Acta Cryst. (1993). D**49**, 193–212

Entropy Maximization Constrained by Solvent Flatness: a New Method for Macromolecular Phase Extension and Map Improvement

BY SHIBIN XIANG AND CHARLES W. CARTER JR

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

GÉRARD BRICOGNE

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

AND CHRISTOPHER J. GILMORE

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

(*Received* 2 *June* 1991; *accepted* 10 *August* 1992)

Abstract

A practical generally applicable procedure for exponential modeling to maximum likelihood of macromolecular data sets constrained by a moderately large basis set of reliable phases and a molecular envelope is described, based on the computer program *MICE* [Bricogne & Gilmore (1990). *Acta Cryst.* A**46**, 284–297]. Procedures were first tested with simulated data sets. Exact and randomly perturbed amplitudes and phases were generated, together with a known envelope for solvent-free protein and for protein in an electron-dense crystal mother liquor typical of many real protein crystals. These experiments established useful guidelines and values for various parameters. Tests with basis sets chosen from the largest amplitudes indicate that exponential models with considerable correct extrapolated phase and amplitude information can be constructed from as few as 16% of the total number of reflections, with mean phase errors of about 30°, at resolution limits of either 5 or 3 Å. When the shape of the solvent channels in macromolecular crystals is known, it offers an important additional source of information. *MICE* was, therefore, adapted to average the density outside the molecular boundary defined by an input envelope. This flattening process imposes a uniform density distribution in solvent-filled channels as an additional constraint on the exponential model and is analogous to the treatment of solvent in conventional solvent flattening. Experimental data for cytidine deaminase, a structure recently solved by making extensive use of conventional solvent flattening, provides an example of the performance of maximum-entropy methods in a real situation and a compelling comparison of this method to standard procedures. Exponential models of the electron density constrained by the most reliable phases obtained by multiple isomorphous replacement with anomalous scattering (MIRAS) (figure of merit > 0.7, representing 34% of the total number of reflections) and by the envelope give rise to centroid electron-density maps which are quantitatively superior by numerous statistical criteria to conventionally solvent-flattened density. Similarity of these maps to the $2F_{obs} - F_{calc}$ map calculated with phases obtained after crystallographic refinement of the model implies that maximum-entropy extrapolation provides better phases for the remaining 66% of the reflections than the original centroid MIRAS distributions. Importantly, the solvent-flattened electron density, although it did permit interpretation of the map which was not readily accomplished with the MIRAS map, contains substantial errors. It is proposed that errors of this sort may account for previously noted deficiencies of the solvent-flattening method [Fenderson, Herriott & Adman (1990). *J. Appl. Cryst.* **23**, 115–131] and for the occasional tendency of incorrect interpretations to be 'locked in' by crystallographic refinement [Brändén & Jones (1990). *Nature (London)*, **343**, 687–689, and references cited therein]. Solvent flattening with combined maximization of entropy and likelihood represents a phase-refinement path independent of atomic models, using the experimental amplitudes and the most reliable phases. It should, therefore, become a valuable and generally useful procedure in macromolecular crystal structure determination.

1. Introduction

Exponential modeling and likelihood maximization are computer algorithms for applying maximum-entropy and Bayesian inference methods to the phase problem in X-ray crystallography (Bricogne, 1988*b*, 1992). Their effectiveness in *ab initio* phase determination has been demonstrated in successful applications involving small-molecule and especially powder crystal structures (Gilmore, Bricogne & Bannister, 1990; Bricogne, 1991; Gilmore, K. Henderson & Bricogne, 1991) and in electron

0907-4449/93/010193-20$06.00

microscopy (Dong *et al.*, 1992). Likelihood ranking has been shown to be a powerful discriminator of phase sets generated by traditional direct-methods algorithms for small protein data sets (Gilmore, A. N. Henderson & Bricogne, 1991). There are compelling arguments for believing that the greatest utility of these methods may lie in macromolecular crystal structure determination (Bricogne, 1988*a*, 1993). Yet, despite these successes and expectations, they have not yet proved to be of practical value when applied jointly to protein crystal structures. There are several possible reasons. At best these methods are only beginning to be understood by practising crystallographers. At worst, they are misunderstood and/or viewed in unrealistic ways, so their great promise has engendered high expectations that have not, as yet, been realized in practice. Although the foundations for these algorithms were established nearly a decade ago, they have remained intimidating from a mathematical standpoint and their intuitive rationalization (Bricogne, 1988*b*) has not been widely appreciated. Finally, development and distribution of a user-friendly computer program from that originally described for use with small molecules (Bricogne & Gilmore, 1990) has been slowed by a lack of experience with handling the unique aspects of macromolecular data sets. Consequently, the widespread interest in the methods is often accompanied by skepticism regarding their value.

Much of the misunderstanding surrounding these methods stems from a preoccupation with the long-range goal of *ab initio* phase determination (Sjölin, Prince, Svensson & Gilliland, 1991), which remains exceedingly difficult. A brief and perceptive analysis of this preoccupation is given by Lemarechal & Navaza (1991). The theoretical justifications for the methods are not limited to the *ab initio* problem, however, and they provide considerable motivation to utilize the methods in ways appropriate for dealing with familiar problems in macromolecular structure determination. It is therefore important to clarify intuitively what combined entropy and likelihood maximization entails, what the methods can and cannot do, and to explore their effectiveness in conventional contexts. Here we address ourselves to these goals with a brief review of the theory and a description of computational procedures, followed by a summary of results.

2. Review of theory

Although the theoretical foundations for the work described herein have been presented in previous publications (Bricogne, 1984, 1988*a*,*b*), and although these provided conclusive validation for previous anticipation of a possible role for information theory in the X-ray phase problem (Bricogne, 1982; Narayan & Nityananda, 1982; Britten & Collins, 1982; Wilkins, Varghese & Lehmann, 1983; Piro, 1983), it is useful to restate some of the key ideas in less mathematical terms, in order to facilitate a better understanding of how and why the methods work, and to lay a foundation for the computational algorithms and results which are presented later.

2.1. *Partially phased structures and conditional probability*

Statistical approaches to the phase problem in crystallography (known as 'direct methods') are based on the idea of a probability distribution describing where atoms might be located in the (as yet unknown) structure contained in the crystallographic unit cell. A probable distribution of atomic positions is something like an electron-density map, with probabilities, not electron densities, at each pixel. Initially, this distribution is assumed to be uniform because we know nothing about where the atoms are. The probable distribution of atomic positions changes, however, as phases and other aspects of the crystal structure become known. Owing to the mathematical relation between atomic positions and structure factors, any knowledge about one of these changes the probable distributions of the other. Thus, as phases become known, the probable distribution of atomic positions changes and so, too, does the probable distribution of the remaining phases. When enough known phase information has accumulated the probable distribution of the remaining phases becomes so sharp as to actually determine their values with great accuracy, and the crystal structure is solved.

A brief illustration is useful. The *uniform* distribution, which has a constant value everywhere, is the safest assumption if there is no information regarding the atomic positions. However, once X-ray data have been collected, we are no longer totally ignorant about the distribution of atomic positions, and the uniform distribution no longer applies. Knowing the experimental data, one can usually choose phases for at least two or three strong reflections, to fix the origin of the unit cell. Doing so changes the atomic probability distribution law irreversibly; suddenly some regions of the unit cell are more likely than others to contain atoms. The new probability distribution is said to be *conditional*, in the sense that it is valid only if given the specific phase choices that have already been made. Associated with this is also a new conditional probability distribution for the remaining unknown phases.

What are these new conditional distributions of atomic positions and phases? The problem of choosing new distribution laws for the random atomic positions and unknown phases, once some of the reflections are phased, is the quintessence of the phase problem in crystallography. One approximation to the probable distribution of atomic positions is the Fourier synthesis calculated from the structure factors of the phased reflections. As a probability distribution, this map is not very realistic: it may have regions of very high probability, matched with regions with 'negative' probability, an ill-defined concept. It is also a bad approximation because the missing reflections all have amplitudes of zero, whereas one has actually measured them to be non-zero. Thus, the Fourier transform

of any small subset of phased reflections, *by itself*, is never a good guess regarding the distribution of atomic positions conditional on these phase assumptions.

The two defects in the Fourier transform of phased reflections are closely connected, because we can modulate the peaks and eliminate the holes of the map if, and only if, we introduce reasonable estimates for the phases and amplitudes of the missing reflections. Clearly, the probable distribution of atomic positions and the electron-density map converge to very similar functions when the structure is solved. However, even with refined high-resolution structures, significant numbers of reflections may not have been included in the data set. So, one way or another, making optimal estimates for the conditional distribution of atomic positions can contribute to a structure solution in different ways, depending on how much phase information is available. Various contexts are summarized in Table 1.

2.2. *The basis set and maximum-entropy extrapolation*

We can divide the X-ray data into two parts - a *basis set* of phased reflections, $\{\mathbf{H}\}$, and the complementary set, $\{\mathbf{K}\}$, for which phases are unknown. Given any basis set, $\{\mathbf{H}\}$, choosing an appropriate probability distribution for the atomic positions is equivalent to choosing amplitudes and phases for the reflections $\{\mathbf{K}\}$, outside the basis set, to modulate the defects in the simple Fourier transform of $\{\mathbf{H}\}$. The resulting conditional probability distributions for the phases in $\{\mathbf{K}\}$ are obtained by a weighting procedure that compares the estimated amplitudes with their corresponding observed values. This amounts to extrapolating phase information from reflections in $\{\mathbf{H}\}$ to those with unknown phases in $\{\mathbf{K}\}$.

2.3. *Exponential modeling*

So far we have not described any specific algorithms for providing either the most-probable distribution of atoms or the centroid phases in $\{\mathbf{K}\}$. This problem has been recognized and discussed by many authors (Hauptman & Karle, 1953; Luzzati, 1955; Klug, 1958). A comprehensive description is given by Bricogne (1984), who provided a non-trivial proof that the desired conditional distribution of atoms can be approximated uniquely by means of a mathematical device called the saddlepoint (SP) method. This approximation involves constructing an exponential model for the distribution of atoms,

$$q^{\mathrm{ME}}(\mathbf{x}) = \exp \sum_{\mathbf{h}\in\mathbf{H}} \zeta_{\mathbf{h}} \exp(-2\pi i \mathbf{x}.\mathbf{s}_{\mathbf{h}}) \qquad (1)$$

whose Fourier transform, $\{U^{\mathrm{ME}}\}$, matches the amplitudes and phases of reflections in $\{\mathbf{H}\}$ (Collins, 1982). Construction of this model can be carried out by a process called 'exponential modeling', described below. It results in a map, $q^{\mathrm{ME}}(\mathbf{x})$, that has maximum entropy, $S = -\sum_{i=1}^{M} q_i \log q_i$, subject to the constraints in $\{\mathbf{H}\}$. The maximum-entropy distribution restricts where atoms may be located only to the extent implied by the amplitudes and phase choices in $\{\mathbf{H}\}$. Therefore it will optimally modulate the undesirable features of peakiness and negativity associated with the electron-density map. Concomitantly, as noted above, it will also make the best possible estimates for amplitudes and phases in $\{\mathbf{K}\}$, given those in $\{\mathbf{H}\}$. Far from being an arcane transplant from statistical mechanics, the concept of a constrained maximum-entropy distribution is highly intuitive in the context of conditional probability. It offers no more nor less than the shrewdest statistical guesses about reflections in $\{\mathbf{K}\}$ permitted by the current phase choices represented by the constraints, $\{U^{*}_{\mathbf{H}}\}$. The corresponding conditional probability distribution of structure factors in K is then given by a multivariate Gaussian distribution centered around the distribution of extrapolated structure-factor values.

2.4. *The global log-likelihood gain criterion*

The pivotal element of the exponential model is thus that its Fourier transform provides estimates for structure factors beyond the basis set, in the complementary set $\{\mathbf{K}\}$. Useful estimates for the conditional probability distributions associated with these extrapolated structure factors can be constructed with a multivariate Gaussian centered around the vector of structure factors, $\{U^{\mathrm{ME}}\}$, and whose covariance matrix is approximated by a diagonal matrix with elements, $1/N(\varepsilon_h)$, where ε_h are the statistical weights for the space group. These conditional distributions can be integrated over the unknown phases to produce the conditional marginal probability distributions of the directly observable structure-factor amplitudes, $\{|U_{\mathbf{k}}|\}$. There are slightly different expressions for acentric and centric reflections. In terms of unitary structure factors, U

$$P^{\mathrm{SP}}_{\mathrm{diag}}(|U_{\mathbf{k}}|\,|\,U_{\mathbf{H}}=U^{*}_{\mathbf{H}}) = (2N/\varepsilon_{\mathbf{k}})|U_{\mathbf{k}}|\exp[-(N/\varepsilon_{\mathbf{k}})(|U_{\mathbf{k}}|^2 + |U^{\mathrm{ME}}_{\mathbf{k}}|^2)] \times I_0[(2N/\varepsilon_{\mathbf{k}})|U_{\mathbf{k}}||U^{\mathrm{ME}}_{\mathbf{k}}|], \qquad (2)$$

for an acentric reflection $\mathbf{k}$ and

$$P^{\mathrm{SP}}_{\mathrm{diag}}(|U_{\mathbf{k}}|\,|\,U_{\mathbf{H}}=U^{*}_{\mathbf{H}}) = (2N/\pi\varepsilon_{\mathbf{k}})^{1/2}\exp[-(N/2\varepsilon_{\mathbf{k}})(|U_{\mathbf{k}}|^2 + |U^{\mathrm{ME}}_{\mathbf{k}}|^2)] \times \cosh[(N/\varepsilon_{\mathbf{k}})|U_{\mathbf{k}}||U^{\mathrm{ME}}_{\mathbf{k}}|],$$

for a centric reflection $\mathbf{k}$.

Since the amplitudes $\{|U_{\mathbf{k}}|\}$ are known from experimental measurements, these expressions can be evaluated quantitatively and used to compare them with those obtained from maximum-entropy extrapolation. The logarithm of the ratio between these probabilities and the corresponding expressions for the Wilson distribution based on the hypothesis of a uniform distribution of atomic positions, given by $P^{\mathrm{SP}}_{\mathrm{diag}}(|U_{\mathbf{k}}|\,|\,U_{\mathbf{H}} = 0)$, is called the log-likelihood gain (Bricogne, 1988*b*)

Table 1. *Representative methods of structure solution*

	Small basis set	Moderate basis set	Large basis set
Size of the basis set	$\{\mathbf{H}\} \ll \{\mathbf{K}\}$	$\{\mathbf{H}\} \gtrsim \{\mathbf{K}\}$	$\{\mathbf{H}\} \gtrsim \{\mathbf{K}\}$
Uses for exponential modeling	Tree-directed search for new phases	Density modification to improve maps	Super-resolution of maps
Uses for likelihood	Good figure of merit for different nodes	Indicator for convergence	Not applicable – too few F_{obs} for reflections in $\{\mathbf{K}\}$
Role in phase determination	*Ab initio* phase determination	Refinement of experimental phases	Phase extension?
Crystal structure status	Unsolved	May be solved	Solved

$$L(U^{\mathbf{K}}) = \sum_{\substack{\mathbf{k}\in\mathbf{K} \\ \mathbf{k}\ \text{acentric}}} \{\log I_0[(2N/\varepsilon_{\mathbf{k}})|U_{\mathbf{k}}^{\text{obs}}||U_{\mathbf{k}}^{\text{ME}}|] - N/\varepsilon_{\mathbf{k}}|U_{\mathbf{k}}^{\text{ME}}|^2\}$$

$$L(U^{\mathbf{K}}) = \sum_{\substack{\mathbf{k}\in\mathbf{K} \\ \mathbf{k}\ \text{centric}}} \{\log\cosh[(N/\varepsilon_{\mathbf{k}})|U_{\mathbf{k}}^{\text{obs}}||U_{\mathbf{k}}^{\text{ME}}|] - N/2\varepsilon_{\mathbf{k}}|U_{\mathbf{k}}^{\text{ME}}|^2\}. \quad (3)$$

Summing over all reflections gives the global log-likelihood gain. This function is called a log-likelihood *gain* because it measures how much the probability of the data (the observed structure-factor amplitudes for reflections in $\{\mathbf{K}\}$) has been enhanced by choosing the current phases in $\{\mathbf{H}\}$. Those familiar with the Sim probability distribution for structure factors obtained for a portion of the atoms in the unit cell (Sim, 1959) may recognize the first term of the summands in the argument of the Sim weighting factor, with $U_{\mathbf{h}}^{\text{ME}}$ replacing F^{partial}. Qualitatively, this function increases as the quantitative agreement between the observed and extrapolated structure-factor amplitudes improves.

2.5. *Centroid maps*

The maximum-entropy distribution, $q^{\text{ME}}(\mathbf{x})$, is not, strictly speaking, an electron-density map. Minimum variance estimates for the electron density itself are obtained from $q^{\text{ME}}(\mathbf{x})$ by analogy with the arguments of Blow & Crick (1959) and also Bricogne & Gilmore (1990). Centroid estimates for the structure factors in $\{\mathbf{K}\}$, involving the observed amplitudes together with Sim-like weights, are obtained from the first moment of the conditional probability distributions:

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

$$\langle U_{\mathbf{k}}\rangle = |U_{\mathbf{k}}^{\text{obs}}|[I_1(X_{\mathbf{k}})/I_0(X_{\mathbf{k}})]\exp(i\varphi_{\mathbf{k}}^{\text{ME}})$$

with

$$X_{\mathbf{k}} = (2N/\varepsilon_{\mathbf{k}})|U_{\mathbf{k}}^{\text{obs}}||U_{\mathbf{k}}^{\text{ME}}|$$

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

$$\langle U_{\mathbf{k}}\rangle = |U_{\mathbf{h}}^{\text{obs}}|\tanh(X_{\mathbf{k}})\exp(i\varphi_{\mathbf{k}}^{\text{ME}})$$

with

$$X_{\mathbf{k}} = (N/\varepsilon_{\mathbf{k}})|U_{\mathbf{k}}^{\text{obs}}||U_{\mathbf{k}}^{\text{ME}}|.$$

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

Finally, the likelihood, (3), is a very sensitive criterion for the correctness of the constraint phases used to construct $q^{\text{ME}}(\mathbf{x})$ (Bricogne, 1984). This has two important consequences for the phase problem. Firstly, as demonstrated by Gilmore, A. N. Henderson & Bricogne (1991) the likelihood is a very good discriminator between exponential models based on different phase sets. As a result, it can be used effectively in tree-directed searches for *ab initio* phases, as is done for small molecules (Woolfson, 1980). Secondly, since the likelihood statistic tends to assume a maximum for the correct basis-set phases, it provides a way to refine the basis-set phases.

To summarize, expressions (1)–(3) offer a novel and very powerful addition to existing methods for solving crystal structures. Table 1 shows three different representative situations, from *ab initio* phasing to high-resolution phase extension. It is known that the phase problem is usually generously over-determined, and that the unknown phases are coupled mathematically to the observed amplitudes (Sayre, 1952*b*; Bricogne, 1974; Rothbauer, 1980). The real strength of exponential modeling and likelihood methods is that they exploit the phase information contained in these complex coupling patterns. The likelihood (3) serves to evaluate quantitatively how well the exponential model predicts amplitudes in $\{\mathbf{K}\}$, and hence the correctness of the basis-set phases. Bricogne has reviewed the contexts in which these two devices may be useful, including a comprehensive plan for *ab initio* phasing (Bricogne, 1988*a*, 1993).

2.6. *Relevance of statistical direct methods to macromolecular crystallography*

Although the long-term goal of *ab initio* phase determination is appealing, construction of a maximum-entropy exponential model (1) from a set of known phases is of more immediate importance to protein crystallography. Since it represents the optimal way to couple reliable knowledge about reflections in $\{\mathbf{H}\}$ to the set of reflections for which phases are less certain, it should provide a way to insure that the phases used to solve a structure are as consistent as possible with the observed amplitude data, *i.e.* that the best possible phases are used. Since the relations that couple amplitudes and phases arise from conditional probability, they depend explicitly on the quality of the known phases, and hence on the size and accuracy of the basis set. It therefore makes sense to explore their use first in contexts involving a considerable foundation of reliable phase information. Several aspects of pro-

tein crystal structure determination and refinement suggest that these methods should be very beneficial, for example, with problems suggested in the second column of Table 1, where the primary phase determination has produced phases sufficient to solve the structure, but where existing refinement methods are weak:

(i) Numerous structure solutions rely heavily on density-modification methods to improve the poor quality of isomorphous replacement phases (Podjarny, Bhat & Zwick, 1987). The most popular method involves 'solvent flattening' (Wang, 1985), a procedure that repeatedly strives to eliminate features of density outside the molecular boundary where the density should be uniform and, thereby, correct errors within the boundary caused by poor quality phases. This process is inefficient, especially so if the starting isomorphous replacement phases are of poor quality (Fenderson, Herriott & Adman, 1990). Moreover, solvent-flattened maps invariably appear 'improved' in a subjective sense. But as noted in §4, we have found recently that conventional solvent flattening can introduce significant errors into the flattened map.

(ii) When the available phase information is insufficient to solve a structure, even with extensive solvent flattening, the added phasing power afforded by the amplitudes themselves could provide the difference between solving these structures and not solving them at all.

(iii) Occasionally, and perhaps more frequently than is realized, errors are made interpreting electron-density maps which are never corrected by subsequent refinement (Brändén & Jones, 1990, and references cited therein; Backes *et al.*, 1991). Crystallographic refinement of atomic models against the data affords little protection against such errors (Howard, Lorsbach, Ghosh, Melis & Stout, 1983; Stout, Turley, Sieker & Jensen, 1988; Stout, 1988) because statistical phase indications are intrinsically multimodal (Bricogne, 1984), and the model itself introduces such a strong bias that quite good *R* factors can be obtained with models corresponding to incorrect local minima. For this reason it is important to use the best possible map for the initial interpretation.

Exponential modeling and likelihood maximization have important advantages in these situations:

(i) Additional phase information comes from the amplitudes for the native structure. These are normally the highest quality data involved in the phase determination.

(ii) Native amplitudes have no systematic errors due to lack of isomorphism.

(iii) Phasing power arising from the native amplitudes is model independent and unbiased.

3. Computational procedures and X-ray data sets

We have modified the computer program, *MICE* (*M*aximum entropy *I*n a *C*rystallographic *E*nvironment; Bricogne & Gilmore, 1990) to work with protein data sets. *MICE* integrates exponential modeling and likelihood evaluation, and has been thoroughly tested with a variety of X-ray data sets involving small-molecule structures (Bricogne & Gilmore, 1990; Gilmore, Bricogne & Bannister, 1990; Bricogne, 1991; Gilmore, A. N. Henderson & Bricogne, 1991; Gilmore, K. Henderson & Bricogne, 1991; Dong *et al.*, 1992). Our purpose here has been to enable *MICE* to work with larger data sets and, in particular, to make use of a known molecular envelope and then to evaluate its performance in the conventional context of solvent flattening.

3.1. *Algorithms*

Algorithms for exponential modeling and log-likelihood calculation have previously been designed to implement formulas (1) and (3) (Bricogne & Gilmore, 1990). A brief description follows.

3.1.1. *Exponential modeling.* The Fourier coefficients of the desired exponential model, $\{\zeta_{\mathbf{H}}\}$, are parameters, which must be built iteratively by fitting the model to the constraints, $\{U_{\mathbf{H}}\}$. The fitting process is illustrated in Fig. 1. Each cycle begins with calculation of the exponential model based on the current parameters by Fourier transformation of $\{\zeta_{\mathbf{H}}\}$, followed by exponentiation of that map. This gives the current estimate for $q^{\mathrm{ME}}(\mathbf{x})$. Shifts to bring the exponential model into closer agreement with the constraint values on the next cycle, $\{\Delta\zeta_{\mathbf{H}}\}$, are estimated from the difference Fourier coefficients by back transformation and map division.

Convergence is achieved either when the entropy of successive iterations is stationary, or when the constraints are fitted to within a reduced χ^2-like statistic, C, equal to 1.0. The C statistic is given by

$$C = \tfrac{1}{2} \sum_{\substack{\mathbf{h}\in\mathbf{H} \\ \mathbf{h}\ \text{acentric}}} (|U_{\mathbf{h}} - U^*_{\mathbf{h}}|^2)/(\mathbf{p}\varepsilon_{\mathbf{h}}\Sigma_a + \sigma^2_{\mathbf{h}}) + \tfrac{1}{2} \sum_{\substack{\mathbf{h}\in\mathbf{H} \\ \mathbf{h}\ \text{centric}}} (|U_{\mathbf{h}} - U^*_{\mathbf{h}}|^2)/(\mathbf{p}\varepsilon_{\mathbf{h}}\Sigma_c + \sigma^2_{\mathbf{h}}),$$

where $U^*_{\mathbf{h}}$ is the constraint value from the phased $\{U^{\mathrm{obs}}_{\mathbf{h}}\}$, $U_{\mathbf{h}}$ is the current value, $\epsilon_{\mathbf{h}}$ is the usual statistical weight of reflection **h**, Σ_a and Σ_c are refinable variance parameters representing the reciprocal of the number of effective scatterers (see Bricogne & Gilmore, 1990, §§2.3.3 and

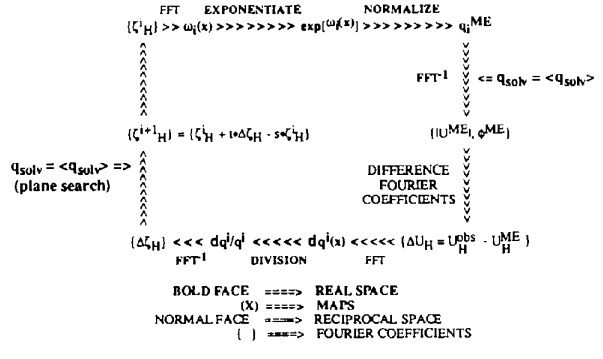


Fig. 1. Exponential modeling.

2.4), and $\sigma^2_{\mathbf{h}}$ is the (experimental) variance of $U^{obs}_{\mathbf{h}}$. An adjustable parameter, **p**, determines the relative weights given to the Σ variance parameters and the experimental $\sigma^2_{\mathbf{h}}$. Softening the constraints in this manner helps to prevent overfitting of the model when the basis-set phases are incorrect. In practice, the best indicator of convergence seems to be a maximum in the log-likelihood gain, and adjustment of the **p** parameter is often necessary either to soften or tighten the constraints according to the ability of the exponential model to fit the $\{U^{obs}_{\mathbf{h}}\}$ while still increasing the likelihood.

3.1.2. *The plane search.* The current iterate of the model can be moved into better agreement with the constraints by a large variety of different shifts, with varying impact on the entropy. The shifts actually used are calculated from the $\{\Delta\zeta_{\mathbf{H}}\}$ after an algorithm, known as the plane search, is used to find the best compromise between the constraint and entropy-search directions (the parameters t and s; Bricogne & Gilmore, 1990). In the plane search, a decision is first made regarding the appropriate force of the constraints - how closely the Fourier transform of the exponential model should agree with the constraints after the next iteration. The parameter s then subtracts a portion of the current $\{\zeta_{\mathbf{H}}\}$ in order to flatten the map and increase its entropy. The behavior of the exponential model over a patch of the (s,t) plane surface is approximated by a bicubic function, which is then contoured at the targeted level of the constraint value to determine the s and t parameters giving maximum entropy (Bricogne & Gilmore, 1990).

In the process of building the parameters of the exponential model (the ζ's) terms of the Fourier series for the electron density outside the basis set are built up in a statistically meaningful way, *via* $\{U^{ME}\}$ until, at convergence, there is a set of extrapolated values for the phases and amplitudes of reflections outside the basis set. This process of extrapolation is central to the use of exponential modeling in the determination and refinement of phases.

3.2. *Adaptations for protein data sets*

MICE has been modified for work with protein data sets. The necessary modifications included minor changes in array dimensions and inclusion of new functions for inputting a molecular envelope and using it to reset the solvent regions to their average value.

Array sizes in *MICE* were, for the most part, designed for small molecules at very high resolution; protein data sets require larger arrays, because the number of observed data is very much larger than for small molecules. Also of importance is the size of maps calculated by Fourier transformation. For small-molecule problems *MICE* is generally run with maps in space group *P*1, sampled at 0.2 Å intervals, or less. Since the size varies as the cube of the linear dimension of the unit cell, this resolution cannot be used for protein data sets on current computers. Aliasing errors can occur if the maps are not sufficiently finely sampled (Shannon, 1949; Sayre, 1952*a*), and we have, therefore, studied this possibility in order to determine the best compromise between resolution and aliasing. We have looked for several kinds of indicators that might suggest aliasing problems: substantial extrapolation beyond the second neighborhood of the basis set, lack of agreement between *P*1 structure factors related by space-group symmetry, and the overall appearance of centroid maps. We have found no evidence of this problem working close to the Shannon limit at three times the effective resolution of the data. This means that working with a 3.0 Å data set requires Fourier transforms be carried out on maps sampled only at 1.0 Å intervals. This size is quite manageable with current RISC architecture workstations, and it permits useful improvements in many protein electron-density maps.

Protein data sets do, however, present several unique problems that had to be addressed before these methods could be applied directly. Preparation of the structure-factor amplitudes required a special scaling treatment to estimate the F_{000}, and the *MICE* program had to be modified to utilize a molecular envelope in an effective manner.

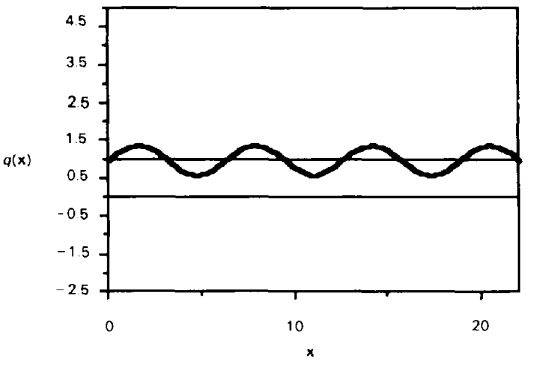
3.2.1. *Data preparation.* Normally, structure-factor amplitudes are prepared for use in direct-methods programs by a process known as normalization, in which the Lorentz- and polarization-corrected intensities are put on absolute scale, (optionally) sharpened with an overall temperature factor, and then sharpened again by division by the mean-square amplitude in shells of resolution (Rogers, 1980). The resulting normalized structure-factor amplitudes, representing the root-mean-square deviation from the mean, or variance, of the scattering amplitudes, are taken to represent the amplitudes from a structure of point atoms at rest. The advantage of using normalized structure factors is considered to be that they simplify expressions of the reciprocal space marginal joint probability distributions for crystals with heterogeneous atomic composition by isolating the trigonometric part of the structure-factor expression. The utility of normalization in estimating real-space approximations to non-uniform joint probability distributions has been questioned (Bricogne, 1988*a*; Bricogne & Gilmore, 1990), and it has been pointed out that in these cases normalization is both ineffectual in dealing with heterogeneous populations of atoms and unnecessary if the heterogeneity can be represented by a multichannel maximum-entropy approach (Bricogne, 1988*a*). The large solvent channels present in protein crystals represent an important source of such heterogeneity, and they alter the distribution of amplitudes significantly at low resolution, thereby posing difficulties with the estimation of normalized structure factors from the measured amplitudes (Harker, 1953; Carter, Crumley, Coleman, Hage & Bricogne, 1990). For these reasons, $|E|$ values are not used in this work.

MICE uses unitary structure factors, $\{U_{\mathbf{h}}\}$, where $|U_{\mathbf{h}}| = |F^{obs}_{\mathbf{h}}|/F_{000}$ and U_{000} is equal to 1.0. Under these circumstances, only one scaling parameter needs to be applied to

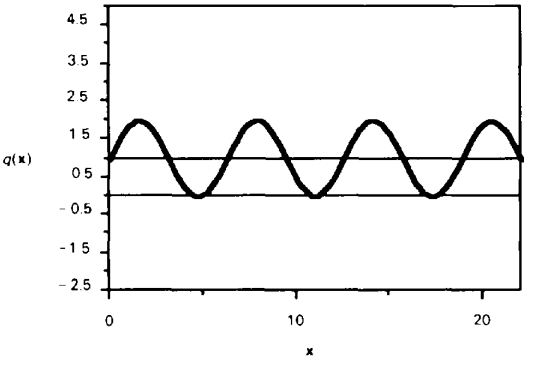
$|F_{\mathbf{h}}^{\text{obs}}|$, namely, the $|F_{000}|^{-1}$ on the same scale as the data. Use of an appropriate scale factor is important, since scale factors that are either too large or too small will adversely affect the fitting process. These difficulties can be visualized with the aid of a simple one-dimensional illustration (Fig. 2). For unitary structure factors on the correct scale, the fitted exponential model will have a minimum value equal to (or slightly greater than) 0.0 (Fig. 2*a*). If they are too small, then addition of U_{000} = 1.0 will increase this minimum value, decreasing the dynamic range of q^{ME} and its Fourier transform (Fig. 2*b*). Fitting of the $\{\zeta_{\mathbf{H}}\}$ will cease prematurely in this case, with weakened extrapolation into the set $\{U_{\mathbf{K}}\}$, because the reduced dynamic range of the constraining Fourier coefficients implies too smooth a distribution to be fitted. Alternatively, if the scale factor is too high, then the initial Fourier map will have excessively large negative regions (Fig. 2*c*), contradicting the assumption of a positive distribution of atoms. Truncation of these negative regions during exponential modeling will corrupt the shape of the distribution to be fitted. Fitting will terminate prematurely with a decrease in log-likelihood gain, this time because the exponential model cannot reproduce the constraint structure factors, and maximum-entropy extrapolation is incorrect.

In either case, the correct scale constant should lead to a higher likelihood than scale constants that are either too small or too large. We have, therefore, developed a rapid search procedure, as suggested by Bricogne (1988*a*), using the log-likelihood gain on the first cycle of exponential modeling to search for the absolute scale (and hence the U_{000}) that gives maximum likelihood. This procedure is far from rigorous, because it does not pursue each trial exponential model to maximum likelihood. Nevertheless, we have carried out numerous tests without finding exceptional cases where the first cycle of exponential modeling did not turn out to be a good indicator of the ultimate likelihood maximum. This approach has worked quite well for the cytidine deaminase experiments described below, where Wilson scaling was used first to find the approximate absolute scale of the data.

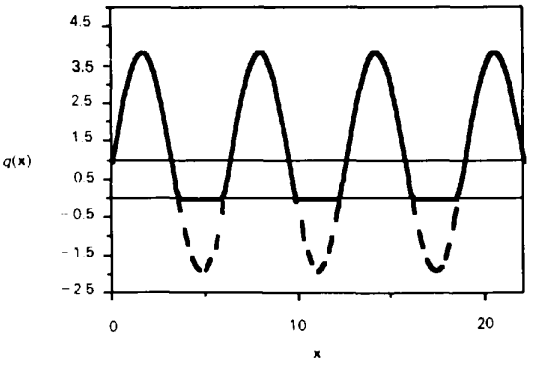
3.2.2 *Use of the molecular envelope.* When the shape of the solvent channels is known, it offers an important additional source of information. The usefulness of this information was previously investigated by Prince, Sjölin & Alenljung (1988). That work utilized exact data and a known envelope, much as we describe in our initial model studies, but should not be confused with the present work as they maximized the entropy of maps after conventional solvent flattening by fitting the entire (F_{obs} - F_{calc}) data sets exactly, whereas we have found that use of maximum likelihood as a stopping criterion is essential to prevent overfitting of imperfectly phased structure-factor data. Anecdotal reference is made to work with experimental data, but details of these extensions appear not to have been published. Incorporation of information about the envelope as a constraint in the process of expo-



(*a*) $U' < U_o$



(*b*) Correct scale: $U' = U_o$



(*c*) $U' > U_o$

Fig. 2. Difficulties associated with incorrect choices for the F_{000} term used in calculating unitary structure factors. If the unitary structure-factor amplitudes in the constraint reflections are too small relative to $|U_{000}|$ = 1.0 (*a*), the minimum value of the distribution $q(\mathbf{x})$ will be greater than 0.0, the constraints will be too soft, and the fitting process will cease before it has the opportunity for strong maximum-entropy extrapolation. For the correct scale (*b*), $q(\mathbf{x})$ achieves its minimum value nearly at 0.0, and the extrapolation at maximum likelihood is as strong as possible. If the unitary structure-factor amplitudes in the constraint reflections are too large relative to $|U_{000}|$ = 1.0 (*c*), $q(\mathbf{x})$ will have negative values. In this case, truncation of the distribution in constructing the exponential model will distort its true shape and, as the constraints are fitted the maximum-entropy extrapolation will be incorrect. The log-likelihood gain in case (*b*) should therefore always exceed that for either case (*a*) or (*b*).

Table 2. *Exponential modeling of kallikrein*

The basis set, {**H**}, consisted of the 237 out of 3163 reflections with the largest amplitudes to 5 Å resolution. The correlation coefficients are calculated using a 5 Å target map which is a fast Fourier transform of 3163 reflections with correct phases within 5 Å. $L(H) - L(0)$ is the global log-likelihood gain defined in §2.4; S is the map entropy; R_{basis} and R_{extr} are the (unscaled) crystallographic R factors for basis-set and extrapolated reflections, respectively; and CC(qme) and CC(cent) are the Fisher correlation coefficients for the q^{ME} distribution and centroid maps, respectively, as defined in §3.3, with the target map. Maximum likelihood is indicated by bold type here and in Tables 3 and 5.

p	Cycle No.	$L(H) - L(0)$	S	R_{basis}	R_{extr}	N_{extr}	CC(qme)	CC(cent)
0.5	1	171	0.0197	0.713	0.820	1422	0.766	
	3	196	0.0192	0.703	0.817	1554	0.768	
	5	222	0.0043	0.687	0.814	1707	0.771	
	7	239	0.0027	0.676	0.811	1801	0.772	
	9	252	−0.0012	0.667	0.808	1878	0.773	
	11	267	−0.0024	0.658	0.806	1953	0.774	
	13	279	−0.0041	0.650	0.804	2016	0.775	
	15	292	−0.0055	0.640	0.802	2102	0.776	
	17	309	−0.0068	0.631	0.800	2172	0.777	
	19	323	−0.0088	0.622	0.798	2243	0.778	
	21	336	−0.0108	0.613	0.796	2300	0.779	
	23	350	−0.0127	0.604	0.795	2369	0.780	
	25	360	−0.0175	0.598	0.793	2407	0.781	
	27	368	−0.0181	0.594	0.792	2445	0.781	
	29	381	−0.0197	0.587	0.791	2486	0.782	
0.1	31	389	−0.0213	0.580	0.791	2550	0.782	
	33	397	−0.0230	0.574	0.790	2607	0.783	
	35	414	−0.0393	0.563	0.789	2716	0.784	
	37	**429**	**−0.0422**	**0.550**	**0.787**	**2818**	**0.785**	**0.801**
	39	428	−0.0422	0.551	0.787	2813	0.785	

nential modeling was anticipated in the version of *MICE* described in Bricogne & Gilmore (1990), but the uniform distribution of the solvent atoms had not been exploited. One way to incorporate this information is to use two different 'channels' in which the initial distributions of atoms have different properties (Bricogne, 1988*a*). *MICE* does not implement multichannel maximum-entropy algorithms. It can, however, make use of a prior distribution of atoms, in this case represented by the envelope map. This distribution, $m(\mathbf{x})$, is used as originally described [Bricogne, 1984, §3.3.1 (ME1)] in construction of the exponential model.

For data sets involving solvent-free protein atoms, $\{F^{PROT}\}$, this procedure is adequate; there is only one type of atom (the protein atom) and it is found only where $m(\mathbf{x})$ assumes non-zero values. The presence of electron-dense solvent regions increases considerably the difficulty of constructing optimal exponential models. For real data sets where the solvent regions are electron dense, but are occupied by atoms with very high temperature factors, the multiplicative $m(\mathbf{x})$ filter provided in *MICE* is unable to accommodate the qualitatively different behavior of the two types of atoms present in the crystal. A correct statistical description of the joint distribution of structure factors requires the multichannel formalism (Bricogne, 1988*a*), in which the protein and the solvent atomic positions have separate probability distributions, updated by means of separate exponential models.

In this work we chose to implement an approximation to this formalism. Its basis is hinted by the remark (in §2.3 of Bricogne, 1988*a*) that, because the scattering factors of the two types of atoms intervene multiplicatively in the exponents of their separate exponential models, the rapid fall-off of the scattering factors of the solvent atoms will keep their distribution flat and featureless, forcing the high-resolution detail to appear exclusively within the macromolecular envelope. A fully fledged implementation of this approach would provide a statistical technique for enforcing solvent flatness in advance, which would be useful for *ab initio* phase generation. Since here, however, we were interested primarily in phase extension and refinement from a substantial amount of prior phase information we settled for an approximation in which solvent flatness outside the envelope is imposed after the calculation of q^{ME} before calculating U^{ME}, and for each of the four trial maps used by the plane-search algorithm to estimate structure-factor shifts on each cycle of iteration as indicated on the right- and left-hand sides of Fig. 1, respectively.

3.3. *Data analysis and interpretation*

We used several statistical criteria, together with subjective evaluations of electron-density maps to monitor the behavior of exponential modeling, maximum-entropy extrapolation, and the consequent modification of the centroid electron-density maps. The overall strategy was to use simulated exact data first, to establish that various indicators work as expected, then to compare their behavior with perfect and noisy simulated data to estimate roughly the radius of convergence for useful improvements in maps based on experimental data. *MICE* provides four useful statistics that need to be considered together:

(i) The relative log-likelihood gain (3) for the exponential model itself should increase.

(ii) This increase in likelihood should be achieved without unduly decreasing the entropy.

(iii) The likelihood should ideally continue to increase until the constraints are fitted to within the variance of the experimental data. This is reflected in the value of the reduced χ^2 statistic, C.

(iv) Crystallographic R factors, $R = (\sum_{\mathbf{h}=0}^{N} ||U_{\mathbf{h}}^{\text{obs}}| - |U_{\mathbf{h}}^{\text{ME}}||)/\sum_{\mathbf{h}=0}^{N} |U_{\mathbf{h}}^{\text{obs}}|$ are calculated for both basis-set and extrapolated reflections at maximum likelihood, *without scaling* U^{obs} *and* U^{ME}. They reflect the ability of exponential modeling to fit the constraints without decreasing the likelihood. Although these statistics are similar to conventional R factors, the $|U_{\mathbf{h}}^{\text{obs}}|$ and $|U_{\mathbf{h}}^{\text{ME}}|$ are not scaled together, and hence they are almost meaningless by themselves. They are quite useful when taken in the context of the behavior of the log-likelihood gain, where they provide a valuable indication of how tightly the constraints can be imposed on the exponential model without causing the log-likelihood gain to plummet.

These statistics indicate the behavior of the exponential modeling, and for example, the likelihood gain proves to be a better indicator of convergence than either the entropy

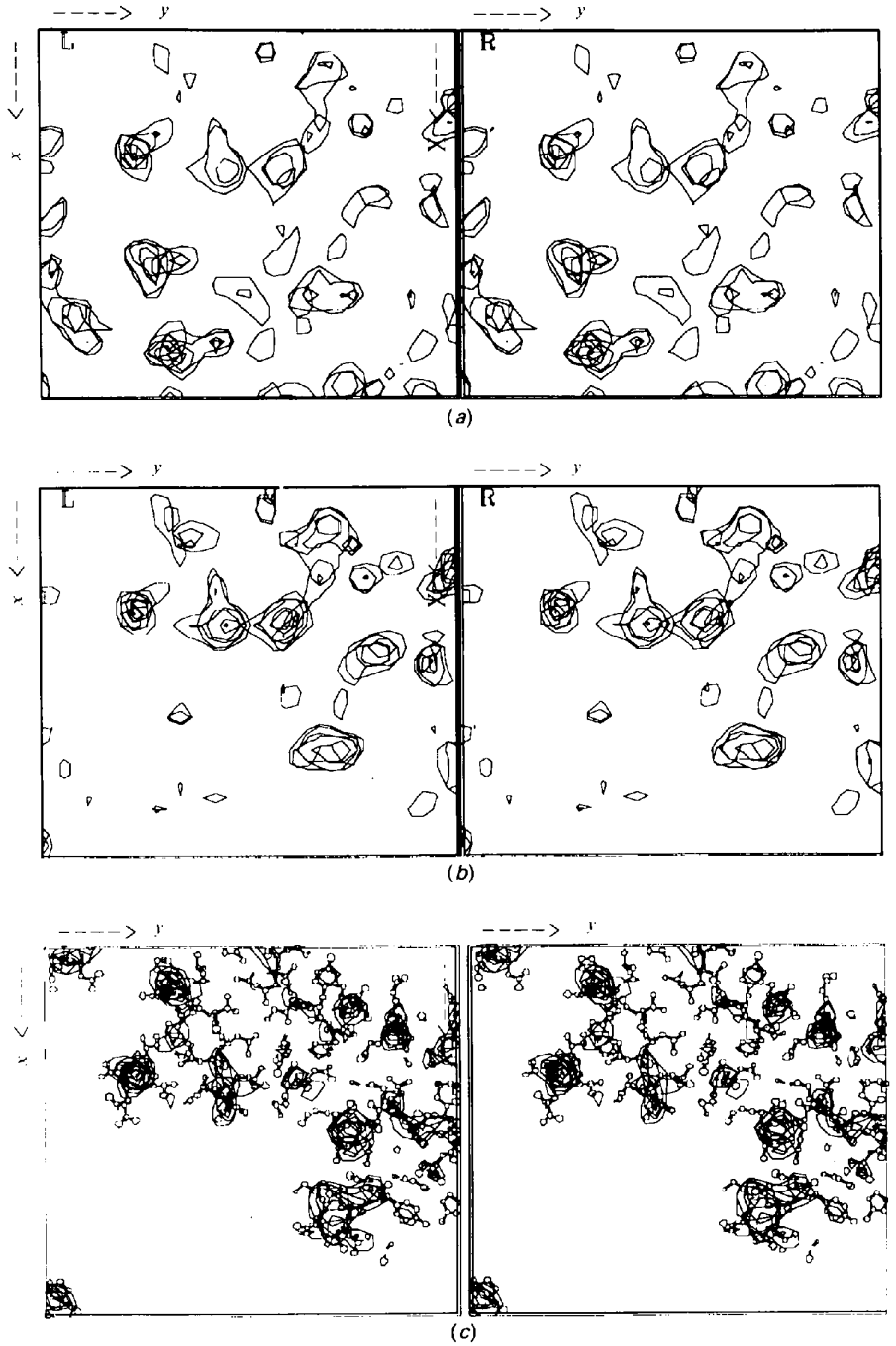


Fig. 3. Comparison of (*a*) initial, (*b*) centroid and (*c*) target maps for 5 Å exponential modeling with simulated data for protein surrounded by an electron-dense solvent, with random phase errors averaging 30°. The basis set consisted of approximately 16% of the reflections in the data set. A six section composite of each map has been contoured at 1.8σ and the target map is superimposed on the correct structure.

shift or C. We monitor the likelihood closely, stopping the fitting at maximum likelihood, and then examine the other statistics.

To evaluate quantitatively the performance of exponential modeling as a density-modification algorithm, we have compared the centroid maps at maximum likelihood (Bricogne & Gilmore, 1990, §1.6) with other maps using the Fisher correlation coefficient over the two discrete electron-density vectors, $\rho(\mathbf{x})$ and $\sigma(\mathbf{y})$ (Read, 1986):

$$R_{xy} = \{\sum_{i=1}^{M}[\rho(X_i) - \langle\rho(X)\rangle][\sigma(Y_i) - \langle\sigma(Y)\rangle]\}/ \{\sum_{i=1}^{M}[\rho(X_i) - \langle\rho(X)\rangle]^2[\sigma(Y_i) - \langle\sigma(Y)\rangle]^2\}^{1/2}.$$

Depending on the context, we have compared the centroid map to the starting (simulated or MIRAS phases), solvent-flattened maps, and to the final target maps calculated with the entire set of known correct phases and amplitudes or with coefficients $[(2|F_{obs}| - |F_{calc}|)\exp(i\varphi_{calc})]$. Comparisons are also made between the mean phase errors, $\langle\Delta\varphi\rangle$ for different maps. Phase errors were evaluated with respect to structure factors calculated for the refined model of cytidine deaminase, with solvent contributions included as described in §3.4 for the simulated data sets. In both cases (R_{xy} and $\langle\Delta\varphi\rangle$) we look for the exponential model to introduce features into the Fourier transform of $\{\mathbf{H}\}$ that bring it into better agreement with the target map.

3.4. *X-ray data sets*

The full potential of entropy maximization and likelihood ranking will probably be realized by stepwise application to problems of increasing difficulty. We have decided to focus initially on situations where considerable phase information is already available from other sources, that is with moderately sized basis sets (Table 1). In these cases it is likely that the methods will work, we can evaluate readily whether or not they do work and they are most likely to have immediate practical utility. We have carefully selected two protein data sets for use in the preliminary characterization and testing of *MICE*. Briefly, these data sets are:

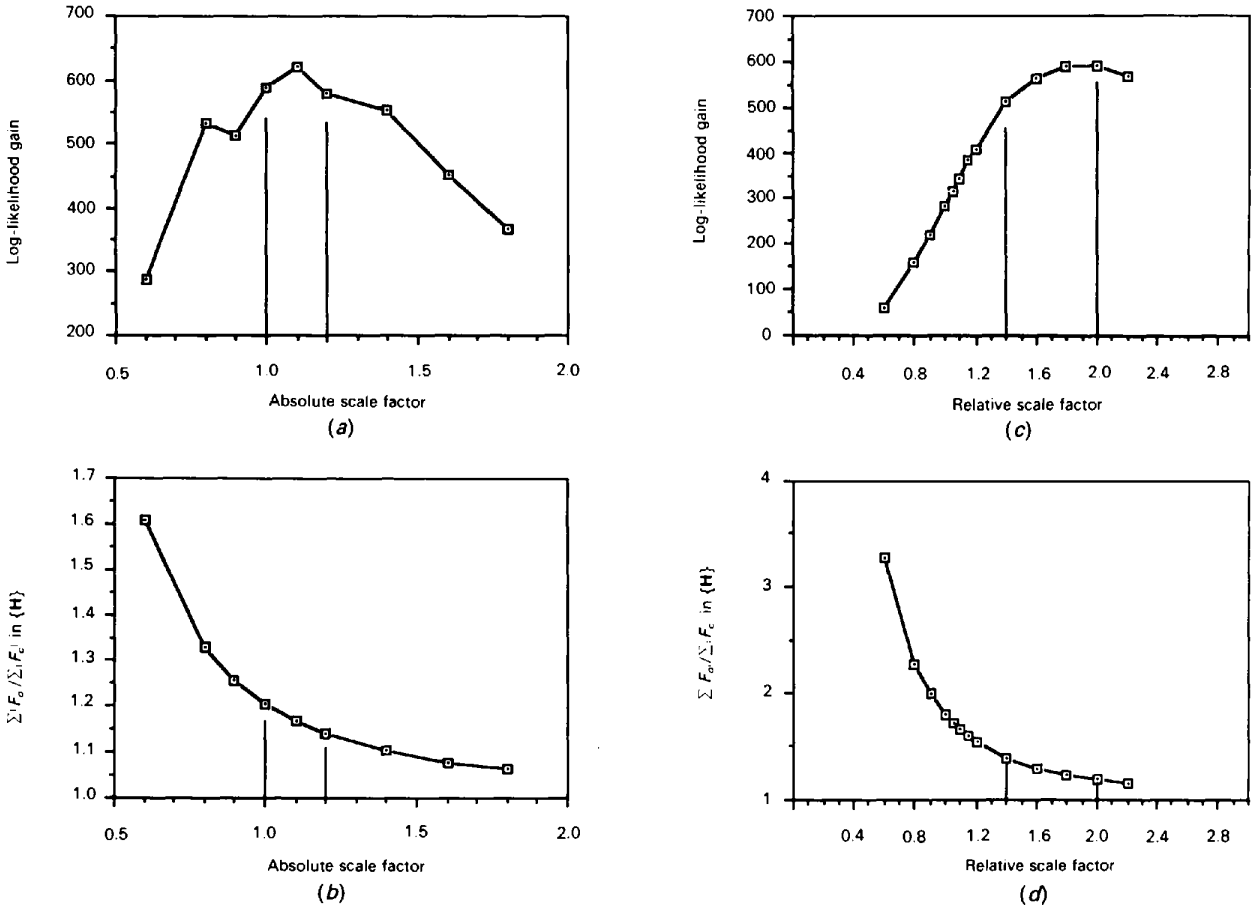


Fig. 4. Maximum-likelihood determination of the scaling parameter, $|F_{000}|$, for converting $|F_{\mathbf{h}}^{obs}|$ to $|U_{\mathbf{h}}^{obs}|$. The log-likelihood gain and ratio of the total $|U_{\mathbf{h}}^{obs}|$ to the total $|U_{\mathbf{h}}^{calc}|$ are plotted against the absolute scaling parameter for exact simulated data in (*a*) and (*b*), and against the scale factor relative to that derived from a Wilson plot for the experimental data for cytidine deaminase in (*c*) and (*d*), respectively.

(i) A simulated data set assembled from coordinates for a known structure taken from the database. The protein coordinates selected for these simulations were those of kallikrein (Bode *et al.*, 1983). This crystal form has a typical solvent content of about 55%, and was considered representative of most protein crystals, and found not to present an unusually favorable geometric redundancy from large solvent regions. Calculated structure factors, including F_{000}, based on refined coordinates were generated by Fourier transformation of electron densities, on an absolute scale, for the protein in the absence of solvent, $\{F^{\mathrm{PROT}}\}$; the envelope function, $\{\mathbf{G}\}$; and the internal fluctuations of the protein density around its mean value, $\{\Delta\}$, subject to the defining relationship $F^{\mathrm{PROT}} = \mathbf{G} + \Delta$, as described elsewhere (Carter, Crumley, Coleman, Hage & Bricogne, 1990, §1.1). Simulated $\{F_{\mathrm{obs}}\}$ structure factors and phases representing a real experiment were obtained by setting the electron density outside the molecular boundary to 0.40 e Å^{-3}, equivalent to about 75% saturated ammonium sulfate. The molecular boundary in the electron-density map simulated for solvent-free protein was chosen to include a mean protein electron density of 0.418 e Å^{-3}. In the immediate neighborhood of the protein, the density was smoothed by a simple seven-point pixel averaging procedure, constrained to satisfy the analytical relationship that the electron density of the unit cell was equal to the sum of the contents of the protein region and the region volume, $\rho(V) = \rho[\chi(U)] + \rho[\chi(V - U)]$ (Carter, Crumley, Coleman, Hage & Bricogne, 1990). Either the amplitudes or the phases, or both were then varied randomly, to simulate noise. The maximum and mean phase errors introduced were 60 and 30°, and the maximum amplitude error was $(F_{\mathrm{obs}})^{1/2}$.

(ii) Experimental $\{F_{\mathrm{obs}}, \varphi_{\mathrm{MIRAS}}\}$ data for *E. coli* cytidine deaminase, a structure recently solved and partially refined to 2.8 Å (Betts & Carter, 1991; Betts, 1991), with an envelope determined during solvent flattening.

4. Results

We have tested *MICE* with the adaptations to accommodate molecular envelopes as an alternative to using solvent flattening. Experiments with simulated data using calculated structure factors for the solvent-free protein $\{F^{\mathrm{PROT}}\}$, the solvated protein $\{F^{\mathrm{obs}}\}$, and the envelope $\{\mathbf{G}\}$ show that the process works well with protein data, that the likelihood is an excellent figure of merit for the correct basis-set phases, that the use of an envelope as an initial $m(\mathbf{x})$ map aids considerably in convergence and in overall likelihood gain, and that structure factors $\{U_{\mathbf{K}}\}$ generated by maximum-entropy extrapolation represent a considerable amount of correct new phase information. These trials also defined reasonable limits for the requisite size and quality of the basis set. Experiments with our cytidine deaminase data show remarkable improvement over the solvent-flattened map.

Table 3. *Exponential modeling of cytidine deaminase*

MIR phases for 4174 reflections to 3.2 Å resolution with an MIR figure of merit larger than 0.7 were selected for the basis set from a total of 12093 reflections in the data set.

p	Cycle No.	$L(H) - L(0)$	S	R_{basis}	R_{extr}	N_{extr}
0.5	1	935	−0.015	0.770	0.768	390
	3	1238	−0.036	0.734	0.746	652
	5	1555	−0.059	0.699	0.727	936
	7	1902	−0.086	0.662	0.705	1289
	9	2280	−0.118	0.625	0.681	1633
	11	2655	−0.158	0.589	0.659	2020
	13	3124	0.223	0.546	0.635	2468
	15	3568	0.275	0.506	0.612	2954
	17	4027	−0.330	0.466	0.589	3431
	19	4436	−0.388	0.431	0.571	3879
0.5	21	4799	0.461	0.402	0.554	4258
	23	5128	−0.531	0.376	0.539	4584
	25	5332	−0.586	0.361	0.530	4796
	27	5537	−0.634	0.345	0.519	4969
	29	5743	−0.684	0.329	0.510	5147
	32	5967	−0.767	0.313	0.501	5361
	34	6070	−0.836	0.305	0.495	5455
	36	6148	−0.882	0.299	0.491	5528
	38	**6214**	**−0.935**	**0.294**	**0.487**	**5609**
	40	6193	−0.934	0.296	0.488	5593

Table 4. *Phase combination of cytidine deaminase*

Phase combination started with 4174 MIR phases of figure of merit ≥ 0.7; new combined phases were used in the next *ab initio* exponential modeling.

	No. of reflections	⟨Phase shift⟩	⟨Old f.o.m.⟩	⟨New f.o.m.⟩
MIR phases in constraints				
Basis set	4174	6.466°	0.867	0.914
Extrapolated	5870	31.808°	0.336	0.467
Total	10044	21.276°	0.557	0.653
Recombined phases in constraints				
Basis set	4174	8.029°	0.867	0.920
Extrapolated	5888	33.194°	0.335	0.478
Total	10062	22.755°	0.556	0.661

Table 5. *Exponential modeling of cytidine deaminase after phase recombination*

Recombined phases for the same 4174 reflections used in Table 2 were used in the basis set.

p	Cycle No.	$L(H) - L(0)$	S	R_{basis}	R_{extr}	N_{extr}
0.5	1	979	−0.009	0.762	0.774	386
	3	1313	−0.036	0.722	0.750	676
	5	1679	−0.060	0.682	0.726	1005
	7	2076	−0.090	0.641	0.701	1375
	9	2519	−0.123	0.567	0.674	1795
	11	2995	−0.165	0.553	0.648	2253
	13	3480	−0.213	0.509	0.624	2749
	15	4018	−0.285	0.463	0.598	3318
	17	4523	−0.346	0.420	0.576	3846
	19	5004	−0.401	0.380	0.555	4300
	21	5405	0.462	0.348	0.536	4640
	23	5776	−0.529	0.320	0.519	4950
	25	6051	−0.580	0.300	0.508	5188
	27	6243	−0.645	0.287	0.501	5351
	29	6387	−0.692	0.278	0.494	5458
	31	6507	−0.729	0.270	0.489	5535
	33	6648	−0.765	0.259	0.483	5632
	35	6722	−0.816	0.255	0.480	5686
	37	**6769**	**−0.838**	**0.251**	**0.478**	**5729**
	39	6722	−0.858	0.258	0.480	5698

4.1. *Ideal error free data*

Test experiments with exact data used the envelope map, $m(\mathbf{x})$, as a non-uniform initial prejudice as originally described by Bricogne [1984, §3.3.1 (ME1)] in constructing the exponential model, q^{ME}. Under these conditions, *MICE* would reconstruct electron density essentially correctly, working with a basis set of only 16% of the

strongest reflections to 3.0 Å resolution. These experiments constitute a 'positive control' or a 'best case', in the sense that structure-factor amplitudes ($|F^{PROT}|$ for solvent-free protein) are maximal for reflections sensitive to the transform of the molecular envelope (Carter, Crumley, Coleman, Hage & Bricogne, 1990), the F_{000} and absolute scale factor are known exactly, as is the envelope, which serves as a very strong constraint. Exponential modeling behaves according to expectation:

(i) The likelihood continues to increase until χ^2 = 0.01, permitting almost arbitrarily strong imposition of the constraints.

(ii) The R factor for basis-set reflections at maximum likelihood is about 0.038.

(iii) The R factor for extrapolated reflections is 0.42.

(iv) Correlation coefficients between the q^{ME} maps, the centroid maps, and the target map were calculated at each cycle, and they too reach maxima. The q^{ME} map itself

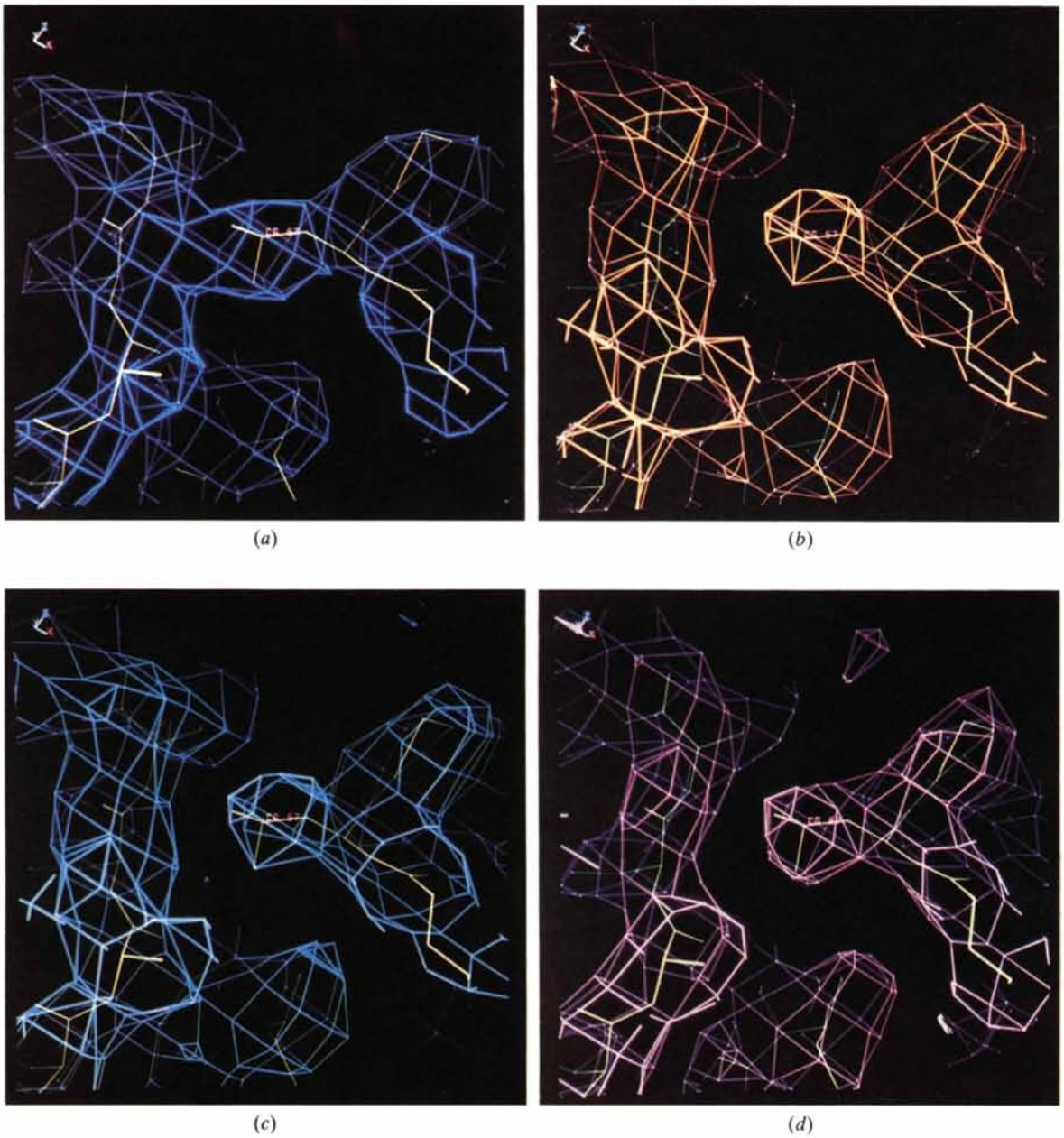
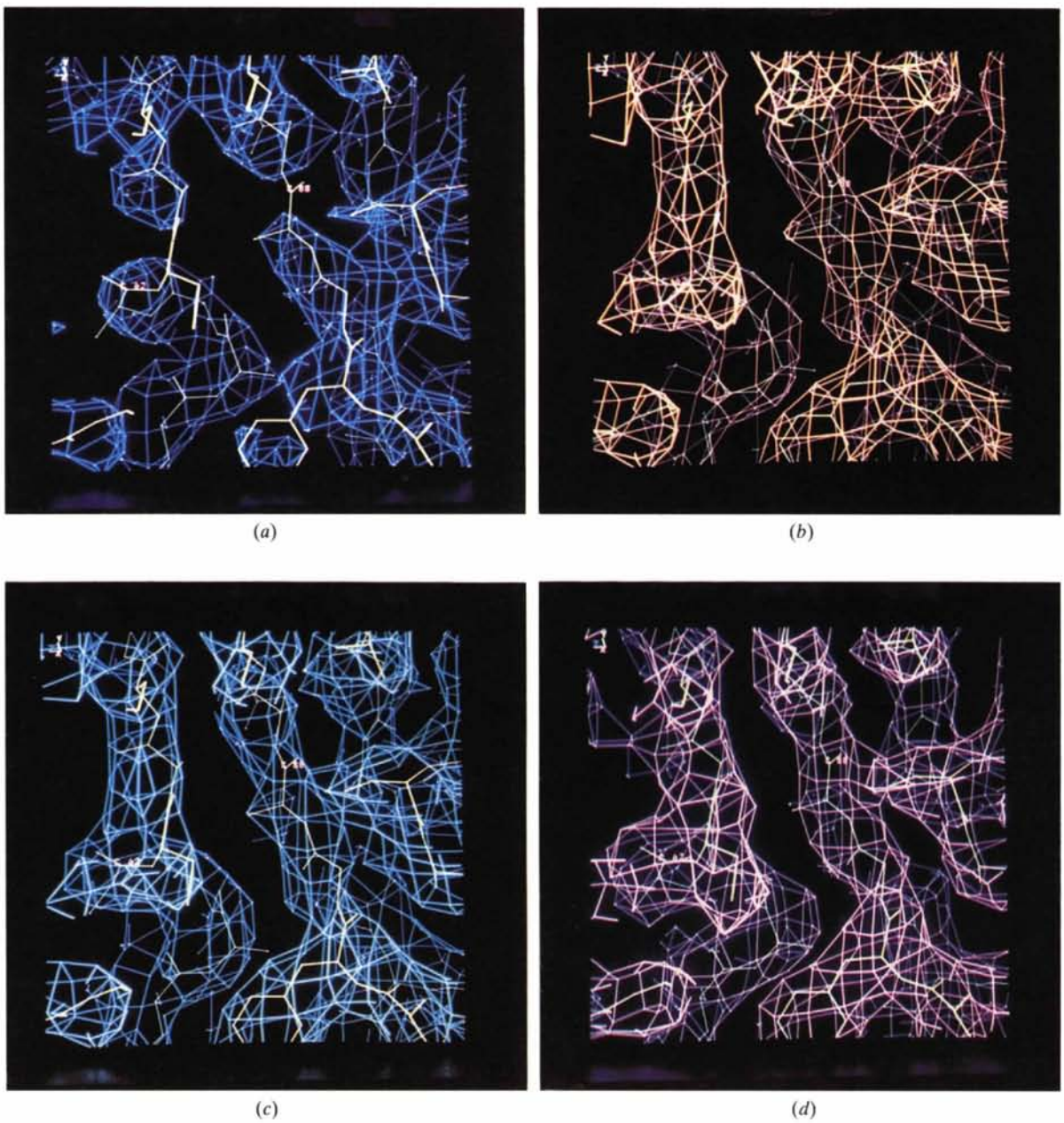


Fig. 5. Density modification of the cytidine deaminase map, based on the initial MIRAS phases in the region surrounding Leu 57. The final solvent-flattened map (*a*) fails to differentiate the density of Leu 57 from that of a neighboring main-chain segment. Two successive maximum-likelihood centroid maps, before (*b*) and after (*c*) phase recombination with the MIRAS phase probability distributions both strongly resemble the final $2|F^{obs}| - |F^{calc}|$ map (*d*) after structure refinement.

reaches a maximum correlation with the target map earlier than the centroid map, the log-likelihood gain, the R factors, or reduced χ^2 reach their optima. For the q^{ME} map, this correlation coefficient is about 0.89, while for the centroid map it is 0.93, showing that the latter is, as expected, a better representation of the electron density than is q^{ME} itself. Both the maximum-entropy extrapolation and the ability to fit the constraints therefore continue to improve in parallel beyond the point at which the exponential model reaches its maximum correlation with the target map. These results show that under ideal conditions the algorithms work exceedingly well. Maximum-entropy extrapolation predicts the amplitudes of 84% of the data to within an R factor of 0.42, quite an acceptable figure



(*a*) (*b*)

(*c*) (*d*)

Fig. 6. Density modification of the cytidine deaminase map, based on the initial MIRAS phases in the region containing Ala 62–Cys 63 and Gly 87–Asn 89. The final solvent-flattened map (*a*) is interrupted in the main-chain continuity for these two segments, one an α-helix on the left, the other a β-sheet on the right. Two successive maximum-likelihood centroid maps, before (*b*) and after (*c*) phase recombination with the MIRAS phase probability distributions both show strong, continuous main-chain density in these regions and resemble the final $2|F^{obs}| - |F^{calc}|$ map (*d*) after structure refinement. There is also improved density for the side chain of Phe 86 in the lower central region of the figure.

for structure-factor calculations from an unrefined model. Clearly, under these circumstances, the basis set is redundant, even at only 16% of the data.

4.2. *Simulated data with random errors*

Exponential modeling was next carried out in the same way using the same basis-set reflections, with exact amplitudes, but with random phase errors averaging 30°. A 'starting map' calculated with only basis-set reflections had a correlation coefficient of 0.725 with the target map. A map calculated with all 3.0 Å reflections had a correlation coefficient of 0.83 with the target map, illustrating the effects of the 30° phase errors. The centroid map at maximum likelihood had a correlation coefficient of 0.88 with the starting map, and 0.775 with the target map. Thus, with only 16% of the reflections and with random phase errors, the centroid map at maximum likelihood moved closer to the target map (0.725 → 0.775), while remaining faithful to the starting map. This phenomenon recurs in experiments using experimental data, demonstrating that exponential modeling is indeed a very conservative density-modification algorithm; in keeping with the criterion of maximum entropy, it moves only minimally away from the starting map.

4.2.1. *The log-likelihood gain is a maximum for the correct basis-set phases.* The maximum likelihood in this second case was lower by a factor of about five (1799 *versus* 8338), and the constraints could only be satisfied to within an R factor of 0.37 for basis-set reflections and 0.57 for extrapolated reflections. This is an encouraging demonstration of the fact that the likelihood is a maximum for the correct phases, and showed that this procedure could be used to test phase-refinement algorithms. It also underscores the central contribution made by the maximum-likelihood criterion in preventing overfitting of the exponential model.

4.2.2. *Accuracy of maximum-entropy extrapolation.* The accuracy of phase extrapolation was also encouraging. Distributions of phase errors for extrapolated centric and acentric reflections revealed that nearly 70% of the centric reflections were correctly extrapolated, while extrapolated acentric reflections had average, r.m.s. and U-weighted r.m.s. phase errors of 63, 80 and 59°, respectively for all reflections with Sim weights > 0.1. This distribution included 9706 of the remaining 11 491 reflections to 3.0 Å. Electron-density maps based on the phases in the basis set, the centroid map and the target map show that the centroid map has a strong resemblance to the target map where the map based on the constraints alone is very poorly defined.

4.3. *Simulated data with solvent contrast matching*

Use of simulated amplitude data, representing the protein embedded in a solvent of average density close to the mean value of protein, provided a more realistic test of the algorithms. A number of experiments were carried out using these data at different resolution limits, with different choices of strong basis-set reflections, and making different uses of the molecular envelope.

Exponential modeling under the constraint of solvent flatness, described in §3.2.2, proved to work very well with both exact and noisy simulated data. A useful illustration of this performance is shown in Table 2. Here, a basis set of the strongest 237 of 3163 (7.5%) reflections to 5 Å resolution, with random average phase errors of 30°, was used to build an exponential model using the solvent-averaging procedure. The log-likelihood gain indicates that despite the relatively small basis set, the extrapolation is quite strong. This is confirmed by the increase in the correlation coefficients between q^{ME} and centroid maps and the target map calculated with all the 5 Å reflections and the correct phases. Nearly all of the remaining reflections outside the basis set have significant extrapolation. Nevertheless, the exponential model itself leaves much to be desired. Although the entropy is high, the ability to fit the constraints is rather poor, as indicated by the R factor for basis-set reflections. Further fitting to the constraints results in a dramatic decrease in the likelihood (not shown). This example is typical of what might be encountered in a low-resolution structure determination with weak isomorphous replacement phase information. Fig. 3 presents a comparison of the starting, centroid and target maps for this experiment.

4.4. *Experimental data for cytidine deaminase*

Cytidine deaminase crystallizes with a monomer of about 30 kdalton in the asymmetric unit and a solvent volume about 70% of the unit cell. The crystal structure has been solved and refined at 2.8 Å resolution using *X-PLOR* (Brünger, 1990) to an R factor of about 24% without individual atomic B factors or water molecules (Betts & Carter, 1991; Betts, 1991). Native intensities were measured using a multiwire area detector, and they are of high quality (R_{merge} = 5.0% for 12 348 reflections to 2.8 Å resolution). This problem is useful and interesting for at least three reasons:

(i) The structure solution was based on a 3.2 Å double isomorphous replacement phase determination with anomalous scattering, and involving extensive electron-density modification by means of the solvent-flattening algorithm of B. C. Wang and subsequently by the modified isomorphous pseudo-derivative procedure (Zelwer, 1988; Romanaora, Arnoux & Zelwer, 1991). Two heavy-atom derivatives were used in the phase determination, but they were very similar, having several sites in common. This situation was, therefore, scarcely more favorable than a single isomorphous replacement phase determination with anomalous scattering. Therefore, it provided a useful standard against which to measure the performance of *MICE* as a density-improvement algorithm.

(ii) The solvent-flattened map was difficult to interpret, and refined $2F_{obs} - F_{calc}$ and $F_{obs} - F_{calc}$ difference maps indicate that there are still areas where the model should be improved by rebuilding. These properties of the solu-

tion suggest that solvent flattening did not fully correct the MIRAS phase errors, and that there is room for improvement in processes subsequent to the primary phase determination.

(iii) Significant structural questions with unclear answers at the current state of refinement, might be answerable if a better set of experimental phases could be obtained. Among these is the configuration of a bound ligand, 5-fluoropyrimidine-2-one riboside. The electron-density map with the current set of phases is ambiguous about whether the pyrimidine ring is *syn* or *anti*, relative to the ribose. It may be possible, even at 2.8 Å resolution, to resolve this ambiguity using omitted-fragment difference maps. Unfortunately, the best phases currently available are those of the model, which have a built-in bias. We are therefore interested in whether or not exponential modeling can provide a better set of phases than those obtained from solvent flattening, and which could unambiguously settle this question. (We do not address this issue further here.)

4.4.1. *Preparation of unitary structure factors.* Wilson scaling provided a good initial estimate for the scaling parameter, $|F_{000}|^{-1}$. In order to optimize this value for density modification, we carried out a series of studies varying the scale constant for the simulated data, and determining the behavior of the log-likelihood gain on the first cycle, and at convergence of the exponential modeling. We also observed that at the correct scale, the ratio $F_{\text{oscl}} = \sum_{\mathbf{h}\in\mathbf{H}}^{N}|U_{\mathbf{h}}^{\text{obs}}|/\sum_{\mathbf{h}\in\mathbf{H}}^{N}|U_{\mathbf{h}}^{\text{ME}}| \simeq 1.2$, which provided an additional guide to the appropriate scale.

These experiments, together with similar studies with the cytidine deaminase $\{|F_{\mathbf{h}}^{\text{obs}}|\}$ are illustrated in Fig. 4. Using *MICE* without making reference to the $m(\mathbf{x})$ file, the scale of the input unitary structure factors was systematically varied, and the log likelihood, the log-likelihood gain and the ratio F_{oscl} were determined for the first cycle of an exponential modeling run constrained by the basis-set phases alone, with a uniform prior distribution for the random atomic positions, *i.e.* without imposing solvent flatness. For the simulated data, the points near the maximum were taken to maximum likelihood with subsequent cycles of exponential modeling. The ultimate maximum likelihood, as a function of the scale parameter, closely matched that observed for the first cycle. In the plots shown in Fig. 4, ranges within which the maximum log likelihood and, for the simulated data, the final maximum likelihood were observed and hence the ranges of appropriate scale factors are indicated by vertical lines for the log-likelihood gain and for the ratio, F_{oscl}. For cytidine deaminase data the F_{000} term for converting $\{|F^{\text{obs}}|\}$ amplitudes to unitary structure factors was estimated initially from Wilson scaling and optimized by using one cycle of exponential modeling to estimate the log-likelihood gain for each of a series of values of the scaling factor. That factor which maximized this first cycle log-likelihood gain was then used throughout the subsequent procedures.

4.4.2. *Density modification by exponential modeling.* Various other parameters for the cytidine deaminase data were chosen based on experience with the simulated data. It was clear that significant electron-density improvements were realized with constraints having a mean phase error of about 30°. This error corresponds roughly to a mean figure of merit of 0.867. We used, therefore, a figure of merit threshold of 0.7 to select basis-set phases from the original cytidine deaminase MIRAS phases, having a mean figure of merit of 0.87. This afforded a basis set of 4174 reflections to 3.2 Å resolution, a somewhat larger fraction of the total number of reflections (12 348) than we had used with the simulated data.

The envelope defined by the final round of solvent flattening was used as a mask for solvent averaging, resetting the density outside the envelope to its average value on each cycle of exponential modeling. Exponential modeling under these circumstances has fulfilled or exceeded most of the expectations stated above (Tables 3–5):

(i) The log-likelihood gain reached a value of 6214 after 40 cycles of fitting. At this stage the R factor was 0.29 for basis-set reflections and 0.49 for extrapolated reflections.

(ii) Throughout the electron-density maps (Figs. 5*a*, 5*b*, 6*a*, 6*b*, 7*a* and 7*b*) there were regions where the centroid map was clearly superior to the solvent-flattened map, more nearly resembling the final $2F_{\text{obs}} - F_{\text{calc}}$ map. Side-chain densities connected to neighboring main-chain density in the solvent-flattened map were clearly separated (Fig. 5). Regions where connectivity of the main chain was broken in the solvent-flattened map were clearly and correctly delineated in the centroid map (Fig. 6). Places with no side-chain density in the solvent-flattened map had developed appropriate density in the centroid map (Fig. 7).

(iii) The centroid map obtained from exponential modeling is clearly superior to that obtained by solvent flattening with reference to the same envelope.

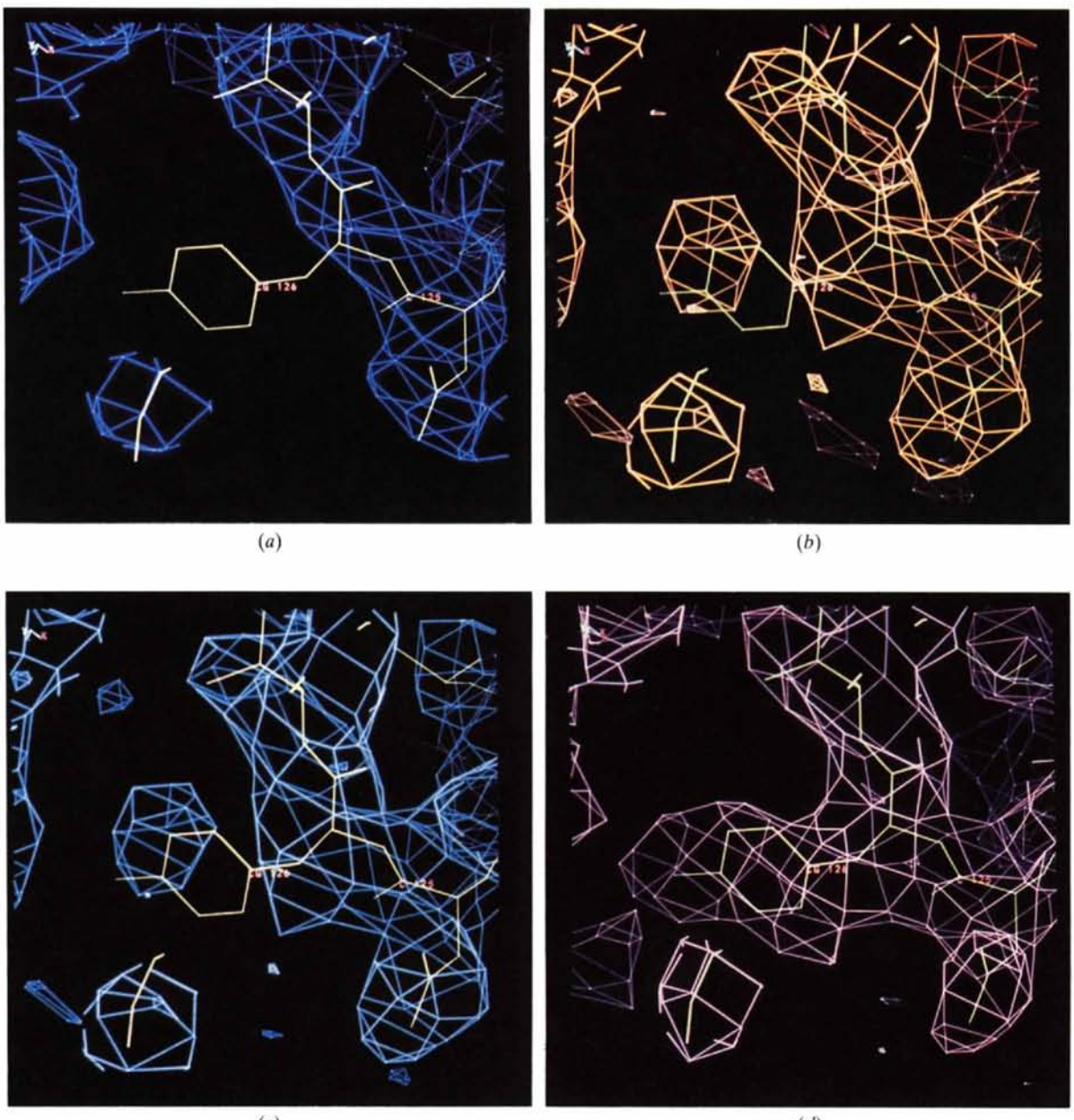
4.5. *Recombination of centroid phases from exponential modeling with MIRAS phases*

At this point, the Sim phase probability distribution from exponential modeling of cytidine deaminase was recombined with the original MIRAS probability distribution (Table 4). The mean figure of merit for all reflections rose from 0.55 to 0.66, and that for the basis-set reflections had risen from 0.867 to 0.914. An even more significant improvement was observed for reflections outside the basis set, and for which *MICE* generated extrapolated phases. For these reflections, the mean figure of merit rose from 0.34 to 0.47. The mean phase changes in the recombined phases were 6.4° for basis-set reflections, 32° for extrapolated phases and 22° overall. Judging from the experiments with simulated data, these changes are largely made in the right direction and represent improvements in the phases for most reflections. More exhaustive study will be required in order to determine whether the recombined phases are actually better than the centroid U^{ME} phases

for reflections having poor MIRAS figures of merit. We are now investigating this question by rebuilding the cytidine deaminase model in conjunction with phase refinement and extension *via* phase permutation and likelihood evaluation.

Recombined phases for the original basis-set reflections were input to a second refinement by *MICE*. The resulting centroid maps showed modest additional improvement, (Figs. 5*c*, 6*c* and 7*c*), even though they already strongly resembled the final $2F_{obs} - F_{calc}$ map. New figures of merit for basis-set and extrapolated phases, together with the exponential modeling statistics indicated significant improvement in the basis-set phases:

(i) Fitting the exponential model continued to improve the likelihood beyond the point at which the previous refinement had converged, as measured by the *R* factor



(*a*) (*b*) (*c*) (*d*)

Fig. 7. Density modification of the cytidine deaminase map based on the initial MIRAS phase in the region surrounding Tyr 126. The final solvent flattened map (*a*) fails to provide any density for the side chain of Tyr 126. Two successive maximum-likelihood centroid maps, before (*b*) and after (*c*) phase recombination with the MIRAS phase-probability distributions both indicate density in a region translated from the refined position of the side chain and resembling an extension of the density for this side chain in the final $2|F^{obs}| - |F^{calc}|$ map (*d*) after structure refinement.

between observed and ME structure-factor amplitudes at maximum likelihood (0.25 for basis set; 0.48 for extrapolated reflections).

(ii) Remarkably, this improvement in the fit of the exponential model at maximum likelihood is also accompanied by a modest *increase* in the entropy, from -0.94 to -0.85, even though the constraints were more strictly enforced, as indicated by the decrease in the crystallographic R factor. This suggests that in the neighborhood of the correct phases the entropy itself may be a useful secondary criterion. This importance of the entropy as a potential criterion for the correct phases was also apparent with the simulated data.

(iii) Phase recombination continues to improve the basis set phases. Ongoing work shows that phase recombination also improves phases outside the basis set, suggesting that these too can be recruited into a new basis set. Thus, the whole calculation can be iterated, in a process that potentially could accurately determine all phases. This implies that all experimental electron-density maps should be amenable to considerable improvement by this process prior to model building.

4.6. *Comparison with cytidine deaminase results using traditional methods*

Solvent flattening in combination with exponential modeling to maximum likelihood has turned out to have important advantages over traditional density-modification methods. Several aspects of the improvement in the cytidine deaminase maps are worth special emphasis.

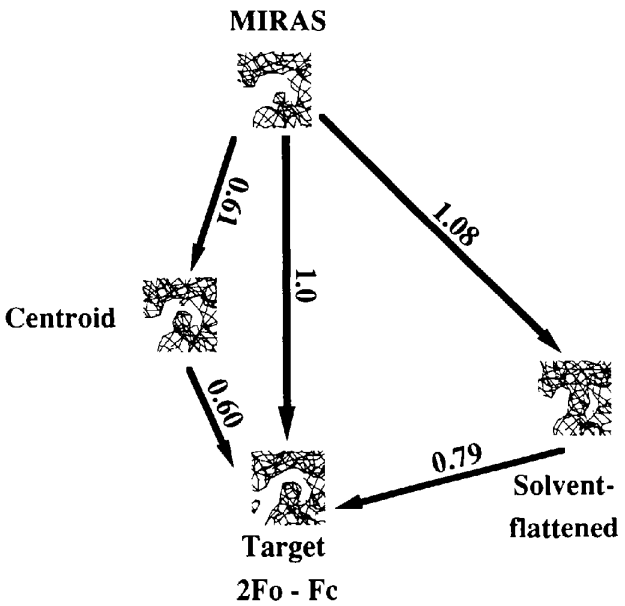


Fig. 8. Global distances between the initial MIRAS, the target ($2F_o$ - F_c), the centroid and the solvent-flattened maps. Distances are estimated as described in the text. The centroid map lies close to a direct path between the starting and target maps.

4.6.1. *The centroid cytidine deaminase map lies nearly on a direct path from the MIRAS map to the target map.* The MIRAS map for cytidine deaminase has a correlation coefficient of 0.48 with the target $2F_{obs}$ - F_{calc}. The centroid map after the first round of exponential modeling retains high correlation coefficients with both the MIRAS (0.68) and starting (0.78) maps, while having an improved correlation coefficient with the target map (0.69). In marked contrast, the solvent-flattened map, although it does improve the agreement with the target map (0.59), also shows a pronounced deviation from the MIRAS map (0.44). *This deviation is reflected in the errors in connectivity and side-chain density illustrated in Figs.* 5–7. The improvement in the centroid maps is not difficult to identify; it is dramatic. The centroid map obtained using only 35% of the data is superior to the final solvent-flattened map both subjectively and by every quantitative criterion after only one cycle of exponential modeling and without phase recombination. Subsequent phase recombination and reconstruction of the optimal exponential model produces additional improvements in the map; the correlation coefficient with the MIRAS map remains at 0.68, while that with the target map increases to 0.70. Maps (*b*) and (*c*) in Figs. 5–7 are both nearly indistinguishable from the target $2F_{obs}$ - F_{calc} map.

Of particular relevance from the experiments with cytidine deaminase is that the centroid maps (Figs. 5*b*, 5*c*, 6*b*, 6*c*, 7*b* and 7*c*) are comparatively free from the artifacts introduced by solvent flattening. This phenomenon is a global feature of the maps themselves, as is illustrated in Fig. 8. Global distances, D_{xy}, between maps were estimated from the correlation coefficients, R_{xy}, *via*

$$D_{xy} = (1 - R_{xy})/(1 - R_{\text{target,MIRAS}}).$$

From this comparison it is evident that the centroid map based on the first exponential model lies more nearly on the direct path from the MIRAS map to the target $2F_{obs}$ - F_{calc} map, and that the solvent-flattened map interpreted to construct the atomic model represents a rather significant and essentially unpredictable excursion from both starting and target maps. This important phenomenon may help to explain why incorrect models (Brändén & Jones, 1990) are occasionally built based on solvent-flattened maps.

4.6.2. *Maximum-entropy extrapolation is responsible for most of the map improvement.* Conventional solvent flattening and maximum-entropy solvent flattening both improve average phase errors. However, several interesting trends emerge from consideration of the distributions of phase errors in the two cases (Fig. 9). Maximum-entropy solvent flattening reduces phase errors significantly more than does conventional solvent flattening, particularly for strong reflections (Figs. 9*b* and 9*d*). Conventional solvent flattening seems to be particularly ineffective for low-resolution reflections (Figs. 9*a* and 9*c*), where maximum-entropy solvent flattening shows a major improvement. These trends are even more evident for reflections in {**K**}

(Figs. 9*c* and 9*d*). Hence, maximum-entropy phase extension produces a more significant phase improvement for those reflections which are least well determined by isomorphous replacement, and these reflections are the least improved by conventional solvent flattening. This may reflect the fact that conventional solvent flattening tends not to improve phases that are initially poor (Fenderson, Herriott & Adman, 1990). Most of the improvement of the maps seen in Figs. 5–7 can be attributed therefore to the improvement in the phases of reflections in {**K**}, or those arising from maximum-entropy extrapolation.

4.6.3. *Maximum-entropy solvent flattening only requires reasonable computing resources.* Exponential modeling to maximum likelihood with solvent averaging requires comparable computing resources to those required for solvent flattening. A single cycle of fitting (resulting in Figs. 5*b*, 6*b* and 7*b*) required only about 12 h of CPU time on a DEC station 5000/200. The method is entirely compatible with conventional phase combination, and may be iterated to convergence, involving re-definition of the envelope. It can therefore be used in exactly the same context as conventional solvent flattening.

These algorithms are so effective for the following reasons. Both solvent-flattening algorithms remove peaks from the solvent regions that are considered to result from errors in the phases. Iteration by combining observed amplitudes with phases modified by recombination of MIRAS phases with those from the solvent-flattened map tends to restore scattering represented by the peaks removed from the solvent by introducing new density inside the envelope. Conventional solvent flattening gives no guidance whatsoever regarding where inside the envelope to put this new density. In contrast, construction of a constrained exponential model for the solvent-flattened density assures that new features inside the envelope are optimally consistent with the observed amplitudes and the

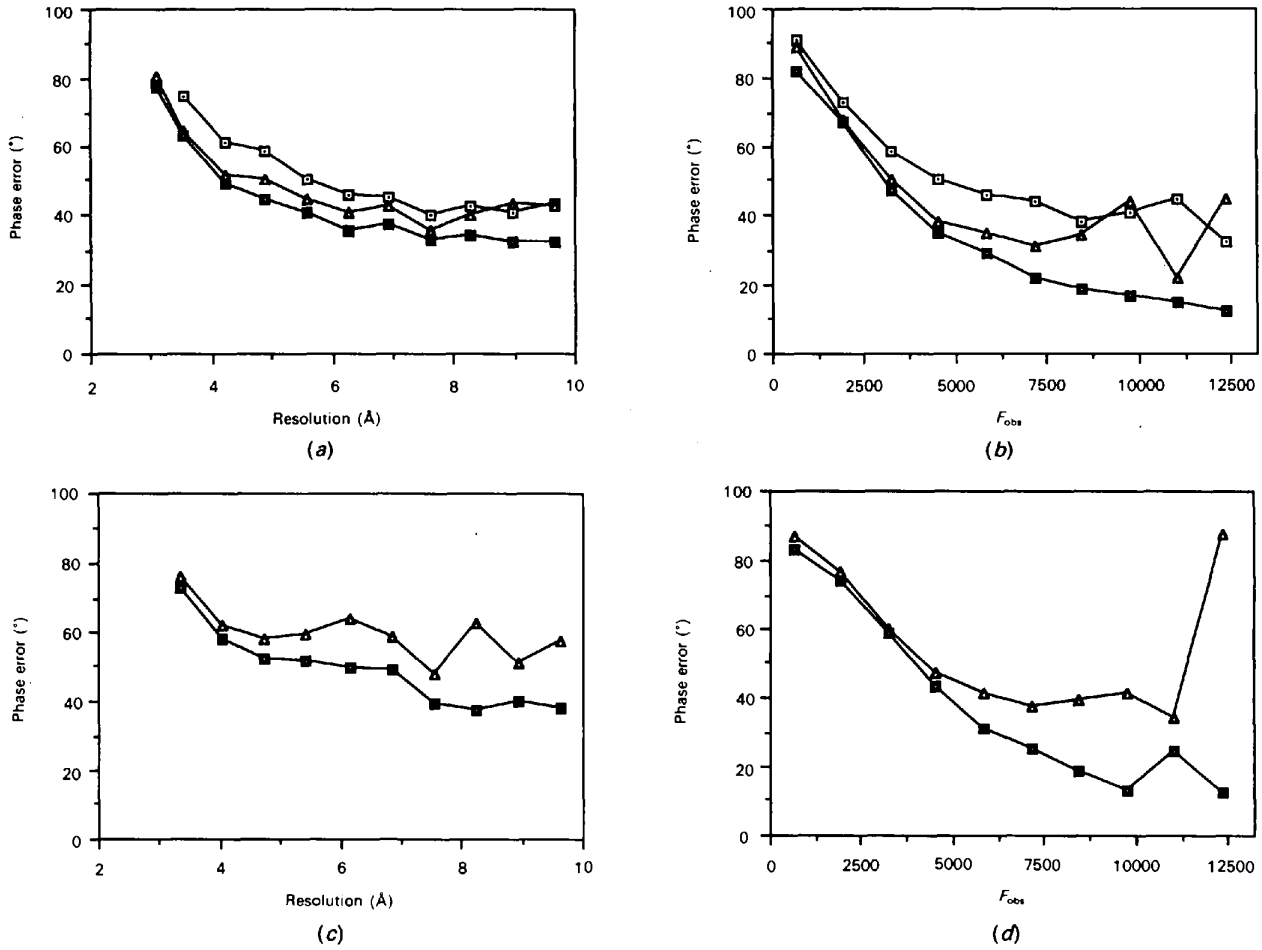


Fig. 9. Mean phase error distributions for MIRAS, conventionally solvent-flattened and centroid phases. (*a*) Overall mean phase error as a function of resolution. (*b*) Overall mean phase error as a function of $|F_{obs}|$, on an arbitrary scale. (*c*) Mean phase errors for extrapolated reflections (reflections in {**K**}) as a function of resolution. (*d*) Mean phase errors for extrapolated reflections (reflections in {**K**}) as a function of $|F_{obs}|$. Symbols for all plots: (open squares) MIRAS phase set, (filled triangles) solvent-flattening phase set and (filled squares) centroid phase set from maximum-entropy solvent flattening.

current phases in the basis set by maximizing the map entropy. In contrast, the errors in part (*a*) of Figs. 5–7 are *low-entropy* features because there are relatively fewer ways to construct them consistently using the basis-set structure factors. They were introduced into the map by the unguided solvent-flattening procedure, as indicated by the lack of correlation between the solvent flattened map and the initial MIRAS map (0.44).

4.7. *Figures of merit*

Both the map entropy and the associated likelihood have been proposed as figures of merit on the correctness of basis-set phases (Bricogne, 1984, 1988*a*; Bricogne & Gilmore, 1990; Gilmore, Bricogne & Bannister, 1990; Gilmore, A. N. Henderson & Bricogne, 1991; Dong *et al.*, 1992). Our experience so far provides compelling evidence regarding the value of the log-likelihood gain. The likelihood seems to be the best indication of both the convergence of the exponential modeling stages of refinement and the quality of the constraints, including the envelope as well as the basis-set phases. It is especially valuable in preventing overfitting of the data, as indicated by the fact that the constraints are fitted to varying degrees, depending on the quality of the constraints (the basis-set phases and the envelope). In the absence of such a criterion and without the envelope constraint, it is possible to fit the constraints exactly, since there are an equal number of parameters in the exponential model and observations contributing to the constraints. Overfitting in this fashion normally leads directly to catastrophic decreases in the global log-likelihood gain.

The entropy is clearly also a useful indicator of the correctness of the basis-set phases under certain conditions. In both experimental and model test situations, even small improvements in the basis-set phases lead to increases in the entropy at maximum likelihood. This suggests that the 'Bayesian score', combining the log-likelihood gain with the entropy, should be an even more powerful figure of merit on the quality of the constraints (Bricogne, 1988*a*).

R factors for basis-set and extrapolated reflections also provide very useful guides as to the overall quality of the constraints: as the basis-set phases improve, the exponential modeling can be carried further – fitting more tightly to the constraints – as long as the likelihood continues to improve. The *R* factor at maximum likelihood also indicates the quality of the mask, so in situations where the mask is not known *a priori*, the ability to fit to the constraints should also be an indicator of the correctness of the mask.

Taken together, these statistics provide a rational and powerful measure of the quality of all sources of phase information. The ability to fit an exponential model to a set of constraints is, perhaps, the best all round characterization of those constraints. However, none of the individual statistics provides a complete characterization. The entropy itself is actually a rather weak discriminator, compared to the log-likelihood gain. Both entropy and log-likelihood gain are relative quantities, dependent on the contributors in the sets {**H**} and {**K**}, respectively. For this reason, they should be used only to compare comparable configurations of reflections. An example is provided by the comparison between Tables 3 and 5. The improvements seen in Table 5 represent significant improvements in the basis-set phases for cytidine deaminase, because the basis and complementary sets contained the same reflections. The crystallographic *R* factors, on the other hand, reflect the absolute agreement between the observed amplitudes and those from the Fourier transform of the exponential model. As such, they are useful indicators of the overall progress of phase determination. For the purposes of calculations described here, the likelihood is most useful as a safeguard against over- or under-fitting to the current constraints.

4.8. *The minimum size and quality of the basis set*

It is reasonable to ask what the practical limits are for application of this approach as a phase-determination or phase-extension procedure. In other words, how little phase information, and of what quality, must be provided to convert an uninterpretable MIRAS or SIR map into an interpretable one? The minimum size of the starting basis set is probably smaller than 30% of the total number of reflections. It is undoubtedly smaller if a molecular envelope can be determined from other sources. With respect to the question of phase accuracy in the basis set, an important practical conclusion from studies with cytidine deaminase is that phases with a figure of merit > 0.7 are close enough to the correct phases to be within the radius of convergence of the method to the correct structure, provided that an envelope is available. Since even rather poor SIR phase sets often generate phases with a figure of merit > 0.7 for 20–30% of the data set, this suggests that use of these procedures may bring a larger number of partially phased structures into range for full solutions.

5. Summary

We have shown that exponential modeling to maximum likelihood constrained by solvent flatness within a known molecular envelope is a superior way to improve isomorphous replacement electron-density maps before model building. Tests with simulated data involving exact and noisy data provide useful examples of the capabilities of the algorithm, and an essential reference point for calibrating its behavior with experimental data. The method is effective throughout the critical resolution range 5–3 Å which includes most of the macromolecular crystal structure determinations likely to benefit most from density modification. In the presence of a known molecular envelope, maximum-entropy phase extrapolation can generate significantly better phases, on average, than centroid MI-

RAS distributions for as many as 65% of the reflections to 3 Å with figures of merit < 0.70. Perhaps most importantly, exponential modeling is a phase extension and refinement process that does not involve interpretation of the isomorphous-replacement-phased electron-density map, so it is *model independent.* It therefore should also minimize errors in an initial model, thereby preventing them from being 'locked in' by conventional automated refinement algorithms. The Sim-like conditional phase probability distributions obtained from the maximum entropy and experimental structure-factor amplitudes can be recombined with phase information available from other experimental sources in a process that converges to a phase set that is also optimally consistent with the native amplitudes themselves. Hence, it will help make the best use of all experimental data relevant to phase determination before electron-density maps are interpreted.

This work was supported by grant BE-67493 from the American Cancer Society. Support from the North Carolina Biotechnology Center for instrumentation is also gratefully acknowledged. We also thank Bonnie Billard for expert technical support.

References

BACKES, G., MINO, Y., LOEHR, T. M., MEYER, T. E., CUSANOVICH, M. A., SWEENEY, W. V., ADMAN, E. T. & SANDERS-LOEHR, J. (1991). *J. Am. Chem. Soc.* **113**, 2055-2064.

BETTS, L. (1991). PhD thesis, Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

BETTS, L. & CARTER, C. W. JR (1991). Am. Crystallogr. Assoc. Abstr. Ser. 2, No. 19, p. 76.

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

BODE, W., CHEN, Z., BARTELS, K., KUTZBACH, C., KASENER, G. S. & BARTUNIK, H. (1983). *J. Mol. Biol.* **164**, 237-282.

BRÄNDÉN, C.-I. & JONES, T. A. (1990). *Nature (London)*, **343**, 687-689.

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

BRICOGNE, G. (1982). *Generalised Density Modification Methods.* In *Computational Crystallography*, edited by D. SAYRE, pp. 258-264. New York: Oxford Univ. Press.

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

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

BRICOGNE, G. (1988*b*). *Maximum Entropy Methods in the X-ray Phase Problem.* In *Crystallographic Computing* 4, edited by N. W. ISAACS & M. R. TAYLOR, pp. 60-79. IUCr/Oxford Univ. Press.

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

BRICOGNE, G. (1992). *The X-ray Crystallographic Phase Problem.* In *Maximum Entropy in Action*, edited by B. BUCK & V. A. MACAULAY, pp. 187-216. Oxford: Clarendon Press.

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

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

BRITTEN, P. L. & COLLINS, D. M. (1982). *Acta Cryst.* A**38**, 129-132.

BRÜNGER, A. T. (1990). *X-PLOR Manual.* Version 2.1. Yale Univ., New Haven, USA.

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

COLLINS, D. M. (1982). *Nature (London)*, **298**, 49-51.

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

FENDERSON, F. F., HERRIOTT, J. R. & ADMAN, E. T. (1990). *J. Appl. Cryst.* **23**, 115-131.

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

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

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

HARKER, D. (1953). *Acta Cryst.* **6**, 731-736.

HAUPTMAN, H. & KARLE, J. (1953). In *The Solution of the Phase Problem: I. The Centrosymmetric Crystal*, Vol. 3. Pittsburgh: Polycrystal Book Service.

HOWARD, J. B., LORSBACH, T. W., GHOSH, D., MELIS, K. & STOUT, C. D. (1983). *J. Biol. Chem.* **258**, 508-522.

KLUG, A. (1958). *Acta Cryst.* **11**, 515-532.

LEMARECHAL, C. & NAVAZA, J. (1991). *Acta Cryst.* A**47**, 631-632.

LUZZATI, V. (1955). *Acta Cryst.* **8**, 795-806.

NARAYAN, R. & NITYANANDA, R. (1982). *Acta Cryst.* A**38**, 122-128.

PIRO, O. E. (1983). *Acta Cryst.* A**39**, 61-83.

PODJARNY, A. D., BHAT, T. N. & ZWICK, M. (1987). *Annu. Rev. Biophys. Biophys. Chem.* **16**, 351-373.

PRINCE, E., SJÖLIN, L. & ALENLJUNG, R. (1988). *Acta Cryst.* A**44**, 216-222.

READ, R. (1986). *Acta Cryst.* A**42**, 140-149

ROGERS, D. (1980). *Definition of Origin and Enantiomorph and Calculation of |E| Values.* In *Direct Methods*, edited by M. F. C. LADD & R. A. PALMER, pp. 23-92. New York: Plenum Press.

ROMANAORA, E., ARNOUX, B. & ZELWER, C. (1991). *J. Appl. Cryst.* **24**, 936-941.

ROTHBAUER, R. (1980). *Acta Cryst* A**36**, 27-32.

SAYRE, D. (1952*a*). *Acta Cryst.* **5**, 843.

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

SHANNON, C. E. (1949). *Proc. Inst. Radio Electron. Eng.* **37**, 10-21.

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

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

STOUT, C. D. (1988). *J. Biol. Chem.* **263**, 9256-9260.

STOUT, G. H., TURLEY, S., SIEKER, L. & JENSEN, L. H. (1988). *Proc. Natl Acad. Sci. USA*, **85**, 1020-1022.

WANG, B. C. (1985). *Methods Enzymol.* **115**, 90-112.

WILKINS, S. W., VARGHESE, J. N. & LEHMANN, M. S. (1983). *Acta Cryst.* A**39**, 47-60.

WOOLFSON, M. M. (1980). *Methods of Solving for the Phases.* In *Computational Crystallography*, edited by D. SAYRE, pp. 110-125. Oxford: Clarendon Press

ZELWER, C. (1988). *Acta Cryst.* A**44**, 485-495.