electronic reprint

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Incorporation of Prior Phase Information Strengthens Maximum-Likelihood Structure Refinement

Navraj S. Pannu, Garib N. Murshudov, Eleanor J. Dodson and Randy J. Read

Acta Cryst. (1998). D**54**, 1285–1294

Incorporation of Prior Phase Information Strengthens Maximum-Likelihood Structure Refinement

Navraj S. Pannu,[a]† Garib N. Murshudov,[b] Eleanor J. Dodson[b] and Randy J. Read[c]*

[a]*Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta T6G 2H7, Canada,* [b]*Chemistry Department, University of York, Heslington, York, England,* [c]*CLRC, Daresbury Laboratory, Daresbury, Warrington WA4 4AD, England, and* [d]*Department of Medical Microbiology and Immunology, University of Alberta, Edmonton, Alberta T6G 2H7, Canada. E-mail: rjr27@cam.ac.uk*

(*Received 1 December 1997; accepted 16 March 1998*)

Abstract

The application of a maximum-likelihood analysis to the problem of structure refinement has led to striking improvements over the traditional least-squares methods. Since the method of maximum likelihood allows for a rational incorporation of other sources of information, we have derived a likelihood function that incorporates experimentally determined phase information. In a number of different test cases, this target function performs better than either a least-squares target or a maximum-likelihood function lacking prior phases. Furthermore, this target gives significantly better results compared with other functions incorporating phase information. When combined with a procedure to mask 'unexplained' density, the phased likelihood target also makes it possible to refine very incomplete models.

1. Introduction

Notation used is given in Table 1. The advent of a maximum-likelihood approach to crystallographic structure refinement has yielded substantial improvements. Test refinements using maximum-likelihood target functions in the programs *BUSTER* (Bricogne & Irwin, 1996), *REFMAC* (Murshudov *et al.*, 1997) available from the Collaborative Computational Project, Number 4 (1994), *CNS*, *TNT* and *X-PLOR* (Pannu & Read, 1996) show clearer electron-density maps and reduction of phase error over those carried out against a least-squares target function. Recent tests have shown that the combination of a maximum-likelihood target function and simulated-annealing optimization parameterized in torsion-angle space (Rice & Brünger, 1994) further enhances refinement (Adams *et al.*, 1997). Using refinement with the Rice distribution as a target function and updating atoms, for example using *ARP* (Lamzin & Wilson, 1993), also improves phases (Z. Dauter, personal communication).

For many structure solutions some experimental phase information is available before refinement. Using this information should, in principle, increase the power of refinement, since the ratio of observations to parameters is greater. There have been previous attempts to exploit this information in refinement. For example, Lunin & Urzhumtsev (1985) suggested adding $\log[P(\alpha_c)]$ to the refined residual, where $P(\alpha)$ is a prior phase probability distribution and α_c is the phase of the structure factor calculated from the model.

Arnold & Rossmann (1988) suggested minimizing

$$\sum w|\mathbf{F_o} - \mathbf{F_c}|^2 \quad (1),$$

where $\mathbf{F_o}$ and $\mathbf{F_c}$ are vectors and w is the figure of merit of the phases. Another approach would be to carry out the refinement in real space as suggested by Diamond (1971). In this approach, cycles of refinement and phase recombination with the new model enhance the quality of the target map.

However, in all the above cases the addition of the phase information requires new assumptions or techniques. The maximum-likelihood formulation can implicitly incorporate prior phase information (Bricogne & Irwin, 1996; Murshudov *et al.*, 1996, 1997).

$$P(F_o; F_c) = \begin{cases} \dfrac{F_o}{\pi\varepsilon\sigma_\Delta^2}\exp\left(-\dfrac{F_o^2 + D^2F_c^2}{\varepsilon\sigma_\Delta^2}\right)\displaystyle\int_0^{2\pi} P(\alpha) \\ \quad \times \exp\left[\dfrac{2F_oDF_c}{\varepsilon\sigma_\Delta^2}\cos(\alpha - \alpha_c)\right]\mathrm{d}\alpha & \text{acentric} \\ \left(\dfrac{1}{2\pi\varepsilon\sigma_\Delta^2}\right)^{1/2}\exp\left(-\dfrac{F_o^2 + D^2F_c^2}{2\varepsilon\sigma_\Delta^2}\right)\displaystyle\sum_{l=0}^{1} P(\alpha_l) \\ \quad \times \exp\left[\dfrac{F_oDF_c}{\varepsilon\sigma_\Delta^2}\cos(\alpha_l - \alpha_c)\right] & \text{centric.} \end{cases} \quad (2)$$

(For centrosymmetric reflections integration is replaced by summation over the two possible phases.)

Different assumptions about the prior phase probability $P(\alpha)$ generate different forms of $P(F_o; F_c)$.

† Present address: Department of Haematology, University of Cambridge, Cambridge CB2 2QH, England.

Assuming that phases are exactly known, then $P(\alpha)$ is Dirac's delta function, and $P(F_o; F_c)$ becomes

$$P(F_o; F_c) = \begin{cases} \dfrac{F_o}{\pi\varepsilon\sigma_\Delta^2}\exp\left[-\dfrac{(F_o - DF_c)^2}{\varepsilon\sigma_\Delta^2}\right] & \text{acentric,} \\ \dfrac{1}{(2\pi\varepsilon\sigma_\Delta^2)^{1/2}}\exp\left[-\dfrac{(F_o - DF_c)^2}{2\varepsilon\sigma_\Delta^2}\right] & \text{centric.} \end{cases} \tag{3}$$

There is an obvious similarity between (3) and (1). Extremely accurate phases can be derived from rich non-crystallographic symmetry (NCS) and in such a case the use of (1) could be appropriate.

The other extreme case is when no prior phase information is available [*i.e.* $P(\alpha)$ is constant]. In this case $P(F_o; F_c)$ becomes the Rice distribution which, as previously mentioned, has already been shown to be a powerful tool for refinement.

In many cases it is convenient to express the available phase probability in terms of Hendrickson–Lattman coefficients (Hendrickson & Lattman, 1970) and this paper discusses a likelihood function (MLHL) using these, or the symmetric unimodal phase probability distribution based on knowledge of phase and figure of merit alone. Even at the end stages of refinement, it can be shown that using this probability distribution for phases can improve refinement behaviour.

Bricogne & Gilmore (1990) and Murshudov *et al.* (1997) have suggested using the experimental uncertainties of the structure-factor amplitudes to increment $\varepsilon\sigma_\Delta^2$. This contribution is not included in the above equations, but is included in the *REFMAC* implementation of MLHL. Alternatively, it can be incorporated by assuming that the experimental errors in the structure-factor amplitudes are distributed as Gaussian as outlined in *Appendix A*.

Results of test refinements to demonstrate the power of the MLHL target function are described in §5.

2. MLHL: a likelihood function incorporating prior phase information

Under the assumption that individual reflections are independent, the principle of maximum likelihood states that the best parameters for a model are obtained by maximizing the following likelihood function (L),

$$L = \prod_{hkl} P(F_o; F_c) \tag{4}$$

or, equivalently, minimizing the minus log likelihood [$\mathcal{L} = -\log(L)$]. In the above expression $P(F_o; F_c)$ denotes the probability distribution of the observed structure-factor amplitude given the calculated structure-factor amplitude.

Hendrickson & Lattman (1970) have shown that the prior probability distribution of a phase (α) can be represented in the following form:

$$P(\alpha) = N\exp[A_{hl}\cos(\alpha) + B_{hl}\sin(\alpha) + C_{hl}\cos(2\alpha) + D_{hl}\sin(2\alpha)], \tag{5}$$

where A_{hl}, B_{hl}, C_{hl} and D_{hl} are Hendrickson–Lattman coefficients and N is a normalization constant. This form can generate a bimodal probability distribution. If C_{hl} and D_{hl} are zero the probability distribution is unimodal.

In the acentric case, multiplication of the density $P_a(F, \Delta\alpha; F_c)$ with the prior probability distribution $P(\alpha)$ gives the joint probability distribution $P_a(F, \Delta\alpha, \alpha; F_c)$. Integrating the true phase out of this joint probability distribution gives the required distribution.

$$P(F; F_c) = \begin{cases} \dfrac{NF}{\pi\varepsilon\sigma_\Delta^2}\exp\left(-\dfrac{F^2 + D^2F_c^2}{\varepsilon\sigma_\Delta^2}\right)\displaystyle\int_0^{2\pi}\exp[A'_{hl}\cos(\alpha) \\ \quad + B'_{hl}\sin(\alpha) + C_{hl}\cos(2\alpha) + D_{hl}\sin(2\alpha)]\,\mathrm{d}\alpha & \text{acentric} \\ N\dfrac{1}{(2\pi\varepsilon\sigma_\Delta^2)^{1/2}}\exp\left(-\dfrac{F^2 + D^2F_c^2}{2\varepsilon\sigma_\Delta^2}\right) \\ \quad \times \displaystyle\sum_{l=0}^{1}\exp\left[A'_{hl}\cos(\alpha_l) + B'_{hl}\sin(\alpha_l)\right] & \text{centric,} \end{cases} \tag{6}$$

where $A'_{hl} = A_{hl} + X\cos(\alpha_c)$, $B'_{hl} = B_{hl} + X\sin(\alpha_c)$, and $X = 2F_oDF_c/\varepsilon\sigma_\Delta^2$ for the acentric case or $F_oDF_c/\varepsilon\sigma_\Delta^2$ for the centric case. A sample surface plot of (6) for the acentric case is shown in Fig. 1, using parameters taken from a reflection in a test case. This figure demonstrates that the power of phased likelihood comes from the model reinforcing a phase choice consistent with the experimental phase-probability distribution.

Taking the minus logarithm of (6), removing all terms that are constant, and summing over all reflections gives the desired target function.

$$\begin{aligned} \mathcal{L} = &\sum_{hkl\ \text{acentric}}\left\{\frac{F^2 + D^2F_c^2}{\varepsilon\sigma_\Delta^2} - \log\int_0^{2\pi}\exp[A'_{hl}\cos(\alpha) \right. \\ &\left. + B'_{hl}\sin(\alpha) + C_{hl}\cos(2\alpha) + D_{hl}\sin(2\alpha)]\,\mathrm{d}\alpha\right\} \\ &+ \sum_{hkl\ \text{centric}}\left\{\frac{F^2 + D^2F_c^2}{2\varepsilon\sigma_\Delta^2} - \log\sum_{l=0}^{1}\exp[A'_{hl}\cos(\alpha_l) \right. \\ &\left. + B'_{hl}\sin(\alpha_l)]\right\}. \end{aligned} \tag{7}$$

Hendrickson & Lattman (1970) derived a series representation for the integral in the acentric case (6). Unfortunately this series exhibits numerical instabilities for particular arguments, so the above integral is eval-

Table 1. *Notation*

F_o	Experimental amplitude of the structure factor
$\sigma_F = \sigma_{F;\exp}$	Experimental uncertainty in amplitude of the observed structure factor
$\mathbf{F_c} = F_c \exp(i\alpha_c)$	Calculated structure factor
$\alpha_{\exp}$	Centroid phase from prior experimental phase probability distribution
$A_{hl}, B_{hl}, C_{hl}, D_{hl}$	Hendrickson–Lattman coefficients for phase-probability distribution
$\mathbf{s}$	Vector of position in reciprocal space; $s = \lvert\mathbf{s}\rvert = 2\sin\theta/\lambda$
$\Delta\mathbf{x}$	Error in position of atoms
D	$\langle\cos(\Delta\mathbf{xs})\rangle$
σ_Δ^2	$\sum_N - D^2 \sum_P$
$\sum_N = \sum f_j^2$	Sum of form factors squared for all atoms in crystal (Wilson, 1949)
$\sum_P = \sum f_j^2$	Sum of form factors squared for all atoms in model
ε	Expected intensity factor of diffracting plane
m_{comb}	Figure of merit of combined phase
α_{comb}	Combined phase
$X = 2F_oDF_c/\varepsilon\sigma_\Delta^2$ or $F_oDF_c/\varepsilon\sigma_\Delta^2$	For acentric and centric reflections, respectively
$I_0(x),\ I_1(x)$	Zero- and first-order modified Bessel functions of the first kind
$P(A, \ldots; B, \ldots)$	Conditional probability distribution of $(A, \ldots)$ when $(B, \ldots)$ are known
$P(A) = \int_B P(A, B)dB$	marginal probability distribution of A

For experimental observations $(A, \ldots)$ and parameters $(B, \ldots)$ to be estimated using these, the minus log likelihood function, $\mathcal{L}$, will be $\mathcal{L} = -\log P(A, \ldots; B, \ldots)$. The maximum-likelihood estimation of parameters $(B, \ldots)$ is achieved by minimization of this function.

uated numerically in the general case of non-zero A_{hl}, B_{hl}, C_{hl} and D_{hl} Hendrickson–Lattman coefficients. However, in the special case when C_{hl} and D_{hl} are both zero (always true for centric reflections), an analytical form exists,

$$\mathcal{L} = \sum_{hkl\ \text{acentric}} \frac{F^2 + D^2F_c^2}{\varepsilon\sigma_\Delta^2} - \log\left\{I_0[(A'_{hl})^2 + (B'_{hl})^2]^{1/2}\right\} + \sum_{hkl\ \text{centric}} \frac{F^2 + D^2F_c^2}{2\varepsilon\sigma_\Delta^2} - \log\left\{\cosh[(A'_{hl})^2 + (B'_{hl})^2]^{1/2}\right\}. \quad (8)$$

From this equation, it is easy to calculate the cosine, the sine and the figure of merit (FOM) of the combined phases,

$$\text{FOM} = \begin{cases} \dfrac{I_1[(A'_{hl})^2 + (B'_{hl})^2]^{1/2}}{I_0[(A'_{hl})^2 + (B'_{hl})^2]^{1/2}} & \text{acentric} \\ \tanh[(A'_{hl})^2 + (B'_{hl})^2]^{1/2} & \text{centric} \end{cases} \quad (9)$$

and

$$\cos(\alpha_{\text{comb}}) = [(A'_{hl})^2 + (B'_{hl})^2]^{-1/2} \times [X_{\exp}\cos(\alpha_{\exp}) + X\cos(\alpha_c)] \quad (10)$$

$$\sin(\alpha_{\text{comb}}) = [(A'_{hl})^2 + (B'_{hl})^2]^{-1/2} \times [X_{\exp}\sin(\alpha_{\exp}) + X\sin(\alpha_c)] \quad (11)$$

where $X_{\exp} = (A_{hl}^2 + B_{hl}^2)^{1/2}$ is a measure of the prior phase probability, $\alpha_{\exp}$ is the centroid of the prior phase probability distribution, X is a measure of the quality of the model, α_c is the calculated phase and α_{comb} is the combined phase. If $X_{\exp}$ is too large, *i.e.* if it underestimates the experimental phase error, the refinement will try to reach an unrealistic target and will be rather unstable. This highlights the importance of the reliability of the prior phase probability distribution and the need to obtain less biased $X_{\exp}$ values. At the beginning of refinement when $X < X_{\exp}$, using prior phase information will influence the model more than at the end stages when $X > X_{\exp}$, where the calculated phases will dominate. But since $X_{\exp}$ and $\alpha_{\exp}$ always contain independent information about the crystal it seems appropriate to use these prior phases at all stages, possibly with some adjustment of the weighting to correct for bias, and possibly after improvement with some other procedures such as density modification.

3. Blurring of the phase-probability distribution

In many cases the phase-probability distribution does not reflect the true distribution of phases. This may arise from a correlation between heavy-atom sites (Terwilliger & Berendzen, 1997) or from density-modification procedures. At present the best treatment for obtaining a reliable phase probability distribution is that developed by Fortelle & Bricogne (1997) and incorporated into the program *SHARP*.

The phase-probability distributions for phases obtained after density modification have been derived assuming that the phases come from partial atomic models, not from electron density. Even more vexingly, much of the power of density-modification procedures relies on combining the new density-modification phases with the original experimental phases. These sources of phase information are not independent because the modified map will retain many of the features of the original map. Two approaches to overcoming this problem of the lack of independence of the phases have been devised (Cowtan & Main, 1996; Abrahams, 1997). While these approaches improve the situation substantially, further study will be required to determine the accuracy of the phase-error estimates.

To compensate for the overestimation of experimental phase accuracy, *REFMAC* provides an option to 'blur' the phase-probability distributions,

$$A_{\rm new} = S\exp(-B|s|^2)A_{\rm old},$$
$$B_{\rm new} = S\exp(-B|s|^2)B_{\rm old},$$
$$C_{\rm new} = S\exp(-B|s|^2)C_{\rm old},$$
$$D_{\rm new} = S\exp(-B|s|^2)D_{\rm old},$$

where $A_{\rm old}$, $B_{\rm old}$, $C_{\rm old}$, $D_{\rm old}$ are the current Hendrickson–Lattman coefficients, $A_{\rm new}$, $B_{\rm new}$, $C_{\rm new}$ and $D_{\rm new}$ are the modified coefficients, and S and B are scale and B values for blurring. By modifying the blurring factors for a particular set of experimental phases, it is possible to optimize the model improvement, which can be monitored through $R_{\rm free}$. This type of modification of the phase-probability distribution is only a stop-gap solution; it would be more proper to remove the bias as the phase-probability distributions are derived. At that point the data underlying the distributions are available, whereas the refinement programs usually have no access to this information, and hence can hardly perform satisfactory bias removal.

4. Refinement of partial structures

Often only part of a structure can be built into the first experimentally phased map. To refine this partial structure Murshudov *et al.* (1997) proposed assuming that there are two components of the structure: one comprising the modelled atoms, and the other comprising the unexplained part of the electron density, which can be transformed to give a component of the total $\mathbf{F_c}$ vector. Practical experience has shown that it is better to modify this unexplained part of the electron density. The simplest useful modification is:

$$\rho_{\rm new} = \begin{cases} \rho_{\rm old} & \text{if } \rho_{\rm old} > \beta \text{ r.m.s.} \\ 0 & \text{otherwise} \end{cases},$$

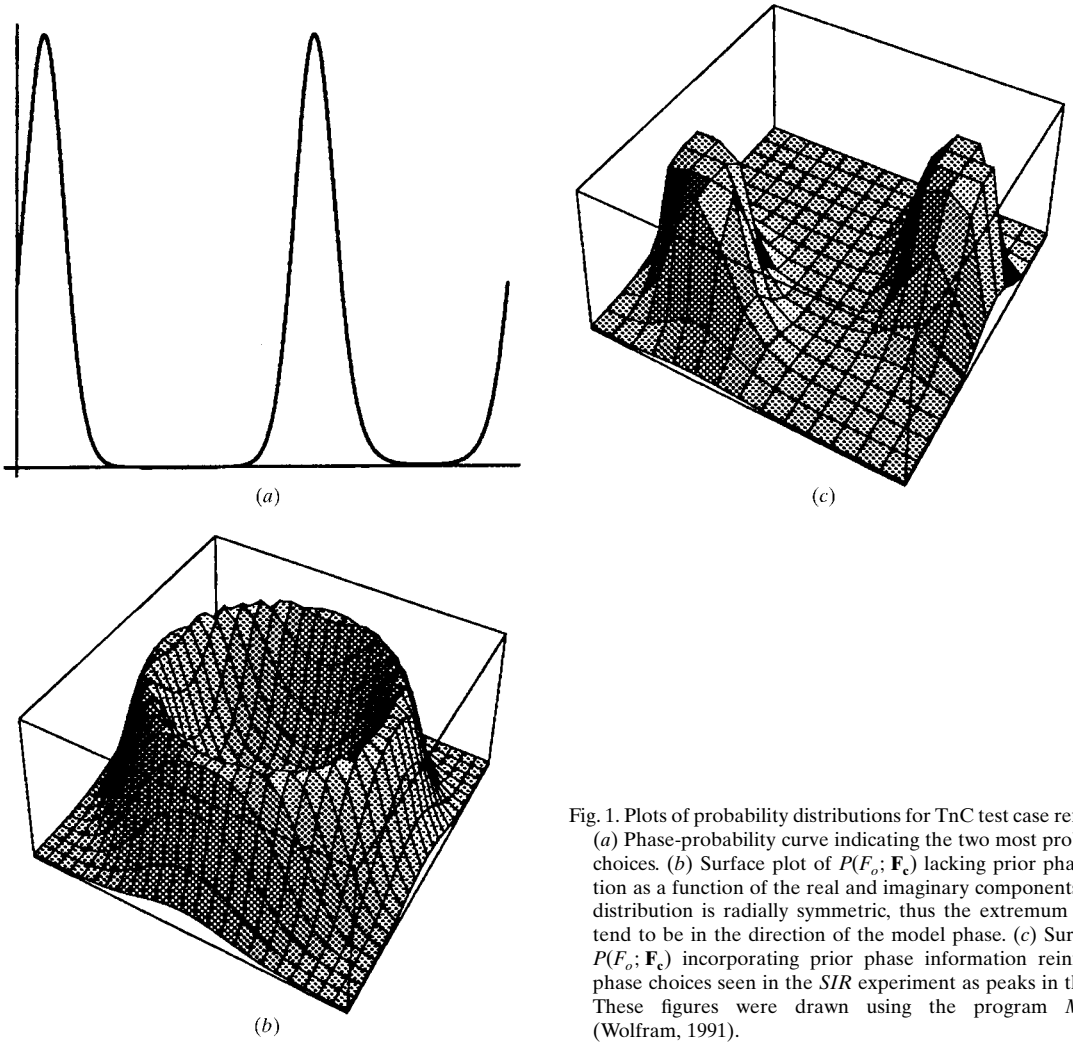


Fig. 1. Plots of probability distributions for TnC test case reflection 116. (*a*) Phase-probability curve indicating the two most probable phase choices. (*b*) Surface plot of $P(F_o; \mathbf{F_c})$ lacking prior phase information as a function of the real and imaginary components of $\mathbf{F_c}$. This distribution is radially symmetric, thus the extremum chosen will tend to be in the direction of the model phase. (*c*) Surface plot of $P(F_o; \mathbf{F_c})$ incorporating prior phase information reinforces both phase choices seen in the *SIR* experiment as peaks in the function. These figures were drawn using the program *Mathematica* (Wolfram, 1991).

Table 2. *Refinement statistics for the TnC test case*

	Start	Least squares	MLF	MLI	Vector	Mixed	MLHL
R factor	0.557	0.380	0.362	0.348	0.464	0.374	0.340
$R_{\rm free}$	0.544	0.518	0.440	0.434	0.493	0.456	0.404
Mean phase error (°)	73.7	67.6	50.4	48.6	57.8	46.8	38.5
Mean cos (phase error)	0.21	0.29	0.51	0.53	0.42	0.56	0.66
Mean map correlation	0.369	0.477	0.709	0.731	0.579	0.739	0.818

where $\rho_{\rm old}$ is the electron density calculated using the available phases, r.m.s. is the root-mean-squared deviation from the mean for $\rho_{\rm old}$ and β is a constant. To avoid bias towards prior phase information, it is necessary to exclude the cross-validation reflections while calculating $\rho_{\rm old}$. The following procedure for the refinement of partial structures can be used:

(i) build part of the structure;

(ii) screen out the interpreted density, and modify the remaining part as described above or by some other method such as skeletonization (Greer, 1974) or pseudo-atom addition (Isaacs & Agarwal, 1978; Perrakis *et al.*, 1997);

(iii) calculate structure factors from this modified electron density;

(iv) use these structure factors as a partial structure contribution;

(v) after a few cycles of refinement go to step (ii);

(vi) if the map is good enough go to step (i).

To refine the partial structure, $\mathbf{F_c}$ is equated to $D_1\mathbf{F_{c1}} + D_2\mathbf{F_{c2}}$ and σ_Δ^2 is replaced by $\sum_N - D_1^2\sum_{p1} - D_2^2\sum_{p2}$ to take into account the separation of the two different parts of structure and their different expected errors. For more details, see Murshudov *et al.* (1997). Alternatively, an overall difference in the errors of $\mathbf{F_{c1}}$ and $\mathbf{F_{c2}}$ can be accounted for by refining overall B factors for the two contributions. A difference in an overall B factor corresponds to a difference in an overall Gaussian coordinate error for each partial model (Read, 1990). If this approach is taken, D and σ_Δ^2 can be estimated from the combined $\mathbf{F_c}$ in the conventional manner.

As long as the modified unexplained density bears some resemblance to the missing structure, it will be better to include its contribution than to leave it out. The inclusion of a contribution from this density in the total $\mathbf{F_c}$ vector leads to a smaller variance in the probability distribution of the true structure factor given the calculated structure factor and thus further increases the power of the refinement target.

5. Test refinements

The maximum-likelihood target MLHL has been implemented in the programs *REFMAC* (Murshudov *et al.*, 1996), *CNS* (Brünger *et al.*, 1998), *TNT* (Tronrud *et al.*, 1987) and *X-PLOR* (Brünger *et al.*, 1987). Results from tests in *REFMAC* and *CNS* will be discussed here. The first test example shows the superiority of phased likelihood refinement over other types. In the second and third examples the effect of 'bad' and 'good' phases are analysed for two types of problems often arising in a macromolecular structure solution.

5.1. *Troponin-C*

In this test, a 'scrambled' (Rice & Brünger, 1994) starting model was refined using only poor *SIR* phases to supplement the likelihood function. The test protein was troponin-C (TnC), which was originally solved at 2.8 Å resolution using multiple isomorphous replacement (MIR) phasing from 11 derivatives (Herzberg & James, 1985). Of these 11, a single derivative ($TmCl_3$) was chosen. The heavy-atom parameters for this derivative were further refined by *MLPHARE* (Otwinowski, 1991), which subsequently generated the Hendrickson–Lattman coefficients used by MLHL, and the 'best' phase and figure of merit used by the vector and mixed (Fujinaga, 1993) residuals. These phases were relatively poor; *MLPHARE* reported a mean figure of merit of

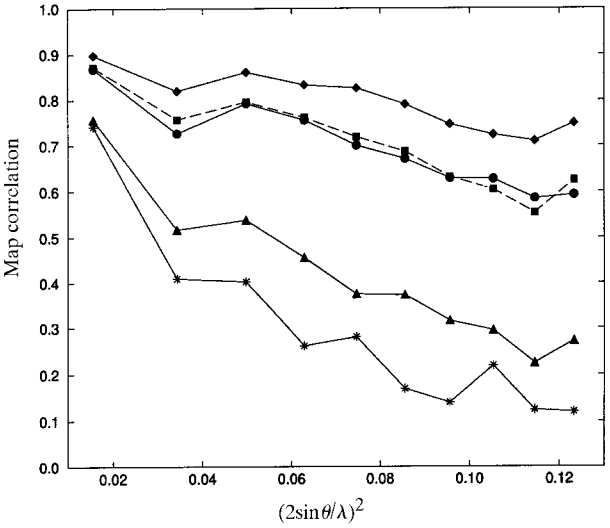


Fig. 2. Map correlations after the TnC test refinements. Stars correspond to the starting model, triangles to the least-squares model, circles to the MLI model, squares to the mixed residual model and diamonds to the MLHL model. Statistics for this graph, as well as all of the reflection files used in the TnC refinement, were generated by the program *SFTOOLS* (B. Hazes, unpublished work).

0.39, while the mean cosine of the phase difference from phases calculated from the final published structure was 0.29.

A starting model was generated by 'scrambling', *i.e.* performing a molecular dynamics run using a target function without reference to X-ray information. The starting model generated in this way had a root-mean-squared deviation for all atoms of 2.28 Å from the published structure. Of the 3868 observed native reflections at 2.8 Å resolution, 496 were flagged as cross-validation data for σ_A estimation (Read, 1997) and $R_{\rm free}$ calculation (Brünger, 1992). The test refinement involved 600 cycles of conjugate-gradient refinement in *CNS* using MLHL, the mixed residual, the vector residual, MLI, MLF and least-squares.

Results from this test are shown in Table 2 and Fig. 2. The results compare refinements using different target functions against the published final structure of TnC refined at 2.0 Å (Herzberg & James, 1988). This final structure has an R factor of 0.155 within the 10–2.0 Å resolution range for intensities $I \geq 2\sigma(I)$. As indicated by the map correlation with this final model, MLHL clearly performed better than any other target function. Additionally, MLHL gave the lowest $R_{\rm free}$ value. Figs. 3 and 4 show regions of TnC in which MLHL accomplished a major shift towards the final model unmatched by any other target function.

5.2. *Cytochrome c′ – starting refinement from a very poor molecular-replacement model.*

The structure of cytochrome *c′* was solved by Baker *et al.* (1995) through a process of intensive model building and refinement. The starting model was based on a molecular-replacement (MR) solution where the model used had only ∼25% homology to cytochrome *c′*. The structure was refined using *TNT* to an R factor of 16.7% computed on data from 20.0–2.15 Å resolution. In the final structure, 96.3% of the residues are within the most-favoured region of the Ramachandran plot as defined by *PROCHECK* (Laskowski *et al.*, 1993).

Very poor MIR phases extending to 3 Å resolution were available. Density modification by the program *DM* was used to refine and extend this set. Tests were performed using both the MIR phases and those generated by *DM*. Although the mean phase difference between the initial MR phases and those calculated from the final model was 89°, *i.e.* close to random, Fig. 5(*a*) shows that for the few low-resolution reflections it was about 65°.

Refinement of the initial MR model without phases, using the maximum-likelihood residual and the sparse-matrix method of minimization within *REFMAC*, failed and improved neither the model nor the phases. When the poor MIR phases were included, the refinement procedure yielded better combined phases, and when *DM* phases extended to 2 Å were used results improved even more. The mean FOM for the MIR phases to 3 Å was 0.36 and for the *DM* phases extending to 2 Å it was 0.45. After several trials, a blurring factor with a scale of 0.7, and a 'temperature factor' of 30 Å^2 was chosen for application to the phase probabilities from *DM*. The geometry was quite loosely restrained, with a weighting ratio of 1.0 between the X-ray and geometric restraint contributions to the residual. 200 cycles were run, with atomic shifts 'damped' by a factor of 0.1. This severe damping factor slowed down the rate of convergence, but avoided the generation of large meaningless shifts. The results of the phased refinement for this extreme case are summarized in Table 3. The behaviour of the phase difference over the resolution range (Fig. 5) shows that the use of the MIR phases improved the low-resolution phases substantially, much more than the high-resolution phases. Using *DM*, phases with the blurring factor gave improvement across the whole

Table 3. *Cytochrome c′ – very poor MR model*

Data 70% complete

	⟨Δα⟩†	Resolution
MIR‡	70.1	3.0
Initial§	89.7	2.0
Model 1¶	80.4	2.0
Combined 1††	73.3	2.0
DM‡‡	73.3	2.0
Model 2§§	70.5	2.0
Combined 2¶¶	68.3	2.0

† Average absolute phase difference between phases calculated from the deposited model. ‡ MIR experimental phases extending to 3 Å. § Phases calculated from initial coordinates. ¶ Phases calculated from coordinates refined using these MIR experimental phases. †† Phases obtained by combining the model phases and the MIR experimental phases. ‡‡ *DM* experimental phases refined and extended to 2 Å §§ Phases calculated from coordinates refined using these *DM* experimental phases. ¶¶ Phases obtained by combining the model phases and the MIR experimental phases.

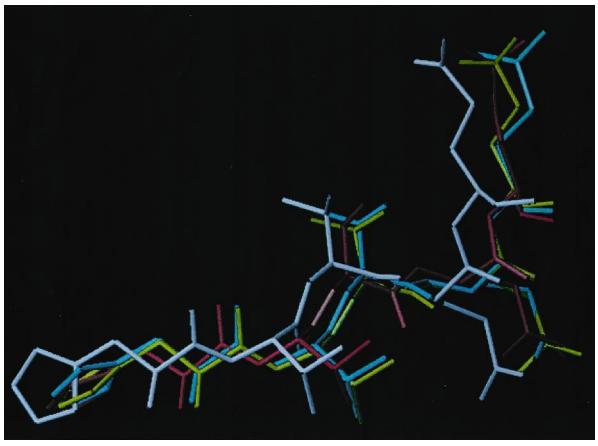
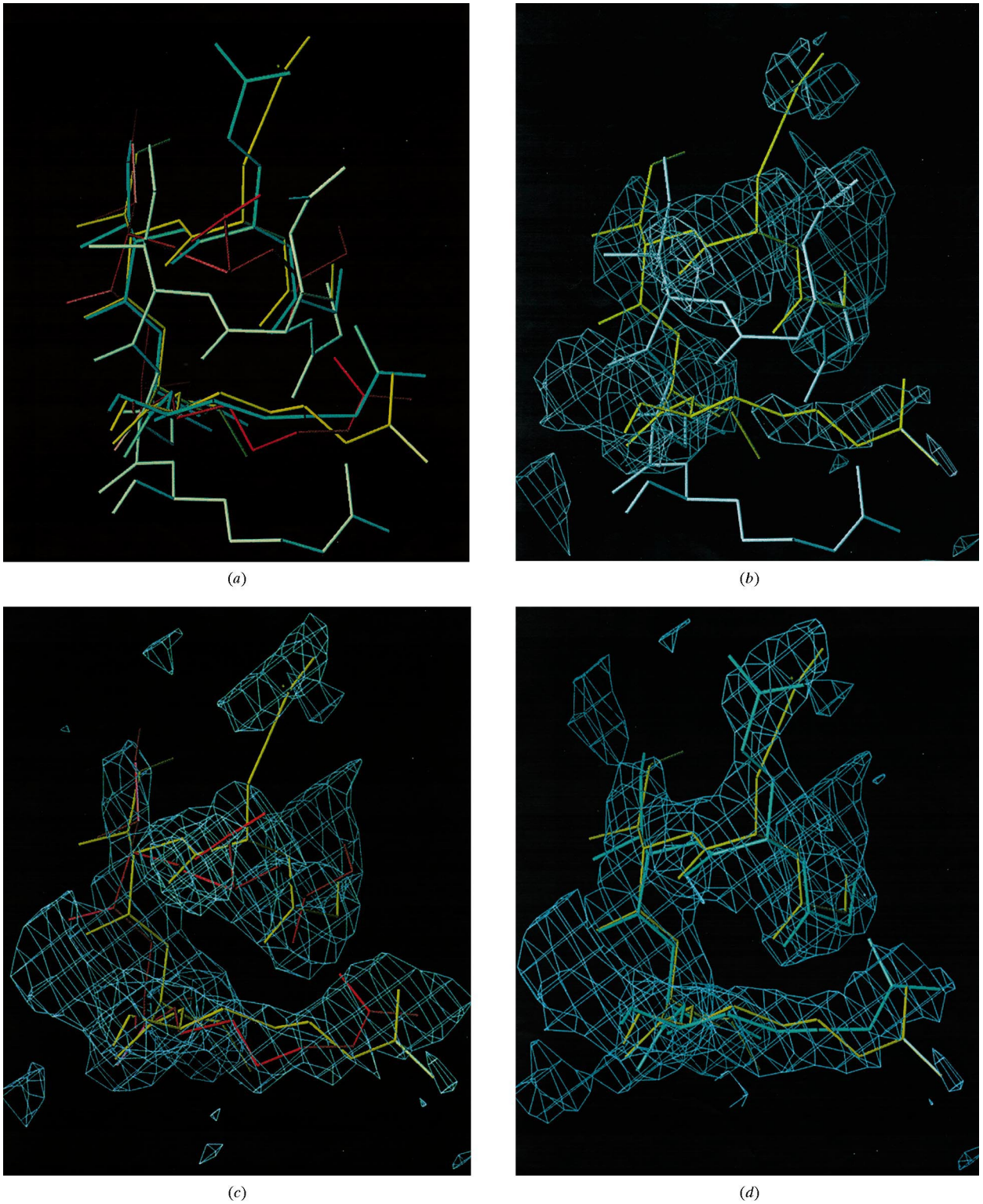


Fig. 3. Rigid-body shift in the TnC test case. In this region of TnC, only the MLHL function (blue) was able to make a full shift from the starting model (white) to the final model (yellow). The result of the refinement of the mixed residual is shown in (red).



(*a*) (*b*)

(*c*) (*d*)

Fig. 4. (*a*) In this region of TnC, none of the refinements were able to fully converge from the starting model (shown in white) to the published structure (shown in yellow). The result of the refinement using the mixed residual and MLHL are shown in red and blue, respectively. (*b*) Combined-phase *SIGMAA* map (Read, 1997) of the starting model and the SIR phases used in refinement. (*c*) Combined-phase *SIGMAA* map of the mixed residual model and the SIR phases used in refinement. (*d*) Combined-phase *SIGMAA* map of the MLHL model and the SIR phases used in refinement. This figure and Fig. 3 were drawn using the program *O* (Jones *et al.*, 1991).

range. The combined phases are better in each range than either the experimental or calculated phases.

5.3. *Oestrogen receptor – refinement of partial structure and effect of 'bad' and 'good' phases*

This structure has been solved by a combination of MIR, multi-crystal NCS averaging and phased refinement (Brzozowski *et al.*, 1997). The R factor for the deposited coordinates (Protein Data Bank, deposition code 1ERR) is 21.8% and the $R_{\rm free}$ is 29.8%. 94.2% of the residues are within the most favoured region of the Ramachandran plot as defined by *PROCHECK*. In this case, the geometry was more tightly restrained, with a weighting ratio of 0.2 between X-ray and geometry restraint contributions to the residual. Shifts were 'damped' by a factor of 0.5.

Two highly correlated derivatives were available and *MLPHARE* gave unrealistically high figures of merit (mean FOM = 0.49 to 3 Å). There were two copies of the molecule in the unit cell and two other crystal forms were available. *DM* was used to carry out twofold averaging within the unit cell and *DMMULTI* was used for multi-crystal averaging. After *DM* the mean FOM was 0.45, and after *DMMULTI* it was 0.58. After several trials, a blurring factor with a scale of 0.7 and a 'temperature factor' of 20 Å^2 were applied to both the MIR and *DM* phase probabilities.

The refinement gave improvement for 20 cycles using the *DM* phases, but improvement continued for 100 cycles using the better *DMMULTI* phase set.

To analyse the effect of 'bad' and 'good' phases for this paper, an intermediate stage of refinement was chosen as a starting point. This intermediate structure had been built using only single-crystal averaging performed with an inadequate mask. Only 45% of the atoms were built and there was considerable error in their positions. The refinement was carried out using the procedure described in §4. The unmodelled part of the electron density was modified with β set to 1.8, and used to calculate a partial structure factor for the unexplained part of the asymmetric unit. This partially built structure was subjected to phased refinement using both the original MIR phases and those obtained after multi-crystal averaging. In both cases blurring factors were used. The results are summarized in Table 4. This shows that the refinement using the more reliable multi-crystal averaged phases gives much better results, although there is still improvement after using MIR phases.

6. Conclusions

The application of the MLHL target function to these test cases has yielded promising results. MLHL performed significantly better than any other target function resulting in clearer electron density and improved phase quality. Furthermore, the difference between the working and free R factors is smaller with the MLHL target, because the inclusion of prior phase information provides more observations for the refinement.

There still remain problems in using prior phase information, one of them being the bias in the available

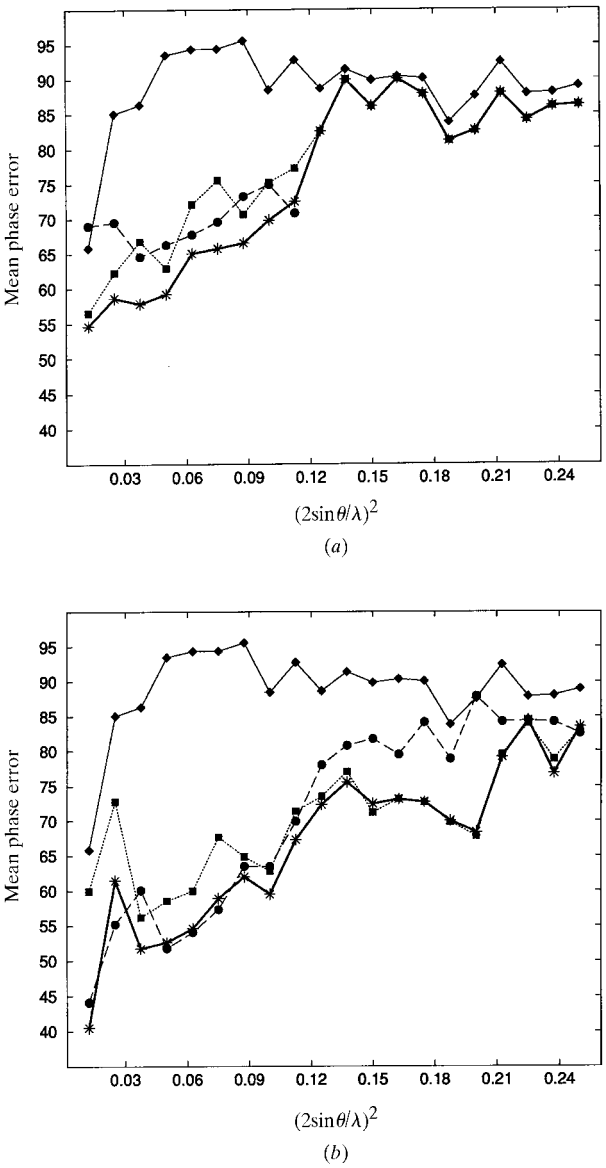


Fig. 5. Plots of absolute mean phase differences for cytochrome c'. All values are given relative to phases calculated from the deposited model. Diamonds correspond to differences from phases calculated from the initial model coordinates, circles to differences from the initial experimental phases, squares to differences from phases calculated from coordinates refined using these experimental phases and stars to differences from the combined phases. (*a*) Using MIR experimental phases to 3 Å. (*b*) Using *DM*-modified experimental phases. The density modification both refines the initial set and extends the set to cover the full resolution range. A blurring factor was applied to the prior phase probabilities.

Table 4. *Oestrogen-receptor (45% complete) model*

	⟨Δα⟩†	Resolution
MIR‡	72.3	3.0
Initial§	68.9	2.6
Model 1¶	61.2	2.6
Combined 1 ††	61.4	2.6
DM averaged‡‡	45.8	2.6
Model 2§§	50.5	2.6
Combined 2¶¶	38.4	2.6

† Average absolute phase difference between phases calculated from the deposited model. ‡ MIR experimental phases extending to 3 Å. § Phases calculated from initial partly built model. ¶ Phases calculated from coordinates refined using these MIR experimental phases. †† Phases obtained by combining the model phases and the MIR experimental phases. ‡‡ *DM* experimental phases after averaging between two crystals, and twofold single-crystal averaging. §§ Phases calculated from coordinates refined using these *DM* experimental phases. ¶¶ Phases obtained by combining the model phases and the MIR experimental phases.

prior phase probability distributions. Using the suggested 'blurring' factors has been shown to improve the performance of MLHL. Future work will include automatic optimization of these blurring factors. Often atomic coordinates are derived from available experimental phases so there will be correlation between the phase error from the current model and that for the 'prior' phases. Taking this into account could improve the behaviour of refinemement.

Another general problem addressed is the refinement of a partial model. Adding a contribution to the calculated structure factor from the modified electron density for the unmodelled part of the crystal has been shown to help. For a better solution to this problem, the probability distribution used should be modified to include off-diagonal terms of the multivariate distribution of structure factors in refinement. However, this may require huge computer resources.

The landscape of the MLHL target function with geometric restraints still contains local minima, suggesting that further improvements can be accomplished with a global optimization scheme such as simulated annealing parameterized in torsion-angle space (Rice & Brünger, 1994). Preliminary tests indicate that the MLHL function in combination with torsion-angle molecular dynamics yields promising results (P. Adams and A. Brünger, personal communication).

In our implementations of MLHL, the prior phase information was assumed to be one dimensional and corresponded to the logarithm of the first two terms of the Fourier series of the actual phase distribution. A more rigorous derivation of the required distribution would take into account the various sources of errors associated with an MIR or MAD experiment, as used in the heavy-atom refinement program *SHARP* (Fortelle & Bricogne, 1997), and combine this information with the errors associated with the current model. This treatment would undoubtedly give a better theoretical account of the sources of errors, and would be suitable for the joint refinement of the structure and the heavy-atom derivatives. However, such a likelihood function would require a large computational cost. The derivation outlined above provides a fast approximation to this more rigorous treatment.

APPENDIX *A*
Derivation of a prior phased likelihood function including measurement errors

In order to derive a likelihood function incorporating prior phase information that includes the effect of measurement error of the native structure-factor amplitude, the joint probability distribution, $P(F, \Delta\alpha, \alpha; F_c)$, must be multiplied by a probability distribution of the observed structure-factor amplitude given the true structure-factor amplitude, $P(F_o; F)$. The resulting expression is the joint probability distribution $P(F_o, F, \Delta\alpha, \alpha; F_c)$. The required distribution is obtained by integrating out the true structure-factor amplitude and phase,

$$P(F_o; F_c) = \int_0^{2\pi}\int_0^{\infty} P(F, \Delta\alpha, \alpha; F_c)P(F_o; F)\mathrm{d}F\mathrm{d}\alpha.$$

In this derivation, a Gaussian probability distribution of the observed structure-factor amplitude given the true structure-factor amplitude will be assumed. As well, only acentric reflections will be considered here, but similar equations can be derived for the centric case. The required integral for the acentric case is

$$P(F_o; F_c) = \frac{1}{(2\pi^3)^{1/2}\sigma_F\varepsilon\sigma_\Delta^2} \times \int_0^{2\pi} P(\alpha)\int_0^{\infty} F \exp\left\{-F^2\left(\frac{1}{2\sigma_F^2}+\frac{1}{\varepsilon\sigma_\Delta^2}\right) + F\left[\frac{F_o}{\sigma_F^2}+\frac{2DF_c\cos(\Delta\alpha)}{\varepsilon\sigma_\Delta^2}\right]\right\}\mathrm{d}F\mathrm{d}\alpha.$$

The true structure-factor amplitude can be integrated out of this expression (Gradshteyn & Ryzhik, 1980), leaving only a numerical integration of the true phase,

$$P(F_o; F_c) = \frac{\sigma_F}{(2\pi^3)^{1/2}(\sigma_F^2+\varepsilon\sigma_\Delta^2)} \times \exp\left(-\frac{F_o^2}{2\sigma_F^2}-\frac{D^2F_c^2}{\varepsilon\sigma_\Delta^2}\right) \times \int_0^{2\pi} P(\alpha)\left\{1+\nu(\pi)^{1/2}\exp(\nu^2)\mathrm{erfc}(-\nu)\right\}\mathrm{d}\alpha,$$

where

$$\nu = \frac{F_o\varepsilon\sigma^2 + 2DF_c\cos(\alpha-\alpha_c)\sigma_F^2}{\sigma_F}\left(\frac{\varepsilon\sigma_\Delta^2+2\sigma_F^2}{2\varepsilon\sigma_\Delta^2}\right)^{1/2}.$$

NSP and RJR thank Osnat Herzberg for permitting us to use the TnC test data, Marie Fraser for assistance in retrieving it, Paul Adams and Bart Hazes for useful discussions and testing the MLHL function and Maxwell Cummings and Michael Ellison for the allocation of computing resources. NSP was supported by the Natural Sciences and Engineering Research Council of Canada, the Alberta Heritage Foundation for Medical Research and a Walter H. Johns Graduate Fellowship. RJR was a Senior Scholar of the Alberta Heritage Foundation for Medical Research and an International Research Scholar of the Howard Hughes Medical Institute. GNM is supported by a BBSR post doctoral fellowship awarded to CCP4 (grant B05273). Some of the work was supported by EU BIOTECH grant BIO2CT-92-0524 (1992-96). EJD is supported by the Medical Research Council (grant G94 13078). GNM and EJD thank A. Brzozowski and A. Pike for the use of their data, the members of the Protein Structure Group of the Department of University of York for cheerfully testing the program and in particular A. Pike, K. Verschuren, M. Isupov and S. Cutfield for intensive testing and valuable suggestions.

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.

Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1997). *Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.

Arnold, E. & Rossmann, M. G. (1988). *Acta Cryst.* A**44**, 270–282.

Baker, E. N., Anderson, B. F., Dobbs, A. J. & Dodson, E. J. (1995). *Acta Cryst.* D**51**, 282–289.

Bricogne, G. & Gilmore, C. J. (1990). *Acta Cryst.* A**46** 284–297.

Bricogne, G. & Irwin, J. (1996). *Macromolecular Refinement: Proceedings of the CCP4 Study Weekend*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–474.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Brzozowski, A. M., Pike, A. C. W., Dauter, Z., Hubbard, R. E., Bonn, T., Engströöm, O., Öhman, L., Greene, G. L., Gustafsson, J.-A. & Carlqulst, M. (1997). *Nature (London)*, **389**, 753–758.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cowtan, K. D. & Main, P. (1996). *Acta Cryst.* D**53**, 43–48.

Diamond, R. (1971). *Acta Cryst.* A**27**, 436–452.

Fortelle, E. de la & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

Fujinaga, M. (1993). *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, edited by W. van Gunsteren, P. Weiner & A. Wilkinson, pp. 371–381. Leiden: ESCOM Science Publishers B.V.

Gradshteyn, I. S. & Ryzhik, I. M. (1980). *Tables of Integrals, Series and Products: Corrected and Enlarged Edition.* San Diego: Academic Press.

Greer, J. (1974). *J. Mol. Biol.* **82**, 279–301.

Hendrickson, W. A. & Lattman, E. E. (1970). *Acta Cryst.* B**26**, 136–143.

Herzberg, O. & James, M. N. (1985). *Nature (London)*, **313**, 653–659.

Herzberg, O. & James, M. N. (1988). *J. Mol. Biol.* **203**, 761–779.

Isaacs, N. W. & Agarwal, R. C. (1978). *Acta Cryst.* A**34**, 782–791.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Lunin, V. Yu. & Urzhumtsev, A. G. (1985). *Acta Cryst.* A**41**, 327–333.

Murshudov, G. N., Dodson, E. J. & Vagin, A. A. (1996). *Macromolecular Refinement: Proceedings of the CCP4 Study Weekend*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Otwinowski, Z. (1991). *Isomorphous Replacement and Anomalous scattering: Proceedings of the CCP4 Study Weekend*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

Perrakis, A., Sixma, T. K., Wilson, K. S. & Lamzin, V. S. (1997). *Acta Cryst.* D**53**, 448–455.

Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.

Read, R. J. (1997). *Methods Enzymol.* **227**, 110–128.

Rice, L. M. & Brünger, A. T. (1994). *Proteins Struct. Funct. Genet.* **19**, 277–290.

Terwilliger, T. C. & Berendzen, J. (1997). *Acta Cryst.* A**53**, 571–579.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.

Wolfram, S. (1991). *Mathematica: A system for doing mathematics by computer*, 2nd ed. Reading, MA: Addison–Wesley.