Acta Cryst. (1993). A**49**, 493-501

A Multisolution Method of Phase Determination by Combined Maximization of Entropy and Likelihood. VI. Automatic Likelihood Analysis *via* the Student *t* Test, with an Application to the Powder Structure of Magnesium Boron Nitride, Mg_3BN_3

By K. Shankland and C. J. Gilmore

Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland

G. Bricogne

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

and H. Hashizume

Research Laboratory of Engineering Materials, Tokyo Institute of Technology,
Nagatsuta, Midori, Yokohama 227, Japan

(*Received* 1 *June* 1992; *accepted* 10 *November* 1992)

Abstract

The multisolution method to solve powder structures *ab initio* from their X-ray diffraction data developed by Bricogne and Gilmore [Bricogne (1991). *Acta Cryst.* A**47**, 803-829; Gilmore, Henderson & Bricogne (1991). *Acta Cryst.* A**47**, 830-841] has been further tested by redetermination of the structure of the low-pressure phase of magnesium boron nitride, Mg_3BN_3, on which a previous attempt using the maximum-entropy (ME) procedure devised by Gull, Livesey & Sivia [*Acta Cryst.* (1987), A**43**, 112-117] had failed. In the successful application of the ME method presented here, the data were normalized using both overlapped and nonoverlapped reflections in the program *MITHRIL* [Gilmore (1984). *J. Appl. Cryst.* **17**, 42-46; Gilmore & Brown (1988). *J. Appl. Cryst.* **21**, 571-572]. After definition of the origin by the phase of a single reflection, seven reflections selected by a criterion of optimum second-neighbourhood enlargement were given permuted phases, thus generating 128 nodes of a phasing tree. Each node was subjected to constrained entropy maximization followed by the evaluation of a log-likelihood gain incorporating both overlapped and nonoverlapped reflections. These log-likelihood gains were analysed with the Student *t* test in which single-, double- and triple-phase indications were tested. Eight nodes survived the tests at the 2% significance level to give a subset of preferred nodes; the member of this subset with the highest log-likelihood gain gave a centroid map that revealed the positions of all the Mg, B and N atoms. Detailed examination of the phasing tree confirmed previous observations that the log-likelihood gain, not the entropy, is the most reliable criterion on which to base a multisolution phasing procedure.

0. Introduction

The problems involved in the solution of crystal structures *ab initio* from their X-ray or neutron powder diffraction data using conventional direct methods are well known: in powder diffraction, a three-dimensional data set is projected into one dimension where it is spherically averaged and, as a result, reflections that would otherwise be separately measured overlap; this degree of overlap increases with $(\sin\theta)/\lambda$, thus often reducing the effective resolution of the data to 1.3 Å or less. This overlap can arise accidentally from the diffraction geometry or systematically as a consequence of point-group symmetry. There have, of course, been many successes. [See, for example, McCusker (1988) and Hiraguchi, Hashizume, Fukunaga, Takenaka & Sakata (1991) and the references cited therein for a list of these.] However, no generally applicable technique to solve powder structures has yet emerged in spite of a recent surge of activity in this field (David, 1990; Cascanaro, Favia & Giacovazzo, 1992; Estermann & Gramlich, 1992; Jansen, Peschar & Schenk, 1992).

The maximum-entropy (ME) method as formulated by Bricogne (1984, 1988, 1991*a*, *b*, *c*) has already been successfully applied to powder structures (Gilmore, Henderson & Bricogne 1991*b*; Gilmore & Bricogne 1991), as well as to single-crystal data (Gilmore, Bricogne & Bannister, 1990), protein data sets (Gilmore, Henderson & Bricogne, 1991*a*) and electron crystallography data (Dong, Baird, Fryer, Gilmore, MacNicol, Bricogne, Smith, O'Keefe & Hovmöller, 1992; Gilmore, Shankland & Fryer, 1993) and we feel that these varied applications have provided strong practical support to the original theoretical arguments in favour of this technique as a rational

0108-7673/93/030493-09$06.00

method of structure determination, especially in cases where the data are limited in resolution. We present here a retrospective investigation of the *ab initio* structure determination of the low-pressure phase of magnesium boron nitride, Mg_3BN_3 (Hiraguchi, Hashizume, Fukunaga, Takenàka & Sakata, 1991). Originally, this structure was solved by means of heavy-atom methods. Standard direct methods were also successful but there are, however, several reasons to investigate the behaviour of our mode of structure solution on this problem.

(i) The maximum-entropy method devised by Gull, Livesey & Sivia (1987) was tried on this data set and was unsuccessful in spite of the exceptional and almost single-crystal quality of the intensity data.

(ii) Examination of the phasing tree demonstrates quite clearly that entropy alone is not a good indicator of correct phase sets under these circumstances. This is in direct contradiction with the claims made by Sjölin, Prince, Svensson & Gilliland (1991) in their study of the protein recombinant bovine chymosin.

(iii) A new method to analyse log-likelihood gains by means of node partitioning coupled with the Student *t* test is presented. This overcomes any notion of subjectivity when selecting nodes on a phasing tree by taking just those with the largest log-likelihood gains, which is the method we have used previously. The underlying theory of this method is discussed in § 1.

(iv) The ease with which this structure was solved here, where a virtually noise-free map was produced that showed the positions of all the Mg, B and N atoms, makes a compelling case for our method.

(v) We have devised and implemented an efficient computational strategy to exploit the inherent parallelism of the tree search on clusters of workstations, which we expect to be of great value in subsequent applications of our method to powder structure determination or, indeed, in any parallel computation in crystallography.

§ 1 describes the underlying theory and in particular the application of statistical tests of significance to log-likelihood gains. § 2 outlines the crystal structure, the data preparation and the ME results. § 3 discusses the results and, finally § 4 presents a summary with concluding remarks.

1. Theory and methods

The rationale of the theory underlying this work and its quantitative aspects have been described in many publications (see, for example, the references given in the *Introduction*). The multisolution strategy itself consists of exploration of the space of hypothetical phase sets in a hierarchical fashion by building a search tree; each phase set is ranked according to a statistical criterion, the log-likelihood gain (LLG), which acts as an heuristic function in the determination of the subsequent growth of the tree. This criterion measures the extent to which the observed pattern of the unphased intensities has been rendered more likely by the phase choices made for a basis set of reflections (with a hypothesis H_1, specified) than they were under the null hypothesis, H_0 (which leads to Wilson's statistics).

The LLG is defined 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 will, in general, yield different values of the LLG. This intrinsic randomness of the LLG results in the possibility that $L(H_0)$ may be greater than $L(H_1)$ even if $L(H_1)$ is true, because of the random fluctuations in $L(H_0)$ and $L(H_1)$. It is, therefore, of the utmost importance to compare the observed value of the LLG with the statistical distribution of its fluctuations so as to gauge the level of significance of any indication of preference for H_1 over H_0. This significance level is defined as the probability that the observed LLG be due to a statistical fluctuation in $L(H_0)$ and not to the change of distribution associated with the alternative hypothesis, 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 tree) should only be carried out on the basis of a significance test found to be conclusive at a pre-set significance level. When working on a known structure, this discipline affords the only means to guarantee that no use is made of this prior knowledge; even then, it remains possible that such knowledge may exert an indirect or subconscious influence on certain strategic choices!

In this work, we have not attempted to calculate the theoretical variance of the LLG and thus to ascertain absolute significance levels. Rather, we have adopted an empirical standpoint, viewing the set of scores attached to the various phase assumptions as the results of a designed experiment (Cochran & Cox, 1957) in which the signs chosen for the *n* coordinates of the $\mathbf{U}^*$ vector assume the roles of treatments and the scores are considered as yields. The statistical analysis then consists of the detection of those combinations of treatments (if any) that have a significant effect on the yields. An effect associated with an individual sign choice is called the main effect of that sign; an effect associated with combined choices of more than one sign is said to result from an interaction of these signs.

The simplest test consists of the detection of the main effect associated with a single sign. This is done by calculation of the average μ^+ (μ^-) and the variance V^+ (V^-) of the scores attached to the nodes where this sign is $+$ ($-$) and a test to find whether the contrast $\mu^+ - \mu^-$ is significant by means of a

Student *t* test [see, for example, Press, Flannery, Teukolsky & Vetterling (1986)]. This test defines the significance level of this contrast as the probability that it could arise solely from the fluctuations measured by V^+ and V^- even if the two distributions of scores had the same theoretical mean μ. The same procedure allows one to test whether there are significant effects associated with the interactions of several signs, the averages and variances being computed for the two subsets of scores where the product of signs is + or −.

If the set of permuted phases is the full grid of 2^n sign choices, this procedure is none other than an *n*-fold tensor product of two-point Fourier transforms. For acentric reflections, this determines separately the signs of the real and imaginary parts of the structure factors; this is equivalent to the computation of a four-point transform from samples taken at the points of a quadrant permutation. It is possible to develop this Fourier-analysis viewpoint further and a complete description is given by Bricogne (1993) together with a report of its implementation and use over the past two years in a computer program (*BUSTER*) for direct *ab initio* phasing of biological macromolecules. The simpler procedure described here, which was one sign per centric and two signs per acentric reflection, has the advantage of being simple to implement and has been found to work satisfactorily on several test structures beside the one reported here. It also generalizes in a natural fashion to powder diffraction when the method of hyperoctant permutation (Bricogne, 1991*a*) has been used.

2. Data preparation and structure solution

The low-pressure phase of magnesium boron nitride, Mg_3BN_3, crystallizes in the hexagonal space group $P6_3/mmc$ with $a = 3.54453\,(4)$, $c = 16.0353\,(3)$ Å and $Z = 2$ (Hiraguchi, Hashizume, Fukunaga, Takenaka & Sakata, 1991). The intensity data were collected on a good laboratory instrument.

The data were first normalized using *MITHRIL*91 (Gilmore, 1984; Gilmore & Brown, 1988) to give unitary structure factors $|U_h|^{\rm obs}$. The normalization included the overlapped reflections. The data were partitioned into two disjoint sets $\{N\}$ and $\{O\}$, which are the nonoverlapped and overlapped data, respectively. The variances, $\sigma^2(|U_h|^{\rm obs})$, were also computed. There was a total of 69 reflections in the data set, including two overlap sets, each comprising two reflections, so the overlaps play a minor role in this structure determination. The 2θ range of the data was 0–120°, giving an effective resolution of 0.9 Å, which is very good for a powder data set. A temperature factor of 0.8 $Å^2$ was suggested by the Wilson plot and used in the normalization. One feature is of importance here, however. The unit cell of Mg_3BN_3 comprises six Mg, six N and two B atoms. This number of atoms is small and, in consequence, the values of $|U_h|^{\rm obs}$ were very large with a maximum value of 0.78. The exponential modelling algorithm that we use can accommodate such large magnitudes and remain stable but problems are encountered with phase-permutation methods as some phase combinations may give rise to negative Toeplitz determinants. The associated nodes will then not permit an adequate fit between $|U_h|^{\rm obs}$ and $|U_h^{\rm ME}|$ in the sense of making the χ^2 statistic equal to 1.0 (Bricogne & Gilmore, 1990). To overcome this problem, the unit-cell contents were multiplied by five, which gave a maximum *U* magnitude of 0.35. There were no problems with the χ^2 statistic for any node under these circumstances. We will in due course introduce checks for negative determinants prior to node evaluation to avoid the need for such *ad hoc* readjustments but we have checked that the use of a phased likelihood function (Bricogne, 1993) and the use of *U* magnitudes with the correct cell contents, which takes into account the impossibility of reaching $\chi^2 = 1.0$ for certain constraints, leads to essentially the same result as those described below.

The remaining calculations were carried out using the *MICE* maximum-entropy program (Gilmore, Bricogne & Bannister, 1990). The origin was defined by reflection 107, $|U_h^{\rm obs}| = 0.153$, $d =$ 1.8 Å, which belonged to set $\{N\}$ and satisfied the usual rules and criteria appropriate for the space group. (A new version of the ME program now permits the use of set $\{O\}$ here if necessary.) It was selected as the reflection with the greatest ability to allow optimal enlargement of the second neighbourhood of the basis set when extra reflections are added later. This single reflection defined the basis set $\{H\}$; the remaining non-basis-set reflections were assigned to set $\{K\}$. This generated the root node (node 1).

With the concept of optimal second-neighbourhood enlargement again invoked, seven reflections were added to the basis set and given permuted phases (0 or π since the space group is centrosymmetric). This generated 128 new nodes numbered 2 to 129. Entropy maximization was carried out on each node using exponential modelling; a line search was employed for two cycles then the slower plane-search algorithm was used with bicubic modelling of both the entropy and the constraint functions to hold χ^2 at 1.0 (Bricogne & Gilmore, 1990, § 2.3). For each node, the log-likelihood gain was calculated with use of the likelihood expression described for single-crystal data by Bricogne (1984) and later by Bricogne & Gilmore (1990) and generalized by Bricogne (1991*a*) to include overlapped reflections. Likelihood considerations also permit the refinement of the Σ parameter, which is related to the effective number, $N_{\rm eff}$, of atoms in the unit cell by $N_{\rm eff} = 1/\Sigma$. Its value reflects both the quality and the resolution of the data. $N_{\rm eff}$ tends to increase as the data resolution increases

Table 1. *The phasing tree for* Mg_3BN_3

Node	To node	Entropy	LLG (No overlaps)	LLG (Overlaps)	Σ
1	0	−0.5	0.01	0.00	0.01429
2	1	−0.88	2.75	2.81	0.00705
3	1	−0.75	7.58	7.89	0.00298
4	1	−0.89	1.50	1.56	0.00758
5	1	−0.67	1.72	1.72	0.00722
6	1	−0.93	0.65	0.66	0.00855
7	1	−0.86	1.54	1.57	0.00746
8	1	−0.83	0.78	0.84	0.00820
9	1	−0.66	4.39	4.53	0.00509
10	1	−0.83	1.59	1.64	0.00773
11	1	−1.32	9.11	9.40	0.00340
12	1	−0.83	2.55	2.66	0.00685
13	1	−0.82	2.82	2.80	0.00660
14	1	−0.69	1.12	1.17	0.00823
15	1	−1.07	3.31	3.41	0.00591
16	1	−0.79	−0.38	−0.39	0.00869
17	1	−1.23	7.19	7.37	0.00398
18	1	−0.76	−0.42	−0.49	0.00982
19	1	−1.08	3.60	3.82	0.00456
20	1	−0.81	−0.52	−0.55	0.00965
21	1	−0.82	2.85	2.93	0.00643
22	1	−0.81	−0.10	−0.16	0.00921
23	1	−1.07	4.70	4.84	0.00520
24	1	−0.74	−0.31	−0.33	0.00937
25	1	−0.71	1.69	1.72	0.00704
26	1	−0.88	−0.60	−0.66	0.00983
27	1	−1.12	3.80	4.02	0.00528
28	1	−0.89	−0.27	−0.26	0.00915
29	1	−0.82	2.45	2.55	0.00616
30	1	−0.75	0.18	0.18	0.00909
31	1	−0.99	5.45	5.74	0.00399
32	1	−0.81	−0.91	−0.96	0.00970
33	1	−0.89	1.84	1.87	0.00737
34	1	−0.95	0.72	0.69	0.00814
35	1	−1.05	5.36	5.67	0.00328
36	1	−0.89	−0.41	−0.54	0.00953
37	1	−0.84	3.03	3.13	0.00603
38	1	−1.03	−0.30	−0.37	0.00901
39	1	−1.09	4.03	4.18	0.00566
40	1	−0.84	−0.64	−0.73	0.00956
41	1	−0.75	2.81	2.95	0.00572
42	1	−0.79	−0.43	−0.51	0.00937
43	1	−1.11	5.05	5.24	0.00461
44	1	−0.80	−0.57	−0.66	0.00933
45	1	−0.76	1.62	1.65	0.00750
46	1	−0.69	−0.52	−0.57	0.00973
47	1	−1.01	3.91	4.13	0.00514
48	1	−0.78	−1.31	−1.44	0.00994
49	1	−0.93	2.00	2.08	0.00721
50	1	−0.80	1.63	1.76	0.00737
51	1	−0.87	3.15	3.38	0.00596
52	1	−0.82	1.18	1.30	0.00741
53	1	−0.92	6.00	6.28	0.00481
54	1	−0.89	1.02	1.09	0.00717
55	1	−1.19	9.79	10.15	0.00257
56	1	−0.78	1.95	2.06	0.00664
57	1	−0.68	1.30	1.38	0.00794
58	1	−0.78	1.06	1.14	0.00800
59	1	−0.75	2.06	2.19	0.00718
60	1	−0.77	1.32	1.44	0.00736
61	1	−0.68	3.38	3.58	0.00620
62	1	−0.72	2.18	2.30	0.00712
63	1	−0.78	7.90	8.22	0.00337
64	1	−0.77	0.26	0.32	0.00776
65	1	−0.67	0.82	0.88	0.00829
66	1	−0.94	2.95	3.11	0.00641
67	1	−0.65	7.38	7.68	0.00301
68	1	−1.02	1.10	1.16	0.00767
69	1	−0.95	0.47	0.48	0.00769
70	1	−0.89	1.91	2.03	0.00744
71	1	−0.70	3.11	3.15	0.00613
72	1	−0.82	1.50	1.53	0.00800
73	1	−1.08	6.09	6.28	0.00396
74	1	−0.83	0.76	0.83	0.00797
75	1	−0.79	7.57	7.84	0.00363
76	1	−0.70	2.64	2.77	0.00686
77	1	−1.10	1.30	1.32	0.00752
78	1	−0.84	0.96	1.01	0.00833
79	1	−0.89	3.90	3.90	0.00516
80	1	−0.82	1.03	1.08	0.00817
81	1	−1.22	6.27	6.49	0.00492
82	1	−0.80	−0.46	−0.49	0.00975
83	1	−0.73	2.98	3.13	0.00499
84	1	−0.83	−0.64	−0.70	0.00971
85	1	−1.03	3.36	3.50	0.00614
86	1	−0.77	−0.20	−0.18	0.00934
87	1	−0.86	4.36	4.51	0.00512
88	1	−0.72	−0.31	−0.33	0.00947
89	1	−1.00	1.40	1.46	0.00713
90	1	−0.88	−0.67	−0.70	0.00970
91	1	−0.90	2.29	2.38	0.00656
92	1	−0.72	−0.25	−0.28	0.00944
93	1	−0.94	2.44	2.57	0.00665
94	1	−0.88	0.28	0.33	0.00879
95	1	−0.85	4.80	5.07	0.00410
96	1	−0.84	−0.77	−0.79	0.00983
97	1	−1.01	1.94	2.07	0.00725
98	1	−0.94	0.25	0.21	0.00851
99	1	−0.78	5.19	5.49	0.00325
100	1	−1.12	−0.75	−0.78	0.00897
101	1	−1.09	2.20	2.36	0.00650
102	1	−0.89	−0.19	−0.27	0.00942
103	1	−0.89	4.89	5.03	0.00481
104	1	−0.85	0.00	−0.02	0.00893
105	1	−1.05	1.92	2.06	0.00633
106	1	−0.84	−1.01	−1.11	0.00974
107	1	−1.07	3.66	3.87	0.00520
108	1	−0.68	−0.39	−0.41	0.00939
109	1	−0.94	1.57	1.65	0.00762
110	1	−0.78	−0.84	−0.95	0.00994
111	1	−0.79	2.74	2.90	0.00586
112	1	−0.75	−0.75	−0.81	0.00971
113	1	−1.00	2.24	2.34	0.00720
114	1	−0.84	1.15	1.24	0.00725
115	1	−0.68	2.24	2.41	0.00657
116	1	−0.90	1.19	1.31	0.00734
117	1	−0.97	6.67	6.98	0.00427
118	1	−0.79	0.43	0.55	0.00759
119	1	−0.91	7.05	7.37	0.00399
120	1	−0.73	1.50	1.65	0.00718
121	1	−0.79	0.98	1.11	0.00762
122	1	−0.81	0.36	0.43	0.00795
123	1	−0.68	1.31	1.39	0.00796
124	1	−0.69	1.09	1.17	0.00800
125	1	−0.77	4.20	4.40	0.00563
126	1	−0.74	0.98	1.10	0.00743
127	1	−0.67	5.26	5.55	0.00472
128	1	−0.73	0.55	0.66	0.00797
129	1	−0.70	1.03	1.14	0.00782

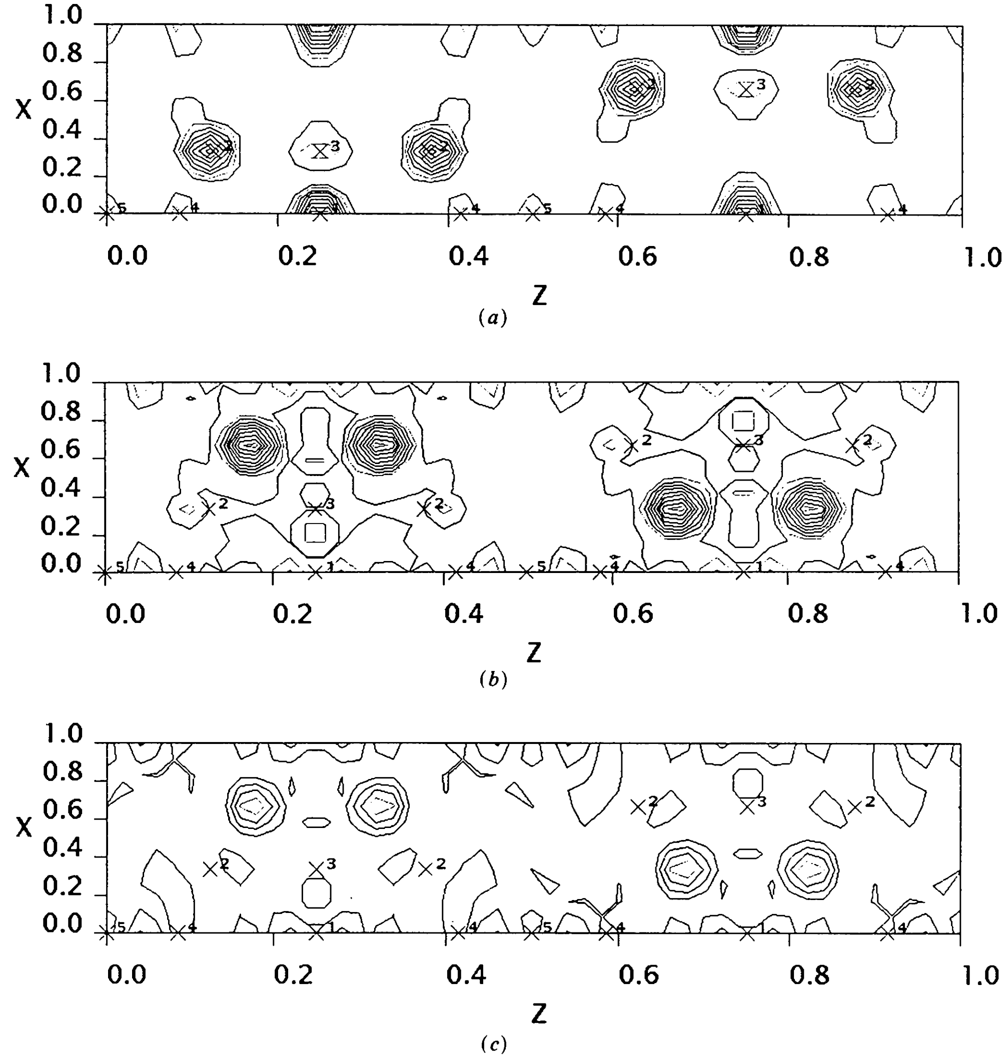
and as the size of the basis set increases and with it the strength of the extrapolation. The behaviour of this parameter for this structure is discussed in § 3. Table 1 summarizes the results of this calculation. An analysis that used the Student *t* test on single-, double- and triple-phase indications at the 2% significance level reduced the tree to eight nodes numbered 3, 11, 55, 63, 67, 75, 119, 127. Of these, node 55 had the largest likelihood. Table 2 summarizes this analysis.

Table 2. *The results of the t test for the permutation of seven reflections*

$s(n)$ refers to the sign of reflection number n. The reflections involved are: (5) 1,0,12 $|U_h|^{\rm obs} = 0.223$; (6) $3\bar{1}4$ $|U_h|^{\rm obs} = 0.210$; (7) 2,0,12 $|U_h|^{\rm obs} = 0.209$; (8) $3\bar{1}7$ $|U_h|^{\rm obs} = 0.196$; (9) $3\bar{1}1$ $|U_h|^{\rm obs} = 0.189$. Reflections (18) $3\bar{1}6$ $|U_h|^{\rm obs} = 0.138$ and (31) $3\bar{1}5$ $|U_h|^{\rm obs} = 0.089$ were also given permuted phases but there were no indications involving them with a significance level of less than 2%

Effect	Sign indication	Significance level
$s(8)$	+	$<10^{-4}$
$s(9)$	−	1.5×10^{-2}
$s(5)s(7)$	+	6.6×10^{-4}
$s(9)s(6)s(5)$	−	4.3×10^{-4}

At this point, all *U* magnitudes had large extrapolated values and the phasing procedure was considered complete. To extract the atomic coordinates, $q^{\rm ME}(\mathbf{x})$ (which is not a map in the traditional sense but a probability distribution of random atomic positions) is used to generate a centroid map (Bricogne & Gilmore, 1990, § 1.6). For this calculation, reflections belonging to both set *H* and set *K* are used and assigned weighted Fourier coefficients; overlapped reflections are also included. Gilmore, Henderson & Bricogne (1991*b*) have shown how the use of this



(*a*)

(*b*)

(*c*)

Fig. 1. centroid maps in projection down the *b* axis for Mg_3BN_3. The atoms labelled 1 and 2 are Mg; 3 and 4 are N; 5 is the B atom. (*a*) The map for the preferred node, 55. (*b*) The map for node 17, which has a high likelihood but was rejected by the Student *t* test. (*c*) The map for node 9, which has the maximum entropy.

formula produces maps that are significantly cleaner than those from which the overlaps are omitted. The *MICE* program allows the computation of both overlapped and non-overlapped maps, with the former as the default.

The centroid map for node 55 is shown in Fig. 1(*a*) in projection down **a**. It is a remarkably clean map, showing all the atomic positions including that of B. The only noise is a slight tail around Mg(2), which can easily be rejected as spurious on chemical grounds.

It is worth emphasizing that these maps are not strictly comparable with those of Hiraguchi *et al.* (1991), whose work was directed primarily towards accurate electron density studies where the maximum-entropy maps were computed for the specific purpose of suppression of series-termination artefacts. Our own study has been limited to investigation of the phase-determination process and the quantitative evaluation of our maximum-entropy maps will be studied separately.

3. Discussion

3.1. *Comments on the phasing tree and the role of the t test*

The ME method worked here automatically, revealing a complete structure. The *t* test is very effective as a method to extract information in a rational way from a set of log-likelihood gains. In the present case, selection of only the node with the highest likelihood would be successful but in many other situations, especially where the tree was being further expanded with new nodes, we have found the use of the *t* test to be essential. Its invocation removes any degree of subjectivity from the decision process as well as making the phasing procedure automatic.

Visual inspection of the maps for the other seven nodes selected by the *t* test shows that none of them quite matches the quality of that from node 55 but that they all contain considerable elements of correctness and probably sufficient information to enable the combination of Rietveld refinement and Fourier synthesis to complete the structure. On the other hand, solutions with relatively high likelihood that are rejected by the *t* test are uniformly poor. Fig. 1(*b*) shows a typical example. This is for node 17, which has a log-likelihood gain of 7.4, marginally higher than that for some of the nodes that have been kept by the *t* test. The map is of very poor quality, with very little contrast; the largest peaks are incorrectly placed and the structure is essentially unsolved at this stage. This illustrates the main point of § 1, *viz* the LLG is an intrinsically random quantity to which a statistical analysis must be applied if reliable phase indications are to result.

The role of the overlaps also deserves a mention, in spite of the fact that the Mg_3BN_3 data used here were of almost single-crystal quality. We have found that, with typical powder diffraction data sets, their inclusion is essential, both as contributors to the LLG (to increase sensitivity) and as Fourier terms in the final centroid map (to minimize distortions caused by the absence of terms for reflections involved in the overlaps). Examples of powder structure determinations where this importance is demonstrated are given by Gilmore, Henderson & Bricogne (1991*b*), Tremayne, Lightfoot, Mehta, Bruce, Harris, Shankland, Gilmore & Bricogne (1992) and Lightfoot, Tremayne, Harris & Bruce (1992).

The behaviour of the Σ parameter is also interesting. For most of the nodes it refines to a value of *ca* 0.009–0.006 but for the correct node it falls dramatically to 0.00257. This behaviour is to be expected: when a great deal of correct phase information has accumulated, the ME extrapolation becomes highly exact as the Toeplitz determinants approach zero (Goedkoop, 1950). In these circumstances, Σ refines to a much smaller value than $1/N$. We have repeatedly observed the combined increase in LLG coupled with a fall in Σ to be an indication of the correct node. It would, however, be unwise to use Σ alone as an indicator of correctness since it depends so critically on data resolution.

3.2. *Is entropy a suitable indicator of phase correctness?*

The question arises of the suitability of entropy as an indicator of phase correctness. Sjölin, Prince, Svensson & Gilliland (1991), for example, claimed to have obtained phases *ab initio* for the structure of recombinant bovine chymosin [previously solved by Gilliland, Windborne, Nachman & Wlodawer (1990) *via* conventional heavy-atom substitution methods] using a combination of phase permutation by means of a fractional factorial design with entropy as the sole indicator of phase correctness. This claim is controversial. Lemaréchal & Navaza (1991) have raised objections to the use of entropy in this way and it is in conflict with the results published by Gilmore, Bricogne & Bannister (1990) and Gilmore, Henderson & Bricogne (1991*a*). The entropy values for the nodes of Mg_3BN_3 range from −0.66 (node 9) to −1.19 (node 55). The correct solution has the *minimum* entropy here and we have found that *t* tests based on entropy instead of likelihood produce maps that are uninterpretable. A typical example is shown in Fig. 1(*c*) for node 9, which has the maximum entropy. The map has one large peak in the asymmetric unit, which is incorrectly placed; indeed, none of the atoms are correctly indicated. It is our experience that this behaviour is quite typical and that, regardless of resolution, entropy alone cannot be considered a viable indicator of phase correctness nor used as a selector of preferred maps. There is no theoretical reason why powder data sets should behave differently in this respect than those of proteins.

3.3. *Large phasing trees - a note of caution*

Given adequate computing resources, it is tempting to generate as many equivalent nodes as possible, particularly during overnight computing jobs. Under these circumstances, the significance level for the *t* test, normally set at 2%, must be greatly reduced. As a typical example for Mg_3BN_3, after origin definition, we permuted the phases of ten reflections to generate 1024 nodes. An analysis of significance at the 2% level rejected all but one of these, leaving a node with a log-likelihood gain of 4.319. The corresponding centroid map was very poor with less than 50% of the structure present. The analysis rejected the solution with the highest likelihood (12.95), which had an associated centroid map that showed the complete structure. The significance level needed to be reduced to 0.001% for the correct node to be included. In our experience, the significance level should be set in such a way as to keep at least eight nodes for further investigation. When this is done in the 1024 node calculation, the best nodes are indeed retained. However, in the interest of efficiency, it is recommended that no more than *ca* 128 equivalent nodes be generated and analysed at any step in the phasing procedure. The need for such an empirical rule is in keeping with the remark made in § 1 that we are determining relative (rather than absolute) levels of significance.

3.4. *The failure of the Gull-Livesey-Sivia algorithm*

The approach to the phase problem using ME methods outlined by Gull, Livesey & Sivia (1987) does not work here. This is not unexpected. The technique used in their approach is to define an origin (and enantiomorph if required) and use these phased reflections as constraints in an entropy-maximization procedure. One then examines the extrapolated structure factors and incorporates those with the largest magnitudes into the basis set using observed structure factors and the extrapolated phase. However, such a procedure traps the entropy maximization in a local optimum from which It becomes impossible to move and potentially incorrect phases are picked up (Gilmore, Bricogne & Bannister, 1990) - all the features of the origin-defining map become exaggerated. The process to collect strong extrapolates also adds nothing new to the calculation since the ME procedure can already predict them. It is the incorporation of those reflections of maximum surprise, *i.e.* $|U_h^{\rm ME}| \simeq 0.0$ while $|U_h|^{\rm obs}$ is large, that optimally enlarges the second neighbourhood of the basis set. This mode of operation is critical if maps are to develop new features as well as to correct wrong ones. It follows in a very natural fashion from the principle of phase sets being ranked according to likelihood.

All this can be seen quite clearly in the attempt to use the Gull-Livesey-Sivia method. The ME map based on a single origin-defining reflection gave three strong extrapolates, which were added to the basis set; two of these had incorrect phases. The next cycle produced one new extrapolate, which had an incorrect phase, and in the next cycle five out of the six strong extrapolates were wrongly phased. This process continued for five cycles, producing a final map that showed only a single large peak that corresponded to the position of an N atom.

3.5. *Computing aspects*

The phasing procedure described here uses more computer time than traditional direct methods. However, given the power of the new generation of workstations, this is not a significant problem for small structures. In addition, such computers are often networked using fast Ethernet connections, which often involve local subnets. Furthermore, there is a high degree of parallelism inherent in the calculations that involve phasing trees. Each node at a given level of phase permutation is an independent entity until the final statistical analysis is carried out, when all such nodes need to be collated. It is possible to exploit both the inherent parallelism of the ME computations and the networking facility simultaneously if the file containing the equivalent nodes is split into a set of individual files each containing a single node. These files and the *MICE* program are centrally mounted on a server (using, for example, the Network File System, NFS, on Unix-based computers or DECnet on DEC machines). A central program on the server farms out nodes to the workstations on the network, records the completion of the ME calculations on these nodes and sends further nodes as required. When all the calculations are complete, the results are collated into a single file. The granularity of such a procedure means that a network of *n* workstations of equivalent power will run the ME calculations *n* times faster than a single machine. The method works best on homogenous computer clusters but can be developed for heterogeneous networks as well, although a different version of the ME program is needed for each different computer architecture. We have developed such an arrangement on a local laboratory network of five SUN workstations (Sparcstation 2 or near equivalents) at Glasgow. With this system, the 129 nodes for Mg_3BN_3 were calculated in just over 34 min with a proportional increase to *ca* $4\frac{1}{4}$ h for the 1024 nodes. Although not as fast as traditional direct methods, ME methods can be competitive for powder structures, given the quality of the final solution.

4. Summary and concluding remarks

We have demonstrated the applicability of the use of combined entropy maximization and likelihood ranking to the determination of the crystal structure of Mg_2BN_3 from its X-ray powder diffraction data collected on a laboratory instrument. This procedure easily produced a map showing the positions of all the atoms in the asymmetric unit, including N and B. The only difficulties experienced concerned the small number of atoms in the unit cell, which gave rise to very large U magnitudes that, although tractable, required some precautions; this inessential complication shows, if anything, that the method has phasing power to spare. In addition, the application of the Student t test to analyse the log-likelihood gains automates the procedure and removes any inherent subjectivity present in the selection of nodes on the criterion of likelihood alone. Care is, however, needed with the significance levels when very large trees are generated. This technique has been applied here to a centrosymmetric structure but it can be readily extended to acentric reflections using quadrant phase permutation and analysis for the signs of both the real and the imaginary parts of the permuted structure factors. Indeed, the structure of formylurea, which crystallizes in space group $Pn2_1a$, has been solved in a routine way from its X-ray powder diffraction data with this method. Entropy alone again turned out to be a very poor indicator of phase correctness.

A further development is now under way to allow the incorporation of overlapped reflections into the basis set when required and to analyse phase permutations of these reflections in the same way using the formalism developed by Bricogne (1991*a*, § 6.3). Such a procedure has been developed as a program and is currently under test. Since the number of permuted hyperphase values can be much greater than that of ordinary phases, the parallelization of node evaluations described in § 3.5 will be of particular value to powder structure evaluation.

Other points arise concerning data quality, completeness and the use of reflection multiplicities when the intensity data are processed. To solve powder structures *ab initio*, the mode of processing and the data quality are of paramount importance. Weak reflections must be accurately measured and included with measured intensity even if considered unobserved. Furthermore, it is important that reflection multiplicities are correctly handled. *MITHRIL* and *MICE* both expect the intensity data from powders to be without multiplicity corrections; other packages will have different criteria. A common fault is to correct nonoverlapped reflections for multiplicity effects, but not to correct the overlaps. In *MICE*, the sophisticated statistical analysis that underlies the likelihood criterion is very vulnerable to such systematic inconsistencies in the data.

This work was supported by BP Research *via* an extra-mural research award and, in part, by the SERC. We are particularly grateful to Dr M. J. Tricker, Professor K. J. Packer, Dr D. White and Dr G. Duckett at BP for their continuing support. Stuart Mackay (Glasgow) provided invaluable assistance with the development of local computer networks to speed up the calculations. GB wishes to thank Trinity College, Cambridge, for the provision of extra working space.

References

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

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

Bricogne, G. (1991*a*). *Acta Cryst.* **A47**, 803-829.

Bricogne, G. (1991*b*). *Crystallographic Computing* 5, edited by D. Moras, A. D. Podjarny & J. C. Thiery, pp. 257-297. Oxford: Clarendon Press.

Bricogne, G. (1991*c*). *Maximum Entropy in Action*, edited by B. Bluck & V. A. Macaulay, pp. 187-216. Oxford: Clarendon Press.

Bricogne, G. (1993). *Acta Cryst.* **D49**, 37-60.

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

Cascanaro, G., Favia, L. & Giacovazzo, C. (1992). *J. Appl. Cryst.* **25**, 310-317.

Cochran, W. G. & Cox, G. M. (1957). *Experimental Designs*, 2nd ed. New York: Wiley.

David, W. I. F. (1990). *Nature (London)*, **346**, 731-734.

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.

Estermann, M. & Gramlich, V. (1992). Accuracy in Powder Diffraction II, Gaithersburg, Abstract 03.1.

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

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

Gilmore, C. J. & Bricogne, G. (1991). *Crystallographic Computing* 5, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 298-307. Oxford: Clarendon Press.

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

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

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

Gilmore, C. J., Henderson, K. & Bricogne, G. (1991*b*). *Acta Cryst.* **A47**, 830-841.

Gilmore, C. J., Shankland, K. & Fryer, J. R. (1993). *Ultramicroscopy.* In the press.

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

Gull, S. F., Livesey, A. K. & Sivia, D. S. (1987). *Acta Cryst.* **A43**, 112-117.

Hiraguchi, H., Hashizume, H., Fukunaga, O., Takenaka, A. & Sakata. M. (1991). *J. Appl. Cryst.* **24**, 286-292.

Jansen, J., Peschar, R. & Schenk, H. (1992). *J. Appl. Cryst.* **25**, 310-317.

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

LIGHTFOOT, P., TREMAYNE, M., HARRIS, K. D. M. & BRUCE, P. G. (1992). *J. Chem. Soc. Chem. Commun.* pp. 1012-1013.
MCCUSKER, L. (1988). *J. Appl. Cryst.* **21**, 305-310.
PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. & VETTERLING, W. T. (1986). *Numerical Recipes*, pp. 484-488. Cambridge Univ. Press.
SJÖLIN, L., PRINCE, E., SVENSON, L. A. & GILLILAND, G. L. (1991). *Acta Cryst.* A**47**, 216-233.
TREMAYNE, M., LIGHTFOOT, P., 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.

Acta Cryst. (1993). A**49**, 501-504

A Rotation Function with Increased Signal Size

BY JORDI RIUS AND CARLES MIRAVITLLES

CSIC, Institut de Ciència de Materials de Barcelona, 08193 *Cerdanyola-Bellaterra, Catalunya, Spain*

(*Received* 12 *May* 1992; *accepted* 15 *October* 1992)

Abstract

The most widespread application of the rotation function is the determination of the relative orientation of a given search fragment in the unit cell of an unknown crystal structure [Rossmann & Blow (1962). *Acta Cryst.* **15**, 24-31]. Here a modification is presented of the rotation function for this specific application, which exploits the information of the intensity data more effectively, thus leading to a higher signal size with the same computing cost.

1. Introduction

Although the primary phasing of intensity data from small equal-atom structures and from macromolecular compounds is normally carried out using either direct methods or the multiple isomorphous-replacement technique, molecular-replacement techniques are increasingly used when a suitable search fragment (or model) is available. Besides the crystal symmetry, the principal factors determining the success of molecular-replacement methods are, on the one hand, the size, the form and the accuracy of the search fragment and, on the other hand, the number and reciprocal-space distribution of the measured intensities. In general, the larger and more accurately known the fragment is, the less drastic are the requirements imposed on the intensity data.

As is well known, the real-space formulation of the rotation function of Rossmann & Blow (1962) for the case where a suitable search model is available is

$$R(\Omega)\propto \int_U P_o(\mathbf{u})P_{\rm model}(\Omega\mathbf{u})\,d\mathbf{u}. \qquad (1)$$

The integral in (1) measures the agreement of the Patterson function of the unknown crystal structure (P_o) with the rotated Patterson function of the isolated search model ($P_{\rm model}$) in a region U around the origin of the unit cell. The symbol Ω denotes a rotation operator that rotates the coordinate system of the search model with respect to that of the unknown crystal structure. $R(\Omega)$ will have a large maximum when the two Patterson functions are brought into maximum coincidence. If U corresponds to the whole unit cell, (1) can be expressed in reciprocal space as the summation

$$\sum_{\mathbf{H}}|F_o(\mathbf{H})|^2|S(\mathbf{H},\Omega)|^2, \qquad (2)$$

where $|F_o(\mathbf{H})|^2$ and $|S(\mathbf{H})|^2$ are, respectively, the Fourier coefficients of the observed and the model Patterson functions (Tollin & Cochran, 1964). $|S(\mathbf{H})|^2$ can be written in the form

$$|S(\mathbf{H})|^2=\sum_{j=1}^{n}\sum_{k=1}^{n} Z_jZ_k\cos(2\pi\mathbf{H}\cdot\mathbf{r}_{jk}) \qquad (3)$$

with n being the number of atoms of the fragment, Z_j being the atomic number of the jth atom and $\mathbf{r}_{jk}$ denoting the difference vector $\mathbf{r}_j-\mathbf{r}_k$, where $\mathbf{r}_j$ is the position vector of the jth atom referred to a fixed local origin.

Inspection of (2) reveals that the contribution to the $\mathbf{H}$ summation of those terms with small $|F_o(\mathbf{H})|^2$ values is not significant. Consequently, it seems reasonable to expect an increased signal size if the rotation function (2) is modified to include additionally the significant contribution of the weak reflections. In practice, this modification can be useful in those cases where only a small intensity data set is available, as is typical for low-resolution X-ray powder diffraction data of organic compounds (Rius & Miravitlles, 1988; Wilson & Wadsworth, 1990; Rius, Miravitlles, Molins, Crespo & Veciana, 1990; Amigó, Ochando, Abarca, Ballesteros & Rius, 1992). Owing to the reduced number of available intensities, the most difficult step is the rotation search. The subsequent fragment positioning is greatly simplified with the combined use of translation and packing functions (Harada, Lifchitz, Berthou & Jolles, 1981; Stubbs & Huber, 1991) as well as with the calculation of the R

0108-7673/93/030501-04$06.00