factors) yielding a 'random soup' model which would completely fail to represent the fact that the two types of atoms are totally segregated.

The multichannel formalism, on the other hand, has no difficulty in providing a satisfactory statistical model once an approximate molecular envelope function is known or assumed. This model provides new insights into the power of the solvent-flatness constraint in macromolecular crystallography (II, §2.3). By structure-factor algebra (§2.1.0) the covariances between contributions emanating from neighbouring reciprocal lattice points can be a substantial fraction of unity, since they are obtained as sample values of the Fourier transform of the molecular envelope near its origin peak. As the number N of atoms increases, the width of this origin peak remains roughly constant in reciprocal lattice units, so that the large covariances due to solvent regions remain close to unity as N increases. In classical direct methods, this would correspond to the existence of $|E|$ values of order $N^{1/2}$ - an unforeseen circumstance which, as argued in §8.3 of (I), indicates that probabilistic methods should be useful for macromolecules even at low or medium resolution.

Entropy maximization under the extra constraint of solvent flatness specified by a molecular envelope can therefore be expected to be *stronger* than in the absence of this constraint, and to produce *more accurate* extrapolated phases since solvent flatness will be enforced *in advance* during phase extension; this is clearly preferable to doing so *a posteriori* by iteratively masking the electron-density map by the molecular envelope function, as is done in conventional solvent flattening (Bricogne, 1974; Schevitz, Podjarny, Zwick, Hughes & Sigler, 1981; Wang, 1985; Leslie, 1987).

A converse to this enhanced extrapolation is that likelihood tests can be applied to the initial estimation or the subsequent refinement of the molecular envelope. The sensitivity of this envelope-detection method will be greatly enhanced if X-ray contrast variation data can be measured (Carter, Crumley, Coleman, Hage & Bricogne, 1990): indeed it was the anticipation of the crucial role which the availability of a precise envelope was likely to play in the direct phasing of macromolecules which motivated the work described in this article. I have derived an expression for the likelihood function attached to a putative envelope, or to a parametrized modification of an existing envelope, and incorporating either native data alone or a series of contrast variation measurements. As pointed out by Carter *et al.* (1990), anomalous scattering by the solvent can be exploited to obtain more information about the envelope. The use of such anomalous scattering at several wavelengths is also a possibility if synchrotron radiation is available, and these measurements can themselves be incorporated into the likelihood function. This work will be reported in greater detail elsewhere.

5.3. *The MIR and MAD methods*

The multichannel formalism offers a natural framework for numerous statistical treatments of these methods, but only those amenable to quasi-immediate implementation will be dealt with here.

I have shown elsewhere (Bricogne, 1991*c*,*f*) that a maximum-likelihood treatment of the heavy-atom parameter refinement problem in the MIR method provides an optimal and definitive answer to a conundrum [the bias problem, see Blow & Matthews (1973)] which had plagued - or at least greatly inconvenienced - most macromolecular crystallographers for the past 25 years. This approach is the only one capable of refining the parameters measuring the lack of isomorphism of the various derivatives in a stable fashion. I have supplemented this study with the design of a method for detecting heavy atoms and checking for minor sites as the parameter refinement proceeds, which will be published separately. This approach can also be applied to the MAD method where it provides a sounder statistical treatment than is currently available, present methods (Pähler, Smith & Hendrickson, 1990) being known to lead to inflated figures of merit.

Once the best possible heavy-atom parameters have been obtained, there remains the problem that the resulting phase indications are most often ambiguous because of varying degrees of *bimodality* in most of the acentric phase probability distributions. The standard solution to this problem is to use centroid structure factors (Blow & Crick, 1959) which minimize the root-mean-square error in the final map and may thus be viewed as a damage-limitation measure.

By contrast the multisolution strategy implemented in *MICE* and *BUSTER* gives access to a much more radical attack on the problem of bimodality.

(1) Resolving the residual phase ambiguity present in a strongly bimodal phase probability density belonging to a large acentric reflexion may be viewed as a *binary choice*; simultaneous choices of modes for several reflexions may be sampled efficiently by the methods of §2.2.2 and used for node expansion within the tree-search mechanism as in §2.2.3.

(2) The ranking of the resulting nodes can be made more effective by using an 'enriched' likelihood function in which the MIR or MAD phase probability densities are used as weights in the integrations of conditional probability distributions over the phases or signs of the second neighbourhood reflexions (§2.1.3). The resulting likelihood measures the extent to which the phases extrapolated from each combination of binary choices of modes in the basis set agree with one of the modes for each second neighbourhood reflexion.

This adaptation of the tree-search mechanism to the systematic resolution of MIR and MAD phase ambiguities should yield a very powerful method indeed. I have derived expressions for the necessary 'enriched' likelihoods, assuming that the phase information available at each stage from the MIR or MAD methods has been encoded by means of the *ABCD* coefficients of Hendrickson & Lattman (1970), and they are being tested by Charles Carter and coworkers. It is worth noting, in passing, that the method proposed by Hendrickson (1971) for producing the best *ABCD* coefficients attached to a phase probability distribution $P(\varphi)$ given numerically, namely a least-squares fit between $\log P(\varphi)$ and $\log P_{ABCD}(\varphi)$ using $P(\varphi)$ as a weighting function, consists precisely of maximizing the relative entropy of $P_{ABCD}(\varphi)$ with respect to $P(\varphi)$.

A last remark concerning the MAD method is in order. Because it exploits only *one* configuration of anomalous scatterers, it is unable to phase reflexions for which the transform of that constellation of scatterers is weak. This gives rise to 'blind regions' in reciprocal space with vanishingly small figures of merit. In the MIR method this problem is mitigated by the fact that several inequivalent configurations of heavy atoms are involved. A corollary to this observation is that, even in the most favourable circumstances, the MAD method would have to rely on some other procedure for propagating phase information from those reflexions where it is strongly indicated to those in the blind region where it is essentially absent. The maximum-entropy extrapolation which accompanies the systematic tree search described above will perform this task in an optimal fashion.

5.4. *The molecular replacement method*

The multichannel formalism provides a means of defining an LLG criterion suitable for the detection of known structural fragments in an unknown crystal (II, §4.1). I have shown (Bricogne, 1992*b*) that the crudest approximation to this LLG yields the Patterson correlation coefficient, which has recently been used with great success by a number of authors (Fujinaga & Read, 1987; Brünger, 1990). This equivalence holds when no phase information is available and when the atoms not belonging to the fragment are assumed to be uniformly distributed in the asymmetric unit. The LLG will be superior to the Patterson correlation coefficient as soon as phase information is available or has been assumed.

The recycling of fragments into the phase determination process can be accomplished by means of conditional distributions of structure factors and of the associated LLG criteria, which can incorporate the exclusion of the remaining random atoms by the fragment.

This statistical formulation also allows a new approach to the fragment detection problem, involving *joint* rotational and translational searches with successive subdivisions of the search grid instead of the conventional *sequential* searches on rotations first, then translations.

If the expected gain in detection sensitivity produced by the use of LLG criteria actually materializes, it may well allow the detection of rather small super-secondary structure motifs, thus offering the possibility of attempting protein structure determination by systematically searching through a library of such motifs derived by a cluster analysis of known structures in the Brookhaven data bank.

5.5. *Non-crystallographic symmetry*

The exploitation of non-crystallographic symmetries (Rossmann & Blow, 1963) by real-space symmetry averaging (Bricogne, 1974, 1976) to assist phase determination has been a substantial advance in macromolecular crystallography. It has been used to solve the structures of a large number of proteins [see Wilson, Skehel & Wiley (1981) for a particularly difficult case], and made it possible to solve the first virus structures to high resolution (Bloomer, Champness, Bricogne, Staden & Klug, 1978; Harrison, Olson, Schutt, Winkler & Bricogne, 1978). Although the method has also succeeded in providing some degree of phase extension in situations of high symmetry (Nordman, 1980; Gaykema, Volbeda & Hol, 1985; Hogle, Chow & Filman, 1985; Arnold, Vriend, Luo, Griffith, Kamer, Erickson, Johnson & Rossmann, 1987), this process is slow and unstable, and does not allow an *ab initio* phase determination.

It was shown in (II, §5) that the real-space treatment of non-crystallographic symmetries given earlier (Bricogne, 1974) could be incorporated into the statistical framework of the saddlepoint method, and thus couple the phase improvement capabilities of the former to the powerful phase extrapolation properties of the latter. The mechanism of this synergy is to be found in the generalized structure-factor algebra (*Appendix* of II) by virtue of which the correlation coefficient between structure factors belonging to reflexions whose orbits under the combination of local and

global symmetries contain a pair of points closer than the spacing between integral lattice points can be arbitrarily close to unity.

The strong modulation of covariances by local symmetries can be used to detect and characterize these symmetries by generalized intensity statistics. I describe in Bricogne (1992*b*) how the rotational parts of these symmetries can be deduced from the distribution of 'spikes' of intensity through reciprocal space, while their translational components are related to the modulation of intensity along the directions of these spikes.

The enrichment of the statistical correlations between structure factors by non-crystallographic symmetries results in conditional distributions for non-basis structure factors with great power of phase extension, enforcing the constraints of non-crystallographic symmetry and solvent flatness *in advance* rather than after the fact as in the iterative procedures used by Nordman (1980), Gaykema *et al.* (1985), Hogle *et al.* (1985) and Arnold *et al.* (1987). The implementation of this combined method may therefore well be the key to the long-awaited *ab initio* determination of virus structures.

Finally it was pointed out in §6 of (II) that this statistical formalism can deal with the occurrence of multiple crystal forms, and could in particular be applied to the exploitation of *non-isomorphous* heavy-atom derivatives for phasing purposes.

5.6. *Model refinement*

I have advocated (Bricogne, 1991*b*, 1992*b*) the use of the LLG as a residual for conducting structure refinement from an atomic model, rather than the residual currently used (the inverse-variance-weighted square of the amplitude difference), as the LLG alone is capable of taking into account the uncertainty on the phases of the calculated structure factors.

This amounts to recommending that the refinement processes be viewed as an instance of LLG maximization, as described in §2.1.3, with a small residual population of random atoms. The maximum-likelihood scaling feature provided by *MLNORM* will be useful in maintaining an appropriate level of discrepancy between calculated and observed amplitudes (thus avoiding bias) when variable numbers of atoms are put in the random atom pool, with strongly non-uniform distributions, for the purpose of calculating omit maps. Similarly, the LLG-gradient maps are proposed as optimal difference maps to replace the usual $2F_o - F_c$ maps.

5.7. *Scope of future applications*

The possibilities of upgrading very substantially each and every aspect of conventional phasing methods through their interface with the multichannel maximum-entropy formalism have been examined separately in the previous sections. Considered together, they give a clear indication that the progressive implementation around the existing *BUSTER* code of the 'Bayesian programme' formulated in (II) should result in a gradual but considerable strengthening of crystallographic methodology as a whole and lead to an increasingly unified, effective and dependable procedure of macromolecular structure determination.

6. Concluding remarks

The specific questions addressed in this work are: (1) can the crystallographic determination of macromolecular structures be turned by new mathematical developments into a routine computational procedure capable of operating from native data alone, as has long been the case with most small molecules? and (2) if routine success cannot be achieved, *e.g.* if exceptional data quality turns out to be an absolute prerequisite, can such computational developments nevertheless yield a substantial increase in the speed and dependability of present methods?

At the time of writing, an affirmative answer to the first question cannot be given. However there are many encouraging signs that the new elements of mathematical technology described here offer considerable fresh potential towards at least a partial solution of the *ab initio* phasing of macromolecules which *BUSTER* has only begun to tap.

As for the second question, it is clear that a stepwise implementation of the Bayesian statistical theory encompassing all phasing procedures formulated in (II) around the central core of optimal mathematical techniques available in *BUSTER* is now possible. This brings most standard macromolecular phasing techniques - particularly isomorphous replacement, molecular replacement and non-crystallographic symmetry averaging - tantalizingly close to radical improvements in their scope and phasing power, with only time and manpower shortage standing in their way.

Both outcomes, in whatever order they occur, will bring about a major enhancement of the overall ability of the X-ray crystallographic machinery to provide essential structural information in fields such as enzymology, immunology, virology, gene expression and cell architecture. But above all it would have a decisive influence in determining whether structural results will be obtained in the future at a sufficient pace to support the full exploitation of the genetic information which will be produced by genome projects during the next two decades.

I wish to thank Dr Chris Gilmore and Professor Charles Carter for innumerable discussions as well as for their staunch commitment to sharing the risks involved in these methodological explorations. I am grateful to IBM UK (Winchester) and to Biostructure SA (Strasbourg) for partial financial support in the course of this work, and to Trinity College, Cambridge, for providing an excellent working environment. I would also like to thank the editor and the referees for their useful and constructive suggestions on ways of presenting some of the material included in this paper.

References

AGMON, N., ALHASSID, Y. & LEVINE, R. D. (1979). *The Maximum Entropy Formalism*, edited by R. D. LEVINE & M. TRIBUS, pp. 207-209. Cambridge: MIT Press.

ALHASSID, Y., AGMON, N. & LEVINE, R. D. (1978). *Chem. Phys. Lett.* **53**, 22-26.

ARNOLD, E., VRIEND, G., LUO, M., GRIFFITH, J. P., KAMER, G., ERICKSON, J. W., JOHNSON, J. E. & ROSSMANN, M. G. (1987). *Acta Cryst.* A**43**, 346-361.

BERTAUT, E. F. (1955*a*). *Acta Cryst.* **8**, 537-543.

BERTAUT, E. F. (1955*b*). *Acta Cryst.* **8**, 544-548.

BERTAUT, E. F. (1955*c*). *Acta Cryst.* **8**, 823-832.

BERTAUT, E. F. (1956*a*). *Acta Cryst.* **9**, 322.

BERTAUT, E. F. (1956*b*). *Acta Cryst.* **9**, 769-770.

BERTAUT, E. F. (1959*a*). *Acta Cryst.* **12**, 541-549.

BERTAUT, E. F. (1959*b*). *Acta Cryst.* **12**, 570-574.

BERTAUT, E. F. & DULAC, J. (1956). *Acta Cryst.* **9**, 322-323.

BERTAUT, E. F. & WASER, J. (1957). *Acta Cryst.* **10**, 606-607.

BLOOMER, A. C., CHAMPNESS, J. N., BRICOGNE, G., STADEN, R. & KLUG, A. (1978). *Nature (London)*, **276**, 362-368.

BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794-802.

BLOW, D. M. & MATTHEWS, B. W. (1973). *Acta Cryst.* A**29**, 56-62.

BRAGG, W. L. & LIPSON, H. (1936). *Z. Kristallogr.* **95**, 323-337.

BRICOGNE, G. (1974). *Acta Cryst.* A**30**, 395-405.

BRICOGNE, G. (1976). *Acta Cryst.* A**32**, 832-847.

BRICOGNE, G. (1982). *Computational Crystallography*, edited by D. SAYRE, pp. 258-264. Oxford: Clarendon Press.

BRICOGNE, G. (1984*a*). *Acta Cryst.* A**40**, 410-445.

BRICOGNE, G. (1984*b*). *Methods and Applications in Crystallographic Computing*, edited by S. R. HALL & T. ASHIDA, pp. 141-151. Oxford: Clarendon Press.

BRICOGNE, G. (1988*a*). *Acta Cryst.* A**44**, 517-545.

BRICOGNE, G. (1988*b*). *Crystallographic Computing* 4, edited by N. W. ISAACS & M. R. TAYLOR, pp. 60-79. Oxford: Clarendon Press.

BRICOGNE, G. (1991*a*). *Maximum Entropy in Action*, edited by B. BUCK & V. A. MACAULAY, pp. 187-216. Oxford Univ. Press.

BRICOGNE, G. (1991*b*). *Acta Cryst.* A**47**, 803-829.

BRICOGNE, G. (1991*c*). *Crystallographic Computing* 5, edited by D. MORAS, A. D. PODJARNY & J. C. THIERRY, pp. 257-297. Oxford: Clarendon Press.

BRICOGNE, G. (1991*d*). *Mathematics and Molecular Biology* II: *DNA Sequence to Protein Structure*. Santa Fe, New Mexico, USA, 24-29 March 1991. Abstract S2.

BRICOGNE, G. (1991*e*). *Second European Workshop on the Crystallography of Biological Macromolecules*. Como, Italy, 13-16 May 1991. Abstract 25.

BRICOGNE, G. (1991*f*). *Isomorphous Replacement and Anomalous Scattering. Proceedings of the CCP4 Study Weekend 25-26 January* 1991, edited by W. WOLF, P. R. EVANS & A. G. W. LESLIE, pp. 60-68. Warrington: SERC Daresbury Laboratory.

BRICOGNE, G. (1992*a*). *Fourier Transforms in Crystallography: Theory, Algorithms and Applications*. In *International Tables for Crystallography*, Vol. B, edited by U. SHMUELI, pp. 23-106. Dordrecht: Kluwer Academic Publishers.

BRICOGNE, G. (1992*b*). *The Molecular Replacement Method. Proceedings of the CCP4 Study Weekend 31 January-1 February* 1992, edited by W. WOLF, E. J. DODSON & F. GOVER. Warrington: SERC Daresbury Laboratory.

BRICOGNE, G. & GILMORE, C. J. (1990). *Acta Cryst.* A**46**, 284-297.

BRÜNGER, A. T. (1990). *Acta Cryst.* A**46**, 46-57.

CARTER, C. W. JR, CRUMLEY, K. V., COLEMAN, D. E., HAGE, F. & BRICOGNE, G. (1990). *Acta Cryst.* A**46**, 57-68.

COCHRAN, W. G. & COX, G. M. (1957). *Experimental Designs*, 2nd ed. New York: John Wiley.

COCHRAN, W. & DOUGLAS, A. S. (1953). *Nature (London)*, **171**, 1112-1113.

COCHRAN, W. & DOUGLAS, A. S. (1955). *Proc. R. Soc. London Ser. A*, **227**, 486-500.

COCHRAN, W. & DOUGLAS, A. S. (1957). *Proc. R. Soc. London Ser. A*, **243**, 281-288.

COLLINS, D. M. & PRINCE, E. (1991). *Crystallographic Computing* 5, edited by D. MORAS, A. D. PODJARNY & J. C. THIERRY, pp. 308-316. Oxford: Clarendon Press.

CONWAY, J. H. & SLOANE, N. J. A. (1988). *Sphere Packings, Lattices and Groups*. New York: Springer-Verlag.

CROWFOOT, D., BUNN, C. W., ROGERS-LOW, B. W. & TURNER-JONES, A. (1949). *The X-ray Crystallographic Investigation of the Structure of Penicillin*. Oxford Univ. Press.

DONG, W., BAIRD, T., FRYER, J. R., GILMORE, C. J., MACNICOL, D. D., BRICOGNE, G., SMITH, D. J., O'KEEFE, M. A. & HOVMOLLER, S. (1992). *Nature (London)*, **355**, 605-609.

FUJINAGA, M. & READ, R. J. (1987). *J. Appl. Cryst.* **20**, 517-521.

GAYKEMA, W. P. J., VOLBEDA, A. & HOL, W. G. J. (1985). *J. Mol. Biol.* **187**, 225-275.

GILMORE, C. J., BRICOGNE, G. & BANNISTER, C. (1990). *Acta Cryst.* A**46**, 297-308.

GILMORE, C. J., HENDERSON, A. N. & BRICOGNE, G. (1991). *Acta Cryst.* A**47**, 842-846.

GILMORE, C. J., HENDERSON, K. & BRICOGNE, G. (1991). *Acta Cryst.* A**47**, 830-841.

GOEDKOOP, J. A. (1950). *Acta Cryst.* **3**, 374-378.

GOOD, I. J. (1954). *Acta Cryst.* **7**, 603-604.

HARRISON, S. C., OLSON, A. J., SCHUTT, C. E., WINKLER, F. K. & BRICOGNE, G. (1978). *Nature (London)*, **276**, 368-373.

HENDRICKSON, W. A. (1971). *Acta Cryst.* B**27**, 1472-1473.

HENDRICKSON, W. A. & LATTMAN, E. E. (1970). *Acta Cryst.* B**26**, 136-143.

HOGLE, J. M., CHOW, M. & FILMAN, D. J. (1985). *Science*, **229**, 1358-1365.

IWASAKI, H. & ITO, T. (1977). *Acta Cryst.* A**33**, 227-229.

LESLIE, A. (1987). *Acta Cryst.* A**43**, 134-136.

LIPSON, H. & COCHRAN, W. (1968). *The Determination of Crystal Structures*, 3rd ed. London: Bell.

LONSDALE, K. (1929). *Proc. R. Soc. London Ser. A*, **123**, 494.

MACWILLIAMS, F. J. & SLOANE, N. J. A. (1977). *The Theory of Error-Correcting Codes*. Amsterdam: North-Holland.

MAIN, P. (1977). *Acta Cryst.* A**33**, 750-757.

MAIN, P. (1978). *Acta Cryst.* A**34**, 31-38.

NEYMAN, J. & PEARSON, E. (1933). *Philos. Trans. R. Soc. London Ser. A*, **231**, 289-337.

NORDMAN, C. E. (1980). *Acta Cryst.* A**36**, 747-754.

PÄHLER, A., SMITH, J. L. & HENDRICKSON, W. A. (1990). *Acta Cryst.* A**46**, 537-540.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. & VETTERLING, W. T. (1986). *Numerical Recipes*. Cambridge Univ. Press.

PRINCE, E. (1989). *Acta Cryst.* A**45**, 200-203.

RICE, S. O. (1944). *Bell Syst. Tech. J.* **23**, 283-332 (parts I and II); (1945). **24**, 46-156 (parts III and IV). Reprinted in (1954). *Selected Papers on Noise and Stochastic Processes*, edited by N. WAX, pp. 133-294. New York: Dover Publications.

RICHE, C. (1970). *C. R. Acad. Sci.* **271**, 396-398.

RICHE, C. (1973). *Acta Cryst.* A**29**, 133-137.

ROBERTSON, J. M. & WHITE, J. G. (1945). *J. Chem. Soc.* pp. 607-617.

ROSSMANN, M. G. & BLOW, D. M. (1963). *Acta Cryst.* **16**, 39-45.

SAYRE, D. (1982). Editor. *Computational Crystallography*. New York: Oxford Univ. Press.

SCALES, L. E. (1985). *Introduction to Non-Linear Optimization*. London: Macmillan.

SCHEVITZ, R. W., PODJARNY, A. D., ZWICK, M., HUGHES, J. J. & SIGLER, P. B. (1981). *Acta Cryst.* A**37**, 669-677.

SHANKLAND, K., GILMORE, C. J., BRICOGNE, G. & HASHIZUME, H. (1992). *Acta Cryst.* A**48**. Submitted.

SHANNON, C. E. & WEAVER, W. (1949). *The Mathematical Theory of Communication*. Urbana: Univ. of Illinois Press.

SHELDRICK, G. M. (1990). *Acta Cryst.* A**46**, 467-473.

SJÖLIN, L., PRINCE, E., SVENSSON, L. A. & GILLILAND, G. L. (1991). *Acta Cryst.* A**47**, 216-223.

STEWART, J. M. & KARLE, J. (1976). *Acta Cryst.* A**32**, 1005-1007.

STEWART, J. M., KARLE, J., IWASAKI, H. & ITO, T. (1977). *Acta Cryst.* A**33**, 519.

TSOUCARIS, G. (1970). *Acta Cryst.* A**26**, 492-499.

VAND, V. & PEPINSKY, R. (1956). *Z. Kristallogr.* **107**, 202-224.
WANG, B. C. (1985). In *Methods of Enzymology*, Vol. 115, *Diffraction Methods for Biological Macromolecules*, edited by H. WYCKOFF, C. W. HIRS & S. N. TIMASHEFF. New York: Academic Press.
WHITE, P. S. & WOOLFSON, M. M. (1975). *Acta Cryst.* A**31**, 53–56.
WILSON, I. A., SKEHEL, J. J. & WILEY, D. C. (1981). *Nature (London)*, **289**, 366-373.
WOOLFSON, M. M. (1954). *Acta Cryst.* **7**, 65-67.
WOOLFSON, M. M. & YAO, J. (1990). *Acta Cryst.* A**46**, 409-413.
XIANG, S., CARTER, C. W. JR, BRICOGNE, G. & GILMORE, C. J. (1993). *Acta Cryst.* D**49**, 193-212.