

## Likelihood-enhanced fast translation functions

Airlie J. McCoy,<sup>a</sup> Ralf W. Grosse-Kunstleve,<sup>b</sup> Laurent C. Storoni<sup>a</sup> and Randy J. Read<sup>a\*</sup>

<sup>a</sup>Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 2XY, England, and

<sup>b</sup>Lawrence Berkeley National Laboratory, One Cyclotron Road, Building 64R0121, Berkeley, California 94720-8118, USA

Correspondence e-mail: rjr27@cam.ac.uk

This paper is a companion to a recent paper on fast rotation functions [Storoni *et al.* (2004), *Acta Cryst. D* **60**, 432–438], which showed how a Taylor-series expansion of the maximum-likelihood rotation function leads to improved likelihood-enhanced fast rotation functions. In a similar manner, it is shown here how linear and quadratic Taylor-series expansions and least-squares approximations of the maximum-likelihood translation function lead to likelihood-enhanced translation functions, which can be calculated by FFT and which are more sensitive to the correct translation than the traditional correlation-coefficient fast translation function. These likelihood-enhanced translation targets for molecular-replacement searches have been implemented in the program *Phaser* using the *Computational Crystallography Toolbox* (*cctbx*).

### 1. Introduction

Macromolecular structure solution by molecular replacement is usually a two-step process. Firstly, a rotation function is used to find the orientation of the search model. Secondly, the position of the (oriented) search model is found using a form of translation function (Rossmann, 1972; Machin, 1985). Less commonly, full  $6n$ -dimensional searches are carried out using either systematic (Sheriff *et al.*, 1999) or stochastic (Kissinger *et al.*, 1999; Glykos & Kokkinidis, 2000) algorithms.

Many translation-search functions have been described in the literature. They fall into two general categories: those that are evaluated at each sampled translation point in real space in a brute-force search and those that are calculated by FFT and therefore generate values for all points on the Fourier grid in real space simultaneously. The FFT methods have the advantage of being several orders of magnitude faster than the brute-force searches. In the brute-force category are *R*-factor searches (Dodson, 1988), correlation searches on amplitude or intensity (Fujinaga & Read, 1987) and full maximum-likelihood searches (Bricogne, 1992, 1997; Read, 2001). In the FFT category are the overlap function (Crowther & Blow, 1967) and variations (Tickle, 1985, 1992), which measure the overlap between the observed and calculated Patterson functions, and the fast correlation coefficient on intensity (Navaza & Vernoslova, 1995). When there is prior phase information, either from experimental phases or a partial model, FFT-based phased translation functions can be used (Colman *et al.*, 1976; Read & Schierbeek, 1988; Navaza, 2001).

The correlation coefficient on intensity (CORR) is currently the most successful fast translation function and is widely used in molecular-replacement software [*e.g.* *AMoRe* (Navaza, 1994), *MolRep* (Vagin & Teplyakov, 1997) and *CNS* (Brünger *et al.*, 1998)]. If  $\mathbf{x}$  is the translation of the oriented search model, then CORR is given by

Received 21 December 2004

Accepted 17 January 2005

$$\text{CORR}(\mathbf{x}) = \frac{\sum_{\mathbf{h}} M_{\mathbf{h}} (I_{\mathbf{h}}^{\text{obs}} - \overline{I_{\mathbf{h}}^{\text{obs}}}) [I_{\mathbf{h}}^{\Phi}(\mathbf{x}) - \overline{I_{\mathbf{h}}^{\Phi}(\mathbf{x})}]}{\left[ \sum_{\mathbf{h}} M_{\mathbf{h}} (I_{\mathbf{h}}^{\text{obs}} - \overline{I_{\mathbf{h}}^{\text{obs}}})^2 \right]^{1/2} \left\{ \sum_{\mathbf{h}} M_{\mathbf{h}} [I_{\mathbf{h}}^{\Phi}(\mathbf{x}) - \overline{I_{\mathbf{h}}^{\Phi}(\mathbf{x})}]^2 \right\}^{1/2}}, \quad (1)$$

where  $\mathbf{h}$  is the Miller index of a reflection,  $M_{\mathbf{h}}$  is its multiplicity,  $I_{\mathbf{h}}^{\text{obs}}$  is the intensity of the observed data,  $\overline{I_{\mathbf{h}}^{\text{obs}}}$  is its mean value,  $I_{\mathbf{h}}^{\Phi}(\mathbf{x})$  is the square of the amplitude of the sum of the phased fixed (*i.e.* known) and moving (*i.e.* search) structure-factor contributions and  $\overline{I_{\mathbf{h}}^{\Phi}(\mathbf{x})}$  is its mean value.

CORR is not as reliable in identifying the correct translation as the maximum-likelihood translation function, which is the same as the Rice function used for structure refinement (Read, 2001). The expression presented previously is rearranged here in order to make the approximations that will be developed more intuitive. To maximize numerical stability, we compute the log of the likelihood, which has its maximum for the same values of the parameters as the likelihood. If the reflections are assumed to be independent, the total log likelihood for a translation  $\mathbf{x}$  in the Rice approximation is given by the sum of the reflection log likelihoods. The likelihood for a single reflection is given by

$$L_{\mathbf{h}}[I_{\mathbf{h}}^{\Phi}(\mathbf{x})] = \frac{2(I_{\mathbf{h}}^{\text{obs}})^{1/2}}{\varepsilon \Sigma_T} \exp \left[ -\frac{I_{\mathbf{h}}^{\text{obs}} + I_{\mathbf{h}}^{\Phi}(\mathbf{x})}{\varepsilon \Sigma_T} \right] I_0 \left\{ \frac{2[I_{\mathbf{h}}^{\text{obs}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})]^{1/2}}{\varepsilon \Sigma_T} \right\}$$

for acentric reflections, where  $I_0$  is the modified Bessel function of order zero, and by

$$L_{\mathbf{h}}[I_{\mathbf{h}}^{\Phi}(\mathbf{x})] = \left( \frac{2}{\pi \varepsilon \Sigma_T} \right)^{1/2} \exp \left[ -\frac{I_{\mathbf{h}}^{\text{obs}} + I_{\mathbf{h}}^{\Phi}(\mathbf{x})}{2\varepsilon \Sigma_T} \right] \times \cosh \left\{ \frac{[I_{\mathbf{h}}^{\text{obs}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})]^{1/2}}{\varepsilon \Sigma_T} \right\}$$

for centric reflections. These likelihoods are defined in terms of the probability of measuring an amplitude  $|F_{\mathbf{h}}^{\text{obs}}| [= (I_{\mathbf{h}}^{\text{obs}})^{1/2}]$ .

The contribution of acentric reflections to the log-likelihood is therefore given by

$$\text{LL}_{\mathbf{h}}[I_{\mathbf{h}}^{\Phi}(\mathbf{x})] = \ln \left[ \frac{2(I_{\mathbf{h}}^{\text{obs}})^{1/2}}{\varepsilon \Sigma_T} \right] - \frac{I_{\mathbf{h}}^{\text{obs}} + I_{\mathbf{h}}^{\Phi}(\mathbf{x})}{\varepsilon \Sigma_T} + \ln \left( I_0 \left\{ \frac{2[I_{\mathbf{h}}^{\text{obs}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})]^{1/2}}{\varepsilon \Sigma_T} \right\} \right) \quad (2a)$$

and for centric reflections by

$$\text{LL}_{\mathbf{h}}[I_{\mathbf{h}}^{\Phi}(\mathbf{x})] = \frac{1}{2} \ln \left( \frac{2}{\pi \varepsilon \Sigma_T} \right) - \frac{I_{\mathbf{h}}^{\text{obs}} + I_{\mathbf{h}}^{\Phi}(\mathbf{x})}{2\varepsilon \Sigma_T} + \ln \left( \cosh \left\{ \frac{[I_{\mathbf{h}}^{\text{obs}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})]^{1/2}}{\varepsilon \Sigma_T} \right\} \right). \quad (2b)$$

In these equations,

$$I_{\mathbf{h}}^{\Phi}(\mathbf{x}) = |D_{\text{move}} F_{\mathbf{h}}^{\text{move}}(\mathbf{x}) + D_{\text{fix}} F_{\mathbf{h}}^{\text{fix}}|^2,$$

$$\Sigma_T(\mathbf{h}) = \Sigma_{N'}(\mathbf{h}) - D_{\text{fix}}^2 \Sigma_P^{\text{fix}} - D_{\text{move}}^2 \Sigma_P^{\text{move}},$$

$$\Sigma_{N'}(\mathbf{h}) = \Sigma_N(\mathbf{h}) - \sum_{j_f} D_{j_f}^2 F_{j_f}^2(\mathbf{h}) - \langle D_{j_f}^2 F_{j_f}^2 \rangle,$$

$$\Sigma_P^{\text{fix}} = \langle I^{\text{fix}} / \varepsilon \rangle, \quad \Sigma_P^{\text{move}} = \langle I^{\text{move}} / \varepsilon \rangle.$$

The subscripts  $j_f$  refer to any fixed (*i.e.* non-translating) molecules that have an unknown origin relative to the moving molecule. Each  $F_{j_f}$  thus represents a structure-factor component with unknown relative phase compared with other components (for example, from fixing the orientation but not the position of a molecule) and may represent the sum of a number of molecular transforms with known relative phase. In contrast,  $F_{\mathbf{h}}^{\text{fix}}$  is a fixed contribution with known phase relative to the contributions of symmetry-related copies of the moving molecule.  $\Sigma_N$  is the bare variance of the Wilson (1949) distribution, in which nothing is known apart from the unit-cell content.  $\Sigma_T$  is a variance that takes into account the acquisition of extra information from the contributions of the fixed and moving molecules.  $\Sigma_{N'}$  accounts for the part of the extra information that arises from the  $F_{j_f}$  contributions with unknown relative phase. The factor  $\varepsilon$  accounts for the statistical effect of symmetry on the expected intensity and is equal to the number of symmetry operations that, when applied to  $\mathbf{h}$ , leave it unchanged. The  $D$  factors are the fractions of the calculated structure-factor components that are correlated with the true values (Luzzati, 1952). To account for the effect of errors in measuring the observed amplitudes, an observational variance contribution is added to  $\Sigma_N$ , as performed for experimental phasing (Green, 1979; de La Fortelle & Bricogne, 1997) and structure refinement (Murshudov *et al.*, 1997).

The maximum-likelihood translation function is time-consuming to compute and this problem is one that can affect success in finding the correct solution. In difficult molecular-replacement solutions, the correct orientation may be a long way down the sorted list of potential orientations in the results from the rotation function, and it may only be possible to identify the correct orientation by the high translation-function score that it generates. If the translation function is too time-consuming to compute then, in practice, the number of potential orientations that can be tested may be limited and the correct orientation may be missed by the search. Thus, developing an approximation to the full-likelihood translation function that retains its superior ability to discriminate correct from incorrect solutions, but that may be calculated by FFT, is important to the practical success of a maximum-likelihood molecular-replacement program. We follow the strategy used in *AMoRe* (Navaza, 1994), in which fast methods are used to generate a list of plausible solutions that is then rescored by a better but computationally more expensive target. In our case, we rescore potential solutions using the translation likelihood target (Read, 2001).

We showed recently (Storoni *et al.*, 2004) that likelihood-enhanced fast rotation functions are an excellent compromise between the high quality but slow full-likelihood rotation-function target and the lower quality but much faster traditional Crowther FFT-based search methods, as they provide better discrimination between correct and incorrect orientations than the Crowther function but at the same speed. Here, we use series approximations to the full maximum-likelihood

Rice translation function to derive several likelihood-enhanced FFT translation functions. These are of higher quality and as fast or faster than CORR.

## 2. Series approximations of maximum-likelihood translation function

The fast correlation coefficient algorithm (Navaza & Vernoslova, 1995) provides an efficient method to compute translation targets expressed through linear and quadratic terms in  $I_{\mathbf{h}}^{\Phi}(\mathbf{x})$ . We have examined two methods to construct such series approximations of the maximum-likelihood translation function. Firstly, we have used Taylor-series expansions to the first and second order. Secondly, we have fitted least-squares linear and quadratic approximations to the likelihood function.

### 2.1. Taylor-series expansions

To compute Taylor-series expansions, we require the derivatives of the function with respect to the expansion variable. Starting from (2), the first derivative of  $LL_{\mathbf{h}}[I_{\mathbf{h}}^{\Phi}(\mathbf{x})]$  with respect to  $I_{\mathbf{h}}^{\Phi}(\mathbf{x})$  is given by

$$LL'_{\mathbf{h}}[I_{\mathbf{h}}^{\Phi}(\mathbf{x})] = \frac{1}{w_{\mathbf{h}}\varepsilon\Sigma_T} \left\{ \frac{m_{\mathbf{h}}(I_{\mathbf{h}}^{\text{obs}})^{1/2}}{[I_{\mathbf{h}}^{\Phi}(\mathbf{x})]^{1/2}} - 1 \right\} \quad (3)$$

and the second derivative is given by

$$LL''_{\mathbf{h}}[I_{\mathbf{h}}^{\Phi}(\mathbf{x})] = \frac{1}{w_{\mathbf{h}}^2\varepsilon\Sigma_T I_{\mathbf{h}}^{\Phi}(\mathbf{x})} \left\{ \frac{(1 - m_{\mathbf{h}}^2)I_{\mathbf{h}}^{\text{obs}}}{\varepsilon\Sigma_T} - \frac{m_{\mathbf{h}}(I_{\mathbf{h}}^{\text{obs}})^{1/2}}{[I_{\mathbf{h}}^{\Phi}(\mathbf{x})]^{1/2}} \right\}, \quad (4)$$

where for acentric reflections

$$m_{\mathbf{h}} = \frac{I_1\{2[I_{\mathbf{h}}^{\text{obs}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})]^{1/2}/\varepsilon\Sigma_T\}}{I_0\{2[I_{\mathbf{h}}^{\text{obs}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})]^{1/2}/\varepsilon\Sigma_T\}}, \quad w_{\mathbf{h}} = 1$$

and for centric reflections

$$m_{\mathbf{h}} = \tanh\{[I_{\mathbf{h}}^{\text{obs}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})]^{1/2}/\varepsilon\Sigma_T\}, \quad w_{\mathbf{h}} = 2.$$

The first-order Taylor series expansion of the Rice function, centred at  $I_{\mathbf{h}}^{\Phi}(\mathbf{x}) = \chi_{\mathbf{h}}$ , is therefore given by

$$\begin{aligned} LL_{\mathbf{h}}^1[I_{\mathbf{h}}^{\Phi}(\mathbf{x})] &= LL_{\mathbf{h}}(\chi_{\mathbf{h}}) + LL'_{\mathbf{h}}(\chi_{\mathbf{h}})[I_{\mathbf{h}}^{\Phi}(\mathbf{x}) - \chi_{\mathbf{h}}] \\ &= C_{\mathbf{h}}^1 + LL'_{\mathbf{h}}(\chi_{\mathbf{h}})I_{\mathbf{h}}^{\Phi}(\mathbf{x}), \end{aligned} \quad (5)$$

where  $C_{\mathbf{h}}^1$  is a constant not dependent on  $\mathbf{x}$ .

Similarly, the second-order Taylor series expansion of the Rice function, centred at  $I_{\mathbf{h}}^{\Phi}(\mathbf{x}) = \chi_{\mathbf{h}}$ , is given by

$$\begin{aligned} LL_{\mathbf{h}}^2[I_{\mathbf{h}}^{\Phi}(\mathbf{x})] &= LL_{\mathbf{h}}(\chi_{\mathbf{h}}) + LL'_{\mathbf{h}}(\chi_{\mathbf{h}})[I_{\mathbf{h}}^{\Phi}(\mathbf{x}) - \chi_{\mathbf{h}}] \\ &\quad + \frac{1}{2}LL''_{\mathbf{h}}(\chi_{\mathbf{h}})[I_{\mathbf{h}}^{\Phi}(\mathbf{x}) - \chi_{\mathbf{h}}]^2 \\ &= C_{\mathbf{h}}^2 + [LL'_{\mathbf{h}}(\chi_{\mathbf{h}}) - \chi_{\mathbf{h}}LL''_{\mathbf{h}}(\chi_{\mathbf{h}})]I_{\mathbf{h}}^{\Phi}(\mathbf{x}) \\ &\quad + \frac{1}{2}LL''_{\mathbf{h}}(\chi_{\mathbf{h}})I_{\mathbf{h}}^{\Phi}(\mathbf{x})^2 \end{aligned} \quad (6)$$

where  $C_{\mathbf{h}}^2$  is a constant not dependent on  $\mathbf{x}$ .

The expansions provide good estimates of the values of the likelihood function over only a restricted range of values of  $I_{\mathbf{h}}^{\Phi}(\mathbf{x})$  close to the point of expansion. Optimal results thus require a good choice of the region to be approximated. We have chosen to centre the Taylor-series expansions on the expected value of  $I_{\mathbf{h}}^{\Phi}(\mathbf{x})$ , so that they are most accurate over

the range of values likely to be sampled during the translation search. The expected value takes account of the fixed contribution, if any, and the amplitudes of the molecular transforms of symmetry copies  $k$  of the moving molecule. This leads to

$$\chi_{\mathbf{h}} = \langle I_{\mathbf{h}}^{\Phi}(\mathbf{x}) \rangle = D_{\text{fix}}^2 |\mathbf{F}_{\mathbf{h}}^{\text{fix}}|^2 + D_{\text{move}}^2 \sum_k |\mathbf{F}_k(\mathbf{h})|^2. \quad (7)$$

We have tested the effect of computing the expected value of  $I_{\mathbf{h}}^{\Phi}(\mathbf{x})$  using less of the available information, *i.e.* by taking account only of the scattering power of the moving molecule but ignoring the amplitudes of the molecular-transform contributions. As expected, this approximation works less well (results not shown). In addition, we have tested the use of Taylor expansions centred on zero, which degrades the results significantly (results not shown).

### 2.2. Least-squares approximations

Least-squares approximations are computed by fitting either a line or a parabola to values of the likelihood function sampled over the range likely to be spanned by  $I_{\mathbf{h}}^{\Phi}(\mathbf{x})$ , weighted by the probability of encountering each value of  $F_{\mathbf{h}}^{\Phi} = [I_{\mathbf{h}}^{\Phi}(\mathbf{x})]^{1/2}$ . The probability distribution for  $F_{\mathbf{h}}^{\Phi}$  is computed by analogy with the Sim-like rotation likelihood function (Read, 2001),

$$p(F_{\mathbf{h}}^{\Phi}) = \frac{2F_{\mathbf{h}}^{\Phi}}{\varepsilon\Sigma_S} \exp\left(-\frac{F_{\mathbf{h}}^{\Phi^2} + |\mathbf{F}_{\text{big}}|^2}{\varepsilon\Sigma_S}\right) I_0\left(\frac{2F_{\mathbf{h}}^{\Phi} |\mathbf{F}_{\text{big}}|}{\varepsilon\Sigma_S}\right)$$

for acentric reflections and

$$p(F_{\mathbf{h}}^{\Phi}) = \left(\frac{2}{\pi\varepsilon\Sigma_S}\right)^{1/2} \exp\left(-\frac{F_{\mathbf{h}}^{\Phi^2} + |\mathbf{F}_{\text{big}}|^2}{2\varepsilon\Sigma_S}\right) \cosh\left(\frac{F_{\mathbf{h}}^{\Phi} |\mathbf{F}_{\text{big}}|}{\varepsilon\Sigma_S}\right)$$

for centric reflections, where

$$\Sigma_S = \langle I_{\mathbf{h}}^{\Phi}(\mathbf{x}) \rangle - |\mathbf{F}_{\text{big}}|^2$$

and  $\mathbf{F}_{\text{big}}$  is the largest term in the sum contributing to  $\langle I_{\mathbf{h}}^{\Phi}(\mathbf{x}) \rangle$ .

The linear least-squares approximation is defined by determining the coefficients  $B_{\mathbf{h}}^L$  and  $C_{\mathbf{h}}^L$  that minimize the residual

$$\int p(F_{\mathbf{h}}^{\Phi}) \{LL_{\mathbf{h}}[I_{\mathbf{h}}^{\Phi}(\mathbf{x})] - [C_{\mathbf{h}}^L + B_{\mathbf{h}}^L I_{\mathbf{h}}^{\Phi}(\mathbf{x})]\}^2 dF_{\mathbf{h}}^{\Phi}. \quad (8)$$

Similarly, the quadratic least-squares approximation is defined by determining the coefficients  $A_{\mathbf{h}}^Q$ ,  $B_{\mathbf{h}}^Q$  and  $C_{\mathbf{h}}^Q$  that minimize the residual

$$\int p(F_{\mathbf{h}}^{\Phi}) \{LL_{\mathbf{h}}[I_{\mathbf{h}}^{\Phi}(\mathbf{x})] - [C_{\mathbf{h}}^Q + B_{\mathbf{h}}^Q I_{\mathbf{h}}^{\Phi}(\mathbf{x}) + A_{\mathbf{h}}^Q I_{\mathbf{h}}^{\Phi}(\mathbf{x})^2]\}^2 dF_{\mathbf{h}}^{\Phi}. \quad (9)$$

In practice, we find that it is sufficient to compute the residual with a sum over as few as five points spanning the range of  $F_{\mathbf{h}}^{\Phi}$ ; *Phaser* uses seven points for stability.

## 3. Likelihood-enhanced translation functions

For calculating the optimal position of a search model given a particular orientation, the translation-independent constants could be ignored as they only change the mean of the search-function scores. However, we have chosen to retain them so

that the scores for different orientations can be compared. The first-order Taylor-series expansion of the Rice function, combining (5) and (7), then gives what we call the likelihood-enhanced translation function 1 (LEFTF1),

$$\text{LEFTF1}(\mathbf{x}) = \sum_{\mathbf{h}} C_{\mathbf{h}}^1 + \text{LL}'_{\mathbf{h}}(\langle I_{\mathbf{h}}^{\Phi} \rangle) I_{\mathbf{h}}^{\Phi}(\mathbf{x}). \quad (10)$$

The second-order Taylor-series expansion of the Rice likelihood target, combining (6) and (7), gives the likelihood-enhanced translation function 2, or LEFTF2,

$$\begin{aligned} \text{LEFTF2}(\mathbf{x}) = \sum_{\mathbf{h}} C_{\mathbf{h}}^2 + [\text{LL}'_{\mathbf{h}}(\langle I_{\mathbf{h}}^{\Phi} \rangle) - \langle I_{\mathbf{h}}^{\Phi} \rangle \text{LL}''_{\mathbf{h}}(\langle I_{\mathbf{h}}^{\Phi} \rangle)] I_{\mathbf{h}}^{\Phi}(\mathbf{x}) \\ + \frac{1}{2} \text{LL}''_{\mathbf{h}}(\langle I_{\mathbf{h}}^{\Phi} \rangle) I_{\mathbf{h}}^{\Phi}(\mathbf{x})^2. \end{aligned} \quad (11)$$

The linear least-squares approximation of the Rice likelihood target, using coefficients determined by minimizing (8), gives the linear likelihood-enhanced translation function, or LETFL,

$$\text{LETFL}(\mathbf{x}) = \sum_{\mathbf{h}} C_{\mathbf{h}}^L + B_{\mathbf{h}}^L I_{\mathbf{h}}^{\Phi}(\mathbf{x}). \quad (12)$$

Finally, the quadratic least-squares approximation of the Rice likelihood target, using coefficients determined by minimizing (9), gives the quadratic likelihood-enhanced translation function, or LEFTFQ,

$$\text{LEFTFQ}(\mathbf{x}) = \sum_{\mathbf{h}} C_{\mathbf{h}}^Q + B_{\mathbf{h}}^Q I_{\mathbf{h}}^{\Phi}(\mathbf{x}) + A_{\mathbf{h}}^Q I_{\mathbf{h}}^{\Phi}(\mathbf{x})^2 \quad (13)$$

Information from fixed parts of the model is introduced into the coefficients of the fast translation targets in two ways. Phased structure-factor contributions are incorporated directly through  $I_{\mathbf{h}}^{\Phi}(\mathbf{x})$  and through the contribution to the variance term  $\Sigma_p^{\text{fix}}$  in  $\Sigma_T$ . Those parts of the structure for which the orientation but not the position are known also contribute to the variance through  $\Sigma_{N'}$  in  $\Sigma_T$ .

#### 4. Implementation

The target functions CORR, LEFT1, LEFT2, LETFL and LEFTFQ described above were implemented in the program *Phaser* using the *Computational Crystallography Toolbox* (Grosse-Kunstleve *et al.*, 2002). For convenience, the calculations are performed in terms of  $E$  values, normalized by dividing the structure factors by  $(\varepsilon \Sigma_N)^{1/2}$ . At the same time the variances, such as  $\varepsilon \Sigma_T$  in (2), are divided by  $\varepsilon \Sigma_N$ .

The fast translation function of Navaza & Vernoslova (1995) was factored into functions that compute  $\sum_{\mathbf{h}} A_{\mathbf{h}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})^2$  and  $\sum_{\mathbf{h}} B_{\mathbf{h}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})$  given the Miller indices  $\mathbf{h}$ , the coefficients  $A_{\mathbf{h}}$  and  $B_{\mathbf{h}}$ , the observed data  $(I_{\mathbf{h}}^{\text{obs}})^{1/2}$ , the fixed components  $\mathbf{F}_{\mathbf{h}}^{\text{fix}}$  of the calculated structure factor and the molecular transform (in  $P1$ ) of the moving molecule before translation. With these functions, all of the LEFTF functions can be computed. For computing  $\sum_{\mathbf{h}} B_{\mathbf{h}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})$ , the run time scales with the second power of the number of symmetry operations. For computing  $\sum_{\mathbf{h}} A_{\mathbf{h}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})^2$ , the run time scales with the fourth power of the number of symmetry operations. For centred cells, the summations can be carried out using only the symmetry operations corresponding to the null centring (the 'primitive'

subset) to minimize the run time (*e.g.* for  $F$ -centred cells, this decreases the run time by a factor of  $4^4 = 256$  for the summations involving the square of the calculated intensities). The same computational saving can also be achieved by transforming the reflection data and coordinates to a primitive setting. This slightly more involved approach has the additional advantage of reducing the memory requirements (*e.g.* by a factor of 4 for  $F$ -centred cells).

The coefficients of the FFT to compute  $\sum_{\mathbf{h}} B_{\mathbf{h}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})$  involve terms to twice the data resolution and those to compute  $\sum_{\mathbf{h}} A_{\mathbf{h}} I_{\mathbf{h}}^{\Phi}(\mathbf{x})^2$  involve terms to four times the data resolution (Navaza & Vernoslova, 1995). It follows from Langs (2002) that the fast translation functions may be evaluated using a grid coarser than the Shannon sampling corresponding to the terms involved. In principle, to preserve all the details, the grid spacing should be at least as fine as the Shannon sampling:  $d_{\text{min}}/4$  for doubled resolution and  $d_{\text{min}}/8$  for quadrupled resolution. Numerical tests show that a grid spacing of  $d_{\text{min}}/4$  is optimal for the first-order approximations (LEFTF1 and LETFL). The results for the second-order approximations (LEFTF2 and LEFTFQ) do not improve much when the grid is made finer than  $d_{\text{min}}/5$  and they are usually acceptable with a grid of  $d_{\text{min}}/4$ .

Note that in our implementation of CORR, the components of  $I_{\mathbf{h}}^{\Phi}(\mathbf{x})$  are weighted by Luzzati (1952)  $D$  values reflecting the expected coordinate errors of the models. This improves the results over those obtained without weights.

#### 5. Test cases

Results from three tests are shown below. These examples were chosen to illustrate the performance of the fast translation function targets in a variety of circumstances, not because the use of the new targets is essential to solving these structures. Earlier work (Read, 2001) has already demonstrated that the likelihood targets are more sensitive to the correct solution than traditional targets such as CORR. In *Phaser*, the top translations from the fast translation search are rescored with the full translation likelihood target; the better the fast search predicts the top peaks, the shorter the list for rescoring can be. We use scatter plots and the correlation coefficient between the fast and slow (LLG) scores to evaluate how well the fast scores approximate the slow score and thus predict the order of the rescored peaks. The test cases below were chosen to assess the fast translation scores in cases where an accurate model accounts for either a small proportion or a large proportion of the total structure factor and also in cases where the model is less accurate.

No low-resolution cutoffs were applied to the available data in any of the tests.

##### 5.1. $\beta$ -Lactamase and $\beta$ -lactamase inhibitor protein complex

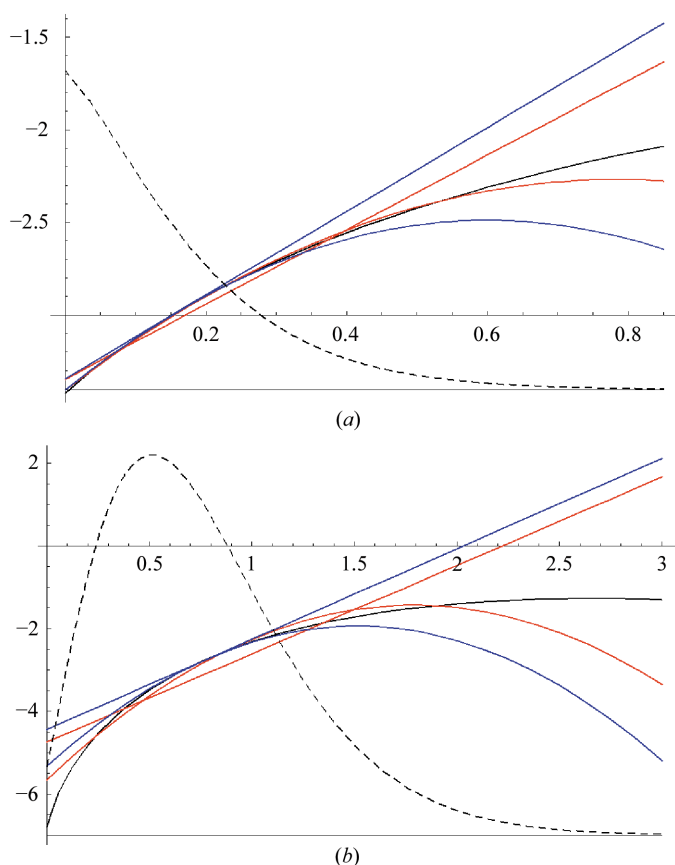
The structure of the complex between  $\beta$ -lactamase (BETA) and  $\beta$ -lactamase inhibitor protein (BLIP) has served as a test structure for maximum-likelihood molecular replacement (Read, 2003; Storoni *et al.*, 2004) because the original structure

**Table 1**

Correlation coefficients between peaks of fast translation maps and LLG values from rescoring in three sets of test calculations.

	Translate BLIP alone, 6 Å	Fix BETA, translate BLIP, 3 Å	Translate 1d0d model of TOXD
CORR	0.752	0.714	0.803
LETF1	0.931	0.920	0.949
LETFL	0.936	0.922	0.950
LETF2	0.969	0.946	0.971
LETFQ	0.981	0.987	0.984

determination using traditional molecular-replacement techniques was difficult, even though good models for BETA and BLIP were available (Strynadka *et al.*, 1996). The difficulty arose in the search for the BLIP component, especially in determining its orientation, as the BETA component is easily found by traditional (and maximum-likelihood) methods. BLIP was difficult to find by traditional methods for two main reasons. Firstly, the BLIP component of the structure



**Figure 1**

Plots of Rice log-likelihood function and its approximations (vertical axis) for a single acentric reflection from the BLIP test case, as a function of  $I_h^\Phi(\mathbf{x})$ . The data are normalized, so that  $I_h^\Phi(\mathbf{x})$  has been divided by  $\Sigma_N$ . The log-likelihood function ( $LL_h$ ) is shown in black, the linear (LETF1) and quadratic (LETF2) Taylor-series approximations are shown in blue and the linear (LETFL) and quadratic (LETFQ) least-squares approximations are shown in red. A dashed line shows the probability distribution of  $I_h^\Phi(\mathbf{x})$ , superimposed using an arbitrary scale and origin. (a) BLIP alone. (b) BLIP with BETA fixed.

**Table 2**

Peak-to-noise discrimination in translation searches.

Results are expressed as  $Z$  scores, *i.e.* r.m.s. deviations above the mean score.

	Translate BLIP alone, 6 Å		Fix BETA, translate BLIP, 3 Å		Translate 1d0d model of TOXD	
	Correct	Top incorrect	Correct	Top incorrect	Correct	Top incorrect
CORR	4.49	4.96	24.26	5.12	6.13	5.24
LETF1	5.31	4.85	31.00	7.86	5.98	5.20
LETFL	5.30	4.81	30.55	7.54	5.93	5.15
LETF2	5.45	5.00	27.43	5.98	5.41	4.73
LETFQ	5.45	4.90	30.03	7.18	5.57	4.82
LLG	5.96	4.94	31.25	7.33	5.86	4.97

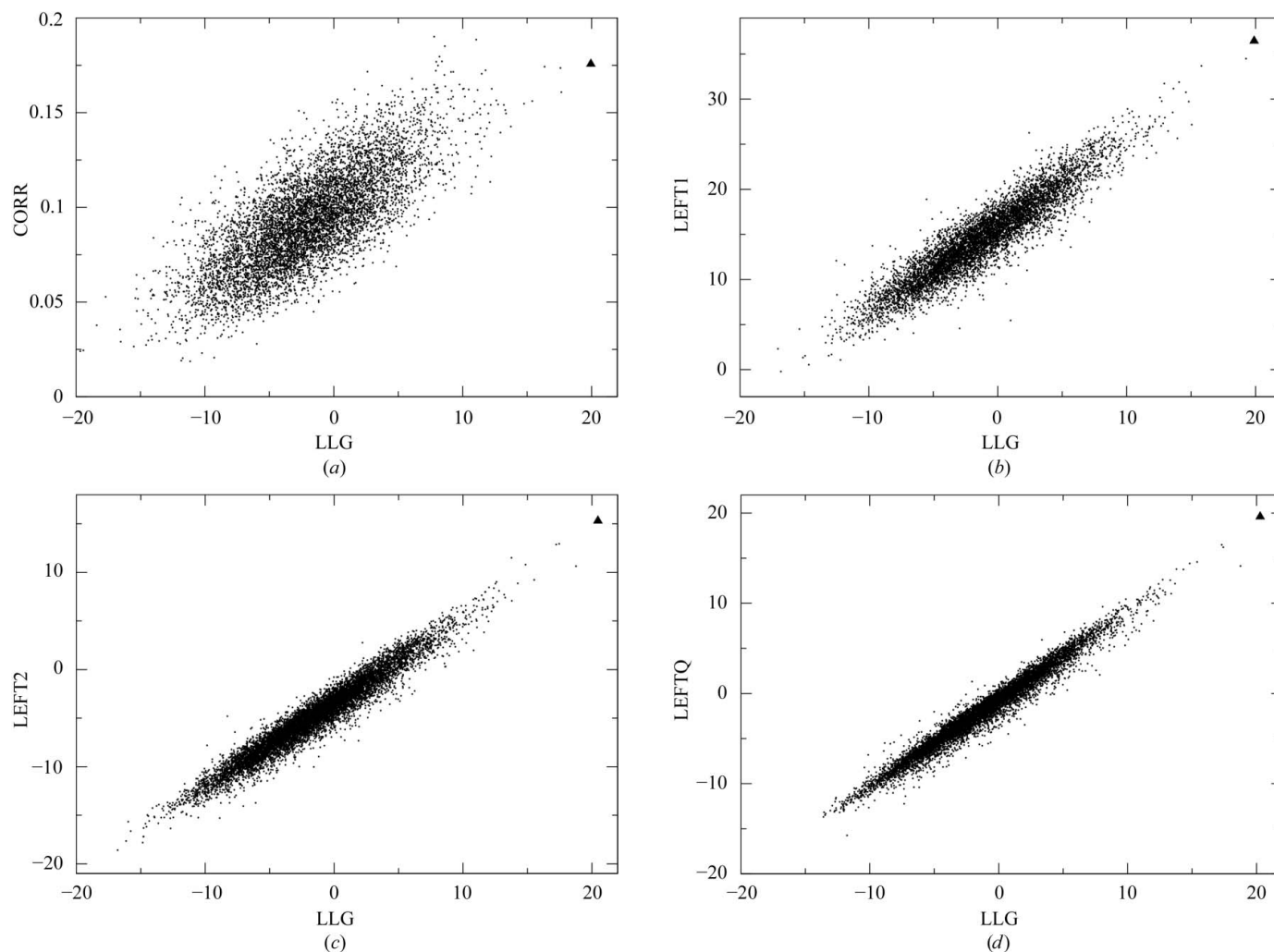
comprises only 38% of the total scattering (the BETA component accounts for the other 62%). Secondly, the data are anisotropic and so there is systematic variation in the structure-factor amplitudes not accounted for by the molecular model, which increases the noise of the search. We have previously shown that full maximum-likelihood molecular replacement (Read, 2003) and the likelihood-enhanced fast rotation functions (Storoni *et al.*, 2004) allow the BLIP component to be found easily. Maximum likelihood overcomes the problems of low scattering and anisotropic data (manuscript in preparation) through better modelling of the structure-factor probabilities and by allowing the information from BETA to be included in the search for BLIP.

**5.1.1. Searching for BLIP alone with restricted resolution.**

The correct orientation for BLIP can be found with a likelihood-based fast rotation search, even when the information about the BETA component is not exploited (Storoni *et al.*, 2004). Once its orientation is known, the translation can be determined easily with any of the fast translation-function scores. To make the translation search more challenging, we have reduced the signal by truncating the resolution of the data to 6 Å. This test illustrates the case where the model predicts only a relatively small component of the structure factor, even if the model is reasonably accurate. As Fig. 1(a) shows, only a small range of values of  $I_h^\Phi(\mathbf{x})$  is spanned for a typical reflection as the model is translated. Over this range the Rice likelihood function is reasonably close to linear. The results in Table 1 demonstrate that all the LETF scores provide a much better prediction of the LLG score than does the CORR score. As one might expect, the higher order approximations provide a better fit to LLG and the least-squares approximations are slightly better than the Taylor-series approximations. The scatter plots in Fig. 2 and the results in Table 2 show that the correct translation receives the top score in all LETF scores, but not with CORR. Nonetheless, the correct translation is near the top of the list even for CORR and would be recovered in this case if the peaks were rescored with the LLG score.

**5.1.2. Searching for BLIP, fixing known BETA contribution.**

In the previous case, the contribution of BLIP accounts for only a small part of the uncertainty in the prediction of the observed structure-factor amplitude, so only a relatively small portion of the Rice-function curve is sampled as the molecule



**Figure 2**

Scatter plots showing correlation between peaks in the fast translation function maps and the LLG values from rescoring. The BLIP component of the BETA–BLIP complex was translated using data restricted to 6 Å resolution and not taking into account the contribution of the BETA component. A triangle indicates the score for the best translation. (a) CORR score. (b) LETF1 score. (c) LETF2 score. (d) LETFQ score.

is translated. To test the case where the translated model accounts for a much greater part of the uncertainty, we carried out tests in which the known contribution of BETA was fixed during the translation search for BLIP using all data to 3 Å resolution. In this case, as shown in Fig. 1(b), a wider range of values of  $I_n^\Phi(\mathbf{x})$  will be sampled and the Rice likelihood function deviates more from a straight line. The results in Table 1 show that, as one might expect, the first-order approximations work somewhat more poorly than in the case with BLIP alone, but the correlation with the LLG score is still very high for all LETF scores. The results in Table 2 show that with the correct orientation this translation problem is trivial for all search targets.

## 5.2. TOXD

In a further test, we used the test data for  $\alpha$ -dendrotoxin (TOXD) distributed with the CCP4 suite (Collaborative

Computational Project, Number 4, 1994). This structure was originally solved by isomorphous replacement (Skarzynski, 1992), but it shares 36% sequence identity with bovine pancreatic trypsin inhibitor. As a model, we have used the structure of bovine pancreatic trypsin inhibitor from PDB entry 1d0d (St Charles *et al.*, 2000). The results in Table 1 demonstrate that the LETF scores are equally good approximations of LLG, whether the model is closely or more distantly related to the target structure.

## 6. Conclusions

The results demonstrate that all four likelihood-based fast translation functions investigated here (LETF1, LETF2, LETFL and LETFQ) are superior to CORR in approximating the full-likelihood target, LLG, and thus in predicting the top solutions. The first-order approximations (LETF1 and

LETFL) have the significant advantage that they only require one FFT, with a map sampled at  $d_{\min}/4$ . The second-order approximations (LETF2 and LETFQ) require only two FFTs compared with the three needed for CORR.

In practice, we prefer the use of the LETF1 fast translation function, which is the program default in *Phaser*. We have not found a molecular-replacement problem in which the second-order targets succeed in finding the correct solution when LETF1 fails. This is probably because molecular-replacement problems become more difficult as the fragment to be found becomes smaller or the model becomes less accurate. In both situations, the proportion of the observed structure-factor amplitude explained by the model decreases and, as illustrated in Fig. 1, the relevant portion of the Rice likelihood-function curve becomes more linear. In addition, the second-order approximations require a second FFT, leading to greater memory requirements. Finally, the calculation of the first derivative needed for LETF1 is simpler and perhaps more reliable than the least-squares fitting required for LETFL.

The translation function is also used in dual-space substructure searches (Grosse-Kunstleve & Adams, 2003), where peaks in the Patterson function are selected. These represent heavy-atom pairs and the heavy-atom pairs are then translated through the unit cell to find the position of the pair. This pair is the basis of a bootstrap procedure to find the rest of the heavy atoms in a substructure. The likelihood-enhanced translation functions described here could also be used for these substructure searches.

The program *Phaser* has been released as part of the *PHENIX* (Adams *et al.*, 2002) software suite and will be released as part of the *CCP4* (Collaborative Computational Project, Number 4, 1994) suite. It is also available from the authors (see <http://www-structmed.cimr.cam.ac.uk/phaser> for details).

We are grateful to Michael James and Natalie Strynadka for supplying the data for the  $\beta$ -lactamase complex test case. This work was funded by NIH/NIGMS under grant No. 1P01GM063210 and by a Principal Research Fellowship from the Wellcome Trust (RJR).

## References

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* **D58**, 1948–1954.  
 Bricogne, G. (1992). *Proceedings of the CCP4 Study Weekend. Molecular Replacement*, edited by W. Wolf, E. J. Dodson & S. Gover, pp. 62–75. Warrington: Daresbury Laboratory.

Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.  
 Brünger, A. T., Adams, P. D., Clore, G. M., 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.* **D54**, 905–921.  
 Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* **D50**, 760–763.  
 Colman, P. M., Fehlhammer, H. & Bartels, K. (1976). *Crystallographic Computing Techniques*, edited by F. R. Ahmed, K. Huml & B. Sedlacek, pp. 248–258. Copenhagen: Munksgaard.  
 Crowther, R. A. & Blow, D. M. (1967). *Acta Cryst.* **23**, 544–548.  
 Dodson, E. J. (1988). *Crystallographic Computing 4: Techniques and New Technologies*, edited by N. W. Isaacs & M. R. Taylor, pp. 80–96. Oxford University Press.  
 Fujinaga, M. & Read, R. J. (1987). *J. Appl. Cryst.* **20**, 517–521.  
 Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* **D56**, 169–174.  
 Green, E. A. (1979). *Acta Cryst.* **A35**, 351–359.  
 Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Acta Cryst.* **D59**, 1966–1973.  
 Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.  
 Kissinger, C. R., Gelhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* **D55**, 484–491.  
 La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.  
 Langs, D. A. (2002). *J. Appl. Cryst.* **35**, 505.  
 Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.  
 Machin, P. A. (1985). Editor. *Proceedings of the Daresbury Study Weekend. Molecular Replacement*. Warrington: Daresbury Laboratory.  
 Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* **D53**, 240–255.  
 Navaza, J. (1994). *Acta Cryst.* **A50**, 157–163.  
 Navaza, J. (2001). *Acta Cryst.* **D57**, 1367–1372.  
 Navaza, J. & Vernoslova, E. (1995). *Acta Cryst.* **A51**, 445–449.  
 Read, R. J. (2001). *Acta Cryst.* **D57**, 1373–1382.  
 Read, R. J. (2003). *Crystallogr. Rev.* **9**, 33–41.  
 Read, R. J. & Schierbeek, A. J. (1988). *J. Appl. Cryst.* **21**, 490–495.  
 Rossmann, M. G. (1972). Editor. *The Molecular Replacement Method*. New York: Gordon & Breach.  
 Sheriff, S., Klei, H. E. & Davis, M. E. (1999). *J. Appl. Cryst.* **32**, 98–101.  
 Skarzynski, T. (1992). *J. Mol. Biol.* **224**, 671–683.  
 St Charles, R., Padmanabhan, K., Arni, R. V., Padmanabhan, K. P. & Tulinsky, A. (2000). *Protein Sci.* **9**, 265–272.  
 Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Cryst.* **D60**, 432–438.  
 Strynadka, N. C., Jensen, S. E., Alzari, P. M. & James, M. N. (1996). *Nature Struct. Biol.* **3**, 290–297.  
 Tickle, I. J. (1985). *Proceedings of the Daresbury Study Weekend. Molecular Replacement*, edited by P. A. Machin, pp. 22–26. Warrington: Daresbury Laboratory.  
 Tickle, I. J. (1992). *Proceedings of the Daresbury Study Weekend. Molecular Replacement*, edited by W. Wolf, E. J. Dodson & S. Gover, pp. 20–32. Warrington: Daresbury Laboratory.  
 Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.  
 Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.