Acta Cryst. (1999). A**55**, 70–83

A multisolution method of phase determination by combined maximization of entropy and likelihood. VI. The use of error-correcting codes as a source of phase permutation and their application to the phase problem in powder, electron and macromolecular crystallography

Christopher Gilmore,[a]* Wei Dong[a] and Gérard Bricogne[b,c]

[a]*Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland,* [b]*MRC Laboratory for Molecular Biology, Hills Road, Cambridge CB2 2QH, England, and* [c]*LURE, Bâtiment 209D, Orsay 91405 CEDEX, France. E-mail: chris@chem.gla.ac.uk*

(*Received 19 March 1998; accepted 4 June 1998*)

Abstract

The use of error-correcting codes as a source of efficient designs of phase permutation schemes is described. Three codes are used, all taken from the Bricogne *BUSTER* program [Bricogne (1993). *Acta Cryst.* D**49**, 37–60]: the Hamming [7, 4, 3], the Nordström–Robinson (16, 256, 6) and the Golay [24, 12, 8] or its punctured [23, 12, 7] form. These are used in a maximum-entropy–likelihood phasing environment to carry out phase permutation of basis-set reflections instead of the usual quadrant permutation or magic integer approaches. The use of codes in this way inevitably introduces some errors in the phase choices, but for most structures this is not significant especially when the gain in sampling efficiency is considered. For example, the Golay [24, 14, 8] allows the permutation of 24 centric phases in such a way that only 4096 phase sets are produced instead of $2^{24} = 16\,777\,216$, and one of these sets has, at most, only four wrong phases. The method is successfully applied to three powder diffraction data sets of increasing complexity, and with increasing degrees of overlap {Mg_3BN_3, Sigma-2 ($[Si_{64}O_{128}]\cdot 4C_{10}H_{17}N$) and the NU-3 zeolite}, a sparse electron diffraction data set for buckminsterfullerene, C_{60}, and the small protein molecule crambin at 3 Å resolution where 42 reflections are phased with a *U*-weighted mean phase error of 58.5°.

1. An introduction to error-correcting codes

We present here an implementation of the use of error-correcting codes as a source of efficient phase permutation in a maximum-entropy (ME) *ab initio* phasing environment, first proposed and used by Bricogne (1993, 1997*b*). Error-correcting codes (e.c.c.'s) are an integral part of late 20th century digital communications. Starting with the classic work of Shannon (1948*a*,*b*), Golay (1949) and Hamming (1947), they are now used everywhere, for example, in CD-ROM devices, in digital telephones and in pictures transmitted from space. From these examples, it may seem strange to see the method used with respect to the crystallographic phase problem, but there is a link between certain e.c.c.'s and experimental designs (Bricogne, 1993, 1997*b*) which can be used as a source of efficient phase-permutation procedures not unlike the use of magic integers (White & Woolfson, 1975; Main, 1977, 1978). To understand how this can be performed, we first need to explore the basic definitions of coding theory. The literature on the subject is vast, but a good introduction may be found by Hill (1993), and the classic book is by MacWilliams & Sloane (1977); there is also an interesting and partly historical introduction by Thompson (1983).

A q-ary code comprises a set of sequences of symbols where each symbol is chosen from a set $F_q = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ of q elements. The set F_q is called the alphabet, and $(F_q)^n$ is the set of all ordered n-tuples of these symbols $a = a_1 a_2 \ldots a_n$, where $a_j \in F_q$. A q-ary code of length n is, therefore, generated as a subset of $(F_q)^n$, *i.e.* $F_q \subseteq (F_q)^n$. A codeword comprises a sequence of symbols, and the set of M codewords defines the code, C. Throughout this paper, we will be concerned only with binary codes; in this case, the alphabet is the binary digits 0, 1, *i.e.* $F_2 = \{0, 1\}$. Binary codes are by far the most important e.c.c.'s, but many others, *e.g.* ternary and quaternary, exist and could be useful as a future source of experimental design in crystallography.

We now need the concept of the minimum distance, d, between the codewords. The (Hamming) distance between two codewords in C is the number of places in which they differ. The minimum distance is the smallest of these distances. Denote three codewords of any code as the vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$; then the Hamming distance (Hamming, 1947) between $\mathbf{x}$ and $\mathbf{y}$ is $d(\mathbf{x}, \mathbf{y})$, and is a legitimate distance function since

(i) $d(\mathbf{x}, \mathbf{y}) = 0$	iff $\mathbf{x} = \mathbf{y}$,
(ii) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$	$\forall\ \mathbf{x}, \mathbf{y} \in (F_q)^n$,
(iii) $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$	$\forall\ \mathbf{x}, \mathbf{y}, \mathbf{z}, \in (F_q)^n$.

A code comprising M codewords of length n and minimum distance d is described as an (n, M, d) code.

ISSN 0108-7673

An e.c.c. can detect up to $d-1$ errors and correct up to $(d-1)/2$ errors in any codeword by assigning the closest codeword (*i.e.* that with the minimum Hamming distance) to that received, and it is this property that makes codes so important in the transmission of signals. The design of an efficient code consists of using the greatest number of codewords for a given n while maximizing d, and is a central problem in coding theory.

As an example, consider the simple binary code {0000, 1100, 0110, 0011, 1001}. There are five codewords of length four characters. The Hamming distances vary between 2 and 4, so that this is a (5, 4, 2) code. It can detect up to one error, but cannot correct any errors.

An important class of codes is the linear codes which are denoted [*n, k, d*]. In this instance, n and d are defined as before, but k defines the number of rows in a generator matrix. The latter is used to generate the codewords by 2^k linear combinations of the words in the matrix under modulo 2 arithmetic in the binary case.

Codes can also be punctured by deleting the last column of the generator matrix for linear codes or the last bit of every codeword in the non-linear case. For example, the (5, 4, 2) code above would puncture to give the code comprising {000, 110, 011, 001, 100}, *i.e.* a (4, 4, 1) code.

1.1. *Error-correcting codes, experimental designs and combinatorics*

The relationship between certain e.c.c.'s and experimental designs is well documented (see, for example, Anderson, 1989, chs. 6 and 7). The classic example of a relationship between a simple experimental design and an error-correcting code is the following example taken from Anderson (1989).

Consider an experiment to assess a new brand of coffee by comparing it with six existing varieties. Define a set $S = \{1, 2, 3, \ldots, 7\}$ whose members are the seven coffee varieties suitably labelled 1–7, one of which is the new brand. To carry out the tasting experiment, a number of crystallographers are chosen to decide the relative merits of these varieties. Not every crystallographer is to taste all seven varieties, however, since it is harder to choose between seven than, say, three varieties, and the experiment is more time consuming. A block design is therefore constructed to reduce the number of trials and yet still allow the necessary information to be measured and assessed. The trials are arranged such that each crystallographer tastes the same number of brands, and each pair of brands is compared by the same number of crystallographers. To do this, blocks are selected which are subsets of S such that:

(i) every block has the same number of elements;

(ii) every pair of varieties is contained in the same number of blocks.

Let us choose the following seven blocks as follows:

$$\{1, 2,4\}, \{2, 3, 5\}, \{3, 4, 6\}, \{4, 5, 7\}, \{5, 6, 1\}, \{6, 7, 2\}, \{7, 1, 3\}. \quad (1)$$

We can see that each crystallographer tastes three coffees, and no-one tastes the same three, and each variety is tested three times. This is a balanced block design of the type (7, 7, 3, 3, 1), *i.e.* there are seven varieties, seven subsets, three elements in each subset, each element appears three times in all the blocks, and each pair of varieties appears only once. The selection of the coffee varieties can thus be seen as a problem in combinatorics, and e.c.c.'s can play a major role here also. The combination of coffee varieties and testers can be represented in binary form using the concept of the incidence matrix, M, of the design

$$M = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}. \quad (2)$$

Each row represents a block and each column represents a variety. The correspondence between this and (1) should be obvious. Now construct a ones complement matrix, M', in which all the zeros in M are replaced by ones and the ones by zeros. Construct a third matrix, N, containing M, M' and with two new rows and a parity bit in the first column,

$$N = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & & & & & & & \\ 1 & & & & & & & \\ 1 & & & & & & & \\ 1 & & & & & & & \\ 1 & & & & M & & & \\ 1 & & & & & & & \\ 1 & & & & & & & \\ 0 & & & & & & & \\ 0 & & & & & & & \\ 0 & & & & & & & \\ 0 & & & & M' & & & \\ 0 & & & & & & & \\ 0 & & & & & & & \\ 0 & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \quad (3)$$

The rows of N define a binary error-correcting (8, 16, 4) code with 16 codewords each containing eight bits. It can detect up to three errors and correct one. Note that only balanced block designs of the type (b, b, r, r, λ) can make e.c.c.'s of this type, but this simple example does show, at least in outline, how codes, designs and combinatorics are interrelated.

2. Error-correcting codes and the crystallographic phase problem

Direct methods of solving the phase problem use the central idea of phase permutation in which a subset $\{H\}$ of strong normalized reflections well linked *via* triplet and possibly quartet relationships are given permuted phases, and input into either the tangent formula, a minimal function (DeTitta *et al.*, 1994), maximum entropy (Bricogne, 1984), the Sayre equation in its many variants (Sayre, 1952) or another phase expansion/refinement formula. Four general types of phase permutation can be identified:

(i) Quadrant permutation or full factorial design. Each of the n_c centric reflections is given both its possible values, *e.g.* 0, π or $\pm\pi/2$, and each of the n_a acentric reflections is assigned a quadrant by assigning the possible values $\pm\pi/4$, $\pm 3\pi/4$. This is a full factorial design generating $2^{n_c}4^{n_a} = 2^{n_c+2n_a}$ phase choices. The total number of degrees of freedom (d.o.f.) is defined as $n_c + 2n_a$, and it can be seen that this soon becomes a combinatorial explosion, *e.g.* permuting the phases of seven acentric reflections would give 14 degrees of freedom and 16384 phase sets.

(ii) Magic integers. These were first used by White & Woolfson (1975) and later refined into a form quite closely related to e.c.c.'s by Main (1977, 1978). Magic integers have some of the properties of codes in the way that they cover phase space, and are an efficient source of phase permutation, and the relationship between magic integers and e.c.c.'s has been explored by Bricogne (1993, 1997*b*). The gain in efficiency compared with the full factorial design is considerable: for example, using the magic integers based on the Fibonacci series would reduce the 16384 phase sets in (i) to 128 with a r.m.s. phase error of 48.0°.

(iii) Random phases. Each of the reflections in the set is given a random phase, and one relies on the power of the phase refinement and extension formulae to refine some of these phase sets such that the mean phase error is sufficiently low to recognize the structure in the associate Fourier map. This is now the technique of preference for large sets of reflections.

(iv) Error-correcting codes (Bricogne, 1993, 1997*b*). Most eccs are unsuitable for this purpose; they need to contain a suitable experimental design that balances both the main reflection phases and the interactions between them, as well as covering the phase space with optimum efficiency. Selecting suitable candidates is a non-trivial task, but those listed below have suitable properties and were first employed in the *BUSTER* computer program (Bricogne, 1993):

(*a*) The Hadamard [8, 4, 4] code or the punctured form which is the Hamming [7, 4, 3]. The former generates 16 phase sets instead of 256 for eight degrees of freedom; one of these will have, at most, two wrong phases (you can, of course, be lucky and have no incorrect phase choices!), whilst the latter gives rise to 128 permutations, and one of these has, at worst, only one wrong phase. The [8, 4, 4] code is also known as the Reed–Muller RM(1, 3) code. The Hamming [7, 4, 3] code appeared indirectly in a phasing environment (it was not named as such) when used by Woolfson (1954) to permute the signs of seven centric reflections in the *ab initio* phase determination of a small organic molecule, fluorene. However, he provided no link of his method with coding theory or e.c.c.'s: the design was presented as an isolated discovery.

(*b*) The Hadamard [16, 11, 4] code or the punctured Hamming [15, 11, 3] form. The former generates 2048 phase sets instead of 65536 for 16 degrees of freedom; one of these will have, at most, two wrong phases. The latter describes 2048 sets instead of 32768, and one set will have, at most, one wrong phase choice. The [16, 11, 4] code can also be described as an RM(2, 4) code. Good (1954) used the [15, 11, 3] form rather like Woolfson (1954) and called the method *substantialization* since one phase choice out of the 2048 was substantially correct. He proved the assertions concerning the minimum number of incorrect phase choices, but again there was no link provided to existing work outside crystallography involving e.c.c.'s or experimental designs. We are not presenting results from this code in this paper since the Golay code [see (*d*) below] provides 23 or 24 d.o.f. with only twice as many phase sets and is therefore much more efficient.

(*c*) The Nordström–Robinson (16, 256, 6) code or the punctured (15, 256, 5) form producing 256 sets instead of $2^{16} = 65536$ or $2^{15} = 32768$ for 15 or 16 d.o.f. One of these will have a maximum of four (for 16 d.o.f.) or three (for 15 d.o.f.) incorrect phase choices.

(*d*) The Golay [24, 12, 8] code or the punctured [23, 12, 7] version producing 4096 phase combinations instead of $2^{24} = 16\,777\,216$ or $2^{23} = 8388608$. One of these will have a maximum of four incorrect phase choices for 24 d.o.f., or three for 23 d.o.f. The gain in efficiency here is quite exceptional, and the code links to the extraordinary Leech lattice and the packing of spheres in 24 dimensions in which each sphere has a contact or kissing number of 196560, and is the densest packing known in any dimension.

To use codes for phase permutation is straightforward:

(i) For centric phases, the binary digit 0 represents one possible choice, and 1 represents the alternative, *e.g.* for a phase restricted to 0 or π, 0 represents a 0° phase angle and 1 represents an angle of π.

(ii) In the acentric situation, two bits are used to assign the quadrant of the phase; one bit describes the sign of the real part of the phase and the second bit describes the imaginary part, *i.e.* $0, 0 = \pi/4$; $1, 0 = 3\pi/4$; $1, 1 = 5\pi/4$; $0, 1 = 7\pi/4$.

3. Incomplete factorial designs

Another source of experimental designs that are suitable for phase permutation are incomplete factorial designs (i.f.d.'s). These have been used extensively by Carter (see, for example, Carter, 1992, 1997) in the design of experiments for screening crystal growth conditions where initial trials do not give useful results. The variables represented by the i.f.d.'s include temperature, pH, ligands *etc.* I.f.d.'s are represented in binary form and can be used like codes in a phasing environment, and several have been constructed by Carter for this purpose, notably, but not exclusively:

(i) Permuting 10 d.o.f. giving 48 phase sets instead of 1024.

(ii) Permuting 12 d.o.f. giving 64 sets instead of 4096.

(iii) Permuting 14 d.o.f. giving 100 nodes instead of 16384.

Although i.f.d.'s may not have all the properties of codes, they can provide efficient designs in the cases that are intermediate between the Hadamard, Hamming and Nordström–Robinson codes and have a high level of efficiency of covering in phase space. They were successfully used in the solution of the tryptophanyl-tRNA synthetase structure using maximum-entropy–likelihood methods (Doublié *et al.*, 1994, 1995). In this case, the unknown phases were those of strong reflections for which the MIR method was not sufficiently reliable, and i.f.d.'s were used to permute subsets of them. We are currently investigating the general uses of i.f.d.'s in a maximum-entropy phasing environment; these results will be reported elsewhere and will not be discussed further.

4. Codes, sphere packings and coverings

We are concerned in this paper with efficient designs for phase permutation. Another way of looking at this is *via* the construction of sphere packings in n-dimension Euclidean space $\mathbf{R}^n$ [see Sloane (1984), for an introduction to this method, and Conway & Sloane (1988), for more detail]. This is discussed in detail by Bricogne (1993, 1997*b*) and, in consequence, only a brief summary will be presented here. A coordinate array for the centres of the spheres is obtained from the n-dimensional $[n, k, d]$ or (n, M, d) code and used as the starting point of the construction for both lattice and non-lattice packings (Conway & Sloane, 1988, ch. 5). At a heuristic level, the packing problem can be seen to be relevant to an efficient phase permutation since we are trying in the latter to sample the phase space as efficiently as possible and this can be visualized as a multidimensional packing problem by invoking the multidimensional Shannon sampling criterion (Bricogne, 1993, Section 2.2.2.2). In this case, we wish to pack Hamming spheres of radius r, where $r = \mathrm{Int}[(d-1)/2]$, as densely as possible in the unit cell of a binary n-dimensional lattice.

An alternative way of viewing the search for optimal codes is to consider the related problem of finding e.c.c.'s with good covering properties, *i.e.* codes with codewords having the property that no codeword is too far from its nearest codeword. The covering radius R of a code is the maximal distance of any n-dimensional vector from the code. Codes specifically designed for optimum covering, *i.e.* small R, are called covering codes (Graham & Sloane, 1985), and it is clear that an efficient code for phase permutation should have a small value of R so that it can be guaranteed that the phase space is effectively sampled without any major voids. All the codes used in this paper have low covering radii. Self-dual codes have the property of both optimum packing and covering, but in general these properties are complementary. The Golay code is self-dual.

5. The maximum-entropy method and the need for experimental designs

In previous papers in this series (Bricogne & Gilmore, 1990; Gilmore, K. Henderson & Bricogne, 1991; Gilmore, A. N. Henderson & Bricogne, 1991), we have described the application of the maximum-entropy (ME) method as a tool for solving crystal structures where the data sets are incomplete, or of < 1.2 Å resolution and/or subject to major measurement errors. We have concentrated especially, therefore, on powder, electron diffraction and macromolecular diffraction data sets. In all these cases, the method works as follows (see, for example, Bricogne, 1984; Gilmore, 1996, Bricogne, 1993, 1997*a*):

(i) The structure factors, $|F_{\mathbf{h}}|^{\mathrm{obs}}$ are normalized to give unitary structure factors $|U_{\mathbf{h}}|^{\mathrm{obs}}$.

(ii) A basis set $\{H\}$ is defined for those reflections whose phase are known either from the rules of origin and enantiomorph definition, from electron micrographs suitably Fourier transformed or from multiple isomorphous replacement/single isomorphous replacement (MIR/SIR) experiments. The disjoint set of unphased reflections is $\{K\}$. The basis set defines the root node of a phasing tree, and the amplitudes and phases of its members are used as the constraints in an entropy maximization.

(iii) Some strong unphased reflections, which optimally enlarge the second neighbourhood of $\{H\}$, $|U_{\mathbf{h}\in K}|^{\mathrm{obs}}$, are given permuted phase values to build the second level of a phasing tree. Each possible phase choice defines a node, each of which is subjected to constrained entropy maximization. Initially, the permutation methods we used were either those of magic integers or quadrant permutation in a full factorial design. Rice-type likelihood functions are used to gauge the correctness of each phase choice; the likelihoods are analysed using simple t-test procedures (Shankland *et al.*, 1993). The best 8–16 nodes are retained and the rest discarded.

(iv) The permutation–likelihood analysis procedure continues until a possible structure emerges or until likelihoods begin to fall, and the structures are completed, if necessary, in the usual way or *via* a ME recycling procedure.

This calculation is carried out using the *MICE* computer program (Gilmore & Bricogne, 1997; Gilmore & Nicholson, 1994). However, it has an obvious computational problem in that if we have n nodes at a given level and m degrees of freedom for the reflections whose phases are being permuted, then any full factorial permutation will generate $n2^m$ phase sets, all of which need to be subjected to constrained entropy maximization. Not only is this a large computational problem, it can also be difficult to interpret the results. It should now be obvious that e.c.c.'s can be used to reduce the scale of this problem dramatically making accessible problems that were hitherto impracticable on the grounds of computing requirements. To do this, the Hadamard, Hamming, Nordström–Robinson and Golay codes described in the previous section have been incorporated as options into *MICE*. They came initially from the Bricogne *BUSTER* program (Bricogne, 1993), but are now used in the form of a code/i.f.d. database.

6. Analysis of the results of phase permutation using e.c.c.'s

In the ME procedure, we assess the likely correctness of a given set of phases using a suitable likelihood function which measures how well we predict the unphased reflection amplitudes from the phased ones. When e.c.c.'s are used, there is also a second factor to consider: the problem of phase errors. In a full factorial design, all phase choices are spanned and one node must have predominantly correct phases depending on how finely the acentric reflection phases are sampled in the range 0–2π, whereas with e.c.c.'s at least one phase must be incorrect, and with the Golay or Nordström–Robinson codes at least three will be wrong by π. We then have a choice: to continue with some wrong phases or to try to reconstruct the full phase space from the sample provided by the code. Bricogne (1993, 1997*b*) has developed a multidimensional Fourier method for the latter, but here we will not attempt this; instead we will work with the incorrectly phased reflections in the expectation that correct phases will predominate. To do this, we need to analyse the log-likelihood gains (LLGs) with care, and the following algorithm has proved successful (Gilmore *et al.*, 1997).

The likelihood evaluation gives rise to a log-likelihood gain into which the concept of a null hypothesis is incorporated (Bricogne & Gilmore, 1990). Since we have no estimates of the variances of the LLGs associated with each node, we cannot simply keep those nodes with the highest values. Instead, tests of significance are used (Shankland *et al.*, 1993; Bricogne, 1993, 1997*b*) in which the LLGs are analysed for phase indications using the Student t test. The simplest example involves the detection of the main effect associated with the sign of a single centric phase. The LLG average, μ^+, and its associated variance, V^+, is computed for those nodes in a given level in the phasing tree in which the sign of this permuted phase under test is '+'. The calculation is then repeated for those sets in which the same sign is '−' to give the corresponding μ^-, and variance V^-. The t statistic is then

$$t = |\mu^+ - \mu^-|/(V^+ + V^-)^{1/2}. \quad (4)$$

The use of the t test enables a sign choice to be derived with an associated significance level s. The calculation is repeated for all the single-phase indications, and is then extended to combinations of two and three phases depending on the code that is used. This problem is discussed in detail by Bricogne (1997*b*, section 2.2). In fact, aliasing and confounding are always problems that need to be addressed when dealing with experimental designs – predictably enough, you cannot get something for nothing (see, for example, Cochran & Cox, 1957, ch. 6*A*). In general, if a code can correct t errors it can be used to retrieve phase interactions up to and including order t without aliasing. Thus the Hamming [7, 4, 3] and Hadamard [8, 4, 4] codes can retrieve only main effects. To see why this is, consider the [8, 4, 4] code as a source of phase permutation for eight centric phases, labelled $\varphi_1 - \varphi_8$. This is shown in Table 1. There are 16 codewords and hence 16 phase sets. To extract the two-phase interaction $\varphi_1 + \varphi_2$, we add columns 1 and 2 modulo 2 to obtain (written as a row rather than a column to save space)

$$1 + 2 = 0011001100110011.$$

The two-phase interaction $\varphi_7 + \varphi_8$ gives exactly the same result, as does $\varphi_3 + \varphi_4$. These three sets of interactions, $\varphi_1 + \varphi_2$, $\varphi_3 + \varphi_4$, $\varphi_7 + \varphi_8$, are aliases of each other and cannot be extracted independently. These interactions are said to be confounded.

The Nordström–Robinson code can retrieve main and two-phase interactions which adds considerably to its power, whilst the Golay code can also include three-phase interactions as well.

When using full factorial designs with good data, only relationships with associated significance levels less than 2% are used in the analysis, but this can be relaxed with sparse diffraction data sets, and when codes are used. Each of the m phase relationships, i, so generated is given an associated weight,

$$w_i = 1 - I_1(s_i)/I_0(s_i), \quad (5)$$

where I_1 and I_0 are the appropriate Bessel functions and s_i is the significance level of the ith relationship from the t test. This weighting function reflects the need for a scheme in which the absolute values of the significance levels are not given undue emphasis since they are

Table 1. *The Hadamard [8, 4, 4] code used as a source of phase permutation for eight centric reflections labelled* φ_1–φ_8

	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7	φ_8
1	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1
3	1	0	1	0	1	0	1	0
4	0	1	0	1	0	1	0	1
5	1	1	0	0	1	1	0	0
6	0	0	1	1	0	0	1	1
7	1	0	0	1	1	0	0	1
8	0	1	1	0	0	1	1	0
9	1	1	1	1	0	0	0	0
10	0	0	0	0	1	1	1	1
11	1	0	1	0	0	1	0	1
12	0	1	0	1	1	0	1	0
13	1	1	0	0	0	0	1	1
14	0	0	1	1	1	1	0	0
15	1	0	0	1	0	1	1	0
16	0	1	1	0	1	0	0	1

themselves subject to errors arising from the nature of the likelihood function used and the lack of error estimates for the LLGs themselves.

Each node n is now given an overall score, s_n,

$$s_n = \mathrm{LLG}_n \sum_{j=1}^{m} w_j, \qquad (6)$$

where the summation spans only those phase relationships where there is agreement between the basis-set phases and the t test derived phase relationships. The scores are sorted and only the top 8–16 nodes are kept; the rest are discarded. New reflections are now permuted and a corresponding new set of ME solutions is generated, and this continues until a recognizable structure or structural fragment appears.

7. The codes in action

In order to demonstrate the versatility of the method, we now describe the use of e.c.c.'s to three powder data sets of varying degrees of overlap, sparse electron diffraction data for C_{60} and a small protein, crambin, at 3 Å resolution using the codes described in §2. These are all structures that have been solved by other methods; this enables us to calculate the phase errors in the phase sets that are generated by the e.c.c.'s. However, the method has also been successfully applied to several unknown structures, and these are discussed briefly in §8.

7.1. *The Hamming [7, 4, 3] code and Mg_3BN_3 powder data*

The ME method has proved to be a powerful tool in the solution of structures from powder diffraction data sets. Bricogne (1991) extended the general theory of the ME–likelihood formalism to encompass overlapped intensity data, and the method was subsequently applied to the Sigma-2 clathrasil and $KAlP_2O_7$ (Gilmore, K. Henderson & Bricogne, 1991), Mg_3BN_3 (Shankland *et al.*, 1993), and the hitherto unsolved structures of $LiCF_3SO_3$ (Tremayne, Lightfoot, Glidewell *et al.*, 1992) and formylurea (Tremayne, Lightfoot, Harris *et al.*, 1992). Dong & Gilmore (1998) have also used likelihood in conjunction with modified t tests to extract accurate intensities of overlapped reflections in favourable cases as well as their associated phases. However, codes are an obvious extension of these ideas, and this paper contains three examples, of increasing complexity, in which the use of codes has led to the routine solution of crystal structure from powder data.

The first of these is the redetermination of the structure of magnesium boron nitride, Mg_3BN_3. This crystallizes in space group $P6_3/mmc$ with $a = 3.54453$, $c = 16.0353$ Å and $Z = 2$. There are 69 reflections in the data set including two overlap sets, each of which contains two reflections. The data resolution is 0.9 Å (Hiraguchi *et al.*, 1991). This is a small structure which is easy to solve, but it provides the perfect vehicle for demonstrating the simplest of the e.c.c.'s we have described: the Hamming [7, 4, 3] code. The first level of the phasing tree was defined *via* the 107 reflection which fixed the origin. The second level of 16 nodes (instead of 128) was generated by permuting the $2\bar{1}0$, $2\bar{1}8$, $3\bar{1}4$, $3\bar{1}7$, 104, 1,0,12 and 300 reflections using the [7, 4, 3] Hamming code. The nodes, their associated entropy, likelihoods and scores are listed in Table 2(*a*). Two likelihood values are quoted for each node: one excludes overlapped reflections and the second includes them. In general, the latter are the most useful, and are always used in the powder structures described in this paper. The results of the t test at the 20% significance level are shown in Table 2(*b*), and Table 2(*c*) lists the phases that are deduced from this analysis. The 20% level is high, but often needed in *ab initio* phasing using this code in the early stages of phasing. Where the t test generates a significant phase indication, the phases are always correctly indicated. Note that, as described in the previous section, only main effects can be extracted using this code. Table 2(*d*) shows the top nodes sorted by scores [from equation (6)]. The number of violations for a given node is the number of phase relationships listed in Table 2(*b*) that are in contradiction to the basis-set phases for that node.

The centroid map computed using basis-set reflections and those extrapolated by the ME process, suitably weighted (Bricogne & Gilmore, 1990), calculated from the node from the highest score, is shown in Fig. 1. All the atoms are clearly indicated.

7.2. *The Nordström–Robinson (15, 256, 6) code and the C_{60} buckminsterfullerene electron diffraction data*

The C_{60} buckminsterfullerene structure can be solved from 45 unique electron diffraction data (Dorset & McCourt, 1994) at room temperature using symbolic

Table 2. *Two-level phasing tree for Mg_3BN_3*

(*a*) The first level has a single node defined by the 107 reflection. The second level of 16 nodes was generated by permuting the $2\bar{1}0$, $2\bar{1}8$, $3\bar{1}4$, $3\bar{1}7$, 104, 1,0,12, 300 reflections using the [7, 4, 3] Hamming code.

Node No.	Connected to node number	Entropy	LLG without overlaps	LLG including overlaps	No. of incorrect phases
1	0	−0.12	0.00	0.00	0
2	1	−1.86	−4.51	−4.58	7
3	1	−0.84	2.66	2.87	3
4	1	−1.47	0.84	1.05	3
5	1	−1.36	−0.30	−0.27	3
6	1	−1.78	−0.66	−0.40	3
7	1	−1.31	0.14	0.16	3
8	1	−1.62	1.79	2.03	3
9	1	−0.69	2.11	2.36	3
10	1	−0.68	1.84	2.06	1
11	1	−1.57	1.56	1.79	4
12	1	−1.32	0.17	0.13	3
13	1	−1.76	−0.56	−0.36	4
14	1	−1.39	−0.40	−0.42	4
15	1	−1.41	0.77	0.91	4
16	1	−0.82	2.21	2.42	4
17	1	−1.87	−4.44	−4.47	4

(*b*) The main effect (single sign) analysis of LLG at the 20% significance level.

Reflection No.	$\langle LLG^+ \rangle$	$\langle LLG^- \rangle$	Significance level	Deduced phase
1	1.162	−0.503	0.135	0
5	1.189	−0.529	0.122	0
2	1.363	−0.703	0.057	0

(*c*) The phase angles of the permuted reflections in (*a*) as deduced by the analysis of the LLGs.

Reflection No.	*h*	*k*	*l*	Deduced phase	Correct?
1	2	−1	0	0	Yes
5	2	−1	8	0	Yes
11	3	−1	4	0 or 180	
21	3	−1	7	0 or 180	
14	1	0	4	0 or 180	
3	1	0	12	0 or 180	
2	3	0	0	0	Yes

(*d*) The resulting scores in order of preference. The number of violations is defined in the text.

Node No.	LLG including overlaps	LLG without overlaps	Entropy	Score	No. of violations
9	2.359	2.110	−0.694	6.783	0
10	2.060	1.841	−0.684	5.924	0
3	2.870	2.662	−0.836	5.521	1
16	2.415	2.209	−0.823	4.647	1
4	1.045	0.842	−1.471	2.017	1
8	2.032	1.791	−1.616	1.933	2
15	0.912	0.768	−1.410	1.760	1
11	1.792	1.559	−1.568	1.705	2
7	0.157	0.137	−1.309	0.148	2
12	0.128	0.168	−1.319	0.121	2
2	−4.576	−4.514	−1.856	0.000	3
17	−4.471	−4.439	−1.869	0.000	3
13	−0.363	−0.560	−1.758	−0.355	2
6	−0.396	−0.657	−1.782	−0.387	2
5	−0.271	−0.304	−1.360	−0.515	1
14	−0.418	−0.399	−1.388	−0.793	1

Table 3. *Phasing C_{60} from electron diffraction data from Dorset & McCourt (1994)*

(*a*) The reflections given permuted phases in a Nordstrom–Robinson (15, 256, 6) code. The origin was defined by the 330 reflection.

Reflection No.	h	k	l	$\|U_{\mathbf{h}}\|^{\rm obs}$
1	2	2	0	0.162
2	8	0	4	0.112
3	6	6	6	0.109
5	6	6	0	0.102
6	4	4	4	0.099
7	5	5	5	0.088
9	4	2	0	0.085
12	6	6	4	0.083
13	1	1	1	0.079
14	8	2	2	0.074
15	4	4	0	0.073
21	6	6	2	0.069
24	7	7	1	0.068
25	9	3	3	0.068
29	7	7	3	0.061

(*b*) LLG analysis at the 5% significance level.

Reflection 1	Reflection 2	$\langle$LLG$^+\rangle$	$\langle$LLG$^-\rangle$	Significance	Sign
13		0.079	−0.119	0.000	+
25		−0.058	0.017	0.041	−
1	15	0.072	−0.112	0.000	+
2	15	−0.068	0.027	0.008	−
3	5	−0.067	0.025	0.011	−
3	21	− 0.113	0.071	0.000	−
5	12	−0.063	0.022	0.019	−
5	14	0.019	−0.060	0.030	+
12	14	0.041	−0.081	0.001	+
24	25	0.022	−0.063	0.019	+

(*c*) The resulting scores. Map cc is the map correlation coefficient; No. + and No. − define the number of extrapolated phases that are 0 or 180°, respectively, and the Map features describe any dominant or unexpected features in the final centroid map.

Node No.	LLG	Entropy	Score	No. of violations	Phase error	No. +	No. −	Map cc	Map features
248	1.388	−0.589	9.691	3	70.2	2	24	0.09	Large origin peak
122	1.133	−0.680	9.034	2	63.7	14	12	0.46	OK
39	0.750	−0.629	6.727	1	64.1	14	12	0.43	OK
186	1.072	−0.645	6.404	4	67.0	2	24	0.09	Large origin peak
41	0.542	−0.506	4.855	1	104.0	8	16	0.05	OK
127	0.482	−0.518	3.842	2	9.1	13	13	0.90	OK
117	0.550	−0.616	3.831	3	81.4	9	17	0.30	Large origin peak
107	0.634	−0.593	3.792	4	60.7	14	12	0.44	OK

addition methods. It is an excellent example of a sparse data set in space group *Fm*3*m* with $a = 14.26$ Å and a resolution of ∼1.4 Å. In our experience, it is not routinely solvable by conventional direct-methods black-box programs; it is, however, simply solved by the ME method using the Nordström–Robinson (15, 256, 6) code.

The origin was defined by the 330 reflection, then 256 nodes (instead of 2^{15}) were generated by permuting the phases of 15 reflections. Table 3(*a*) lists the permuted reflections and their associated U values, and Table 3(*b*) lists the results of the LLG analysis at the 5% significance level. Note that the two-phase interactions can be included in the analysis. Because there are more nodes, the significance level is much reduced from the 20% value used in Mg_3BN_3. Table 3(*c*) outlines the characteristics of the centroid maps derived from the nodes with the highest associated scores; this includes the map correlation coefficients, using the phases listed by Dorset & McCourt (1994). Some maps have a large peak at the origin, and so can be rejected. This left five viable maps; the fourth ranked of these has a mean phase error of 9° and a map correlation coefficient of 0.90. A section of this map is shown in Fig. 2. All map correlation coefficients quoted in this paper use F magnitudes from both the basis set and the extrapolates to the resolution of the basis set, but only including the latter if the associated weight is greater than 0.1.

The preferred map is not ranked first, and this is a common and expected feature of using e.c.c.'s without

trying to correct phase errors. Phase errors in the basis set are almost certain and distort the LLG values and their analysis. However, the correct or at least a viable solution in all the examples cited here still lies in the top eight nodes when ranked by score.

7.3. *The Nordström–Robinson (15, 256, 6) code and the Sigma-2 powder data*

Sigma-2 ($[Si_{64}O_{128}] \cdot 4C_{10}H_{17}N$) is a clathrasil phase for which high-quality synchrotron data are available (McCusker, 1988). It crystallizes in space group $I4_1/amd$ with $a = 10.2387$ and $c = 34.3829$ Å. The cage contains disordered 1-adamantanamine molecules used in the Sigma-2 synthesis. The data set comprises 232 unique non-overlapped data plus 13 pairs of overlaps. The effective data resolution is 1.3 Å. Ignoring the 1-adamantanamine molecule, the asymmetric unit comprises four Si and seven O atoms. The structure readily yields to ME methods using a six-level phasing tree (Gilmore, K. Henderson & Bricogne, 1991), and the data set is used here as a demonstration of the effectiveness of the Nordström–Robinson (16, 256, 6) code. Table 4 summarizes the process: the origin was defined by the 1,0,21 reflection with a U magnitude of 0.276; this was followed by a permutation of 16 centric phases to generate 256 nodes. Table 4(*b*) lists the resulting scores in order of preference following the LLG analysis at the 5% significance level. All these maps show the positions of the Si atoms, and the solution ranked second also gives the positions of five O atoms.

7.4. *The [24, 12, 8] Golay code and the NU-3 powder data set*

McCusker and Baerlocher have reported two forms of the zeolite NU-3 (McCusker, 1993; Baerlocher & McCusker, 1994) which has the LEV-type framework. In one case, 1-adamantanamine (ADAM) was used in the synthesis and in the other *N*-methylquinuclidinium iodide (QUIN) was used. We have used the ADAM form. The space group is $R\bar{3}m$ with $a = 13.2251$ and $c = 22.2916$ Å. There are 373 reflections in total, of which 199 are in 80 overlap sets with up to seven under a single overlap envelope; the maximum resolution is 1.1 Å. The origin was defined by the 107 reflection with a U magnitude of 0.345. The 24 reflections listed in Table 5(*a*) were given permuted phases using the [24, 12, 8] Golay code; the resulting sorted scores following LLG analysis based on values that included the overlapped reflections at the 5% significance level are shown in Table 5(*b*). The centroid map for the top-ranked solution is shown in Fig. 3; the entire zeolite framework and the envelope of the 1-adamantanamine guest molecule is clearly indicated. As in §7.3, the entire calculation was completely routine.

7.5. *The [24, 12, 8] and [23, 12, 7] Golay codes and crambin at 3 Å*

The first use of e.c.c.'s in an *ab initio* macromolecular environment was the location of the Zn atom in avian pancreatic polypeptide (App) (Gilmore & Nicholson, 1994) using only the 3 Å data (although the structure diffracts to 1 Å). A three-level phasing tree was constructed in which the first level defined the origin, the second used the Golay code to generate 4096 new nodes, and the third used the Nordström–Robinson code on child nodes of the best eight from level two. A total of $1 + 4096 + (8 \times 256) = 6145$ nodes were generated, and one of the final nodes clearly indicated the Zn atom to within 0.5 Å of its refined position. The Zn atom could not be located by Patterson methods at this resolution.

Here we are employing a similar strategy in the case of crambin. The data set we used diffracts to ~1.3 Å (Hendrickson & Teeter, 1981) which is an unusually high resolution. We chose a subset of these data with a maximum resolution of 3 Å to more closely simulate more typical protein diffraction data, although this is still atypical in the sense of being relatively accurate and with few systematic errors. The space group is $P2_1$ with $a = 40.96$, $b = 18.65$ and $c = 22.54$ Å, $\beta = 90.7°$. The asymmetric unit comprises 202 × C, 55 × N, 64 × O and 6 × S, and, whereas structures of such complexity are now solvable with 1.0–1.1 Å data, direct methods are unable to produce viable phases at 3 Å.

The origin was partially defined by fixing the phases of two reflections: the $20\bar{5}$ and the $30\bar{4}$. Fixing the origin along **z** and defining the enantiomorph was carried out *de facto* by the process of tree building and analysis. For the second level of the phasing tree, 13 reflections with a total of 24 d.o.f. were given permuted phases *via* the Golay [24, 12, 8] code generating 4096 nodes; the top eight, based on scores, were kept. One

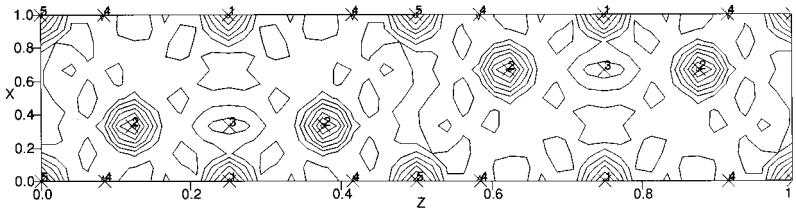

Fig. 1. The centroid map for node 9 for Mg_3BN_3 projected down the *b* axis. Crosses represent atomic positions. 1 and 2: Mg atoms; 3 and 4: N atoms; 5: B atoms.

Table 4. *Phasing the Sigma-2 clathrasil from powder diffraction data from McCusker (1988)*

(*a*) The reflections given permuted phases in a Nordstrom–Robinson (16, 256, 6) code. The origin was defined by the 1,0,21 reflection with a U magnitude of 0.276.

Reflection No.	h	k	l	$\|U_{\mathbf{h}}\|^{\text{obs}}$
1	6	0	0	0.349
2	0	0	24	0.339
3	1	1	6	0.329
4	2	2	4	0.319
5	4	4	8	0.288
6	6	0	12	0.277
8	4	0	18	0.254
9	2	0	8	0.247
13	2	1	1	0.223
16	2	2	20	0.212
18	4	0	14	0.210
20	3	0	1	0.207
21	1	0	17	0.206
22	5	4	9	0.202
27	6	3	1	0.181
33	7	2	3	0.168

(*b*) The resulting scores in order of preference following LLG analysis at the 5% significance level.

Node No.	LLG including overlaps	LLG excluding overlaps	Entropy	Score	No. of incorrect phases in basis set
4	0.835	0.657	−0.521	0.669	7
12	0.629	0.484	−0.550	0.504	4
76	0.601	0.481	−0.485	0.481	6
130	0.552	0.693	−0.711	0.442	5
194	0.471	0.391	−0.496	0.377	7
164	0.445	0.433	−0.535	0.356	5
172	0.418	0.242	−0.478	0.334	4
144	0.389	0.223	−0.572	0.311	7

phase set in this group had a mean phase error of 37.5° and a map correlation coefficient of 0.54. The third level involved permuting the phase of 12 reflections with 23 d.o.f. *via* a Golay [23, 12, 7] e.c.c. Of the top eight nodes, there was a solution with a mean phase error of 46.8° and a correlation coefficient of 0.46. Finally, the top eight nodes were kept and 15 reflections with 24 d.o.f. were given permuted phases using the Golay [24, 12, 8] code. By this time, the accumulation of phase errors begins to bite, and the best mean phase error is 58.5° with a correlation coefficient of 0.33. A total of 42 reflections contribute to the error statistics. The whole process is outlined in Table 6; a total of $1 + 4096 + [2 \times (8 \times 4096)] = 69\,633$ nodes were studied, but using a network of Unix workstations this calculation took less than 72 h in total. The best node (in the sense of minimum phase error) from level 4 is obviously not sufficient to generate maps from which the structure can be extracted, but it is a tribute to the power of the Golay code and the ME formalism that such a calculation is even possible.

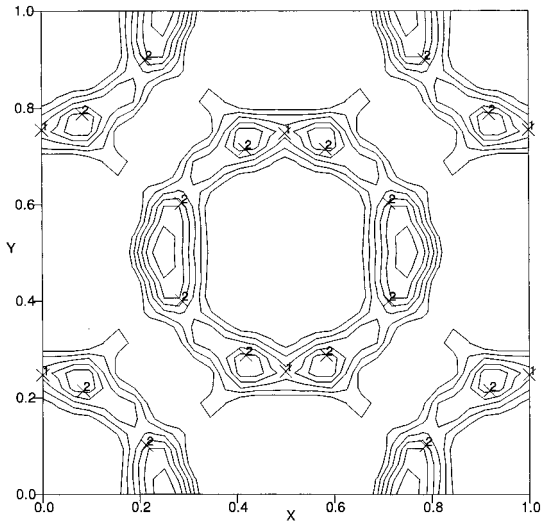

Fig. 2. A typical section for the preferred centroid map for C_{60}. The crosses are C-atom positions taken from the solved structure at low temperature. This data set was collected at room temperature.

8. New structures

We have also solved two unknown structures in this way:

(i) 3-Cyano-4-dimethylaminobiphenyl, $C_{15}H_{14}N$, from electron diffraction data in space group $Pna2_1$ with 118 reflections at a maximum resolution of 1.4 Å, but an effective resolution nearer 2 Å (Voigt-Martin *et al.*,

Table 5. *Phasing the NU-3 zeolite from powder diffraction data from McCusker (1993) and Baerlocher & McCusker (1994)*

(*a*) The reflections given permuted phases in a [24, 12, 8] Golay code. The origin was defined by the 107 reflection with a U magnitude of 0.345.

Reflection No.	h	k	l	$\lvert U_{\mathbf{h}}\rvert^{\text{obs}}$
1	4	0	−2	0.5731
3	5	0	−7	0.4611
4	4	0	16	0.4327
5	10	−3	7	0.3933
7	8	−2	−2	0.3543
8	1	0	−5	0.3318
9	7	−1	−7	0.3128
10	3	−1	−2	0.3045
11	4	−2	0	0.3037
12	9	−4	−5	0.2844
13	2	0	2	0.2808
16	9	−1	−5	0.2629
17	0	0	6	0.2584
19	5	0	−1	0.2504
20	1	0	−14	0.2495
21	8	−2	−5	0.2475
23	8	0	5	0.2352
25	5	0	−4	0.2260
30	5	0	−16	0.2089
33	3	0	0	0.2063
34	8	0	−4	0.2055
35	5	0	−13	0.2020
36	4	0	19	0.2003
40	4	0	−17	0.1899

(*b*) The resulting sorted scores following LLG analysis at the 5% significance level.

Node No.	LLG including overlaps	LLG excluding overlaps	Entropy	Score	No. of incorrect phases in basis set
3349	12.887	9.003	−2.643	0.242	5
1738	13.515	7.171	−2.435	0.233	11
3461	15.502	7.441	−3.047	0.223	7
914	14.025	11.050	−2.281	0.221	8
3433	9.621	9.275	−2.866	0.166	9
1425	6.030	3.353	−2.648	0.133	11
1301	9.005	14.502	−2.780	0.129	5
1627	7.187	9.083	−3.055	0.124	15

1997). This was a difficult problem, and was solved using the Golay and Nordström–Robinson codes. An independent modelling calculation verified the result.

(ii) The metastable alloy phase Al_mFe, commonly found as primary particles in aluminium alloys cast at high cooling rates, has been studied (Gjønnes *et al.*, 1998). The electron diffraction data were collected from eight different projections using a novel precession technique, and supplementary information was obtained *via* the CBED method to check the space-group symmetry and to confirm the proposed structural models. The resulting data set comprised more than 500 unique reflections in space-group $I\bar{4}2m$. The structure was solved totally independently using Patterson, ME and traditional direct methods, all of which indicate a somewhat disordered structure with a network of Al_{10} polyhedra coordinating central Fe atoms. Some correction was made for dynamical scattering, but even without this it is possible to solve this complex inorganic structure.

9. Conclusions and future developments

We have demonstrated the power of e.c.c.'s in a ME–likelihood environment as a source of effective experimental designs for phase permutation, solving both known and unknown structures with some data sets that are resistant to traditional direct methods. Some possible developments and extensions to these ideas that are currently under investigation are as follows:

(i) Errors naturally accumulate with e.c.c.'s when building phasing trees, and phase refinement is a necessary development. In general, the tangent formula is unstable with much of these data because they are sparse or of low resolution. The Bayesian method based on likelihood optimization that we have sometimes

Table 6. *Phasing the crambin data at 3 Å*

(*a*) The reflections given permuted phases. The origin was partially defined by the two reflections in level 1. All subsequent levels used either a [24, 12, 8] or a [23, 12, 8] Golay code.

Level	Number	h	k	l	$\lvert U_{\mathbf{h}}\rvert^{\mathrm{obs}}$
1	11	2	0	−5	0.088
1	24	3	0	−4	0.077
2	1	4	3	−4	0.105
2	4	7	2	4	0.100
2	5	0	2	0	0.097
2	7	9	2	0	0.095
2	8	4	0	4	0.094
2	9	6	1	4	0.094
2	13	9	1	3	0.087
2	14	6	4	−1	0.086
2	16	8	1	0	0.085
2	18	1	3	−4	0.083
2	20	1	1	0	0.080
2	22	4	4	0	0.078
2	26	4	0	0	0.077
3	3	7	3	3	0.102
3	10	12	1	1	0.089
3	15	5	5	0	0.085
3	17	11	2	1	0.085
3	21	5	3	4	0.079
3	27	7	2	−5	0.077
3	29	3	3	−3	0.076
3	35	1	0	5	0.074
3	36	4	5	1	0.072
3	38	10	2	3	0.071
3	52	10	1	3	0.066
3	53	2	1	5	0.066
4	12	11	0	3	0.087
4	28	0	1	5	0.076
4	33	4	0	5	0.074
4	34	7	4	−3	0.074
4	41	2	4	3	0.069
4	43	3	0	0	0.067
4	44	4	0	7	0.067
4	50	1	4	−5	0.066
4	54	0	2	6	0.066
4	55	5	0	−1	0.066
4	59	2	4	−2	0.064
4	61	5	0	0	0.064
4	64	4	3	−3	0.064
4	66	13	1	−2	0.063
4	70	9	3	2	0.062

(*b*) The resulting scores following LLG analysis at the 5% significance level.

Level	No. of reflections in the basis set	Node No.	LLG	Entropy	Score	Phase error (°)	Map cc
1	2		0.00	−0.007			
2	15	3074	1.794	−0.118	39.2	36.5	0.58
2	15	2127	1.773	−0.113	38.8	37.5	0.54
2	15	2774	1.739	−0.118	38.0	87.4	0.03
2	15	3543	1.735	−0.116	37.9	90.9	0.01
2	15	3739	1.762	−0.116	35.1	59.0	0.38
2	15	3918	1.602	−0.119	35.0	70.9	0.13
2	15	2819	1.749	−0.115	34.8	58.6	0.35
2	15	2775	1.454	−0.114	33.2	80.8	0.09
3	27	17933	3.970	−0.204	208.6	70.8	0.20
3	27	23101	3.867	−0.199	206.9	46.8	0.46
3	27	32248	3.987	−0.201	201.6	81.8	0.11
3	27	23455	3.497	−0.196	200.9	55.1	0.39
3	27	11604	4.320	−0.199	197.1	89.1	0.00
3	27	32310	3.813	−0.197	196.6	73.9	0.23
3	27	8131	4.044	−0.194	196.3	79.1	0.09

Table 6 (*cont.*)

Level	No. of reflections in the basis set	Node No.	LLG	Entropy	Score	Phase error (°)	Map cc
3	27	11925	3.938	−0.201	195.2	78.5	0.11
4	42	47716	8.447	−0.265	536.2	67.4	0.21
4	42	51511	7.360	−0.264	503.1	65.8	0.27
4	42	52105	7.588	−0.260	496.2	61.7	0.32
4	42	51087	7.798	−0.274	494.2	63.5	0.31
4	42	49055	6.872	−0.270	491.1	61.4	0.31
4	42	52647	6.718	−0.268	479.8	62.6	0.31
4	42	52591	7.043	−0.265	474.5	62.6	0.29
4	42	50551	7.393	−0.268	468.5	61.9	0.31
4	42	50043	7.281	−0.270	468.4	58.5	0.35
4	42	50578	7.331	−0.269	465.4	71.3	0.22
4	42	52023	7.106	−0.264	465.0	58.8	0.33
4	42	61561	6.991	−0.254	464.4	72.6	0.20
4	42	64121	6.293	−0.265	462.6	64.9	0.29
4	42	39954	7.157	−0.256	462.1	99.8	−0.08
4	42	46940	7.492	−0.272	460.6	70.4	0.19
4	42	45157	7.833	−0.262	459.0	67.6	0.21

employed (Bricogne & Gilmore, 1990; Gilmore *et al.*, 1990) can be useful, but is sometimes unstable or unable to refine phases very far from their initial values. Modifications to this process are being developed.

(ii) In the powder diffraction case, e.c.c.'s can be used as a source of spherical designs for hyperphase permutation (Bricogne, 1991, 1997*b*). This means that both phases and amplitudes of overlapped reflections can be permuted using the appropriate codes, and both can be recovered with suitable analysis of the associated LLGs. Initial studies of the use of ME and likelihood to resolve overlapped intensities without the use of codes have proved to be successful in favourable cases (Dong & Gilmore, 1998), but the use of e.c.c.'s could extend these ideas considerably.

(iii) There has been a great deal of recent activity in using structural fragments translated and rotated through the unit cell to solve organic crystal structures from powder data (see, for example, Harris & Tremayne, 1996; Shankland *et al.*, 1997; Kariuki *et al.*, 1997). These methods use various search procedures, *e.g.* genetic algorithms, to carry out the search. Codes provide an efficient way of defining the initial search parameters (Bricogne, 1997*b*) and could be used to increase the power of these methods by providing more efficient starting points. We have used codes to define molecular envelopes as a starting point in modelling studies of this type with considerable success (Tremayne *et al.*, 1997).

(iv) The literature on coding theory is vast. Other codes, not necessarily binary, may well exist with excellent design and covering properties that could extend these ideas further.

(v) Codes can also be used in conventional direct methods. We have examined the use of the Golay and

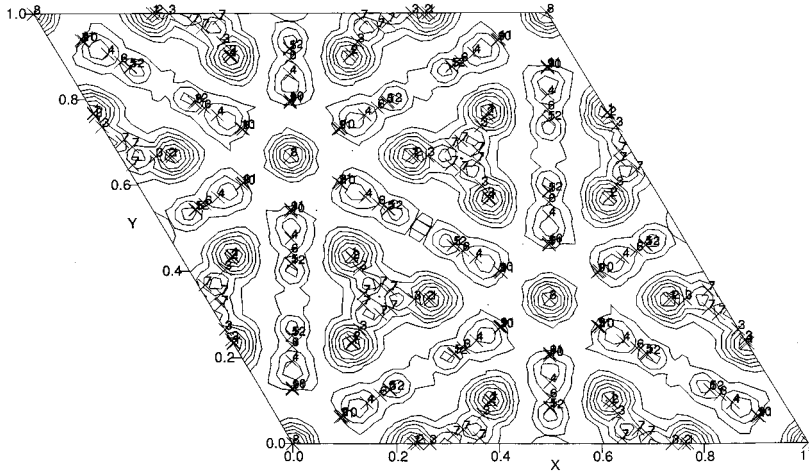

Fig. 3. The NU-3 structure projected down the *c* axis for node 3349. The entire framework is clearly visible.

Nordström–Robinson codes as a starting point in multisolution direct methods using the tangent formula. The results are encouraging, and will be published elsewhere (Gilmore *et al.*, 1999).

We thank Lynne McCusker and Christian Baerlocher for the NU-3 and Sigma-2 data, Doug Dorset for the C_{60} data set, and Professor Hashizume for data from magnesium boron nitride. CG and WD thank the EPSRC for financial support.

References

Anderson, I. (1989). *A First Course in Combinatorial Mathematics.* Oxford University Press.
Baerlocher, C. & McCusker, L. B. (1994). *Stud. Surf. Sci. Catal.* **85**, 391–428.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Bricogne, G. (1991). *Acta Cryst.* A**47**, 803–829.
Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.
Bricogne, G. (1997*a*). *Methods Enzymol.* **276**, 361–423.
Bricogne, G. (1997*b*). *Methods Enzymol.* **276**, 424–448.
Bricogne, G. & Gilmore, C. J. (1990). *Acta Cryst.* A**46**, 284–297.
Carter, C. W. Jr (1992). *Crystallization of Proteins and Nucleic Acids: a Practical Approach*, edited by A. Ducruix & R. Giegé, pp. 47–71. Oxford: IRL Press.
Carter, C. W. Jr (1997). *Methods Enzymol.* **276**, 74–99.
Cochran, W. G. & Cox, G. M. (1957). *Experimental Designs.* New York: John Wiley.
Conway, J. H. & Sloane, N. J. A. (1988). *Sphere Packings, Lattices and Groups.* New York: Springer-Verlag.
DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. (1994). *Acta Cryst.* A**50**, 203–220.
Dong, W. & Gilmore, C. J. (1998). *Acta Cryst.* A**54**, 438–446.
Dorset, D. L. & McCourt, M. P. (1994). *Acta Cryst.* A**50**, 344–351.
Doublié, S., Gilmore, C. J., Bricogne, G. & Carter, C. W. C. Jr (1995). *Structure*, **3**, 17–31.
Doublié, S., Xiang, S., Gilmore, C. J., Bricogne, G. & Carter, C. W. C. Jr (1994). *Acta Cryst.* A**50,** 164–182.
Gilmore, C. J. (1996). *Acta Cryst.* A**52**, 561–589.
Gilmore, C. J. & Bricogne, G. (1997). *Methods Enzymol.* **277**, 65–78.
Gilmore, C. J., Bricogne, G. & Bannister, C. (1990). *Acta Cryst.* A**46**, 297–308.
Gilmore, C. J., Donald, A. & Bricogne, G. (1999). In preparation.
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.
Gilmore, C. J., Marks, L. D., Grozea, D., Collazo, C., Landree, E. & Twesten, R. (1997). *Surf. Sci.* **381**, 77–91.
Gilmore, C. J. & Nicholson, W. V. (1994). *Trans. Am. Crystallogr. Assoc.* **30,** 15–27.
Gjønnes, J., Hansen, V., Berg, B., Runde, P., Cheng, Y. F., Gjønnes, K., Dorset, D. L. & Gilmore, C. J. (1998). *Acta Cryst.* A**54**, 306–319.
Golay, M. J. E. (1949). *Proc. IRE (IEEE)*, **37**, 23–28.
Good, I. J. (1954). *Acta Cryst.* **7**, 603–604.
Graham, R. L. & Sloane, N. J. A. (1985). *IEEE Trans. Inf. Theory*, **31**, 385–401.
Hamming, R. W. (1947). Memorandum 20876. Bell Telephone Laboratories, Murray Hill, NJ, USA.
Harris, K. D. M. & Tremayne, M. (1996). *Chem. Mater.* **8**, 2554–2570.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
Hill, R. (1993). *A First Course in Coding Theory.* Oxford University Press.
Hiraguchi, H., Hashizume, H., Fukunaga, O., Takenaka, A. & Sakata, M. (1991). *J. Appl. Cryst.* **24**, 286–292.
Kariuki, B. M., Serrano-González, H., Johnston, R. L. & Harris, D. M. (1997). *Chem. Phys. Lett.* **280**, 189–195.
McCusker, L. B. (1988). *J. Appl. Cryst.* **21**, 305–310.
McCusker, L. B. (1993). *Mater. Sci. Forum*, **133/136**, 423–434.
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.
Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.
Shankland, K., David, W. I. F. & Csoka, T. (1997). *Z. Kristallogr.* **212**, 550–552.
Shankland, K., Gilmore, C. J., Bricogne, G. & Hashizume, H. (1993). *Acta Cryst.* A**49**, 493–501.
Shannon, C. E. (1948*a*). *Bell Syst. Tech. J.* **27**, 379–423.
Shannon, C. E. (1948*b*). *Bell Syst. Tech. J.* **27**, 623–656.
Sloane, N. J. A. (1984). *Sci. Am.* **250**, 116–125.
Thompson, T. M. (1983). *From Error-Correcting Codes Through Sphere Packings to Simple Groups*, Vol. 21, edited by D. T. Finkbeiner. Washington, DC: The Mathematical Association of America.
Tremayne, M., Dong, W. & Gilmore, C. J. (1997). Proc. Am. Crystallogr. Assoc. Meet., Abstr. SuB06.
Tremayne, M., Lightfoot, P., Glidewell, C., Mehta, M. A., Bruce, P. G., Harris, K. D. M., Shankland, K., Gilmore, C. J. & Bricogne, G. (1992). *J. Solid State Chem.* **100**, 191–196.
Tremayne, M., Lightfoot, P., Harris, K. D. M., Shankland, K., Gilmore, C. J., Bricogne, G. & Bruce, P. G. (1992). *J. Mater. Chem.* **2**, 1301–1302.
Voigt-Martin, I. G., Zhang, Z. H., Kolb, U. & Gilmore, C. J. (1997). *Ultramicroscopy*, **68**, 43–59.
White, P. S. & Woolfson, M. M. (1975). *Acta Cryst.* A**31**, 53–56.
Woolfson, M. M. (1954). *Acta Cryst.* **7**, 65–67.