Acta Cryst. (1994). A**50**, 164–182

Overcoming Non-Isomorphism by Phase Permutation and Likelihood Scoring: Solution of the TrpRS Crystal Structure

BY SYLVIE DOUBLIÉ* AND SHIBIN XIANG

Department of Biochemistry and Biophysics, Campus Box 7260, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7260, USA

CHRISTOPHER J. GILMORE

Department of Chemistry, Glasgow University, Glasgow, Scotland

GÉRARD BRICOGNE

LURE, Université Paris Sud, 91405 Orsay, France, and Department of Molecular Biology, Biomedical Centre, Box 590, University of Uppsala, Uppsala, Sweden

AND CHARLES W. CARTER JR

Department of Biochemistry and Biophysics, Campus Box 7260, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7260, USA

(*Received* 14 *July* 1993; *accepted* 20 *September* 1993)

Abstract

Entropy maximization to maximum likelihood, constrained jointly by the best available experimental phases and by a sufficiently good envelope, can bring about substantial model-independent map improvement, even at medium (3.1 Å) resolution [Xiang, Carter, Bricogne & Gilmore (1993). *Acta Cryst.* D**49**, 193–212]. In the crystal structure determination of the *Bacillus stearothermophilus* tryptophanyl-tRNA synthetase (TrpRS), however, the following had to be dealt with simultaneously: (1) a serious lack of isomorphism in the heavy-atom derivatives, resulting in large starting-phase errors; and (2) an initially poorly known molecular envelope. Because the constraints – both phases and envelope – were insufficiently well determined at the outset, maximum-entropy solvent flattening as previously applied was unsuccessful. Rather than improving the maps, it led to a deterioration of their quality, accompanied by a dramatic decrease of the log-likelihood gain as phases were extended from about 5 Å resolution to the 2.9 Å limit of the diffraction data. This deadlock was broken by the identification of strong reflections, which were initially unphased and which were inaccessible by maximum-entropy extrapolation from the phased ones, and by permutation of the phases of these reflections so as to sample the space of possible electron-density and envelope modifications they represented. Permutation was carried out by successive full and incomplete factorial designs [Carter & Carter (1979). *J. Biol. Chem.* **254**, 12219–12223] for 28 strong reflections selected in decreasing order of their 'renormalized' structure-factor amplitudes. The permuted reflections included one reflection for which the probability distribution from multiple isomorphous replacement with anomalous scattering (MIRAS) indicated an incorrect phase with a high figure of merit and which consequently had a large renormalized structure factor. A similar permutation was carried out for six different binary choices related to the calculation and description of the molecular envelope. Permutation experiments were scored using the log-likelihood gain and contrasts for each main effect were analyzed by multiple-regression least squares. Student *t* tests provided significant and reliable indications for a large majority of the permuted reflections and for all six hypotheses related to the molecular envelope. The resulting phase improvement made it possible to assign positions (hitherto unobtainable) for nine of the ten selenium atoms in an isomorphous difference Fourier map for selenomethionine-substituted TrpRS crystals and hence to solve the structure. Phase-permutation methods continued to be useful in producing improved maps from all the available isomorphous-replacement phase information and therefore played a critical role in solving the structure. This process rescued phases for the tetragonal TrpRS structure (now solved) from an otherwise crippling lack of isomorphism. It represents the first application of a fully fledged Bayesian phase-determination process [Bricogne (1988). *Acta Cryst.* A**44**, 517–545] to the solution of an unknown structure and demonstrates the feasibility of using these methods with low-to-medium-resolution data.

* Present address: European Molecular Biology Laboratory, c/o ILL, 156X, 38042 Grenoble CEDEX, France.

1. Introduction

Tetragonal crystals of the *Bacillus stearothermophilus* tryptophanyl-tRNA synthetase (TrpRS) have resisted structure solution for more than a decade. The reasons for this recalcitrance lie in the fact that derivatization with most heavy-atom reagents leads to a lattice-preserving loss of isomorphism, which, in turn, has prevented us from interpreting difference Patterson maps involving the native amplitudes and hence from using amplitude differences for phase determination. Further substitution of previously derivatized crystals with additional heavy atoms, however, was found to yield doubly derivatized crystals that are considerably more isomorphous with singly derivatized crystals than any of them are with native crystals. Successive derivatization thus provided a strategy for solving the phase problem for TrpRS crystals. Unfortunately, even when the heavy-atom sites were identified and initial phases calculated, the phasing power was low for all derivatives and electron-density maps were of very poor quality because of the errors arising from the remaining non-isomorphism. These errors were reduced to a certain extent through the use of a maximum-likelihood phase-refinement program (Otwinowsky, 1991) but this still did not produce an interpretable map.

Our previous experience (Xiang, Carter, Bricogne & Gilmore, 1993) with maximum-entropy solvent flattening (MESF) led us to expect that this procedure would improve the map sufficiently to produce an interpretable electron-density map. However, in the absence of a reliable envelope, the MESF process was not sufficiently constrained to produce an interpretable map from initial phases of such poor quality. Attempts to use an envelope previously derived by *ab initio* phase determination for structure-factor amplitudes of its indicator function obtained by X-ray contrast variation (Carter, Crumley, Coleman, Hage & Bricogne, 1990) were hindered by the fact that the correct orientation of this envelope was not apparent from cross-rotation functions, so it could not be positioned correctly in the tetragonal unit cell. We were thus left with the task of correcting phase errors while at the same time defining the molecular envelope directly from electron-density maps calculated with poor-quality phases.

This difficulty was eventually overcome by subordinating the MESF process to a phase-permutation process designed to sample simultaneously different phase choices for a number of reflections with unavailable or unreliable MIRAS phases and by monitoring the log-likelihood-gain (LLG) statistic. Previous work had shown the LLG to be a powerful criterion for comparing hypotheses about unknown phases for a small protein structure (Gilmore, Henderson & Bricogne, 1991) and this experience encouraged us to apply the method in the current context of an unknown structure with a limited amount of phase information. In the course of the successful application of this process to phase determination, we found that it was equally effective when we permuted hypotheses about possible modifications of the molecular envelope in the same way (Bricogne, 1988*a*, §2.3), in the sense that the LLG score also provided statistically significant indications to guide the progressive editing of the unknown molecular envelope which ultimately proved to be correct. The resulting phase improvement allowed us first to identify selenium positions in new isomorphous-replacement data from selenomethionyl-TrpRS crystals and subsequently to achieve convergence of the electron density phased by all available derivatives to the correct structure.

Our experience is documented in the rest of this paper and our success provides the first application of the Bayesian phasing methods described by one of us (Bricogne, 1988*a*,*b*, 1991*b*, 1993) to the solution of a new macromolecular crystal structure for which previously available methods were insufficiently powerful.

2. Data and methods

2.1. *Primary sources of phase information*

Tetragonal ($P4_32_12$) crystals of *Bacillus stearothermophilus* TrpRS were grown in 2.1 *M* K_2HPO_4 at pH = 7.5 with 0.0002 *M* tryptophan and 0.01 *M* ATP, as previously described (Carter & Carter, 1979; Carter & Coleman, 1984; Carter, Doublié & Coleman, 1994; Coleman & Carter, 1984), and stabilized for X-ray data collection by soaking in 3.55 *M* $(NH_4)_2SO_4$ with the same ligands. Under these conditions, native crystals have unit-cell dimensions $a = b = 60.6$ and $c = 232.7$ Å, with a monomer of 328 residues ($M_r \simeq 37000$) in the asymmetric unit, and gold-substituted crystals have unit-cell dimensions $a = b = 60.7$ and $c = 232.9$ Å. Heavy-atom substitution therefore preserves the unit-cell dimensions.

Native and derivative data sets for TrpRS are compared in Table 1. They were measured using a multiwire area detector and a rotating anode (one native, 4PAR, and one mersalyl derivative, Hg4) and with phosphor image plates, either with synchrotron radiation (LURE, Orsay, France; one native, NAT2, one crystal soaked in 0.001 *M* gold chloride, Au1, and one double derivative prepared by soaking the gold derivative in 0.0001 *M* mersalyl acid, AuHg3) or with a conventional rotating-anode source (University of Massachusetts Medical Center, Worcester, MA, USA; one native, IVN1, and three selenomethionyl-TrpRS crystal data sets, SMT1, SMT2 and SMT4). Despite the fact that the crystals themselves diffract to about 1.7 Å with synchrotron radiation, the crystal-to-film distance necessary to resolve reflections along $\mathbf{c}^*$ dictated that all data sets measured from a single crystal stop at about 2.9 Å because both image-plate instruments were constrained to a θ angle of 0° and the multiwire

Table 1. *Correlation coefficients (on F) between TrpRS data sets*

Data sets are grouped into distinct sources of phase information, including native crystal data sets (IVN1, 4PAR and NAT2), conventional isomorphous derivatives (Au1, AuHg3 and Hg4) and selenomethionyl-TrpRS data sets (SMT1, SMT2 and SMT4). All data sets are of high quality, with R_{symm} values ranging from 0.044 (NAT2) to 0.079 (Au1). Correlation coefficients for different groups are emphasized by typeface. Those between conventional heavy-atom and native data sets are *italic*, those between the conventional heavy-atom data sets themselves are **bold** and those between native and selenomethionyl-TrpRS are ***bold italic***.

	IVN1	NAT2	4PAR	Au1	AuHg3	Hg4	SMT1	SMT2	SMT4
IVN1	1.0	0.98	0.95	*0.51*			***0.93***	***0.90***	***0.93***
NAT2		1.0	0.99	*0.58*	*0.67*	*0.55*			
4PAR			1.0	*0.66*					
Au1				1.0	**0.94**	**0.77**			
AuHg3					1.0	**0.81**			
Hg4						1.0			
SMT1							1.0	0.97	
SMT2								1.0	0.93
SMT4									1.0

detectors have too small a solid-angle coverage to allow the collection of complete data sets from a single crystal at high θ angle.

Two qualitatively different sources of isomorphous-replacement information for TrpRS are represented in Table 1. The first involved the three conventional heavy-atom-derivative data sets, one of them (Au1) being used as the parent. The second involved the much more nearly isomorphous selenomethionyl-TrpRS data sets collected from crystals of protein produced by a strain auxotrophic for methionine and grown in a medium containing selenomethionine (Doublié & Carter, 1992, 1993). In order to exploit the latter isomorphous differences for phase determination, it was necessary to locate the selenium atoms. As noted below, location of these atoms in difference Fourier maps calculated with improved phases obtained by the process of phase permutation proved to be a crucial step in solving the structure.

2.2. *Bayesian phasing methods and the role of entropy maximization*

In the present context, the Bayesian scheme (Bricogne, 1988*a*, 1993) for extracting missing phase and envelope information from the available structure-factor amplitudes consists of: (*a*) generating multiple hypotheses about the missing information so as to form a representative sample of all available alternatives (*permutation*); (*b*) evaluating the degree of corroboration of each hypothesis by the observed data, as measured by its likelihood (*scoring*); (*c*) combining the *a priori* probability of each hypothesis with its likelihood, by means of Bayes's theorem, to obtain its *a posteriori* probability; (*d*) making decisions, on the basis of these *a posteriori* probabilities, about which hypotheses should be rejected and which should be expanded further (*statistical inference*).

The book-keeping of the multiple hypotheses thus generated is carried out by means of a *phasing tree* (Bricogne, 1984, §8.1; Bricogne & Gilmore, 1990, §3.1), whose *nodes* represent the unique sets of individual hypotheses and whose links reflect the parentage relationships between them.

At each stage of the phase determination, the symmetry-unique non-origin reflections are divided into two sets: a *basis set* $\{H\}$ consisting of those reflections for which explicit phase assumptions have been made; and its complementary set $\{K\}$ of *non-basis* reflections for which only unphased or poorly phased amplitudes are available. Initially, the members of $\{H\}$ were chosen according to their MIRAS figure of merit. In this work, we used a value of 0.6, representing a mean phase error of 53°. This value was considerably greater than the corresponding mean phase error of 30° used previously (Xiang, Carter, Bricogne & Gilmore, 1993) and it reflects the fact that the initial phases were of poorer quality and that in order to define a sufficiently large basis set we had to use a lower threshold.

The optimal evaluation of both the *a priori* probability and the likelihood of a hypothesis in which phase values and a molecular boundary are specified is intimately related to the numerical process of constrained entropy maximization. If $m(\mathbf{x})$ denotes the uniform probability distribution within the protein region, the goal is to construct an exponential model, $q^{\mathrm{ME}}(\mathbf{x})$, for the distribution of atoms:

$$q^{\mathrm{ME}}(\mathbf{x}) = [m(\mathbf{x})/Z(\boldsymbol{\zeta})] \times \exp\Big[\sum_{\mathbf{h}\in H} \zeta_{\mathbf{h}} \exp(-2\pi i \mathbf{h}\cdot\mathbf{x})\Big], \quad (1)$$

by adjusting the complex parameters $\{\zeta_{\mathbf{h}}\}$ until the Fourier transform of $q^{\mathrm{ME}}(\mathbf{x})$, $\{U^{\mathrm{ME}}\}$, matches the unitary structure-factor amplitudes and phases of reflections in $\{H\}$. This process of 'exponential modeling' produces a probability distribution for the random atomic positions in the unit cell that has maximum relative entropy, $S = -\int q(\mathbf{x})\log[q(\mathbf{x})/m(\mathbf{x})]\mathrm{d}^3\mathbf{x}$, with respect to the initial density $m(\mathbf{x})$, subject to the constraints in $\{H\}$ and to that of solvent flatness. We have previously described the various algorithms involved in solving the maximum-entropy (ME) equations under the constraint of solvent flatness (Bricogne, 1988*a*, 1991*b*) using the computer program *MICE* (Bricogne & Gilmore, 1990). To impose the envelope constraint, we used an approximation in which the solvent regions are reset to their average value on each cycle of fitting the exponential model to the basis-set structure-factor constraints (Xiang, Carter, Bricogne & Gilmore, 1993). While this approximation results in a weaker algorithm than a full-blown multichannel entropy maximization, it was shown to be very effective with both simulated and experimental data and was used in this work without significant modification.

In the Bayesian method of phase determination, the role of this entropy maximization is twofold. First, it yields the numerical value of the least entropy loss

that has to be accepted in order for the constraints attached to each node to be fitted; this is related to the *a priori* probability of the hypothesis associated with this node by virtue of the link with the saddlepoint method established by Bricogne (1984). Second and most important, it creates an interaction between the basis and non-basis reflections through the phenomenon of maximum-entropy extrapolation: besides reproducing the unitary amplitudes and phases attached to each node for reflections in $\{H\}$, the maximum-entropy distribution has Fourier coefficients with non-negligible amplitudes $|U_{\mathbf{k}}^{\mathrm{ME}}|$ for many non-basis reflections $\mathbf{k}$ in $\{K\}$ and - crucially - these amplitudes depend on the phase and envelope information attached to the corresponding node. These extrapolated Fourier coefficients have an immediate interpretation in the sense that the conditional probability of structure factors in $\{K\}$ is a multivariate Gaussian centered around these values. The likelihood of the hypothesis represented by a node can then be calculated by integrating this conditional probability over the unknown phases of the reflections in $\{K\}$ over the circle with radius equal to the unitary amplitude $|U_{\mathbf{k},\mathrm{obs}}|$ known from the experimental data. This is a node-dependent score that reflects the ability of the corresponding hypothesis to assign a high probability to the measurements in $\{K\}$ before knowing about them.

The value actually used to score phase-permutation experiments is the increase in conditional probability of the data over that obtained from the null hypothesis that the atomic positions are distributed uniformly in the cell and hence that the data follow Wilson statistics. The logarithm of the ratio of these two probabilities is called the log-likelihood gain. Summing over all reflections gives the global log-likelihood gain (Bricogne, 1988*b*; Bricogne & Gilmore, 1990; Xiang, Carter, Bricogne & Gilmore, 1993):

$$\begin{aligned} L(U^K) = &\sum_{\mathbf{k}\ \mathrm{acentric}\in K} \Big\{ \log I_0\Big[(2N/\varepsilon_{\mathbf{k}})\big|U_{\mathbf{k}}^{\mathrm{obs}}\big|\big|U_{\mathbf{k}}^{\mathrm{ME}}\big|\Big] \\ &- N/\varepsilon_{\mathbf{k}}\big|U_{\mathbf{k}}^{\mathrm{ME}}\big|^2 \Big\} \\ &+ \sum_{\mathbf{k}\ \mathrm{centric}\in K} \log\Big\{ \cosh\Big[(N/\varepsilon_{\mathbf{k}})\big|U_{\mathbf{k}}^{\mathrm{obs}}\big|\big|U_{\mathbf{k}}^{\mathrm{ME}}\big|\Big] \\ &- N/2\varepsilon_{\mathbf{k}}\big|U_{\mathbf{k}}^{\mathrm{ME}}\big|^2 \Big\}. \end{aligned} \tag{2}$$

The log-likelihood gain has a privileged status among statistics of phase choices, by virtue of the Neyman-Pearson theorem, which states that the LLG is a 'most powerful' indicator of the relative correctness of the model, in the sense that it is minimally vulnerable to statistical errors of the second kind (acceptance of the null hypothesis when it should be rejected) at any given level of exposure to errors of the first kind (rejection of the null hypothesis when it should be affirmed) (Bricogne, 1991*a*). Previous tests have demonstrated the superior ability of the LLG to identify correct phase sets from among a large number generated for a small protein structure by conventional direct methods (Gilmore, Henderson & Bricogne, 1991).

In the present work, we do not make use of Bayes's theorem, but use the LLG alone as it tends to dominate in any case and because the relative weighting of entropy loss and LLG entails delicate considerations in view of the non-independence of atoms at non-atomic resolution (Bricogne, 1993, §1.3).

2.3. *Electron-density maps*

At any stage of phase determination, the appropriate representation of the electron density is obtained using centroid estimates for the structure factors in $\{K\}$. These are derived from the first moments of the conditional probability distributions by combining the observed amplitudes with Sim-like weights:

(i) for $\mathbf{k}$ acentric

$$\begin{aligned} \langle U_{\mathbf{k}}\rangle &= \big|U_{\mathbf{k}}^{\mathrm{obs}}\big|[I_1(X_{\mathbf{k}})/I_0(X_{\mathbf{k}})]\exp\big(i\varphi_{\mathbf{k}}^{\mathrm{ME}}\big) \\ X_{\mathbf{k}} &= (2N/\varepsilon_{\mathbf{k}})\big|U_{\mathbf{k}}^{\mathrm{obs}}\big|\big|U_{\mathbf{k}}^{\mathrm{ME}}\big|; \end{aligned}$$

(ii) for $\mathbf{k}$ centric (3)

$$\begin{aligned} \langle U_{\mathbf{k}}\rangle &= \big|U_{\mathbf{h}}^{\mathrm{obs}}\big|\tanh(X_{\mathbf{k}})\exp\big(i\varphi_{\mathbf{k}}^{\mathrm{ME}}\big) \\ X_{\mathbf{k}} &= (N/\varepsilon_{\mathbf{k}})\big|U_{\mathbf{k}}^{\mathrm{obs}}\big|\big|U_{\mathbf{k}}^{\mathrm{ME}}\big|. \end{aligned}$$

Fourier synthesis with these structure factors then gives a 'centroid' electron-density map analogous to the classical centroid map from multiple-isomorphous-replacement phasing.

2.4. *Expansion of an insufficient basis set via phase permutation*

As described below, we determined at a rather early stage of map improvement that our basis sets were insufficient, either because of phase errors or because of uncertainties in the envelope constraint. Phase permutation turned out to be an effective answer to this problem: by exploring untested regions of structure-factor space, we found a reliable path toward better ME extrapolation (increased LLG) and hence towards improved maps. Two requisites for expanding the basis set should be noted. First, appropriate reflections must be identified from $\{K\}$ that will have the greatest impact on the LLG when recruited into the basis set. Second, some means is necessary to sample the many possible phase combinations as efficiently as possible. Fortunately, the necessary tools for dealing effectively with both tasks have been developed.

We had previously learned (Bricogne & Gilmore, 1990, §3.2.7) that examining the pattern of ME extrapolation, or more properly the lack thereof, is the key to making the best choice of reflections to permute. Reflections from $\{K\}$ most likely to change the LLG

are those for which the measured amplitude is strong but ME extrapolation is weak and those that possess strong coupling to other similar reflections by triplet and quartet phase relations, so that they will expand the second neighborhood of the basis set by the greatest number of strong LLG contributors. A procedure (*COBWEB*) for carrying out this analysis has been described elsewhere (Bricogne, 1993). *MICE* provides an approximate implementation of this procedure in a function, *NEXT*, to identify such reflections. However, our basis sets were all such as to include essentially the entire data set in their second neighborhoods, and we initially selected reflections simply from those strongly observed reflections with the largest discrepancy between $|U^{\rm obs}|$ and $|U^{\rm ME}|$. Later, we made this 'maximum surprise' criterion more quantitative by identifying reflections on the basis of the estimated vector difference between $U^{\rm obs}$ and $U^{\rm ME}$, using the renormalized structure-factor expression (Bricogne, 1992):

$$|U^{\rm renorm}| = \Big(|U^{\rm obs}|^2 + |U^{\rm ME}|^2 - 2|U^{\rm obs}||U^{\rm ME}|{\rm Wt}_{\rm sim} \Big)^{1/2} \qquad (4)$$

In this expression, the average cosine of the angle between $U^{\rm obs}$ and $U^{\rm ME}$ is estimated by the same Sim-like weight, $\mathrm{Wt}_{\rm sim}$, that is used in (3).

The direct-phasing problem in crystallography is perhaps the quintessential 'factorial experiment'. One wishes first to determine simultaneously the effects on the electron-density map of a large number of factors (individual reflections) that can, *a priori*, assume many different values or levels (phases). Then, one wishes to infer from these sets of effects which set of level (phase) choices leads to the correct map. The factorial nature of the phase problem, and its relationship to the task of phase permutation in particular, was recognized by Woolfson (1954) and Good (1954). The problem of efficiently sampling large volumes of structure-factor space has been reconsidered from the viewpoint of error-correcting codes by one of us (Bricogne, 1993). In principle, it is possible to exploit the periodic nature of structure-factor phases with 'magic lattice' sampling designs based on coding theory. Factorial experimental design has also been a primary focus of our efforts to screen for crystal-growth conditions for over a decade (Carter & Carter, 1979; Carter, Baldwin & Frick, 1988; Carter, 1990, 1992). In the present work, we have successfully applied these same methods to the problem of phase permutation. We used incomplete factorial designs, rather than magic lattices, for two reasons. First, the number of nodes required for magic-lattice permutations was high enough to prohibit their use on a problem of this size with the computing resources available to us. Second, the sampling efficiency of magic lattices - which were designed for *ab initio* phasing on account of their ability to preserve high-order interactions - was less critical here, since enough phase information had already been developed for the main effects of phases to be the dominant factors.

Table 2. *16-node incomplete factorial design for four centric and three acentric reflections*

Reflections were permuted as described in the text, with quadrant permutation for the acentric reflections. Phase choices are given as the number of the quadrant, beginning at 45 (1) and continuing counterclockwise. This permutation was carried out with basis-set phases derived from phasing group I.

Node	427	5.1.11	614	1.0.25	2.2.19	400	443	LLG
37 (Parent)	0	0	0	0	0	0	0	3000
38	3	2	4	1	1	2	1	2966
39	4	3	4	2	1	1	1	3151
40	1	4	4	1	2	2	2	3111
41	2	4	1	1	1	2	1	3042
42	3	1	1	2	2	1	2	2990
43	1	3	1	2	1	1	1	3140
44	2	2	1	1	2	2	2	2939
45	4	2	2	2	2	1	2	3017
46	3	3	3	2	2	1	2	3010
47	4	1	4	2	1	2	2	3026
48	2	3	2	1	2	2	1	2982
49	1	1	2	1	1	1	1	3009
50	1	2	3	1	1	2	1	3040
51	2	1	3	2	1	1	1	2957
52	3	4	2	2	2	1	2	3026
53	4	4	3	1	2	2	2	3085

Table 3. *24-node incomplete factorial design for four acentric and three centric reflections*

Reflections were permuted as described in the text, with quadrant permutation for the acentric reflections. Phase choices are given as the number of the quadrant, beginning at 45 (1) and continuing counterclockwise. This design was used for nodes 54-122 with basis-set phases from phasing group I and, as shown, for nodes 115-138 with basis-set phases from phasing group III.

Node	6.4.48	7.4.41	10.9.26	9.8.28	2.0.54	0.0.48	9.0.35	LLG
114 (Parent)	0	0	0	0	0	0	0	2109
115	1	3	3	1	1	1	1	2198
116	3	2	3	3	1	2	1	2087
117	2	3	1	4	1	1	2	2064
118	4	1	4	2	1	2	2	2149
119	3	4	4	1	2	2	2	2100
120	2	1	2	3	2	1	1	2034
121	4	4	2	4	2	2	1	2138
122	1	2	1	2	2	1	2	2057
123	3	2	1	3	1	2	1	2063
124	4	1	1	1	1	1	1	2198
125	2	3	4	1	2	1	2	2040
126	1	2	4	4	1	2	2	2128
127	4	3	1	2	2	2	1	2126
128	2	2	3	1	1	1	1	2145
129	1	4	2	2	1	1	2	2116
130	3	1	4	3	2	1	2	2019
131	3	3	3	2	2	2	2	2019
132	1	1	2	4	2	1	1	2113
133	2	1	3	4	2	2	1	2078
134	4	4	3	3	2	2	2	2063
135	3	4	1	4	1	2	1	2120
136	4	2	2	1	2	1	2	2096
137	1	3	2	3	1	1	1	2102
138	2	4	4	2	1	2	2	2104

In an incomplete factorial design, one performs a predetermined number of individual tests, each of which is represented here by a node of the phasing tree, for

Table 4. *Permutation of envelope attributes*

Six different envelopes were prepared with different binary combinations of five different attributes. The permutation was carried out using basis-set phases from phasing group I. 'Mode' refers to the choice of using the envelope directly from the map-editing program or smoothing it by Fourier transformation and spectrum truncation to lower resolution. V_{SOLV} refers to the volume of the solvent, or the stringency of the envelope. The $INCL_{n\ m}$ refer to three regions, indicated by z sections, that were either omitted or included in the envelope as illustrated in Fig. 6.

Node	Mode	V_{SOLV}	$INCL_{6\ 10}$	$INCL_{15\ 20}$	$INCL_{10\ 11}$	LLG
79 (Parent)	0	0	0	0	0	2525
80	1	1	1	1	2	2504
81	2	2	2	2	1	2561
82	2	2	1	1	1	2539
83	1	1	1	2	1	2479
84	1	2	2	1	2	2659
85	2	1	2	2	2	2454

which factor levels are chosen randomly subject to two criteria (Carter & Carter, 1979; Carter, 1992). First, main effects should be balanced, *i.e.*, for each factor, the numbers of nodes evaluated at each factor level should be the same. Second, each two-factor interaction should be as balanced as possible, *i.e.*, for factors taken as pairs, each combination of two levels should be tested with nearly the same number of nodes. The factors in this case are reflections to be permuted and their levels are phase choices. Designs used in this work (the pattern of phase choices for each reflection associated with each node) were generated directly using the program *INFAC* (Carter, 1990). 'Envelope permutation' was carried out in the same fashion, making choices for the inclusion or exclusion of discrete regions of the map and systematically permuting these or other binary choices according to an incomplete factorial design, with likelihood scoring and analysis. Examples of the designs used are given, together with the LLG scores for each 'experiment' in the design, in Tables 2, 3 and 4.

A typical scheme for the permutation experiments we carried out is illustrated in Fig. 1. Centric reflections were tested at their two permitted values and acentric reflections were always permuted at four 'quadrant' positions: 45, 135, 225 and 315°. The experiment described in Fig. 1 was performed using the native amplitudes after all phase information had been combined using *MLPHARE* (Otwinowsky, 1991) and when it was apparent that further map improvement was necessary. From the full factorial design of 2048 nodes, we selected 24 according to the design in Table 3. The analysis for the scores in Table 3 is given in Table 7.

2.5. *Significance testing*

The LLG statistic is actually akin to an experimental measurement, whose value fluctuates as does a random variable. Node-to-node variation of the LLG has many causes. Some are under our control (*i.e.* the values of permuted phases or choices of new features of the molecular envelope) and give rise to a 'signal'. Others are less easy to control (*e.g.* the exact composition of the basis set, the precise stopping point of the exponential modeling and the values of various parameters in the fitting, which cannot always be maintained at precisely the same values for all nodes that are to be compared) and their contribution is akin to 'noise'. Furthermore, it must be recalled that the LLG is calculated from a single sample of observations, whereas its statistical properties of optimality and freedom from biases are guaranteed when it is viewed over an ensemble of samples drawn from the same conditional distribution. This intrinsic fluctuation in the LLG values associated with different phase choices implies, in turn, that inferences about the optimal values for permuted phases can only be drawn in a statistical sense, by comparing LLG averages for all nodes with the same phase for a given reflection. Under these circumstances, it is appropriate to employ standard significance tests, such as the Student t test, to assess the validity of inferences drawn from these averages, relative to the residual noise in the LLG function (Bricogne, 1993, §2.2.4).

We used t tests to evaluate the significance of a quantity called the *contrast*, which is the average difference between those nodes for which the sign bit was (+) and

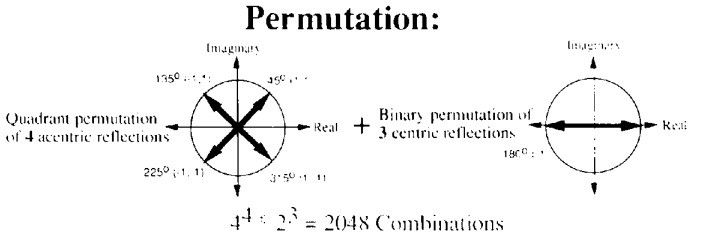


Fig. 1. Experimental paradigm for statistically sampled phase permutation, likelihood scoring and significance testing. A complete factorial experiment involving permutation of four acentric and three centric reflections would involve 2^{11} or 2048 combinations or nodes. Each node has as a basis set the previous basis set (roughly 1500 reflections in this work) plus specific phase choices for the seven permuted reflections. All nodes were subjected to maximum-entropy solvent flattening and scored using the resulting LLG. Scores were analyzed by multiple regression and analysis of variance. Specific illustrations are provided in §3. Full designs corresponding to the permutations illustrated in Table 6 (nodes 54–78) and in Table 7 (nodes 115–138) were sampled with 24 nodes, selected according to an incomplete factorial design. The design matrix for the permutation illustrated in Table 7 is given, together with the LLG scores for each node, in Table 3.

the overall mean for all nodes. Analyses of variance were carried out in conjunction with least-squares multiple-regression analysis using the *MGLH* module of the commercial program *SYSTAT* for the Apple Macintosh (Wilkinson, Hill & Vang, 1992). An appropriate regression model for LLG_{calc} was first selected by using all main effects as a point of departure and evaluating different subsets of coefficients by stepwise multiple regression. Student t tests for individual sign bits were then estimated simultaneously for that model by multivariate regression of the calculated *versus* the observed LLG scores.

A second important point concerns the transposition of the data required for the statistical analysis. The incomplete factorial designs were all constructed by balancing the quadrant permutations of acentric reflections, rather than by separately balancing their real and imaginary components. This decision was confirmed by test experiments in which designs were constructed with two two-level factors for each acentric reflection. In general, analyses of variance for these experiments were less satisfactory than those described here. A possible reason for this is that, because the sampling is so sparse, there is some advantage to balancing the interactions between the complex values of the acentric reflections (*i.e.* their phases) rather than the interactions between their separate real and imaginary components. In full-factorial or magic-lattice designs, this advantage might be less important.

In order to carry out the analysis of the phase permutation designs, it was therefore first necessary to recast the experimental matrices into the fundamental degrees of freedom by giving each acentric reflection two degrees of freedom, one for each of the real and imaginary components, giving each acentric reflection two bits of information. This was done as indicated by the coordinate values beside each quadrant phase in Fig. 1, *i.e.* (1, 1) for 45°; (−1, 1) for 135° and so on, as suggested elsewhere (Bricogne, 1993). This transformation is similar to that described previously, in which different attributes of crystal-growth factors are balanced with each other in the experimental design but are treated as separate binary factors in the subsequent analysis of crystal-growth screening experiments (Carter, 1990, 1992).

3. Results

3.1. *The quality of the initial MIRAS phases*

The quality of phases derived from amplitude differences can be assessed from the difference Patterson maps, as these represent in its entirety the signal owing to the heavy atom in the presence of noise and of errors arising from lack of isomorphism. The sum difference Patterson (Blundell & Johnson, 1976) for the difference amplitudes (AuHg3 - Au1), which was the first of all those we obtained to be interpreted and whose solution ultimately gave rise to the structure determination, is shown in Fig. 2. This derivative proved to have only a single mercury site. The highest peak has a height of about 9.1σ. However, that peak is inconsistent with the mercury site and therefore represents, along with numerous other inconsistent peaks of similar magnitude, the noise level arising from non-isomorphism. The strongest peaks corresponding to the correct solution rank 8th, 11th and 12th overall and have heights of only about 4.5σ. They stand out in only one of the three Harker sections (w = 1/4). In the other two sections, the corresponding peaks are close to the noise level and the section u = 1/2 is an unmitigated disaster. This difference Patterson was solved only by persistent manual examination, after numerous failures with automatic Patterson search and superposition algorithms (Terwilliger, Kim & Eisenberg, 1987). The difficulty experienced in solving the Patterson map for this, the best single-isomorphous pair, is indicative of the marginal quality of

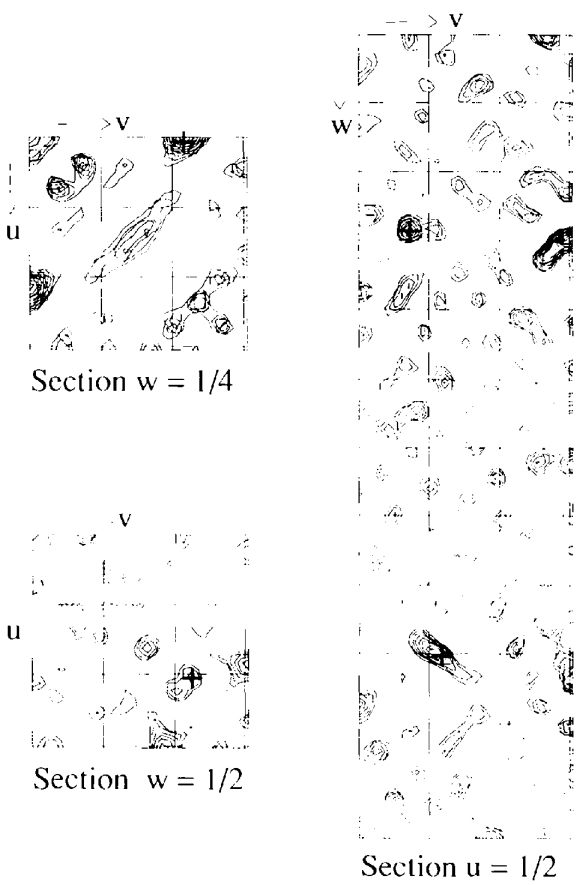


Fig. 2. Harker sections for the combined (isomorphous differences plus anomalous differences) Patterson map for the difference amplitudes (|AuHg3| - |Au1|; double - gold). All sections are contoured from 2σ in intervals of 0.5σ. The consistent peaks for the correct solution, indicated on each section by the symbol +, are at a height of about 3.5σ. The highly corrupted section u = 1/2 is characteristically noisy in all difference Patterson maps obtained for tetragonal TrpRS crystals.

the initial phases from which the structure was ultimately obtained.

3.1.1 *Origin determination for heavy-atom sites.* A common origin for mercury and gold sites was derived by solving the AuHg3 anomalous difference Patterson map, once the mercury site had been determined from the sum difference Patterson, and then identifying cross peaks between gold and mercury sites. A consistent set of Harker peaks in the AuHg3 anomalous difference Patterson map (4.3σ, 3.1σ and 2.4σ relative to a maximum height of 5.8σ) that were unaccounted for by the mercury site (whose peaks were 4.3σ, 3.1σ and 2.7σ) also appeared in the Au1 anomalous difference Patterson map and were subsequently interpreted in terms of the stronger of two gold sites. Difference Fourier maps were also used to confirm the common origin for all sites and to identify the known mercury site plus a weaker secondary mercury site in the Hg4 derivative data set.

3.1.2. *Fixing the enantiomorph.* The use of anomalous differences for primary phase determination necessarily incorporates the assumption that the correct choice has been made for the space-group enantiomorph. In order to fix the enantiomorph, it was useful to recruit the native amplitudes into the phase calculation. Isomorphous difference Patterson maps involving the NAT2 amplitudes, though considerably worse than that shown in Fig. 2, nevertheless had consistent Harker peaks corresponding to the gold and mercury positions. Atomic-parameter refinements to isomorphous differences (Au1 - NAT2) and (AuHg3 - NAT2) for the previously identified sites were initially carried out with *REFINE* for centric reflections only. For each derivative, single isomorphous replacement (SIR) and single isomorphous replacement with anomalous scattering (SIRAS) phases in each enantiomorphic space group were then calculated with *PHARE* (SERC Daresbury Laboratory, 1990) and the peak heights were compared for isomorphous difference Fourier maps calculated using the difference amplitudes for one derivative and the phases from the other derivative (Blundell & Johnson, 1976). For both non-isomorphous derivatives, the peak heights were consistently highest using the SIRAS phases calculated in space group $P4_32_12$. This choice was subsequently confirmed: first by the presence of peaks, rather than holes, in the anomalous difference Fourier maps for both derivatives using multiple isomorphous replacement (MIR) phases from both non-isomorphous derivatives, and finally by the clear presence of right-handed α-helices in the electron-density maps.

Atomic parameters in both isomorphous and non-isomorphous contexts were next refined with *MLPHARE* (Otwinowsky, 1991). Considerable improvement in the phases could be detected with parameters refined by the maximum-likelihood algorithm in this program. This improvement was evident in the enhanced contrast between protein and solvent regions and was essential for

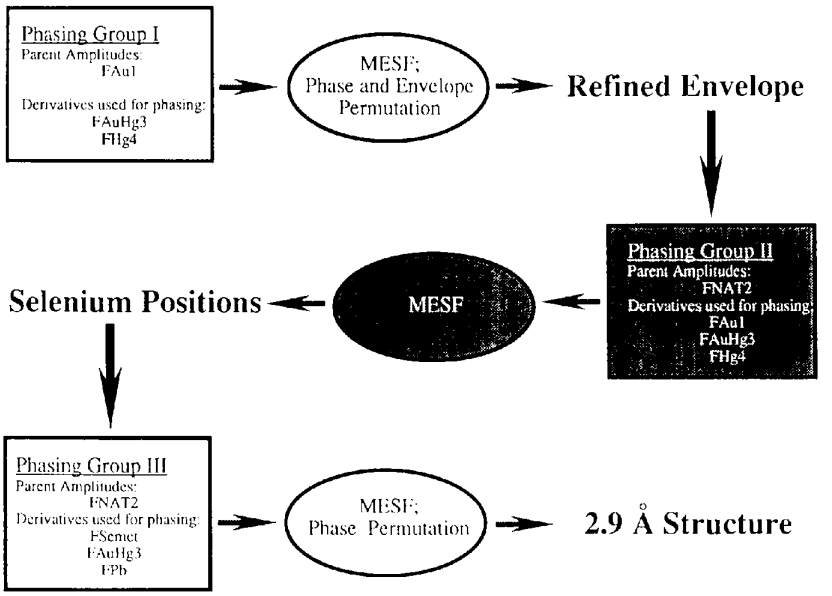


Fig. 3. Overall strategy used to solve the tetragonal *Bacillus stearothermophilus* TrpRS structure. Three different sets of multiple-isomorphous-replacement phases were calculated, as indicated by the three phasing groups. The principal results obtained for each set are indicated in bold typeface at the appropriate end of each line. Phasing groups I (nodes 1-112) and III (nodes 115-139) involved essentially isomorphous comparisons in phase calculation; phasing group II, indicated by shading, involved the highly non-isomorphous comparisons between the heavy-atom derivatives Au1, AuHg3 and Hg4 and the NAT2 native amplitudes. Information transfer from phasing group I to phasing group II was mediated by the envelope and the heavy-atom positions but not the phases. The AuHg3 data set was incorporated into phasing group III because it was the most nearly isomorphous of the three derivatives from phasing group I, as indicated by its correlation coefficient with native amplitudes, and because including it noticeably improved the map. The third derivative, Pb, was obtained after most of this work had been completed as described in the text.

a satisfactory definition of a starting molecular envelope. At this stage, electron-density maps for both sets of phases were compared and it was evident that the best protein map was that calculated using Au1 amplitudes and MIRAS phases from the isomorphous subset of heavy-atom data sets, AuHg3 and Hg4.

3.2. *Overall phasing strategy*

As the correlation coefficients in Table 1 and the difference Patterson in Fig. 2 suggest, we were faced with a plethora of potentially useful but relatively inaccessible phase information (§2.1). Heavy-atom derivatives were only marginally isomorphous with each other, at best, and were severely non-isomorphous with the native crystals. The strategy that finally enabled us to solve the structure is illustrated in Fig. 3. We made use of MESF constrained by phases from three separate 'phasing groups'. Considerable effort was devoted to the use of phasing group I to refine heavy-atom postitions and determine the molecular envelope. Using this envelope, we were able with phasing group II to obtain sufficiently good native phases to locate the selenium atoms in difference Fourier maps. In phasing group III, from which the structure was finally solved, the envelope and the selenium positions were supplemented with data from one of the heavy-atom data sets (AuHg3) and an additional lead acetate derivative. Phasing group III was sufficiently complete to initiate a convergent MESF process which led to the full determination of the structure. In the following section, we describe the role of MESF for each of the phasing groups.

3.3. *Maximum-entropy solvent flattening and phase permutation*

Electron-density maps calculated using (MIRAS) phases from phasing group I were of dubious quality and, although in retrospect the helix and sheet regions of the TrpRS molecule could have been identified, the prospects for solving the structure on the basis of those maps were poor. On the basis of our previous experience, we began to use the maximum-entropy solvent-flattening algorithm in the manner that had previously proved so successful with the cytidine deaminase structure (Xiang, Carter, Bricogne & Gilmore, 1993).

3.3.1. *The folly of following ME extrapolation.* Initiation of the ME solvent-flattening process proved to be non-trivial. Our hope had been to utilize the molecular envelope previously determined for the monoclinic crystal form of TrpRS, which had previously been oriented according to a cross-rotation search using the two native data sets (Carter, Crumley, Coleman, Hage & Bricogne, 1990). Initial results with ME solvent flattening were disappointing. The previously determined envelope, although essentially correct, was not correctly oriented by the molecular-replacement calculations we had done. In retrospect, it might have been possible to periodize the known envelope on a large unit cell and carry out rotation-function analysis from its transform, rather than relying on the cross-rotation peaks of the native data sets; however, the Fourier spectrum of the envelope itself was limited to 18 Å, a resolution range for which the available data from the tetragonal form

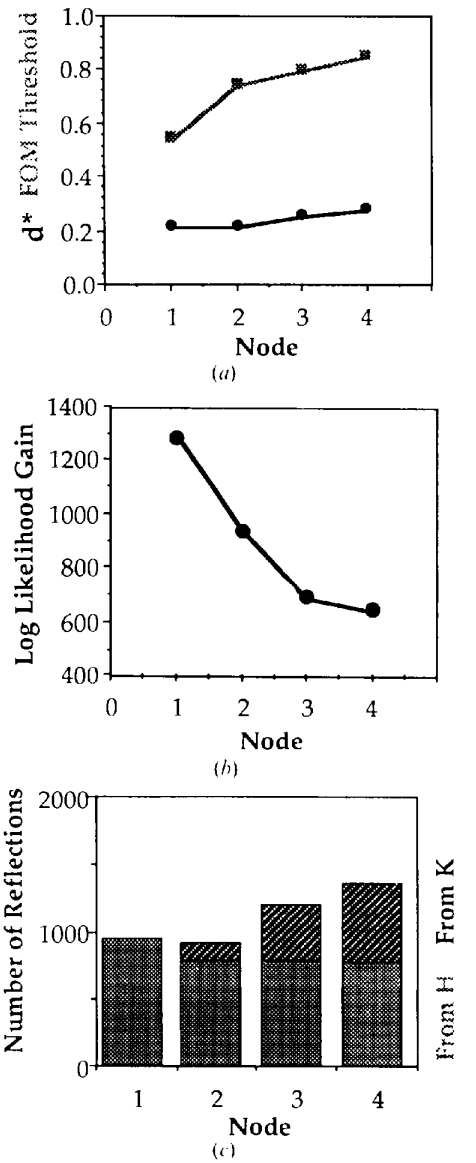


Fig. 4. The consequences of using incorrectly extrapolated phase information in the basis set for exponential modeling. (*a*) Resolution limit (d^*, lower curve) and figure-of-merit threshold (FOM threshold, upper curve) of successive basis sets. (*b*) Extrapolative power of successive nodes estimated by the LLG. (*c*) 'Infiltration' and corruption of basis set by extrapolated reflections. Although the mean phase error estimated by the figure of merit of the phase-probability distribution goes up, the LLG goes down precipitously as the resolution is increased. The reason is the population of the basis set with strongly but incorrectly extrapolated phases, indicated by the hatched regions (from *K*) in the lower histogram. These extrapolated reflections gradually replaced reflections (from *H*) from the original basis set.

Table 5. *Full factorial phase permutation experiments with likelihood ranking*

Each permutation in this table was carried out according to a complete factorial design in which all possible sign combinations were tested for a small number of centric reflections. In this and the following table, the node numbers on the bottom line of each subsection refer to the node evaluated using the optimal phase choices obtained for reflections in that particular permutation experiment.

	Constraints				Sampling strategy	Results				
		Resolution (Å)				Significance tests				
		M threshold	New	Resolution	Number			LLG	LLG	
Node	Envelope	No. of reflections	reflections	(Å)	of nodes	φ ( )	2-tail P	(range)	(recombined)	Comments
1	Original (WL)	4.5 FOM 0.5 1079						1315	1488	Starting node; identified bias introduced by using recombined phases as constraints.
10	Redef* (1)	4.5 FOM 0.5 1082	6,0,10 1,1,32 6,0,30	9.3 7.2 6.2	8 (Full factorial)	90 180 270	**0.001** **0.043** **0.810**	1377 (±33)	1524	
11	Redef (2)	4.5							1708 (1) 1744 (2)	Started using data between ∞ and 14 Å.
13	Redef (3)	4.5 FOM 0.5 1083	2,0,22	10.0	2	0	**0.125**		1783 ± 58	First use of $U_{obs} - U^{MI}$ as selection criterion: Added 2,0,22. Rather modest significance test.
18	Redef (4)	3.6 FOM 0.5 1796	601 506	10.2	4 (Full factorial)	315 0	**0.068** **0.252**	2153 (±53)	2523	Resolution, envelopes not in sequence. These nodes done in parallel with others.
23	Redef (3)	3.9 FOM 0.5 1545	**3,0,21**† 13,0,11	**9.5** 4.6	4 (Full factorial)	225 315	**0.031** **0.613**	2051 (±31)		3,0,21 had a large $U_{obs} - U^{MI}$; Φ(SIR) = 45 ; Φ(ME) = 225 . Permutation confirmed Φ(ME) correct.
28	Redef (5)	3.9 FOM 0.5 1547	3,3,11 4,4,11	11.9 9.6	4 (Full factorial)	270 90	**0.082** **0.028**	2051 (±47)	2118	
37	Redef (6)	3.9 FOM 0.5 1550	605 1,0,13 1,0,15	9.9 17.4 15.0	8 (Full factorial)	135 225 315	**0.008** **0.027** **0.169**	2377 (±33)		

* Redef refers to a redefined envelope based on the centroid map of the previous node, obtained automatically by the procedure of Wang (1985) and Leslie (1988) (WL).

† The 3,0,21 was a basis-set reflection that was nonetheless poorly fitted at maximum likelihood. Its phase was permitted for this reason.

were unreliable. We decided to determine the envelope initially from the electron density itself (Wang, 1985). We tried defining the envelope using first the lower-resolution terms. The MESF map improved, but not to the degree we expected from previous experience with cytidine deaminase. Next, we tried increasing the size of the basis set gradually by including higher-resolution reflections after phase recombination of the ME and MIR probability distributions (Fig. 4*a*). An increased figure-of-merit threshold value was used on each successive step, in an attempt to select only reflections with nearly correct phases. Although the maps appeared to improve, and the overall figures of merit for basis-set reflections also rose, the results were nevertheless disastrous.

The surest clue that we were headed down an incorrect path came from the LLG, which plummeted as we increased the resolution (Fig. 4*b*). The starting phase set was too weak and/or too error ridden to extrapolate properly, and by recruiting new reflections into the basis set from those extrapolated strongly by the exponential modeling we were actually only reinforcing incorrect phase indications introduced by the extrapolation (Fig. 4*c*). This conclusion was confirmed by the fact that a small but important percentage of extrapolated reflections had entered the basis set after the first recombination even without extending the resolution (4.5 Å; node 2 in Fig. 4*c*). Under these conditions, ME extrapolation was behaving like a Trojan Horse, bringing in strongly but incorrectly extrapolated reflections and thereby corrupting the basis set. This behavior had already been observed on small test molecules (Gilmore, Bricogne & Bannister, 1990, §4.2).

3.3.2. *The need for phase permutation.* These observations illustrate a fundamental complication of the phase problem, namely that it is a multiply branched problem (Bricogne, 1984) and that, as the LLG behavior clearly showed, the previous purely iterative process was incapable of dealing with bifurcations correctly. Examination of a list of strongly observed but weakly extrapolated reflections outside the basis set (Fig. 5, uppermost curve) revealed a number of unphased strong reflections that *MICE* could not reach *via* ME extrapolation from the initial phases. Many of these were centrosymmetric low-resolution reflections. The presence of so many unphased low-resolution reflections compounded problems associated with the fact that the envelope was poorly defined. Clearly, the constraints were insufficient to guide the construction of the distribution of atomic positions along the correct lines and the primary obstacle appeared to be the absence of these

Table 6. *Statistically designed phase and envelope permutation with likelihood ranking*

Each permutation in the subsequent selection was performed using an incomplete factorial design, permitting more bits of phase information to be sampled for each design and, in particular, making it possible to include acentric reflections.

		Constraints			Sampling strategy	Results			
		Resolution (Å)				Significance tests		LLG	
Node	Envelope	M threshold No. of reflections	New reflections	Resolution (Å)	Number of nodes	φ ( )	t tests	(range)	Comments
54	Redef* (7)	3.6	1,0,25	9.2	16			3151	Incomplete factorial design provides:
		FOM 0.5	2,2,19	10.7	(Incomplete	270	**0.270**		larger range of values for LLG;
		2004	400	15.2	fractorial)	0	**0.004**	(±106)	more paired comparisons for LLG;
			443	10.6		90	**0.186**		more degrees of freedom for each
			427 (Re)	12.6		345	**0.000**		phase to be determined;
			427 (Im)						more *reliable* phase determination;
			5,1,11 (Re)	10.4					substantially more *efficient* phase
			5,1,11 (Im)			260	**0.001**		determination.
			614 (Re)	9.9		315	**0.001**		
			614 (Im)				**0.120**		
79	Redef (5)	3.9	4,4,44	4.75	24	180	**0.000**	2525	Returned to 3.9 Å basis set to economize
		FOM 0.5	11,0,9	5.40	(Incomplete	225	**0.349**		on computer time. Likelihood is lower
		1808	1,0,57	4.08	factorial)	225	**0.113**	(±87)	for this reason and should be compared
			5,4,14 (Re)	8.25		237	**0.022**		to 2377, which was the resulting LLG
			5,4,14 (Im)				**0.016**		after the last full factorial experiment.
			528 (Re)	10.54		31	**0.000**		
			528 (Im)				**0.000**		
			525 (Re)	10.98		300	**0.015**		
			525 (Im)				**0.001**		
			8,1,38 (Re)	4.75		333	**0.460**		
			8,1,38 (Im)				**0.525**		
86	Inf_1	3.6			6		**0.003**	2659	Five factors sampled with six nodes (Inf_1–6):
	Inf_2	FOM 0.5			(Incomplete		**0.005**		calculation mode (use edited map);
	Inf_3	2064			factorial)		**0.004**	(±103)	solvent volume (use tight boundary);
	Inf_4						**0.003**		region 1 (don't include);
	Inf_5						**0.006**		region 2 (include);
	Inf_6								region 3 (include).
111	Inf_2_opt†	3.6	2,0,54	4.3	23	90	**0.001**	2721	Overall variance of design LLG values
		FOM 0.5	7,0,45	4.4	(Incomplete	225	**0.000**		is reduced, suggesting that the permuted
		2064	447	10.2	factorial)		**0.504**	(±46)	reflections have a smaller impact on the
			7,7,44	4.0			**0.112**		map. The 3.9 Å phase determination is
			2,0,52	4.3		180	**0.002**		nearing completion and the remaining
			6,0,38	5.2			**0.968**		problem is to 'finetune' the envelope.
			1,1,15	14.6		90	**0.013**		
			1,0,24	9.6		90	**0.093**		
112	Inf_5	3.6						2812	Likelihood phase refinement

* Redef is as defined in Table 5.

† Inf_2_opt refers to an envelope incorporating all of the correct choices made on the basis of the envelope permutation.

strong low-resolution reflections from the basis set. We therefore decided to permute phases for several of these and to try using the LLG, in a statistical sense, as a criterion for the correctness of these phase choices. The results of this decision are presented in some detail in Tables 2–7 and documented in Figs. 5, 9 and 10. Nodes are numbered sequentially from 1. The phasing tree is essentially unbranched, exploring only the buds in each successive permutation. Nodes summarized in Tables 5 and 6 were developed with optimal phase choices inferred after each permutation.

3.3.3. *Complete factorial permutation searches with significance testing.* Our first impulse was to proceed cautiously, using full factorial designs for one or several centric reflections representing a small number of bits of new phase information (Table 5). The approach was to use the LLG as an experimental score and to analyze the results of all experiments for the main effects of phase choices (+1 or −1) for each reflection, using conventional analysis of variance (Wilkinson, Hill & Vang, 1992). This approach is a standard extension of what we have used in screening crystal-growth experiments (Carter, 1992) and although it lacks the power of the multidimensional Fourier analysis of the LLG (Bricogne, 1993), it nevertheless captures the intent of that suggestion.

Our use of statistical significance testing of the impact of phase assumptions on the LLG is in marked contrast to previous use of permutation methods in this context (Sjölin, Prince, Svensson & Gilliland, 1991), where no tests of significance were carried out. Clearly, the analysis of variance provides considerably more reliable guidance than does a simple comparison of any score obtained for different nodes. Moreover, in our experience, the entropy is rarely correlated significantly with the correct phase choices.

The immediate results were very encouraging. Using eight different nodes representing all possible combinations of phases for three centric reflections, we got

statistically significant indications for two of them. The LLG increased - something we had previously been unable to bring about by any choice of envelope. Moreover, recombination with the MIRAS phase-probability distributions brought about an even bigger increase in the LLG. Several subsequent factorial designs brought ten correctly phased outlying reflections into the basis set and increased the LLG from 1315 to 2377.

It is worth noting that one reflection that figured in these initial permutations, 3,0,21, was a basis-set reflection, included with its MIRAS phase on account of its high figure of merit. It was nevertheless identified as a possible problem reflection because it failed to fit very well in the exponential modeling and as a result had a very large renormalized structure-factor amplitude. When it was permuted (node 23 of Table 5), it became evident that it had been phased incorrectly by isomorphous replacement.

3.3.4. *Increasing sampling efficiency with incomplete factorial permutation designs.* Once these ten new reflections were phased, we were faced with the problem of permuting acentric reflections, which can in principle take on any value between 0 and 2π. Encouraged by the success of binary permutation of centric reflections by complete factorial designs, we permuted subsequent reflections using incomplete factorial designs as described in §2.4 and Fig. 1. The performance of these designs was superior to that observed for the full factorial designs (Table 6). Owing to the fact that the contrasts were evaluated from a larger number of experiments, the significance tests improved greatly, despite the fact that the number of experiments per bit of phase information was also lower (eight experiments for three bits - second and eighth rows of Table 5 - *versus* 24 experiments for 11 bits of information - second and fourth rows of Table 6). This reflects the familiar dependence of the precision of an estimate on $(N-1)^{-1/2}$ for N observations of a given unknown quantity. For the initial designs, we were comparing averages of as few as two observations, whereas for the 24-node designs we were comparing averages of 12 observations of the LLG, improving the confidence of the estimates for each bit of phase information by a factor of around 3.

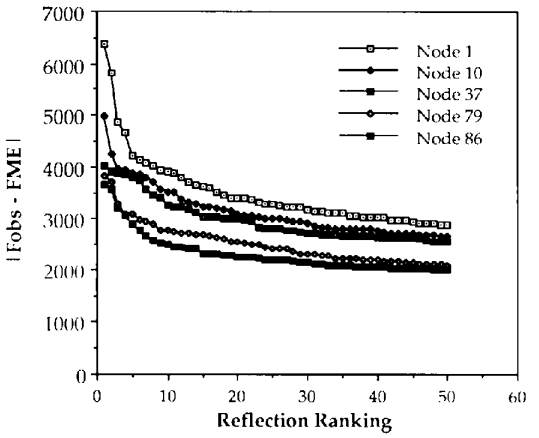


Fig. 5. Reflections least expected by ME extrapolation. The 50 largest values for $|F_{obs} - F^{ME}|$, as estimated using (5) with structure-factor amplitudes rather than unitary structure-factor amplitudes, for some of the nodes represented in Tables 5 and 6. As phase determination proceeds, these decrease significantly in magnitude, indicating that the strength and accuracy of the extrapolation is improving. Node numbers correspond to those in Tables 5 and 6. Plots for nodes 86 and 111 are too similar to be easily distinguished from one another, so only node 86 is shown.

Testing the significant contrasts jointly from a multivariate regression model, rather than individually, enabled us to infer how the LLG score would behave outside the subset of nodes sampled by the incomplete factorial design and hence to make optimal phase choices, in most cases for all permuted reflections. The actual information content provided by these designs is somewhat greater than that estimated purely on the basis of the design matrix. This is because the ratio between the contrasts for the real and imaginary components of a permuted acentric phase can be interpreted as an indication of the actual phase angle, *via* its inverse tangent. In brief, reflections for which the real or imaginary component is dominant probably have phases closer to 0, 90, 180 or 270° than to 45, 135, 225 or 315°. We have verified this expectation by performing a permutation of one such reflection over a 90° range in intervals of 10°. The maximum LLG was obtained very close to the value indicated by the inverse tangent of the real and imaginary component contrasts. This phenomenon probably accounts for the weak significance tests for some, if not most, of the 'relatively insignificant' phase indications in Table 6, where the uncertainty is concentrated in either the real or the imaginary component of the phase. Similar behavior had previously been observed in a wide range of calculations performed by one of us (GB, unpublished results), where these phase estimates were recycled into the multisolution process. We therefore now routinely use the inverse tangents of the two contrasts to phase acentric reflections, assuming the phase to be purely real or purely imaginary when only one contrast is significant.

As an example, the analysis for nodes 114-138 (Table 3) is presented in Table 7. The LLG_{obs} values range from 2019 (nodes 130 and 131) to 2198 (nodes 115 and 124), with a mean value of 2099 (the constant term in the model). The magnitude of the individual contrasts with significant t tests (the coefficients of the linear model) ranges from 5 (the sign of the 0,0,48) to 26 (the real part of the 6,4,48). The latter is around 1% of the average score. Its t test, however, is significant at about one part in 10^7. The predicted LLG_{calc} for the optimized phase choices, obtained by adding the coefficients in the second column of Table 7, was 2229 with a standard error of 10.6. The LLG_{obs} for the 'best' node of the full factorial, obtained by a supplemental node expansion using the indicated signs and quadrants permitted by

Table 7. *Log-likelihood gain scoring gives significant phase indications for seven reflections*

Results of multiple regression analysis for the final phase-permutation (nodes 115–138, Table 3) experiment performed after all sources of phase information had been included and the exponential modeling carried out to convergence, with phase recombination. At this point, there were still strong unphased reflections whose incorporation into the basis set had a consistent significant effect on the dependent variable, which was the LLG_{obs}. The analysis included 24 nodes from the incomplete design plus the parent node, for a total of 25 nodes. The observed LLG_{obs} values in this design ranged from 2019 to 2198 with a mean of 2098.6. The contrasts (coefficients of the regression model) are indicated in bold typeface in column two, as are the corresponding Student t tests and their probability under the null hypothesis that that particular phase bit has no effect on the LLG, in columns six and seven. The optimal phase choices inferred from the contrasts in column two, the final phase calculated from the refined structure and the error of the phase choices appear in columns eight to ten. The regression model for LLG_{calc} had a multiple $R = 0.984$, a squared multiple $R = 0.969$ and an adjusted squared multiple $R = 0.951$. The standard error of the LLG_{calc} estimated from the model was 10.653. The statistics in Tables 5 and 6 were obtained from similar analyses of variance tables for the respective designs.

Variable	Coefficient	Standard error	Standard coefficient	Tolerance	Student t test	(2-tail) P	$\varphi_{Indicated}$ ( )	φ_{calc} ( )	$\Delta\varphi$ ( )
Constant	**2098.612**	2.131	0.000		**984.97**	**0.10E − 14**			
Re 6,4,48	**26.199**	2.485	0.547	0.766	**10.542**	**0.25E − 07**	0	27	27
Re 7,4,41	**10.374**	2.268	0.217	0.920	**4.575**	**0.36E − 03**	0	1	1
Re 10,9,26	**3.050**	2.631	0.064	0.683	**1.159**	**0.264**			
Im 10,9,26	**−7.061**	2.850	0.147	0.582	**−2.477**	**0.026**	66	101	35
Re 9,8,28	**17.351**	2.284	0.362	0.907	**7.597**	**0.16E − 05**			
Im 9,8,28	**17.824**	2.488	0.372	0.764	**7.164**	**0.33E − 05**	46	86	40
Sign 2,0,54	**22.313**	2.337	0.466	0.866	**9.550**	**0.91E − 07**	90	90	0
Sign 0,0,48	**−5.346**	2.571	0.112	0.715	**−2.079**	**0.055**	180	180	0
Sign 9,0,35	**21.243**	2.799	0.443	0.604	**7.589**	**0.16E − 05**	135	135	0

Analysis of variance

Source	Sum-of-squares	DF	Mean-square	F ratio	P
Regression	53382.214	9	5931.357	52.263	0.105066E 8
Residual	1702.352	15	113.490		

the permutation, was 2196. However, that obtained for node 139, evaluated with optimal phase choices, using inverse tangents of contrasts for acentric reflections as described above, was 2230. Statistical inference of the actual phases from the 24-node sample therefore actually provided better choices than simply moving to the node of the full factorial permutation that was indicated as the best.

3.3.5. *Ranking hypotheses about the molecular envelope.* At the outset, we were uncertain not only about the phases but also about the molecular envelope used for imposing solvent flatness. The latter uncertainty was reflected in the fact that the majority of the reflections permuted in Tables 5 and 6 were predominantly strong low-resolution reflections whose most important contribution was in defining the envelope. Consequently, after each round of phase permutation, maximum-entropy solvent flattening and phase recombination, the new centroid map could be used as a template for improving the envelope. This process continued to improve the phases, as indicated by the curves representing histograms of the largest renormalized structure factors [(4), Fig. 5]. We nevertheless realized that a number of attributes of the molecular envelope remained ambiguous and were not being resolved effectively by the recruitment of new phases into the basis set. We reasoned that, since the values of the LLG would be sensitive to the choices made in editing the envelope (Bricogne, 1988*a*, §2.3), these ambiguities could be resolved directly by the same mechanism of hypothesis testing as had been used at the previous stage to infer the values of hitherto unknown phases.

Five possible modifications of the envelope that could be encoded in binary fashion were permuted in the sixth design. In particular, we had experimented with several different algorithms for preparing the envelope map, including direct editing (Minor, 1992) and reciprocal-space weighting of a truncated map (Leslie, 1988) and were uncertain about the molecular volume. We were also uncertain whether or not to include three prominent features of the map, where the electron density, though weak, was nevertheless consistent with protein density (Fig. 6). We therefore carried out an incomplete factorial search for the correct choices for six of these factors (Table 6, resulting in node 86). As with the phase-permutation experiments, the LLG scoring criterion provided statistically significant choices for each factor. Resolution of the three ambiguous regions resulted in considerable improvement in about twenty reflections with moderately strong renormalized structure factors (compare the plots for nodes 79 and 86 in Fig. 5). These three indications regarding redefinition of the envelope were later checked, once the structure was solved and refined, and were found to be correct.

Although we observed nothing to indicate that the process we were following would not readily converge on the correct structure, the map was still difficult to interpret at this stage and we were anxious to incorporate the $\{|F_{selenium}| - |F_{sulfur}|\}$ differences and to use the resulting phase information to help solve the structure. The optimized envelope, OPT1, enabled us to initiate ME solvent flattening of the native map phased with the non-isomorphous derivative MIRAS phases, as described next.

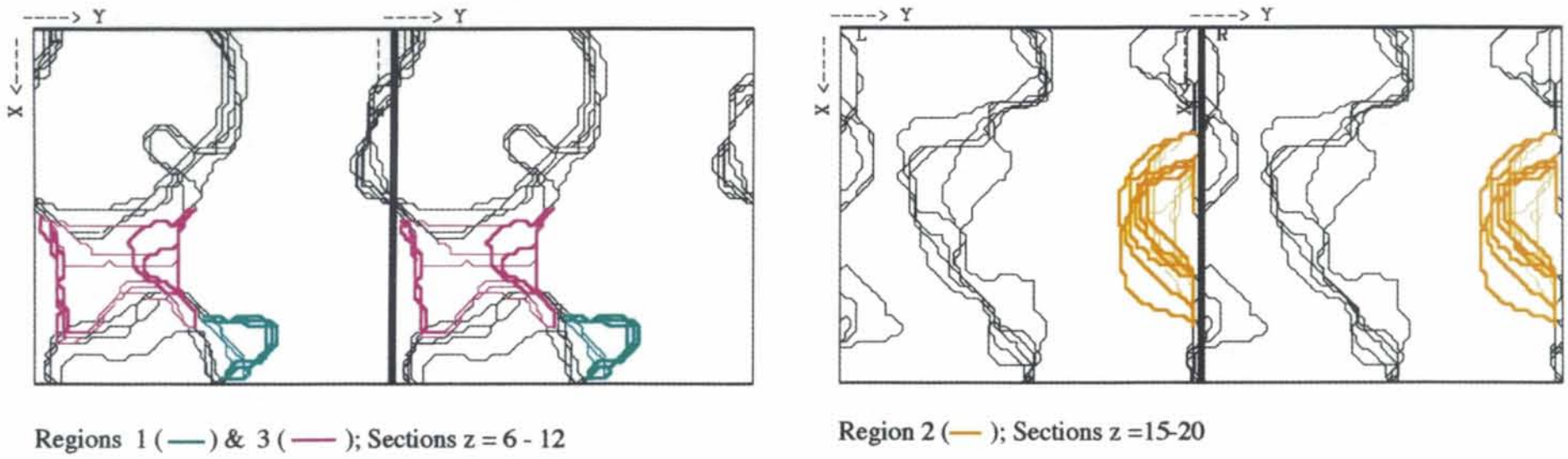


Fig. 6. Regions of the envelope that were subjected to permutation analysis according to the design in Table 4. Z sections (1/240) of an asymmetric unit ($0<x<1$; $0<y<1$) of the envelope maps are shown as stereo pairs as indicated. The enzyme monomer is continuous from the bottom left at $z=0$ to the top left at $z=1/8$. The boundaries of the permuted regions are indicated in color, with the innermost boundaries drawn in two-point contours and the outermost boundaries in four-point contours.

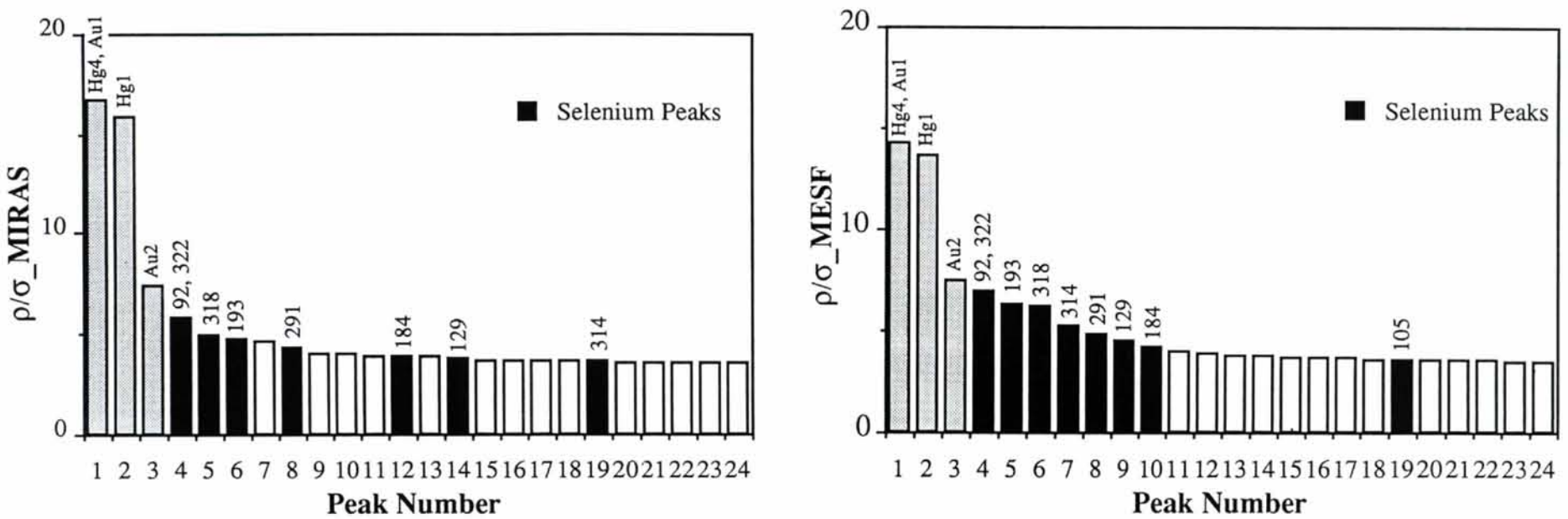


Fig. 7. Relative peak heights in (*a*) the $\{|F_{SeMet}| - |F_{NAT2}|, \varphi^{MIRAS}\}$ and (*b*) the $\{|F_{SeMet}| - |F_{NAT2}|, \varphi^{MESF}\}$ difference Fourier maps. The three largest peaks are 'ghosts' of the heavy-atom sites that could not be eliminated by the various scaling and temperature factors that were applied. Peaks corresponding to the selenium atoms are indicated by residue number and shaded black.

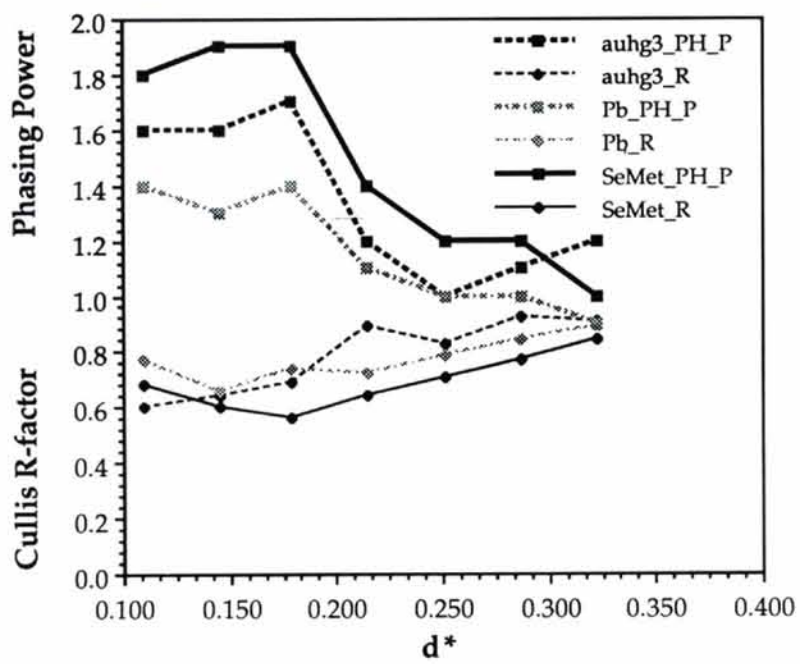
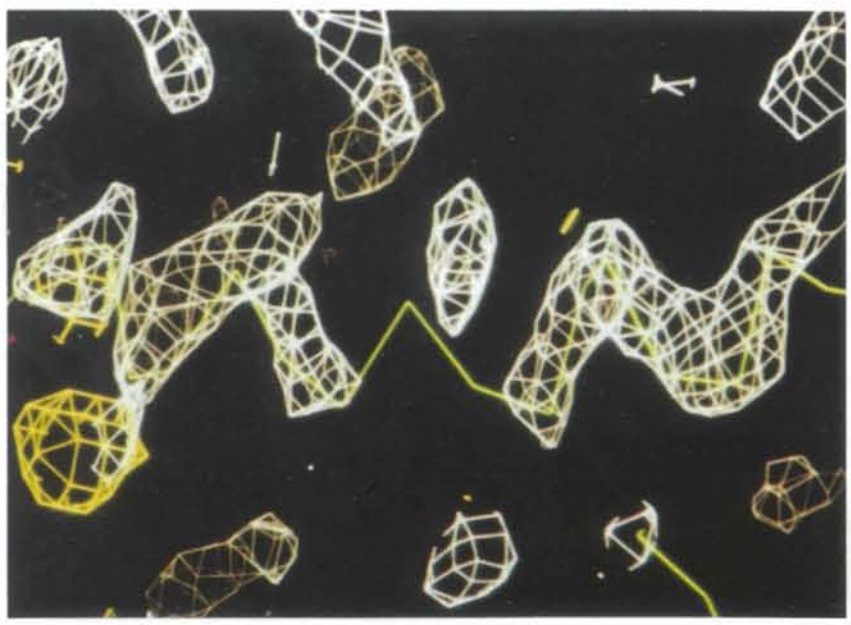
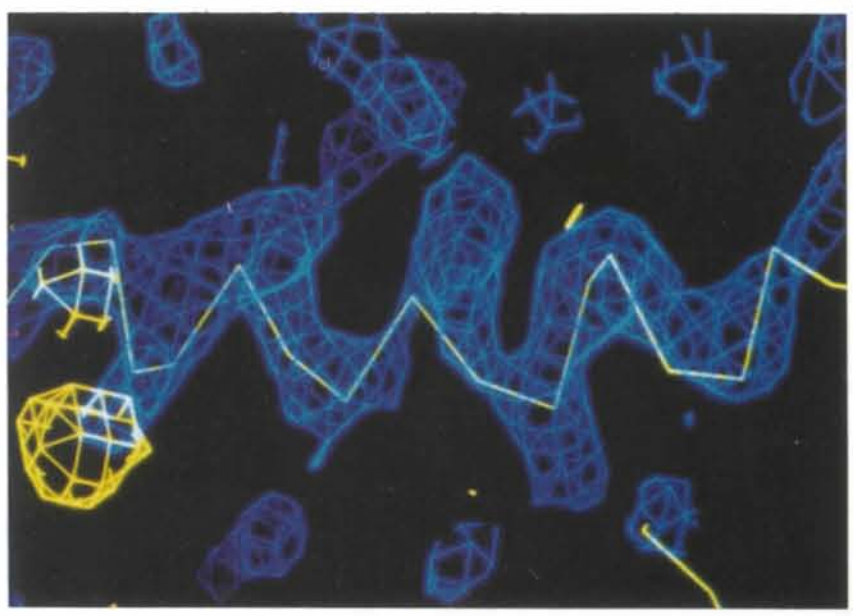


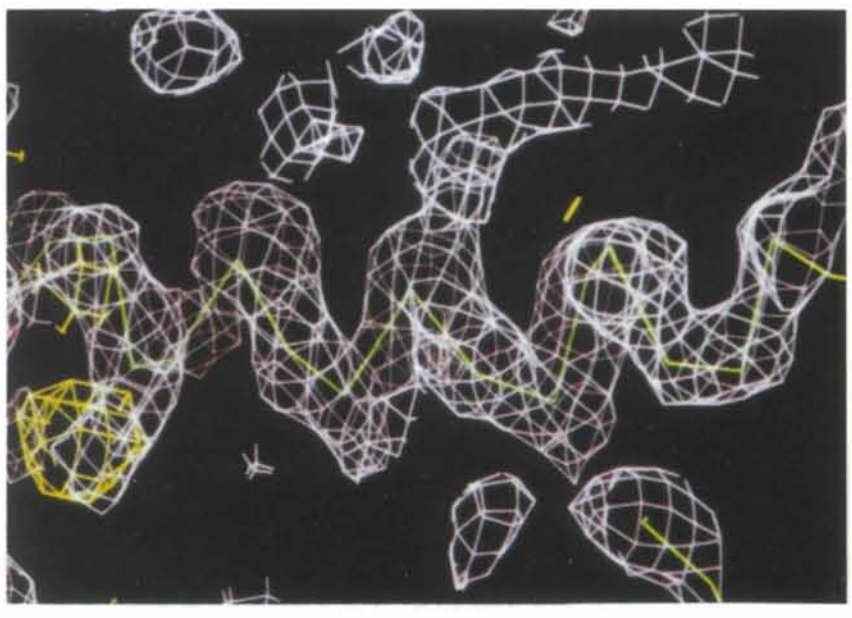
Fig. 8. Primary phasing statistics for phasing group III. The phasing power is estimated from the ratio of the r.m.s. heavy-atom scattering factor to the r.m.s. lack-of-closure error. The Cullis R factor is obtained from the ratio of the mean lack-of-closure error to the mean isomorphous difference.



(*a*)

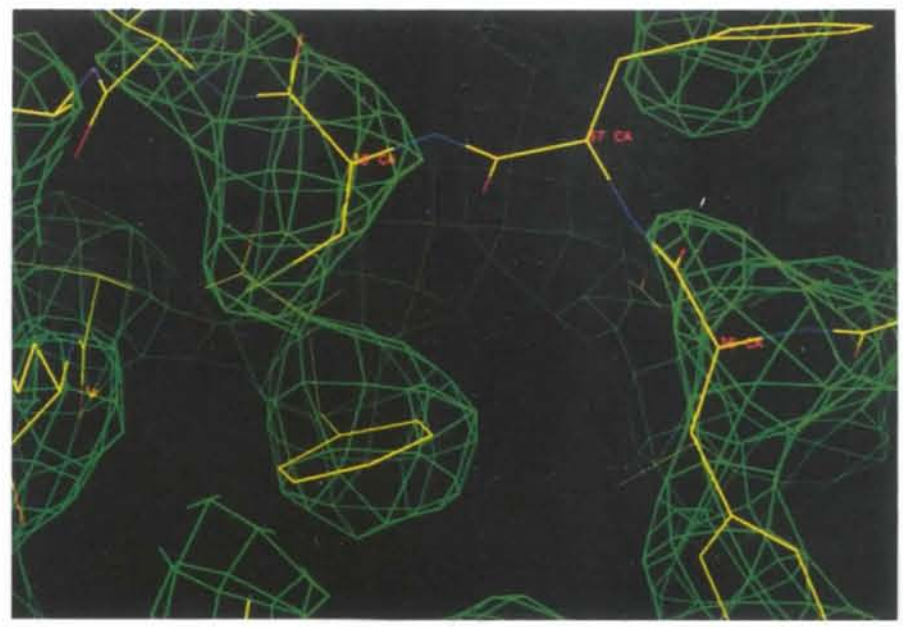


(*b*)

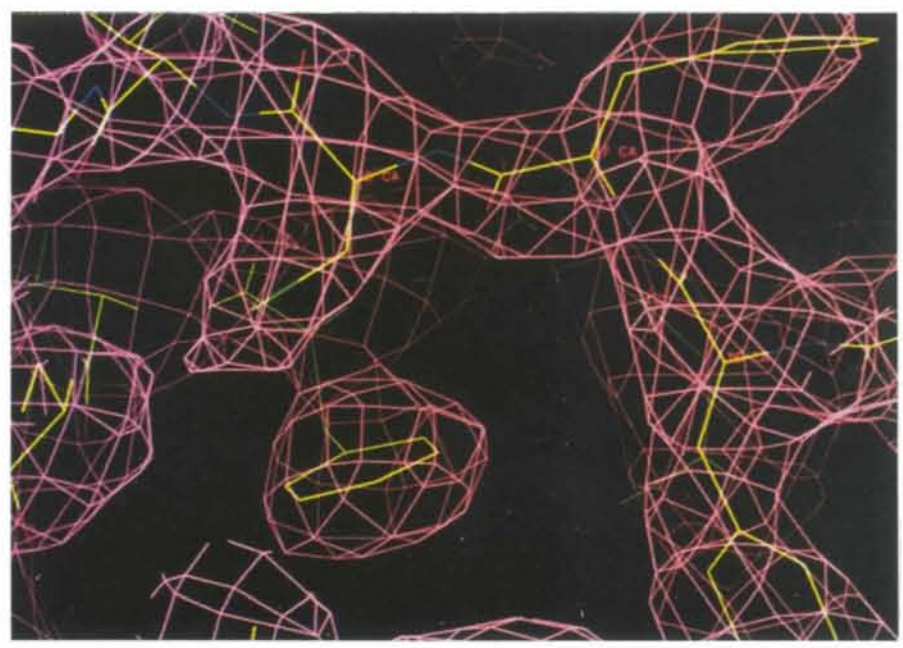


(*c*)

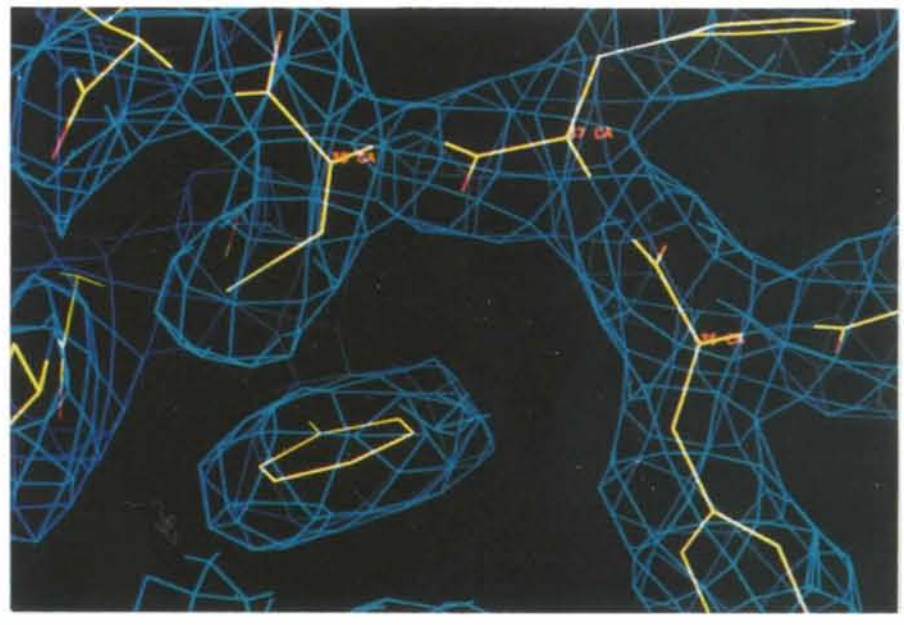
Fig. 9. Electron-density maps for the C-terminal α-helix in the tetragonal TrpRS crystals. All the maps are contoured at 1.5σ. (*a*) MIRAS 3.1 Å map. (*b*) Centroid 3.1 Å map after ME solvent flattening and direct phase determination using the permutation results obtained for phasing group I. Superimposed on each map is the C^{α} tracing of the interpreted map, shown in (*c*), and, in yellow, the ($F_{\text{selenium}} - F_{\text{native}}$) difference Fourier map phased from node 112 of Table 6 (the same difference map is superimposed on all three successive electron-density maps).



(*a*)



(*b*)



(*c*)

Fig. 10. Electron-density maps calculated using NAT2 amplitudes and phases derived from phasing group III, including (*b*) the map that was ultimately interpreted. (*a*) MIRAS map calculated using all derivatives indicated in Fig. 3. (*b*) NAT2 centroid map after MESF refinement and one round of phase permutation (node 139, Table 7). (*c*) NAT2 $\{|2F_{\text{obs}} - F_{\text{calc}}|, \varphi_{\text{calc}}\}$ map obtained for the refined structure at an R factor of 0.18 for all data with $I/\sigma > 2.0$ for reflections between 7 and 2.9 Å.

3.4. *Location of selenium atoms by difference Fourier synthesis using phases determined by MESF using the improved envelope*

As noted earlier, there are two families of data sets for tetragonal TrpRS crystals, which are not isomorphous with each other. The first phases we obtained were for heavy-atom data sets, because we could not locate the selenium atoms from the isomorphous difference Patterson for the ($F_{\text{selenomethionine}} - F_{\text{native}}$) difference amplitudes representing the ($F_{\text{selenium}} - F_{\text{sulfur}}$) differences. TrpRS has 10 methionines and the difference Patterson was too complicated to interpret readily, either manually or by automated methods. Nonetheless, we expected that locating the selenium atoms would greatly enhance the phase quality for the native structure, because the two data sets were so much more nearly isomorphous than were the heavy-atom and native data sets involved in phasing group I. We used the refined atomic positions for the heavy atoms from phasing group I and refined them against difference amplitudes involving the NAT2 native data, which had the best scaling statistics of the three native data sets, giving phasing group II. The selenium positions were very uncertain in the isomorphous difference Fourier map calculated with these phases, as indicated in Fig. 7(*a*). To improve the phases sufficiently to identify them, we began applying the ME solvent-flattening process for the native map calculated with the NAT2 amplitudes and group II phases, constrained by the refined OPT1 envelope.

Centroid phases obtained from ME solvent flattening of the native map after one cycle of phase recombination with the MIRAS probabilities (node 113) provided definitive locations for eight of the ten selenium atoms in a difference Fourier map with coefficients $\{\text{Wt}_{\text{sim}} \times |F_{\text{selenomethionine}} - F_{\text{native}}|, \varphi^{\text{ME}}\}$. Eight spherical peaks from this map (Fig. 7*b*) subsequently refined to give SIR phases for the native amplitudes. One additional selenium position, for Met 105, was then identified from a difference Fourier map calculated with the SIR phases. Residue 105 is in a region of the sequence for which the refined isotropic B values are greater than 40 Å^2 and is immediately followed by a stretch of seven residues for which the density is poorly defined. The 10th selenium-atom position was that for Met 1, which could not be positively identified until the structure was essentially solved. The nine positions of which we could be confident, together with the known amino-acid sequence, provided valuable guidance in interpreting the map. Group III phases were soon thereafter supplemented by a new derivative prepared with trimethyllead acetate (Holden & Rayment, 1991), which independently confirmed the selenium-atom positions. The lead derivative was highly isomorphous but was weakly substituted (occupancies of 0.3–0.4 on an absolute scale) and hence produced only a slight improvement in the map.

It is of some interest to look back (Fig. 8) at the three derivatives from which the final electron-density map was obtained: the double derivative, AuHg3, the selenomethionyl derivative, SMT2, and a trimethyllead acetate derivative, Pb. Phasing-power values were 1.3 for SMT2, 1.2 for AuHg3 and 1.0 for Pb. Cullis R factors for centric reflections were 0.70 for SMT2, 0.83 for AuHg3 and 0.79 for Pb. Moreover, all values deteriorate abruptly beyond 5.0 Å resolution for all three derivatives. These phasing statistics illustrate the marginal quality of the primary phases and demonstrate the critical role of the SMT isomorphous differences and hence of the whole Bayesian phasing procedure that acted as a 'bootstrap' and made the exploitation of these differences possible.

4. Discussion

4.1. *Rescuing phases from non-isomorphous derivatives*

The ability to locate the nine selenium atoms in isomorphous difference Fourier maps after phase improvement with *MICE* represents an unprecedented recovery of useful phase information from an inauspicious set of non-isomorphous differences. Not only is the difference Patterson (Fig. 2) that is typical of the heavy-atom substitution obtained for TrpRS very noisy, but the signal itself is rather weak. The contrast between the relative peak heights for the two sets of phases illustrated in Fig. 7 is convincing evidence for the recovery of substantial phase information from these heavy-atom derivatives.

Post-hoc verification of the phase indications by comparison with phases calculated from the refined structure is possible only for the permutation shown in Table 7 as no refined model yet exists for the non-isomorphous structure represented by the Au1 amplitudes. However, the indications obtained in Table 7 are in good agreement with phases calculated from the model superimposed on the maps in Fig. 10. All three centric indications were correct and the mean phase difference for the four acentric reflections is 26°.

4.2. *Improvement in the electron-density maps*

The evolution of the electron-density maps that accompanied the phase permutation experiments is illustrated in Figs. 9 and 10. Fig. 9 demonstrates the extent to which we have been able to recover from the residual non-isomorphism in the phasing group I data sets. The region shown contains part of the C-terminal α-helix, which contains three of the nine methionines. The map in Fig. 9(*a*) was calculated using the phasing-group-I MIRAS phases to 3.1 Å. The map in Fig. 9(*b*) is the centroid 3.1 Å map after ME solvent flattening combined with the phase and envelope permutations documented in Tables 5 and 6. The map in Fig. 9(*c*) is the 2.9 Å centroid map for native data with phases from group III after ME solvent flattening, using MIRAS constraint phases obtained from both SMT and lead derivatives, followed

by one 24-node 11-bit phase permutation (Table 7). The interpreted α-carbon model and (SMT - native) difference Fourier map are superimposed on all maps, which are contoured at 1.5σ. Fig. 10 shows a similar series of maps for the map improvement achieved with *MICE* for phasing-group-III maps superimposed on the complete model of the refined structure.

The MIRAS map from phasing group I was uninterpretable throughout most of the asymmetric unit. The density of the α-helix illustrated in Fig. 9(*a*) is representative of most of the map. The density is fragmented and many connections are incorrect. The density in Fig. 9(*b*), corresponding to the phases from node 112 of Table 6, which represent the best phases we obtained by MESF from the group-I data sets, is significantly improved. That map might eventually have been interpreted had the other sources of phase information been unavailable to us. More important, however, is the fact that the improvement of the map between Figs. 9(*a*) and (*b*) relative to Fig. 9(*c*) demonstrates that the process we had initiated and carried forward with phasing group I was converging to the correct map. Locating the selenium atoms produced better starting phases and we were therefore able to abandon phasing group I in favor of the much more direct path to the structure provided by phasing group III. In fact, we are now continuing with the group-I phasing path in order to identify *post-hoc* the conformational changes in the heavy-atom derivatives responsible for the loss of isomorphism.

The MIRAS map calculated with phasing-group-III phases was much more readily interpretable than that for phasing group I. However, there were still places where the density was broken (Fig. 10*a*). The MESF map for phasing group III, enhanced by the permutation experiment shown in Table 7 (node 139, Fig. 10*b*), is quite similar in quality to those produced by MESF for cytidine deaminase (Xiang, Carter, Bricogne & Gilmore, 1993). There is little difference between this map and the $\{2|F_{obs}| - |F_{calc}|, \varphi_{calc}\}$ map shown in Fig. 10(*c*).

4.3. *Phase permutation affects all reflections*

An important aspect of these phase permutation studies is that, by enlarging the basis set with the addition of those reflections most unexpected by the exponential model fitting done in *MICE*, the extrapolation improves for all reflections outside the basis set. This is illustrated in Fig. 5, which shows plots of amplitudes for the most unexpected reflections associated with each major node of the phase determination documented in Tables 5 and 6. The pertinent observation is that the overall level of 'surprise' (indicated by the asymptotic behavior of the plots) drops by about 35% for all reflections, not only those new reflections incorporated into the basis set. This is a graphic presentation of the fact that the LLG has increased, showing that the phenomenon is global and not restricted to a small subset of reflections. This 'pleotropic' effect of including new strong reflections into the basis set is one of the most promising aspects of these results, because it shows that phase determination need not be carried out for more than a small subset of the reflections before the centroid map becomes interpretable. Rather, it benefits from strong coupling of phases induced by the various constraints, particularly that provided by the envelope.

4.4. *Sampling efficiency*

The theory of error-correcting codes provides useful insight into the origins of the power of incomplete factorial designs. An error-correcting code is a set of code words in which additional bits are used to separate each code word from the others sufficiently for it to become possible both to detect whenever errors have occurred and to recover the correct code word in cases when the received word does contain an error (Thompson, 1983).

Incomplete factorial experiments are, in a statistical sense, also error-filtering devices. They provide the means of detecting 'errors' in the best of the tested nodes through calculation of the contrasts for each phase bit - the average difference between nodes using one of the two phase choices and those using the other choice. When these choices are made simultaneously for many reflections, the ensemble of contrasts points rather effectively toward the 'correct' node, even though it is not usually among the set of nodes tested in the design. As was suggested heuristically before (Carter, Baldwin & Frick, 1988; Carter, 1990, 1992), these designs have the property that the more efficient they are the more accurate their phase indications. Our results were uniformly more satisfactory when we used incomplete factorial designs involving more nodes with larger numbers of reflections. The significance tests were more decisive, while the sampling was more efficient.

Although this increase in efficiency with the increase in dimensionality may at first seem paradoxical, it derives from a combination of two properties. First, sphere packings can be found in high-dimensional spaces that are much denser than the usual primitive square packing. They can become especially dense in certain 'perfect' lattice designs exemplified by the Golay code, in which each code word is surrounded by approximately 2^{11} nearest-neighbor lattice points that are not code words. Second, higher dimensionality provides for the averaging of increasing numbers of nodes, thereby increasing the precision of the statistical indications regarding which of the nearby untested nodes is the correct one.

The increases in efficiency seen in Table 6 *versus* Table 5 support the conclusion (Bricogne, 1993, §2.2.2) that truly miraculous gains are achievable by using designs based on error-correcting codes. Such designs share with incomplete factorial designs the property that they are based on the geometry of high-dimensional spaces; in addition, they are based on periodic lattices so

that the statistical analysis of scores attached to them can be performed using multidimensional Fourier analysis (Bricogne, 1993, §2.2.4). Coding theory also provides useful criteria [*e.g.* packing density and covering radius (Thompson, 1983)] for evaluating design matrices such as those we use for incomplete factorial experiments, and these additional criteria should help improve the designs themselves.

4.5. *Likelihood phase refinement*

At several stages, we have used a rudimentary implementation of phase refinement based on optimization of the LLG with respect to the basis-set phases (Bricogne, 1984, 1988*a*; Bricogne & Gilmore, 1990, §2.5). An illustration of this use is shown in the fifth row of Table 6. The improvement in LLG obtained by adjustment of the basis-set phases is nearly twice the range of LLG values in the preceding permutation and represents a tangible improvement in the centroid electron density.

5. Summary

These results demonstrate that the value of the log-likelihood gain reached after maximum-entropy solvent flattening provides an accurate and sensitive criterion for correct choices among different hypotheses regarding the constraints. These constraints include both reflections for which experimental phases are missing or incorrect because of lack of isomorphism and features of the unknown molecular envelope. The success of this approach depends on the robustness of the log-likelihood-gain criterion, proper selection of reflections, appropriate sampling procedures for phase permutation and, crucially, on the performance of conventional Student t and F-ratio significance tests on the resulting LLG scores to select reliable indications in an objective fashion.

The incomplete factorial designs used in this study yield a significant increase in screening efficiency over conventional full factorial designs. Designs involving 16-24 nodes repeatedly produced significant correct indications for up to 11 bits of phase information (four acentric and three centric reflections) simultaneously. This already provides a useful degree of efficient sampling for macromolecular data sets, tractable with currently available computing power. The very high sampling efficiency of higher-order phase interactions that can be achieved using 'magic lattice' designs based on coding theory gives the approach some spare power to deal with even worse starting phases when larger computing resources become available.

The results reported here demonstrate the utility of the Bayesian phase-determination methodology for a difficult unknown protein crystal structure of substantial molecular weight. The methods we have tested work as they were predicted to work (Bricogne, 1984, 1988*a*, 1993). Together with our previous work with *ab initio* phasing at low resolution for the monoclinic form of TrpRS crystals (Carter, Crumley, Coleman, Hage & Bricogne, 1990), these new developments bring us one step closer to the point where the Bayesian approach may be able, in at least some cases, to solve protein structures *ab initio*.

We give our sincere thanks to Roger Fourme at LURE and Bill Royer at the University of Massachusetts Medical Center, Worcester, MA, for extensive assistance with imaging-plate data-collection technologies at their respective institutions, without which this work would not have been possible; to Nguyen Xuong and Ron Hamlin at UCSD for assistance in collecting data with their multiwire system; and to Bob Sweet (BNL) for his assistance with our efforts at multiwavelength data collection on selenomethionyl-TrpRS crystals. Considerable assistance was provided at UNC by Frank Hage, Bonnie Billard and Jian Huang and we thank Hengming Ke for his comments on the manuscript. This work was supported at various times and to varying degrees by grants from NIH (26203) and the American Cancer Society (BE-7493) to CWC Jr. GB wishes to thank Trinity College, University of Cambridge, England, for working space and the Swedish Natural Science Research Council for a Tage Erlander Guest Professorship.

References

BLUNDELL, T. L. & JOHNSON, L. N. (1976). *Protein Crystallography*, edited by B. HORECKER, N. O. KAPLAN, J. MARMUR & H. A. SCHERAGA. New York: Academic Press.

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

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

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

BRICOGNE, G. (1991*a*). *Direct Methods of Solving Crystal Structures*, edited by H. SCHENK, pp. 157-175. New York: Plenum Press.

BRICOGNE, G. (1991*b*). *Crystallographic Computing* 5, edited by D. MORAS, A. D. PODJARNY & J. C. THIERRY, pp. 257-297. IUCr/Oxford Univ. Press.

BRICOGNE, G. (1992). *Molecular Replacement*, edited by E. J. DODSON, S. GOVER, & W. WOLF, pp. 62-75. Warrington, England: SERC Daresbury Laboratory.

BRICOGNE, G. (1993). *Acta Cryst.* D**49**, 37-60.

BRICOGNE, G. & GILMORE, C. J. (1990). *Acta Cryst.* A**46**, 248-297.

CARTER, C. W. JR (1990). *Methods: a Companion to Methods in Enzymology*, Vol. 1, pp. 12-24.

CARTER, C. W. JR (1992). *Crystallization of Proteins and Nucleic Acids: a Practical Approach*, edited by A. DUCRUIX & R. GIEGE, pp. 47-71. Oxford: IRL Press.

CARTER, C. W. JR, BALDWIN, E. T. & FRICK, L. (1988). *J. Cryst. Growth*, **90**, 60-73.

CARTER, C. W. JR & CARTER, C. W. (1979). *J. Biol. Chem.* **254**, 12219-12223.

CARTER, C. W. JR & COLEMAN, D. E. (1984). *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **43**, 2981-2983.

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

CARTER, C. W. JR, DOUBLIE, S. & COLEMAN, D. E. (1994). *J. Mol. Biol.* Submitted.
COLEMAN, D. E. & CARTER, C. W. JR (1984). *Biochemistry*, **23**, 381-385.
DOUBLIE, S. & CARTER, C. W. JR (1992). *Crystallization of Proteins and Nucleic Acids: a Practical Approach*, edited by A. DUCRUIX & R. GIEGÉ, pp. 311-317. Oxford: IRL Press.
DOUBLIE, S. & CARTER, C. W. JR (1993). *J. Biol. Chem.* Submitted.
GILMORE, C. J., BRICOGNE, G. & BANNISTER, C. (1990). *Acta Cryst* A**46**, 297-308.
GILMORE, C. J., HENDERSON, A. N. & BRICOGNE, G. (1991). *Acta Cryst.* A**47**, 842-846.
GOOD, I. J. (1954). *Acta Cryst.* **7**, 603-604.
HOLDEN, H. & RAYMENT, I. (1991). *Arch. Biochem. Biophys.* **291**, 187-194.
LESLIE, A. (1988). *Improving Protein Phases*, edited by S. BAILEY, E. DODSON & S. PHILLIPS, pp. 25-31. Warrington, England: SERC Daresbury Laboratory.
MINOR, W. (1992). *MAP_CCP4: a Program for Editing Molecular Envelopes*. Purdue Univ., Indiana, USA.
OTWINOWSKY, Z. (1991). *Isomorphous Replacement and Anomalous Scattering*, edited by W. WOLF, P. R. E. EVANS & A. G. W. LESLIE, pp. 80-86. Warrington, England: SERC Daresbury Laboratory.
SERC Daresbury Laboratory (1990). *CCP4. A Suite of Programs for Protein Crystallography*. SERC Daresbury Laboratory, Warrington, England.
SJÖLIN, L., PRINCE, E., SVENSSON, L. A. & GILLILAND, G. L. (1991). *Acta Cryst.* A**47**, 216-223.
TERWILLIGER, T. C., KIM, S.-H. & EISENBERG, D. (1987). *Acta Cryst.* A**43**, 1-5.
THOMPSON, T. M. (1983). *From Error Correcting Codes Through Sphere Packings to Simple Groups*, Vol. 21, edited by D. T. FINKBEINER. The Mathematical Association of America.
WANG, B. C. (1985). *Methods Enzymol.* **115**, 90-112.
WILKINSON, L., HILL, M. & VANG, E. (1992). *SYSTAT: Statistics*, version 5.2. SYSTAT Inc., 1800 Sherman Avenue, Evanston, IL 60201-6793, USA.
WOOLFSON, M. M. (1954). *Acta Cryst.* **7**, 65-67.
XIANG, S., CARTER, C. W. JR, BRICOGNE, G. & GILMORE, C. J. (1993). *Acta Cryst.* D**49**, 193-212.

Acta Cryst. (1994). A**50**, 182–193

Determination of Quasicrystalline Structures: a Refinement Program using Symmetry-Adapted Parameters

BY L. ELCORO, J. M. PEREZ-MATO AND G. MADARIAGA

Departamento de Física de la Materia Condensada, Facultad de Ciencias, Universidad del País Vasco, Apartado 644, *Bilbao, Spain*

(*Received* 29 *March* 1993; *accepted* 3 *August* 1993)

Abstract

A general program for the refinement of quasicrystalline structures using diffraction data is presented. The program can be used for both icosahedral and polygonal quasicrystals. The refinement process is based on the fitting of the structural model to experimental diffraction data and observed density and chemical composition. Superspace formalism is used for the structure description and the hypersurfaces in superspace describing the atomic positions are assumed to be parallel to the internal space. No additional *a priori* assumption on the form of the atomic hypersurfaces is necessary except that the deviations of the atomic-surface contours from a spherical shape do not contain very short wave components in a significant amount. The contours of each symmetry-independent atomic hypersurface in internal space are parametrized in terms of linear combinations of radial functions (surface harmonic) invariant for the hypersurface point group in internal space. This allows a continuous refinement of the structure in terms of symmetry-adapted parameters consistent with the symmetry restrictions resulting from the postulated superspace symmetry. The program requires an initial very approximate guess of the structure in terms of 'spherical' hypersurfaces of which only the symmetry centres are known with confidence. The continuous parametrization of the hypersurfaces does not *a priori* restrict their form, except in its degree of complexity or fine detail, which is limited by the number of terms considered in the linear expansion of the surface contours. In general, the number of surface harmonics considered should be consistent with the accuracy allowed by the experimental data set. The refinement process can be performed either by a full least-squares method or by means of a simplex algorithm. The physical consistency of the refined hypersurfaces with respect to the predicted density, chemical composition and interatomic distances is controlled by including additional 'penalty functions' in the parameter to be minimized.

1. Introduction

Accurate determination of the structures of quasicrystals (QCs) is still an open problem. The introduction of superspace formalism (Bak, 1985; Janssen, 1986) represented important progress towards achieving a *quasicrystallography* comparable with