Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Likelihood-based refinement. I. Irremovable model errors

V. Y. Lunin, P. V. Afonine and A. G. Urzhumtsev

Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 27 November 2001
Accepted 12 February 2002

Likelihood-based refinement. I. Irremovable model errors

V. Y. Lunin,[a,b] P. V. Afonine[c,b] and A. G. Urzhumtsev[b]*

[a]Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, 142290 Moscow Region, Russia, [b]LCM3B, UMR 7036 CNRS, Université Henri Poincaré, Nancy 1, BP 239, Faculté des Sciences, Vandoeuvre-lès-Nancy, 54506 France, and [c]Centre Charles Hermite, LORIA, Villers-lès-Nancy, 54602 France. Correspondence e-mail: sacha@lcm3b.uhp-nancy.fr

In conventional structure refinement, the discrepancy between the calculated magnitudes and those observed in X-ray experiments is attributed to errors inherent in preliminary assigned values of the model parameters. However, the chosen set of model parameters may not be adequate to describe the structure factors precisely. For example, if some atoms are not included in the current model, then the structure factors calculated from such a partial model contain 'irremovable errors'. These errors cannot be eliminated by any choice of the parameters of the partial structure. Probabilistic modelling suggests a way to take irremovable errors into account. Every trial set of values of the model parameters is now associated with the joint probability distribution of the calculated magnitudes, rather than with a particular set of magnitudes. The new goal of the refinement is formulated as the search for the distribution that is the most consistent with the observed data. The statistical likelihood is a possible measure of the consistency. The suggested quadratic approximation of the likelihood function allows the likelihood-based refinement to be considered as a kind of least-squares refinement that uses appropriate weights and modified targets for the calculated magnitudes. This in turn enables the analysis of tendencies of the likelihood-based refinement in comparison with the classical least-squares refinement.

1. Introduction

Of recent attractive ideas in crystallographic refinement, one is to enhance its power by maximization of a likelihood function instead of the conventional minimization of the least-squares (LSQ) criterion. A special type of this likelihood function, which was used primarily for the evaluation of model quality (Lunin & Urzhumtsev, 1984; Read, 1986, 1990; Lunin & Skovoroda, 1995; Urzhumtsev *et al.*, 1996), was suggested recently as a new goal function for the refinement of atomic models (Pannu & Read, 1996; Bricogne & Irwin, 1996; Murshudov *et al.*, 1997; Adams *et al.*, 1997; Pannu *et al.*, 1998). While the practical use of this approach has demonstrated encouraging progress, the theoretical reasons to change the refinement procedure are still not clear. The likelihood function always has some probabilistic model behind it and the clear understanding of that model and its links to the likelihood-function parameters is necessary to manage the refinement process, which is different from the classical LSQ refinement, as illustrated in §2 below. It is important to stress that the likelihood-based strategy (ML refinement in what follows) changes the course of the work and involves new tendencies in the refinement. To analyse these tendencies, a quadratic approximation of the ML residual is derived and studied in §3. Simple test calculations (§4) illustrate this study. Some technical details are discussed in Appendices *A* and *B*.

In the process of the conventional LSQ refinement, the magnitudes $\{F_{\mathbf{s}}^{\rm calc}\}_{\mathbf{s}\in S}$ calculated from the current values of atomic coordinates and from other model parameters are fitted to the observed structure-factor magnitudes $\{F_{\mathbf{s}}^{\rm obs}\}_{\mathbf{s}\in S}$, minimizing the residual

$$Q_{\rm LSQ} = \sum_{\mathbf{s}\in S} w_{\mathbf{s}}(kF_{\mathbf{s}}^{\rm calc} - F_{\mathbf{s}}^{\rm obs})^2 \qquad (1)$$

or

$$Q_{\rm LSQ}^{(2)} = \sum_{\mathbf{s}\in S} w_{\mathbf{s}}[(kF_{\mathbf{s}}^{\rm calc})^2 - (F_{\mathbf{s}}^{\rm obs})^2]^2. \qquad (2)$$

The weights $\{w_{\mathbf{s}}\}_{\mathbf{s}\in S}$ may reflect the accuracy of the observed magnitudes or other effects. When refining structures of small molecules, the weights $w_{\mathbf{s}} = \sigma_{\mathbf{s}}^{-2}$ are usually used, where $\sigma_{\mathbf{s}}$ reflects the accuracy of the measured $F_{\mathbf{s}}^{\rm obs}$. In protein crystallography, the weights are sometimes ignored and the minimization of (1) or (2) is performed with unit weights. Usually, the scale factor k is calculated for the given values of $\{F_{\mathbf{s}}^{\rm calc}\}_{\mathbf{s}\in S}$ and $\{F_{\mathbf{s}}^{\rm obs}\}_{\mathbf{s}\in S}$ to minimize the chosen criterion.

A likelihood function appears when some probabilistic models are introduced in order to describe structural features

or experimental environments that affect the structure factors but are not reflected *explicitly* in the current model. Such features will be referred to as 'irremovable model errors'. For example, at early stages of the structure solution, approximate atomic coordinates may be known for a part of the model only. In this case, the calculated structure-factor magnitudes do not coincide with the observed ones even when the true coordinates for the partial-model atoms are found. An attempt to force the calculated magnitudes to be as close as possible to the observed ones may move the atoms of such a partial model away from their true positions in the course of refinement (see §4 for examples). A possible way to overcome the obstacle may be to use a probabilistic modelling and to take the missed atoms into account indirectly, for example:

(i) estimate for every trial partial model how large would be the probability to reproduce the observed magnitude values if the model were to be completed randomly by the necessary amount of missed atoms and the structure factors were to be calculated from such a combined model;

(ii) among all possible partial models choose the model that maximizes this probability.

The probability mentioned above is the likelihood value and the suggested approach is the maximal-likelihood principle for the choice of parameters of a probability distribution (in the considered case these parameters are the atomic coordinates of the partial model). Such an approach occupies an intermediate position between the full ignorance of the missed part of the structure and the extension of the set of model parameters by adding new atoms.

In the procedure, usually referred to as ML refinement, the residual (1) is replaced by the negative logarithm of the likelihood. The model-dependent part of this new residual may be represented (see Appendix *A* for details) as

$$Q_{\rm ML} = \sum_{\mathbf{s}\in S} \Psi(F_{\mathbf{s}}^{\rm calc}; F_{\mathbf{s}}^{\rm obs}, \alpha_{\mathbf{s}}, \beta_{\mathbf{s}}) \Rightarrow \min, \qquad (3)$$

with

$$\Psi = \begin{cases} \Psi_a = \dfrac{\alpha_{\mathbf{s}}^2 (F_{\mathbf{s}}^{\rm calc})^2}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}} - \ln\left[I_0\left(\dfrac{2\alpha_{\mathbf{s}} F_{\mathbf{s}}^{\rm calc} F_{\mathbf{s}}^{\rm obs}}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}} \right) \right] \\ \qquad\qquad \text{for acentric reflections,} \\ \Psi_c = \dfrac{\alpha_{\mathbf{s}}^2 (F_{\mathbf{s}}^{\rm calc})^2}{2\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}} - \ln\left[\cosh\left(\dfrac{\alpha_{\mathbf{s}} F_{\mathbf{s}}^{\rm calc} F_{\mathbf{s}}^{\rm obs}}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}} \right) \right] \\ \qquad\qquad \text{for centric reflections.} \end{cases} \qquad (4)$$

Here the parameter $\varepsilon_{\mathbf{s}}$ depends only on the reflection indices and on the particular space group $\Gamma = \{(\mathbf{R}_\nu, \mathbf{t}_\nu)\}_{\nu=1}^n$ and may be calculated as the number of reciprocal-space symmetries $\mathbf{R}_\nu^T$ that when applied to the vector $\mathbf{s}$ leave it invariable, *i.e.* $\mathbf{R}_\nu^T \mathbf{s} = \mathbf{s}$. The notations I_0 (and I_1 below) and cosh (and tanh below) represent the modified Bessel functions and the hyperbolic cosine and tangent, respectively.

The parameters $\alpha_{\mathbf{s}}$ and $\beta_{\mathbf{s}}$ play the key role in the definition of new targets and influence significantly the results of the refinement (Afonine *et al.*, 2001, 2002). These parameters and their values are linked to the probabilistic model used to describe irremovable errors (Lunin & Urzhumtsev, 1984; Read, 1986; Lunin & Skovoroda, 1995; Pannu & Read, 1996). Usually, the parameters $\alpha_{\mathbf{s}}$ and $\beta_{\mathbf{s}}$ may be considered as constant inside thin spherical shells in reciprocal space. To some extent, the values $\{\alpha_{\mathbf{s}}\}$ reflect the scale of irremovable coordinate errors in the model, *e.g.* they may be defined by the mean difference between the coordinates of atoms of the studied object and those of the search model used for rigid-body refinement (see §2.2 below). The values $\{\beta\}_{\mathbf{s}}$ reflect both the irremovable coordinate errors in the model and the amount of scattering density that is not included in the calculation of $\{F_{\mathbf{s}}^{\rm calc}\}_{\mathbf{s}\in S}$ (undetermined part of the structure, bulk solvent *etc.*). Additionally, $\alpha_{\mathbf{s}}$ and $\beta_{\mathbf{s}}$ contain information on the scale factor, which must be applied to the calculated magnitudes to place them on the same scale as the observed values.

There are two main approaches to estimate these parameters. If there exists some probabilistic hypothesis concerning the irremovable errors in the atomic model, then these parameters may be sometimes calculated explicitly [see formulae (17) and (19) below as examples; more examples are given by Urzhumtsev *et al.* (1996)]. Another way is to obtain likelihood-based estimates of these parameters supposing a general form of the distribution and comparing the observed structure-factor magnitudes with those corresponding to the starting atomic model (Lunin & Urzhumtsev, 1984; Read, 1986). Test set reflections only must be used in this case to obtain reliable estimates (Lunin & Skovoroda, 1995; Brünger, 1997; Skovoroda & Lunin, 2000). In what follows, we consider $\alpha_{\mathbf{s}}$ and $\beta_{\mathbf{s}}$ to be known parameters.

By its construction, the likelihood function resulting in the target (3)–(4) is the joint probability distribution of magnitudes of independent complex variables (structure factors) when each of them is distributed according to the two-dimensional Gaussian distribution and has uncorrelated real and imaginary parts (Appendix *A*). Such likelihood functions arise frequently when the probabilistic model considered for irremovable model errors results in a Gaussian distribution for the particular structure factor. To emphasize this common nature of the function (4), we use the most general form of notation (α and β) for the two parameters defining Gaussian distribution (Lunin & Urzhumtsev, 1984). Other notations may be used for these parameters or for their combinations, reflecting the specificity of a particular probability model (Luzzati, 1952; Sim, 1959; Srinivasan & Parthasarathy, 1976; Read, 1986). It must be noted too that (3)–(4) is not the only possible type of likelihood-based target and other more complicated likelihood functions may appear.

The rest of the paper is devoted to a detailed analysis of (3)–(4) and corresponding consequences. Briefly, when being considered as a function of $F_{\mathbf{s}}^{\rm calc}$, any member Ψ in (3) may have quite different behaviour depending on the value of the parameter

$$p = \frac{F_{\mathbf{s}}^{\rm obs}}{(\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}})^{1/2}} \qquad (5)$$

(Fig. 1). If $F_{\mathbf{s}}^{\rm obs} > (\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}})^{1/2}$ (*i.e.* $p > 1$), the function Ψ first decreases from zero to some negative value and then increases monotonically so that its minimum is attained for some positive value $F_{\mathbf{s}}^{*}$, which is different from $F_{\mathbf{s}}^{\rm obs}$ as a rule. If the reflection is relatively weak and $F_{\mathbf{s}}^{\rm obs} \leq (\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}})^{1/2}$ (*i.e.* $p < 1$), then the function Ψ grows monotonically with $F_{\mathbf{s}}^{\rm calc}$ so that its minimum is equal to zero and is attained for $F_{\mathbf{s}}^{\rm calc} = 0$. In this case, the likelihood-based target fits the calculated magnitude to the zero value ($F_{\mathbf{s}}^{*} = 0$) regardless of the particular value of $F_{\mathbf{s}}^{\rm obs}$.

In the vicinity of the point of its minimum, any member in (3) may be approximated by a quadratic function (Lunin & Urzhumtsev, 1999; Afonine *et al.*, 2001, 2002), leading to

$$\widetilde{Q}_{\rm ML} = \sum_{\mathbf{s}\in S} w_{\mathbf{s}}^{*}(F_{\mathbf{s}}^{\rm calc} - F_{\mathbf{s}}^{*})^2. \qquad (6)$$

This new residual has the same form as the classical LSQ residual but the target value for $F_{\mathbf{s}}^{\rm calc}$ is now the modified value $F_{\mathbf{s}}^{*}$ instead of the observed magnitude, and the weight $w_{\mathbf{s}}^{*}$ is defined from the curvature of Ψ at the point of its minimum.

The weights $w_{\mathbf{s}}^{*}$ and the modified target values $F_{\mathbf{s}}^{*}$ in (6) may be represented (see §3) as

$$F_{\mathbf{s}}^{*} = \frac{(\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}})^{1/2}}{\alpha_{\mathbf{s}}}\,\mu\!\left[\frac{F_{\mathbf{s}}^{\rm obs}}{(\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}})^{1/2}}\right], \quad w_{\mathbf{s}}^{*} = c_{\mathbf{s}}\frac{\alpha_{\mathbf{s}}^2}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}}\,\nu\!\left[\frac{F_{\mathbf{s}}^{\rm obs}}{(\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}})^{1/2}}\right], \qquad (7)$$

where $\mu(p)$ and $\nu(p)$ are some uniquely defined functions. The 'attenuating' function $\mu(p)$ is equal to zero for $0 \leq p \leq 1$ and is defined for any $p > 1$ as the unique positive solution of the equation

$$\mu = p\,\frac{I_1(2p\mu)}{I_0(2p\mu)} \quad \text{for acentric reflections} \qquad (8)$$

or

$$\mu = p\tanh(p\mu) \quad \text{for centric reflections} \qquad (9)$$

[some ways of explicit calculation of $\mu(p)$ are discussed in Appendix *B*]. The plots of this function for the centric and acentric cases are shown in Fig. 2.

The 'weighting' function $\nu(p)$ is defined as

$$\nu_a(p) = \begin{cases} 1 - p^2 & \text{for } 0 \leq p \leq 1, \\ 2[1 - p^2 + \mu^2(p)] & \text{for } p > 1, \end{cases} \quad \text{for acentric reflections} \qquad (10)$$

and

$$\nu_c(p) = 1 - p^2 + \mu^2(p), \quad \text{for centric reflections.} \qquad (11)$$

The plot of $\nu(p)$ for the centric and acentric cases is shown in Fig. 3.

The coefficient $c_{\mathbf{s}}$ is defined as

$$c_{\mathbf{s}} = \begin{cases} 1 & \text{for acentric reflections,} \\ \frac{1}{2} & \text{for centric reflections.} \end{cases} \qquad (12)$$

The presentation of the goal function in the form (6) allows one to perform a sort of likelihood-based refinement with the standard LSQ refinement tools. This extends the possibilities to test various probabilistic models for irremovable errors and different sets of parameters of the likelihood function.

The following convention is used below to distinguish real, complex, scalar and vector variables. The italic style is used for structure-factor magnitudes ($F^{\rm obs}$, $F^{\rm calc}$ *etc.*) and real variables and parameters, while complex values of structure factors are shown in bold ($\mathbf{F}^{\rm calc}$, $\mathbf{F}^{\rm part}$ *etc.*). The bold style is used also for three-dimensional vectors of atomic coordinates (**r**, **u** *etc.*) or Miller indexes (**s**) and six-dimensional vectors of rigid-body parameters ($\boldsymbol{\Theta}$). Braces {...} are used to denote a set of

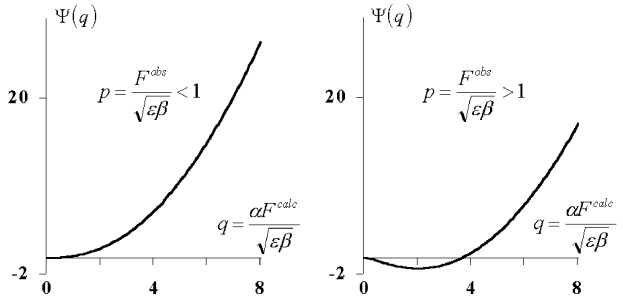

Figure 1
The behaviour of the target function Ψ in the residual (3)–(4) for relatively weak ($p = 0.7$, acentric, left) and relatively strong ($p = 2$, acentric, right) observed magnitudes. The modified observed magnitude p is defined as $p = F^{\rm obs}/\varepsilon\beta$. The dependence on the modified calculated magnitude $q = \alpha F^{\rm calc}/\varepsilon\beta$ is shown.

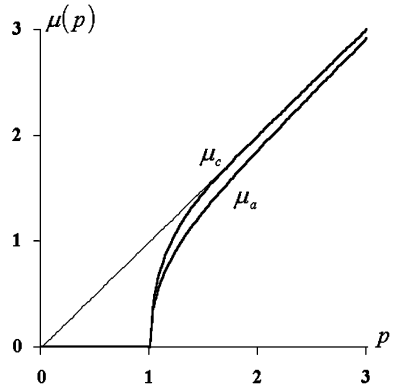

Figure 2
The 'attenuating' function (8)–(9) for centric (μ_c) and acentric (μ_a) reflections.

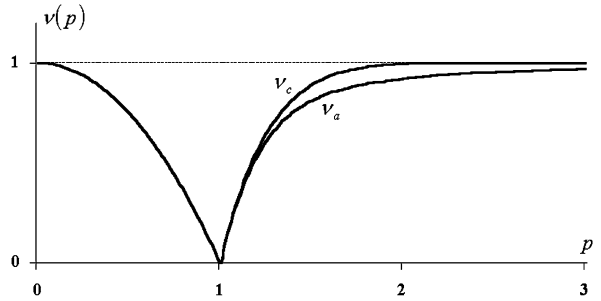

Figure 3
The 'weighting' function (10)–(11) for centric (ν_c) and acentric (ν_a) reflections.

values, *e.g.* $\{F_{\mathbf{s}}^{\rm obs}\}$ means the set of all observed magnitudes, $\{\mathbf{r}_j\}$ means the set of coordinates of all atoms included in the model *etc*. Angular brackets $\langle\ldots\rangle$ are used for mean (expected) values of random variables.

2. Irremovable errors in the modelling and maximal-likelihood principle

There exist several reasons why the structure-factor magnitudes calculated from an atomic model differ from the observed values. The first of them is the presence of errors in the current values of variable parameters of the model; the final goal of the refinement is to remove these errors, *i.e.* to find the exact values for these parameters. In what follows, such errors are called removable. Another, quite different, type of error is that caused by imperfect composition of the model or imperfect algorithms used to calculate structure-factor magnitudes starting from the variable parameters. The simplest example is the presence in the crystal of atoms whose contribution to the diffraction is not included in the calculated magnitudes. Another example is the difference between the true atomic structure and a search model used for the molecular replacement or for the rigid-body refinement. In the latter case, structure-factor magnitudes calculated from this model are different from the experimental values, even when the optimal model parameters are chosen. Changing the variable parameters of the model cannot eliminate such errors; thus we will refer to them as irremovable errors. One more source of discrepancy between the calculated and observed magnitudes is experimental errors. They too may be considered as irremovable errors.

While optimization of parameters is widely discussed in the literature, the second type of error, namely irremovable errors, demands special study. To make our analysis more clear, in the following sections we study independently several idealized situations where only one source of irremovable errors is present at a time. More general questions are discussed in §2.4. An attempt to combine different kinds of information may present extra difficulties and additional approximations are necessary (Pannu & Read, 1996; Read, 2001).

2.1. Free-atom refinement of a partial model

Let us consider a situation where an approximate atomic model (M atoms) for a part of the structure is available while the positions of the rest of the atoms ($N-M$ atoms) are unknown. We denote corresponding coordinates $\{\mathbf{r}_j\}_{j=1}^{M}$ and $\{\mathbf{u}_k\}_{k=M+1}^{N}$. If the observed magnitudes do not contain errors and the complex values $\mathbf{F}_{\mathbf{s}}^{\rm part}(\{\mathbf{r}_j\})$ and $\mathbf{F}_{\mathbf{s}}^{\rm lost}(\{\mathbf{u}_j\})$ represent the partial structure factors calculated separately for the atoms included in the current model and for those that are lost, respectively, then

$$F_{\mathbf{s}}^{\rm obs} = \left|\mathbf{F}_{\mathbf{s}}^{\rm part}(\{\mathbf{r}_j^{\rm true}\}) + \mathbf{F}_{\mathbf{s}}^{\rm lost}(\{\mathbf{u}_k^{\rm true}\})\right|, \tag{13}$$

where $\mathbf{r}_j^{\rm true}$ and $\mathbf{u}_k^{\rm true}$ are the exact values of the atomic coordinates. As a consequence, in the general case,

$$F_{\mathbf{s}}^{\rm obs} \neq \left|\mathbf{F}_{\mathbf{s}}^{\rm part}(\{\mathbf{r}_j^{\rm true}\})\right| \tag{14}$$

and fitting $F_{\mathbf{s}}^{\rm part}$ to $F_{\mathbf{s}}^{\rm obs}$, as suggested by the LSQ criterion (1), may move coordinates $\{\mathbf{r}_j\}_{j=1}^{M}$ away from their exact values. This shows that the comparison of calculated and observed magnitudes is justified only when the contribution $\mathbf{F}_{\mathbf{s}}^{\rm lost}$ of the lost atoms to the structure factor is small enough or if $\mathbf{F}_{\mathbf{s}}^{\rm lost}$ is taken somehow into account. Probabilistic modelling allows one to introduce such a correction.

Let us suppose that the atomic coordinates $\{\mathbf{u}_k\}_{k=M+1}^{N}$ for the lost atoms are chosen randomly (*e.g.* uniformly in the unit cell) and the corresponding partial structure factors $\{\mathbf{F}_{\mathbf{s}}^{\rm lost}\}$ are calculated and added to those for the fixed partial model. The combined magnitudes are now defined as

$$F_{\mathbf{s}}^{\rm comb} = \left|\mathbf{F}_{\mathbf{s}}^{\rm part}(\{\mathbf{r}_j\}) + \mathbf{F}_{\mathbf{s}}^{\rm lost}(\{\mathbf{u}_k\})\right|. \tag{15}$$

Calculated values $\{F_{\mathbf{s}}^{\rm comb}\}$ are different for different choices of the random coordinates $\{\mathbf{u}_k\}_{k=M+1}^{N}$ and in general do not coincide with the observed values. Nevertheless, the question can be posed 'how large is the probability that the magnitudes calculated in (15) will occasionally coincide with $\{F_{\mathbf{s}}^{\rm obs}\}$' or, more appropriately, 'will be close enough to these values'? This probability depends on the fixed partial-model coordinates $\{\mathbf{r}_j\}_{j=1}^{M}$. If these coordinates are exact, then there exists, at least theoretically, a chance that randomly chosen $\{\mathbf{u}_k\}_{k=M+1}^{N}$ values will be close to $\{\mathbf{u}_k^{\rm true}\}_{k=M+1}^{N}$ so that $\{F_{\mathbf{s}}^{\rm comb}\}$ values will be close to the observed magnitudes. On the contrary, if the coordinates $\{\mathbf{r}_j\}_{j=1}^{M}$ are completely incorrect, such a correction of structure factors by (15) may be impossible. The value L of this probability may distinguish poor partial models from the correct one. Going further, the partial model that maximizes $L(\{\mathbf{r}_j\})$ can be searched. The model that maximizes the chance of correct completion by randomly adding the lost atoms may be considered as a new goal of the refinement.

More formally, for every trial partial model $\{\mathbf{r}_j\}_{j=1}^{M}$, we consider the coordinates of the lost atoms as primary random variables and define new random variables $\{F_{\mathbf{s}}^{\rm comb}\}$ *via* (15). For these new variables, we consider their joint probability distribution $P^{\rm comb}(\{F_{\mathbf{s}}^{\rm comb}\};\{\mathbf{r}_j\})$ and define a measure of the quality of the partial model as the value of this function calculated with $F_{\mathbf{s}}^{\rm comb} = F_{\mathbf{s}}^{\rm obs}$, *i.e.* as the probability to obtain the observed magnitudes

$$L(\{\mathbf{r}_j\}) = P^{\rm comb}(\{F_{\mathbf{s}}^{\rm obs}\};\{\mathbf{r}_j\}). \tag{16}$$

In the mathematical statistics, the value $L(\{\mathbf{r}_j\})$ is called the likelihood and the search for the partial model that maximizes the likelihood is nothing but the widely used maximal-likelihood approach to the estimation of parameters $\{\mathbf{r}_j\}$ of the probability distribution $P^{\rm comb}(\{F_{\mathbf{s}}\};\{\mathbf{r}_j\})$.

Various probabilistic models for a distribution of the lost atoms in the unit cell may be considered, and they would lead to different likelihood functions. The realization of the proposed approach depends on the possibility to calculate the value of the likelihood (16) for any trial partial model (see Appendix *A*). If the hypothesis on uniform distribution of the lost atoms in the unit cell is used and the observed data are reduced to the absolute scale, then the maximization of the

likelihood (16) may be replaced by the minimization of the residual (3)–(4) with $\alpha_{\mathbf{s}}$, $\beta_{\mathbf{s}}$ parameters calculated as

$$\alpha_{\mathbf{s}} = 1 \quad \text{and} \quad \beta_{\mathbf{s}} = \sum_{k=M+1}^{N} f_k^2(s). \tag{17}$$

Here $f_k(s)$ are scattering factors of the lost atoms.

2.2. Rigid-body refinement of a full model

Rigid-body refinement is an essential part of the molecular replacement method where an atomic model of a homologous structure (the search model) properly placed in the unit cell is used to calculate approximate values of the structure-factor phases. The search model may be incomplete and imperfect, *i.e.* it may differ from the corresponding part of the model of the macromolecule under study. Likelihood-based residuals allow one to take this into account and thus extend the possibilities of the refinement (Read, 2001). For simplicity, in this section we consider the case of a complete but imperfect search model and suppose that the observed magnitudes do not contain errors [for a more general analysis, see Read (2001)]. The search model is moved as a rigid body, varying its rotation and translation parameters $\mathbf{\Theta}$. If the search model is imperfect, then for any choice of $\mathbf{\Theta}$ its atoms cannot fit precisely together all the atomic positions of the studied structure. This means that the calculated magnitudes $F_{\mathbf{s}}^{\rm calc}(\mathbf{\Theta})$ do not coincide with the observed ones, even for the optimal rotation and translation parameter values.

The coordinate errors remaining in the optimally placed model are irremovable in the frame of the rigid-body refinement, *i.e.* they cannot be reduced to zero by any choice of the rigid-body parameters. Nevertheless, these errors may become removable at the next stage of the structural study when all atoms are allowed to change their positions independently of others (possibly being restrained by some conditions).

Similarly to the case studied above, a probabilistic model can be used to replace unavailable information about differences in atomic positions in the optimally placed search model and the structure under study. Let $\mathbf{\Theta}$ represent current values of the rotation and translation parameters and $\{\mathbf{r}_j^{\rm search}(\mathbf{\Theta})\}_{j=1}^N$ be the atomic coordinates of the search model, rotated and translated correspondingly. As previously, the question can be posed as to how large is the probability that the calculated magnitudes $F_{\mathbf{s}}^{\rm calc}(\{\mathbf{r}_j^{\rm search}(\mathbf{\Theta}) + \Delta\mathbf{r}_j\})$ are equal to the observed ones after random independent corrections $\{\Delta\mathbf{r}_j\}_{j=1}^N$ have been introduced into the search model coordinates. Here the maximal-likelihood choice of the parameters $\mathbf{\Theta}^{\rm opt}$ means the search for such a model position and orientation that maximize this probability:

$$L(\mathbf{\Theta}) = P^{\rm search}(\{F_{\mathbf{s}}^{\rm obs}\}; \mathbf{\Theta}) \Rightarrow \max, \tag{18}$$

where $P^{\rm search}(\{F_{\mathbf{s}}\}; \mathbf{\Theta})$ represents the joint probability distribution of random variables $F_{\mathbf{s}}^{\rm calc}(\{\mathbf{r}_j^{\rm search}(\mathbf{\Theta}) + \Delta\mathbf{r}_j\})$ defined through the primary random variables $\{\Delta\mathbf{r}_j\}_{j=1}^N$.

Similarly to the previous section, the likelihood function depends on the probabilistic model of distribution of errors in the search model. For the case of independent errors possessing an isotropic Gaussian distribution, the maximization of (18) may be reduced to minimization of the function (3)–(4) with the parameters $\alpha_{\mathbf{s}}$ and $\beta_{\mathbf{s}}$ defined as

$$\begin{aligned} \alpha_{\mathbf{s}} &= \langle \cos 2\pi(\mathbf{s}, \Delta\mathbf{r}_j) \rangle = \exp(-\pi^3\omega^2 s^2/4), \\ \beta_{\mathbf{s}} &= (1 - \alpha_{\mathbf{s}}^2) \sum_{j=1}^{N} f_j^2(s). \end{aligned} \tag{19}$$

Here $f_j(s)$ are the scattering factors of atoms of the search model and ω represents the expected mean error in the position of these atoms.

2.3. Errors in the observed magnitudes

One more possible source of irremovable discrepancies between the calculated and the observed magnitudes is experimental errors in the observed magnitudes. In this situation, the calculated magnitudes can be corrected by some random values $\Delta F_{\mathbf{s}}$ in order to simulate the experimental errors:

$$F_{\mathbf{s}}^{\rm corr} = F_{\mathbf{s}}^{\rm calc}(\{\mathbf{r}_j\}) + \Delta F_{\mathbf{s}}. \tag{20}$$

Similarly to the previous sections, for the given model parameters $\{\mathbf{r}_j\}_{j=1}^N$, the question can be posed as to how large is the probability $P^{\rm corr}(\{F_{\mathbf{s}}^{\rm corr}\})$ to make the calculated magnitudes $F_{\mathbf{s}}^{\rm calc}(\{\mathbf{r}_j\})$ equal to the observed ones by these random corrections $\Delta F_{\mathbf{s}}$. The maximal-likelihood choice in this case means the search for the model parameters that maximize

$$L(\{\mathbf{r}_j\}) = P^{\rm corr}(\{F_{\mathbf{s}}^{\rm obs}\}) \Rightarrow \max. \tag{21}$$

Obviously, there is little sense in such a general formulation until it is determined which random corrections of magnitudes may be considered as reasonable or, in other words, until some precise probabilistic model for the experimental errors has been introduced. To define these terms, it is necessary to know something about the accuracy of the data collection, *i.e.* to introduce new information into the problem. If it is known (*e.g.* from multiple measuring of the same reflection or equivalent reflections) that the experimental errors present in the observed magnitudes may be considered as independent ones, distributed in accordance with the Gaussian distribution with mean zero and standard deviations $\sigma_{\mathbf{s}}$, then the likelihood function (21) may be written as

$$L(\{\mathbf{r}_j\}) = \prod_{\mathbf{s}} \frac{1}{(2\pi)^{1/2}\sigma_{\mathbf{s}}} \exp\left\{ -\frac{[F_{\mathbf{s}}^{\rm calc}(\{\mathbf{r}_j\}) - F_{\mathbf{s}}^{\rm obs}]^2}{2\sigma_{\mathbf{s}}^2} \right\}, \tag{22}$$

so that the maximization in (21) may be replaced by the minimization

$$\begin{aligned} -\ln L(\{\mathbf{r}_j\}) = & -\sum_{\mathbf{s}} \ln\left[\frac{1}{(2\pi)^{1/2}\sigma_{\mathbf{s}}} \right] \\ & + \tfrac{1}{2} \sum_{\mathbf{s}} \frac{1}{\sigma_{\mathbf{s}}^2} [F_{\mathbf{s}}^{\rm calc}(\{\mathbf{r}_j\}) - F_{\mathbf{s}}^{\rm obs}]^2 \\ & \Rightarrow \min. \end{aligned} \tag{23}$$

The variable part of (23) is nothing but the conventional target (1) and therefore the conventional crystallographic refinement may also be considered as a ML refinement. This is not

surprising because of the profound links between the likelihood and least-squares methods in mathematical statistics.

2.4. Statistical refinement

The approach suggested above may be generalized as a concept of statistical refinement.

The conventional structure refinement may be described as follows. There exists a formula or a computer algorithm that enables a set of structure-factor magnitude to be calculated starting from a set of atomic parameters (coordinates, temperature factors, occupancies *etc.*). In other words, every set of model parameters is associated with a set of corresponding calculated magnitudes. If a set of experimentally obtained magnitudes is available, the goal of the conventional refinement is formulated as:

To choose the set of atomic parameters for which the corresponding calculated magnitudes are the most consistent with the experimental data.

Different measures may be used to evaluate this consistency numerically, *e.g.* (1) or (2), and they may lead, formally speaking, to different results.

If the formula (algorithm) connecting the model parameters with the structure-factor magnitudes is imperfect and does not allow one to reproduce the experimental data precisely, even for the exact model parameters, then a statistical model for the necessary corrections of the formula may be used. In this way, every set of model parameters is associated with a joint probability distribution of structure-factor magnitudes, rather than with a single set of calculated magnitudes. If it is assumed that the set of observed magnitudes is known, the goal of the new statistical refinement may be formulated as:

To choose the set of atomic parameters for which the corresponding joint probability distribution of magnitudes is the most consistent with the experimental data.

Similar to the LSQ case, different ways to evaluate the consistency numerically may be used and they are the subjects of the mathematical statistics. One of the possible ways is to use the likelihood value as this measure of the consistency, as discussed above. Alternatively, one can search for the probability distribution (*i.e.* for corresponding model parameters) for which the expected values of the structure-factor magnitudes are as close as possible to the observed ones (Adams *et al.*, 1997). Such an approach is close to 'the method of moments' in mathematical statistics. Naturally, other statistical approaches may be tried as well.

It must be emphasized that a new important object appears in the statistical refinement besides the current model parameters and the set of experimental data, namely a probabilistic model for the source of the imperfection in structure-factor formulae. The choice of the probabilistic model for irremovable errors plays the key role. This probabilistic model introduces additional information into the process of refinement and the success of the refinement depends strongly on the correctness of this information. For example, in the case considered in §2.1, we might specify the hypothesis regarding the distribution of the lost atoms. The simplest way is to suppose that these atoms are distributed uniformly in the unit cell. At the first stages of a structure investigation, when side-chain atoms are not included in the model, such a hypothesis seems to be reasonable. However, later, when the bulk solvent atoms only are absent, more detailed hypotheses might be needed. Obviously, different probability hypotheses result in different likelihood functions and in different refined models. Similarly, in the case considered in §2.2, the hypothesis about the distribution of errors in the search model must be specified. Sometimes, extra information may be used for these purposes (Read, 2001). Naturally, such probabilistic information is much weaker than the deterministic information introduced by the conventional extension of the model. Nevertheless, even such weak information may improve the results of the refinement.

After some probabilistic models for irremovable errors have been chosen, the corresponding likelihood must be derived as a function depending on the variable model parameters. A simplification generally used is to neglect the correlation of structure factors and to consider calculated structure factors as independent random variables. In this case, the joint probability distribution may be written as a product of individual distributions and the logarithm of the likelihood becomes a sum of logarithms of these distributions (see Appendix *A*).

3. Local structure of the ML target function

In this section, a quadratic approximation for the function (4) is derived in the case of a centric structure factor. This highlights new tendencies that appear in ML-based refinement in comparison with the standard LSQ refinement. The formulae for acentric reflections are very similar to those corresponding to the centric case; they are derived in Appendix *B*.

3.1. Quadratic approximation of the residual

If dimensionless variables are introduced as

$$x = \frac{\alpha_{\mathbf{s}} F_{\mathbf{s}}^{\rm calc}}{(\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}})^{1/2}} \quad \text{and} \quad p = \frac{F_{\mathbf{s}}^{\rm obs}}{(\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}})^{1/2}}, \tag{24}$$

then the centric term (4) in the residual (3) becomes

$$\Psi = \psi(x; p) = \tfrac{1}{2}x^2 - \ln[\cosh(px)]. \tag{25}$$

Asymptotic expansions for the logarithmic function and hyperbolic cosine lead to

$$\psi(x; p) \simeq \tfrac{1}{2}(1 - p^2)x^2 \quad \text{for small } x \tag{26}$$

$$\psi(x; p) \simeq \tfrac{1}{2}x^2 \quad \text{for large } x. \tag{27}$$

This explains the behaviour of the function Ψ in Fig. 1 and confirms the key role of the parameter p. It follows from (26) that for $p^2 < 1$ the function $\psi(x; p)$ grows starting from $x = 0$ and reaches its minimal value at $x = 0$. For $p^2 > 1$, this function first decreases when x grows from zero, reaches the minimal value at $x = x^* > 0$ and then increases infinitely.

In the case $p^2 < 1$, the quadratic approximation for the function $\psi(x; p)$ in the vicinity of the point of the minimum is given by (26). For $p^2 > 1$, the quadratic approximation may be written as

$$\psi(x; p) \simeq \psi(x^*; p) + \tfrac{1}{2}\psi''(x^*; p)(x - x^*)^2, \tag{28}$$

where x^* represents the point of the minimum of $\psi(x; p)$. This point may be found as the solution of the equation

$$\psi'(x^*; p) \equiv x^* - p\tanh(px^*) = 0, \tag{29}$$

with the additional condition $\psi''(x^*; p) > 0$.

Equation (29) has the trivial solution $x = 0$ for any value of the parameter p. For $p^2 < 1$, this solution is unique and $\psi''(x^*; p) = \frac{1}{2}(1 - p^2) > 0$, so that the conditions of the minimum are satisfied. For $p^2 > 1$, two more solutions of (29) appear, one negative and one positive. The solution $x = 0$ corresponds now to the local maximum. The positive solution, which we denote as $x^* = \mu(p)$, corresponds to the point of the minimum of $\psi(x; p)$. Some methods of practical calculation of $\mu(p)$ are discussed in Appendix *B*. The negative solution corresponds to another local minimum of $\psi(x; p)$ (see Fig. 4) with a negative value of x, *i.e.* with a physically unreasonable value of the structure-factor magnitude.

The curvature in (28) is

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x; p) = 1 - p^2 + p^2\tanh^2(px). \tag{30}$$

At the point of the minimum $x^* = \mu(p)$, (29) is satisfied and

$$\psi''(x^*; p) = 1 - p^2 + \mu^2(p). \tag{31}$$

If $\nu(p)$ is introduced by (11), $c = \frac{1}{2}$ and the function $\mu(p)$ is defined to be equal to zero for $p^2 \leq 1$, then the approximation (28) may be written as

$$\psi(x; p) \simeq \psi(x^*; p) + c\nu(p)[x - \mu(p)]^2. \tag{32}$$

The term $\psi(x^*; p)$ does not depend on x and may be removed from the residual. Coming back *via* (24) to the values $F_{\mathbf{s}}^{\mathrm{calc}}$ and $F_{\mathbf{s}}^{\mathrm{obs}}$, we obtain the residual in the form (6)–(7).

3.1.1. Relative form of the residual. Let the relative magnitudes be defined as

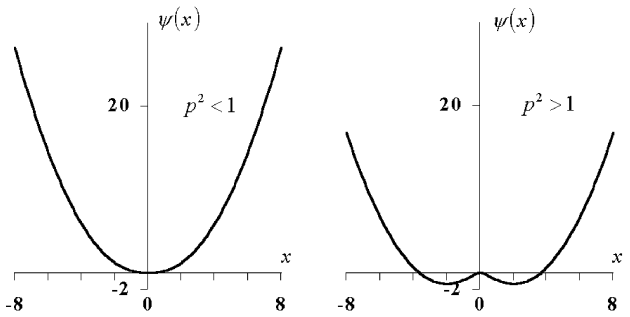

Figure 4
The function (25) for different values of the parameter p: $p = 0.7$ (left) and $p = 2$ (right).

$$\widetilde{E}_{\mathbf{s}}^{\mathrm{obs}} = \frac{F_{\mathbf{s}}^{\mathrm{obs}}}{(\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}})^{1/2}} \quad \text{and} \quad \widetilde{E}_{\mathbf{s}}^{\mathrm{calc}} = \frac{F_{\mathbf{s}}^{\mathrm{calc}}}{(\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}})^{1/2}}. \tag{33}$$

It follows from (6) and (7) that the quadratic approximation of the ML residual (3) is equivalent to

$$\widetilde{Q}_{ML} = \sum_{\mathbf{s}\in S} w_{\mathbf{s}}^*(\alpha_{\mathbf{s}}\widetilde{E}_{\mathbf{s}}^{\mathrm{calc}} - \widetilde{E}_{\mathbf{s}}^*)^2 \tag{34}$$

with

$$\widetilde{E}_{\mathbf{s}}^* = \mu(\widetilde{E}_{\mathbf{s}}^{\mathrm{obs}}) \quad \text{and} \quad w_{\mathbf{s}}^* = c_{\mathbf{s}}\nu(\widetilde{E}_{\mathbf{s}}^{\mathrm{obs}}). \tag{35}$$

This representation highlights new tendencies which appear in ML refinement in comparison with the conventional minimization of (1).

3.2. Tendencies of ML refinement

3.2.1. Normalized magnitudes. The first consequence of the representation (34) is that the ML principle suggests refinement in terms of normalized structure-factor magnitude. Indeed, in the examples considered in §2, the modification (33) differs from the standard procedure of structure-factor normalization essentially by a factor depending on the number of atoms included in the calculation of the denominators. In a particular case when all the atoms have similar scattering factors, (33) can be rewritten as

$$\begin{aligned}\widetilde{E}_{\mathbf{s}}^{\mathrm{obs}} &= \left[\frac{N}{(1-\alpha_{\mathbf{s}}^2)M + (N-M)}\right]^{1/2} E_{\mathbf{s}}^{\mathrm{obs}},\\ \widetilde{E}_{\mathbf{s}}^{\mathrm{calc}} &= \left[\frac{N}{(1-\alpha_{\mathbf{s}}^2)M + (N-M)}\right]^{1/2} E_{\mathbf{s}}^{\mathrm{calc}},\end{aligned} \tag{36}$$

where $E_{\mathbf{s}}^{\mathrm{obs}}$ and $E_{\mathbf{s}}^{\mathrm{calc}}$ are the normalized magnitudes, N is the total number of atoms, M is the number of atoms included in the partial model and $\alpha_{\mathbf{s}}$ reflects the level of irremovable coordinate errors in accordance with (18).

3.2.2. Attenuation of target values. Function $\mu(p)$ defines modified target values $\widetilde{E}_{\mathbf{s}}^*$ for calculated magnitudes. Its main feature is that it assigns zero values to $\widetilde{E}_{\mathbf{s}}^*$ if the observed magnitude $E_{\mathbf{s}}^{\mathrm{obs}}$ is relatively weak:

$$\widetilde{E}_{\mathbf{s}}^* = 0 \quad \text{if} \quad \widetilde{E}_{\mathbf{s}}^{\mathrm{obs}} \leq 1. \tag{37}$$

This means that in the process of ML refinement the calculated values of structure factors are fitted to zero but not to the corresponding $F_{\mathbf{s}}^{\mathrm{obs}}$. It must be emphasized that the cut-off level 1 in (37) is applied to relative magnitudes modified in accordance with (33) and not to the normalized ones.

When some atoms are not included in the model (§2.1), the value of $\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}$ is the mean intensity corresponding to the lost part of the structure and the condition in (37) means that $I_{\mathbf{s}}^{\mathrm{obs}} \leq \langle I_{\mathbf{s}}^{\mathrm{lost}}\rangle$. The ML criterion suggests that such observed intensities be considered as those corresponding completely to the lost atoms and fits the contribution of the partial model to zero.

For intermediate values of relative magnitudes ($1 < \widetilde{E}_{\mathbf{s}}^{\mathrm{obs}} < 1.5$), the modified values $\widetilde{E}_{\mathbf{s}}^* = \mu(\widetilde{E}_{\mathbf{s}}^{\mathrm{obs}})$ are less than $\widetilde{E}_{\mathbf{s}}^{\mathrm{obs}}$ and hence some attenuation of the target values

occurs. For relatively large magnitudes, the values $\widetilde{E}_{\mathbf{s}}^{*}$ are close to the observed ones $\widetilde{E}_{\mathbf{s}}^{\rm obs}$, especially for the centric reflections.

3.2.3. Enhancing of target values. For large p, we have $\mu(p) \simeq p$, so that

$$\widetilde{E}_{\mathbf{s}}^{*} \simeq \widetilde{E}_{\mathbf{s}}^{\rm obs} \quad \text{if} \quad (F_{\mathbf{s}}^{\rm obs})^2 \gg \varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}. \tag{38}$$

In other words, for strong reflections, $F_s^{\rm calc}$ is fitted to $F_{\mathbf{s}}^{\rm obs}$ in the case of an incomplete model (§2.1) and to $\alpha_{\mathbf{s}}^{-1}F_{\mathbf{s}}^{\rm obs}$ in the case of irremovable coordinate errors in the search model (§2.2). The appearance of the factor $\alpha_{\mathbf{s}}^{-1}$ may be explained as follows. If independent random shifts $\Delta\mathbf{r}_j$ are introduced in the model coordinates, then the structure-factor magnitudes for the modified model will be in mean less than the magnitudes for the starting model:

$$\langle F_{\mathbf{s}}^{\rm calc}(\{\mathbf{r}_j + \Delta\mathbf{r}_j\})\rangle_{\Delta\mathbf{r}} = \alpha_{\mathbf{s}}F_{\mathbf{s}}^{\rm calc}(\{\mathbf{r}_j\}) < F_{\mathbf{s}}^{\rm calc}(\{\mathbf{r}_j\}), \tag{39}$$

where

$$\alpha(\mathbf{s}) = \langle \cos 2\pi(\mathbf{s}, \Delta\mathbf{r}_j)\rangle \leq 1. \tag{40}$$

As a consequence, if we want the calculated and observed magnitudes to be close to each other *after* corrections of the search model, then *before* this correction the calculated magnitudes must be slightly larger than is supposed by $F_{\mathbf{s}}^{\rm obs}$, so that

$$\alpha_{\mathbf{s}}F_{\mathbf{s}}^{\rm calc}[\{\mathbf{r}_j^{\rm search}(\Theta^{\rm true})\}] \simeq F_{\mathbf{s}}^{\rm obs}. \tag{41}$$

It is worthy of note that both tendencies, namely the suppression of the target $[\mu(\widetilde{E}_{\mathbf{s}}^{\rm obs})$ instead of $\widetilde{E}_{\mathbf{s}}^{\rm obs}]$ and its enhancement ($\alpha_{\mathbf{s}}\widetilde{E}_{\mathbf{s}}^{\rm calc}$ instead of $\widetilde{E}_{\mathbf{s}}^{\rm calc}$), are present simultaneously. Thus, depending on circumstances, the target for the calculated magnitude may be either less or larger than the corresponding observed magnitude.

3.2.4. Removing reflections from the refinement. The weighting function $\nu(\widetilde{E}_{\mathbf{s}}^{\rm obs})$ (Fig. 3) shows that the ML residual suggests that reflections with $\widetilde{E}_{\mathbf{s}}^{\rm obs} \simeq 1$ should be down weighted (and thus removed from the refinement), *i.e.* the reflections whose observed intensities are close to mean intensities corresponding to the atoms excluded from the refinement.

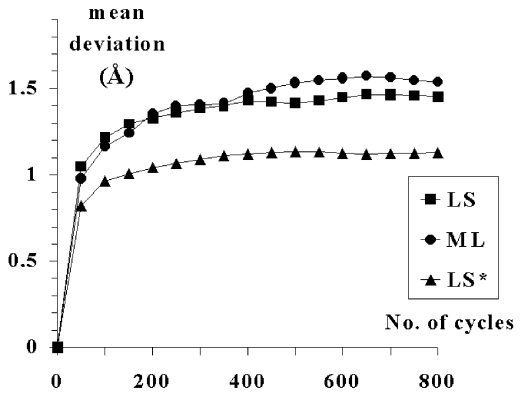

Figure 5
The mean deviation of atomic centres from their true positions in the course of the refinement starting from a 50% complete exact partial model: LS, standard least-squares (1); ML, the criterion marked as ML criterion in the *CNS* program; LS*, modified criterion (6).

4. Numerical tests

The refinement tests were carried out with *CNS* complex (Brünger *et al.*, 1998) using the structure of the Fab fragment of the monoclonal antibody (Fokine *et al.*, 2000). The full model included 439 amino acid residues and 213 water molecules. The crystals corresponded to the space group $P2_12_12_1$ with the unit-cell parameters $a = 72.24$, $b = 72.01$, $c = 86.99$ Å and one Fab molecule per asymmetric unit. To exclude experimental errors from the analysis, in these tests the observed data were simulated by the corresponding values calculated from the complete exact model. As a consequence, the standard σ weighting of the LSQ residual was impossible; corresponding tests with the experimental data will be discussed separately. All refinements were performed in the $d_{\rm min} > 2.2$ Å resolution zone.

The goal of the tests was to study how far the minimization of different crystallographic residuals shifts the atoms of the partial model from their true positions in order to compensate the contribution of the excluded atoms. In all tests, the exact atomic coordinates were used as the starting values. To determine the effect clearly, in all tests all additional restraints, such as stereochemical ones, were excluded and X-ray criteria alone were used for the refinement. In every test about 800 cycles of minimization were used to ensure the convergence. Fig. 5 shows the typical behaviour of the mean coordinate error in the course of the minimization. To check the stability of the perfect model in the course of the refinement without stereochemical restraints, an additional test was performed in which 800 cycles of refinement were performed starting from the complete and exact model. This 'refinement' did not introduce essential errors in the model coordinates, so the

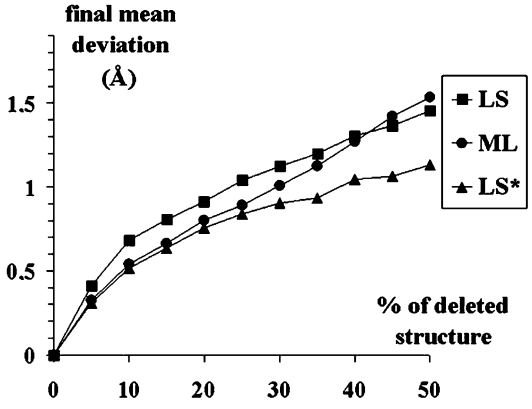

Figure 6
The mean deviation of atomic centres from their true positions after 800 cycles of free-atom refinement is shown as a function of the completeness of the starting model. The deleted atoms were chosen randomly from the whole structure (Fab plus water). The exact coordinates of the rest of the structure were used as the starting values. Different types of residual are compared: LS, standard least-squares (1); ML, the criterion marked as ML criterion in the *CNS* program; LS*, modified criterion (6).

errors that appeared in other tests may be attributed to the incompleteness of the starting model rather than to the instability of the refinement caused by the restraints being switched off.

Three main types of residuals were tested: (i) the least-squares residual (1) with unit weights ('standard crystallographic residual', as defined in *CNS* if the normalizing scale factor is neglected; LS in the figures below); (ii) the criterion called 'ML-target using amplitudes' in *CNS* [ML in the figures below; see Adams *et al.* (1997) for a definition]; (iii) the quadratic approximation (6) of the likelihood-based residual (3)–(4) (LS* in the figures below).

The tests were performed for partial models containing different numbers of atoms. Two types of partial models were generated. Partial models of the first type were obtained by the random deletion of the desired number of atoms from the whole structure (Fab plus water). Every atom might be removed from the model with an equal probability. The models of the second type were obtained by a random deletion of the water atoms only. Figs. 6 and 7 present the results of the minimization of the starting models of different completeness. These results show that the incompleteness of the model can seriously affect the refinement. The more atoms are deleted, the larger are errors in the model that best fits the experimental data. Removal of water molecules has a stronger effect than removal of a similar quantity of atoms randomly in the unit cell. The reason for this may be the following. The water molecules are situated at the surface of the protein but not in the volume. When the same number of atoms are randomly excluded in both tests, in the case of water molecules they are distributed less uniformly in the space, leading to a stronger influence on the structure factors.

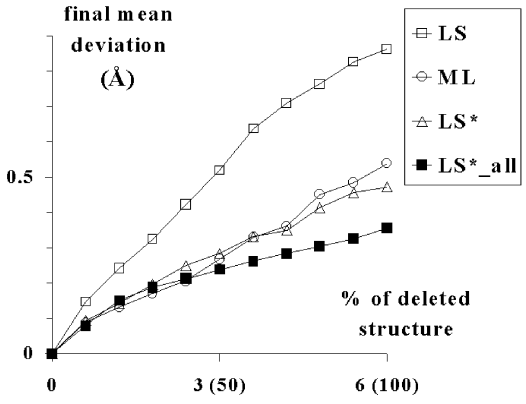

Figure 7
The mean deviation of atomic centres from their true positions after 800 cycles of free-atom refinement is shown as a function of the completeness of the starting model. The deleted atoms were chosen randomly from the water oxygen atoms only. The exact coordinates of the rest of the structure were used as the starting values. The figures in brackets indicate the percentage of deleted atoms with respect to all water oxygen atoms. Different types of residual are compared: LS, standard least-squares (1); ML, the criterion marked as ML criterion in the *CNS* program; LS*, modified criterion (6). A part of the LS* curve from Fig. 6 is shown (LS*_all) for comparison with the case when atoms are deleted from the whole structure and not from water molecules only.

None of the tested residuals guards against the deterioration of an incomplete starting model, so additional restraints (*e.g.* stereochemical ones) are necessary to stabilize the structure. Nevertheless, the modified LS* criterion was found to be essentially less sensitive to the incompleteness of the model than the conventional LS criterion. When compared with the *CNS* ML criterion, it may be mentioned that the LS* criterion gave slightly better results for slightly incomplete models and gave significantly better results for very incomplete ones. Two explanations may be proposed. First, it follows from Adams *et al.* (1997) that the criterion used in *CNS* is a residual that is based on the method of moments for the estimation of the distribution parameters rather than the pure likelihood criterion (3)–(4) (see §2.4). So the tests performed may be considered as a comparative analysis of two ways to define the consistency of probability distributions with the observed data for the considered crystallographic problem. Another reason may be that the LS* criterion, while possessing the same minimum point as the likelihood criterion (3)–(4), may have better minimization properties (*e.g.* convexity), which may be essential in difficult refinement cases.

Additional tests were performed to study which features of the LS* residual are most important for the refinement: the use of the modified target magnitudes F_s^* or the weights w_s^*? For this purpose, a minimization of two mixed criteria was performed:

$$\mathrm{LS}^*1 = \sum_{s\in S} w_s^* (F_s^{\mathrm{calc}} - F_s^{\mathrm{obs}})^2 \qquad (42)$$

and

$$\mathrm{LS}^*2 = \sum_{s\in S} (F_s^{\mathrm{calc}} - F_s^*)^2. \qquad (43)$$

The results of this test (Fig. 8) demonstrate that the use of the weights w_s^* without modification of target magnitudes allows

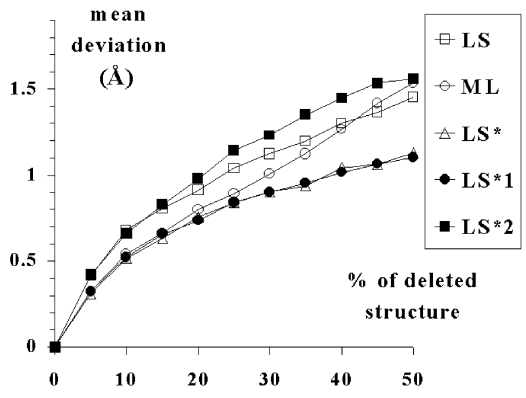

Figure 8
The mean deviation of atomic centres from their true positions after 800 cycles of free-atom refinement is shown as a function of the completeness of the starting model. The deleted atoms were chosen randomly from the whole structure (Fab plus water). The exact coordinates of the rest of the structure were used as the starting values. Different types of residual are compared: LS, standard least-squares (1); ML, the criterion marked as ML criterion in the *CNS* program; LS*, modified criterion (6); LS*1, criterion (42); LS*2, criterion (43).

one to obtain the same results as the use of the full LS* criterion. At the same time, ignoring these weights deteriorates the results significantly. This leads to the conclusion that the main merit of the LS* criterion is based on the inclusion of the proper weighting scheme in the refinement process.

5. Conclusions

Statistical modelling of irremovable errors extends the possibilities of structure refinement. This modelling engages new information concerning the object under study in a weak form of statistical hypothesis. Nevertheless, even such poor information may result in a significant improvement.

In contrast to the conventional refinement, the new strategy associates any set of variable parameters of the model with the joint probability distribution of structure-factor magnitudes rather than with particular values of these magnitudes. The choice of the distribution that is the most consistent with the experimental information substitutes the conventional search for the calculated magnitudes, which are as close as possible to the observed ones.

The analysis of the quadratic approximation of the likelihood function highlights two main tendencies of the statistical refinement that differ from the conventional least-squares refinement. Firstly, new target values appear for the calculated structure-factor magnitudes. These targets are close to the observed values for strong reflections and are different from the observed values for weak reflections. Secondly, the statistical criterion suggests new weights for contributions of different reflections. Simple tests demonstrate that these weights seem to be the most important consequence of the new refinement strategy. When used with the unit weights, the modified targets do not produce a visible improvement in comparison with conventional refinement. At the same time, the refinement with the new weights demonstrates an improvement, even when the non-modified observed magnitudes are used as target values. This is rather typical for statistical approaches: the main result of their application is the proper weights and not a drastic changing of goals.

APPENDIX *A*
From probabilistic modelling of irremovable errors to the likelihood-based residual

A1. Diagonal Gaussian approximation

The majority of likelihood-based approaches in crystallography use the diagonal Gaussian approximation of a likelihood function. This approximation is based on two assumptions. First, it is supposed that every structure factor possesses an uncorrelated two-dimensional Gaussian distribution of real and imaginary parts (Luzzati, 1952; Srinivasan & Parthasarathy, 1976). Second, it is supposed that the joint probability distribution (j.p.d.) of structure-factor magnitudes for a set of structure factors may be approximated by the product of one-dimensional distributions corresponding to individual reflections (Lunin & Urzhumtsev, 1984). The general way to obtain the residual (3)–(4) is described below for acentric reflections.

First, some probabilistic model is introduced that describes uncertainties in the structure so that the structure factors corresponding to the full content of the unit cell become random variables. Then, for any individual structure factor, the joint probability distribution of its real and imaginary parts is studied. It is supposed that this distribution is a particular case of the two-dimensional Gaussian distribution:

$$P(\mathrm{Re}\,\mathbf{F}, \mathrm{Im}\,\mathbf{F}) = (1/\pi\varepsilon\beta)\exp\{-[(\mathrm{Re}\,\mathbf{F} - \alpha F^{\mathrm{calc}}\cos\varphi^{\mathrm{calc}})^2 + (\mathrm{Im}\,\mathbf{F} - \alpha F^{\mathrm{calc}}\sin\varphi^{\mathrm{calc}})^2]/\varepsilon\beta\}. \quad (44)$$

This formula implies that the real and imaginary parts of the structure factor are uncorrelated and have equal variances $\varepsilon\beta$. It follows from (44) that the expected value of the structure factor is proportional to the value of a 'calculated structure factor':

$$\langle\mathbf{F}\rangle = \alpha\mathbf{F}^{\mathrm{calc}}. \quad (45)$$

Such a distribution, which appeared previously in a number of crystallographic applications (Luzzati, 1952; Sim, 1959; Srinivasan & Parthasarathy, 1976; Read, 1986, 1990; Lunin & Skovoroda, 1995), may be rewritten in terms of the structure-factor magnitude and phase:

$$P(F, \varphi) = \frac{F}{\pi\varepsilon\beta}\exp\left[-\frac{F^2 + (\alpha F^{\mathrm{calc}})^2}{\varepsilon\beta}\right] \times \exp\left[2\frac{\alpha F F^{\mathrm{calc}}}{\varepsilon\beta}\cos(\varphi - \varphi^{\mathrm{calc}})\right]. \quad (46)$$

If the only available experimental information is the magnitude, then the marginal distribution for the magnitude of a single structure factor is derived by integrating the $P(F, \varphi)$ over the phase:

$$P(F) = \frac{2F}{\varepsilon\beta}\exp\left[-\frac{F^2 + (\alpha F^{\mathrm{calc}})^2}{\varepsilon\beta}\right] I_0\left(2\frac{\alpha F F^{\mathrm{calc}}}{\varepsilon\beta}\right). \quad (47)$$

The next significant simplification consists of the assumption that for different reflections the corresponding magnitude values are 'almost independent'. In this case, the product of one-dimensional distributions may approximate the joint probability distribution for the set $\{F_{\mathbf{s}}\}_{\mathbf{s}\in S}$ of magnitudes

$$P\left(\{F_{\mathbf{s}}\}_{\mathbf{s}\in\mathbf{S}}\right) \simeq \prod_{\mathbf{s}\in\mathbf{S}} P(F_{\mathbf{s}}). \quad (48)$$

If the experimental values of the magnitudes are known, it is possible to calculate the likelihood value

$$L = \prod_{\mathbf{s}\in S}\frac{2F_{\mathbf{s}}^{\mathrm{obs}}}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}}\exp\left[-\frac{(F_{\mathbf{s}}^{\mathrm{obs}})^2 + (\alpha_{\mathbf{s}}F_{\mathbf{s}}^{\mathrm{calc}})^2}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}}\right] I_0\left(2\frac{\alpha_{\mathbf{s}}F_{\mathbf{s}}^{\mathrm{obs}}F_{\mathbf{s}}^{\mathrm{calc}}}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}}\right), \quad (49)$$

which reflects the probability to reproduce the experimental data $\{F_{\mathbf{s}}^{\mathrm{obs}}\}_{\mathbf{s}\in S}$ by randomly generating $\{F_{\mathbf{s}}\}_{\mathbf{s}\in S}$ values with the distribution (48).

The obtained expression for the likelihood may be used to solve different problems. First, the likelihood maximization allows one to estimate irremovable errors provided that the

calculated magnitudes are fixed (Lunin & Urzhumtsev, 1984; Read, 1986). On the other hand, when considering the calculated magnitudes $\{F_{\mathbf{s}}^{\rm calc}\}_{\mathbf{s}\in S}$ as functions of model parameters (*e.g.* atomic coordinates), the maximization of the likelihood may be used as a tool to refine the values of atomic parameters (Pannu & Read, 1996; Bricogne & Irwin, 1996; Murshudov *et al.*, 1997). In both cases, it is convenient to replace the maximization of (49) by minimization of the negative logarithm of the likelihood

$$-\ln L = -\sum_{\mathbf{s}}\left[\ln\frac{2F_{\mathbf{s}}^{\rm obs}}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}} - \frac{(F_{\mathbf{s}}^{\rm obs})^2}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}}\right] + \sum_{\mathbf{s}}\left\{\frac{(\alpha_{\mathbf{s}}F_{\mathbf{s}}^{\rm calc})^2}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}} - \ln I_0\left(2\frac{\alpha_{\mathbf{s}}F_{\mathbf{s}}^{\rm obs}F_{\mathbf{s}}^{\rm calc}}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}}\right)\right\} \Rightarrow \min. \quad (50)$$

Finally, the members that are independent of variable parameters are excluded from the criterion (50).

For centric reflections, the corresponding formulae are slightly different because (44) becomes a one-dimensional Gaussian distribution and the integration over the phase value is replaced by the summation over two possible values of the structure-factor sign. As a consequence, the formula for the likelihood takes the form

$$L = \prod_{\mathbf{s}\in S}\left(\frac{2}{\pi\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}}\right)^{1/2}\exp\left[-\frac{(F_{\mathbf{s}}^{\rm obs})^2+(\alpha_{\mathbf{s}}F_{\mathbf{s}}^{\rm calc})^2}{2\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}}\right] \times \cosh\left(\frac{\alpha_{\mathbf{s}}F_{\mathbf{s}}^{\rm obs}F_{\mathbf{s}}^{\rm calc}}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}}\right). \quad (51)$$

A2. Non-diagonal joint probability distributions

The approximation (48) does not consider the correlation of magnitudes. This correlation may be taken into account either analytically or with the use of numerical simulation procedures. An explicit representation of the saddle-point approximation for j.p.d. was used to derive the formula for the likelihood in the case of the space group *I*432 (Lunin, 1997; Petrova, Lunin, Lunina & Skovoroda, 1999; Petrova, Lunin & Podjarny, 1999). This representation was constructed on the base of the saddle-point approximation of the j.p.d. suggested by Bricogne (1984).

Monte Carlo type simulation procedures for the calculation of the non-diagonal likelihood were used for low-resolution *ab initio* phasing (Lunin *et al.*, 1998; Petrova, Lunin, Lunina & Skovoroda, 1999; Petrova, Lunin & Podjarny, 1999, 2000). In this approach, the computer-based generation of the set of random parameters is repeated many times, followed by the calculation of corresponding structure factors. The generated set is accepted if the coefficient of correlation between the calculated and observed magnitudes exceeds some specified level. The likelihood value is estimated as the ratio of the number of accepted sets to the total number of generated sets.

APPENDIX *B*
Quadratic approximation for the likelihood-based residual

It was shown in §3 that the study of the likelihood-based residual (3)–(4) could be reduced to the study of the function

$$\psi(x;p) = \begin{cases} x^2 - \ln[I_0(2px)] & \text{for acentric reflections,} \\ \frac{1}{2}x^2 - \ln[\cosh(px)] & \text{for centric reflections.} \end{cases} \quad (52)$$

Formally speaking, the function (52) is defined for all values of the variable x and the parameter p, but as a result of symmetry properties,

$$\psi(-x;p) = \psi(x;p) \quad \text{and} \quad \psi(x;-p) = \psi(x;p), \quad (53)$$

its study may be restricted to the region of physically reasonable non-negative values of x and p. Two different formulae in (52) force one to study the acentric and centric cases separately, and we start from the latter, for which a rigorous mathematical analysis is easier.

B1. Centric reflections

For centric reflections,

$$\psi(x;p) = \tfrac{1}{2}x^2 - \ln[\cosh(px)]. \quad (54)$$

While at the origin and at infinity this function behaves similarly for all values of p, *i.e.*

$$\psi(0;p) = 0 \quad \text{and} \quad \psi(x;p) \to \infty \quad \text{when} \quad x \to \infty, \quad (55)$$

for intermediate values of the argument x, its behaviour strongly depends on the parameter p. The analysis of the derivatives

$$\psi'(x;p) = x - p\tanh(px) \quad (56)$$

and

$$\psi''(x;p) = 1 - \frac{p^2}{\cosh^2(px)} = 1 - p^2[1-\tanh^2(px)] \quad (57)$$

helps to study the function. The second derivative grows monotonically from $1-p^2$ to 1 when x varies from 0 to ∞. If $p^2 < 1$, then $\psi''(x;p)$ is everywhere positive so that the first derivative $\psi'(x;p)$ grows monotonically from 0 to ∞ and thus is always non-negative. This means that $\psi(x;p)$ also grows monotonically from 0 to ∞ when x varies from 0 to ∞ and there exists a single minimum of $\psi(x;p)$ which is attained for $x = 0$ (Fig. 4, left).

Quite differently, for $p^2 > 1$, the second derivative $\psi''(x;p)$ changes its sign once from minus to plus, and thus $\psi'(x;p)$ first decreases from zero to some negative value and then grows to infinity, passing once through the zero value. This means in turn that, for $p^2 > 1$, the function $\psi(x;p)$ first decreases from zero to some negative value and then grows monotonically to infinity. As a result of the relations (53), this means that for $p^2 > 1$ the function $\psi(x;p)$ has a local maximum at $x = 0$ and two symmetric local minima (Fig. 4, right). The coordinates of these points may be found as a solution of the equation $\psi'(x;p) = 0$, *i.e.*

$$x = p\tanh(px). \tag{58}$$

For any p, this equation has a trivial solution, $x_0^* = 0$. This solution is unique when $p^2 < 1$ and corresponds to the minimum of the function ψ. Two additional solutions, $x_+^* > 0$ and $x_-^* = -x_+^* < 0$, of (58) are found at parameter values $p = -1$ and $p = +1$. These new solutions correspond to the minima of $\psi(x; p)$, while $x_0^* = 0$ corresponds now to the local maximum. Fig. 9 shows the bifurcation diagram for the solutions of (58). Let $\mu(p)$ denote the point of the minimum of the function $\psi(x; p)$ in the interval $0 \le x < \infty$:

$$\mu(p) = \begin{cases} 0 & \text{for } p \le 1, \\ x_+^* & \text{for } p > 1. \end{cases} \tag{59}$$

Fig. 2 shows the plot of this function.

B2. Quadratic approximation of ψ

For $p < 1$, the minimum of ψ is attained for $x^* = 0$ and the approximation

$$\psi(x; p) \simeq \tfrac{1}{2}(1 - p^2)x^2 \tag{60}$$

may be obtained directly from the expansion of logarithm and hyperbolic cosine functions into Taylor series in the vicinity of $x^* = 0$.

For $p > 1$, the quadratic approximation in the vicinity of the point $x^* = \mu(p)$ of the minimum is

$$\psi(x; p) \simeq \psi(\mu(p); p) + \tfrac{1}{2}\psi''(\mu(p); p)[x - \mu(p)]^2. \tag{61}$$

At the point of the minimum $x^* = \mu(p)$, (58) is satisfied and the general expression (57) may be simplified to

$$\psi(x; p) \simeq \text{constant} + \tfrac{1}{2}[1 - p^2 + \mu^2(p)][x - \mu(p)]^2. \tag{62}$$

B3. Practical calculation of $\mu(p)$ values

For a given value of $p > 1$, the corresponding value $\mu(p)$ may be calculated by one of iterative procedures for the solution of non-linear equations. Alternatively, an asymptotic formula for this function can be derived as follows.

Suppose that in the vicinity of the point $p = 1$ the function $\mu(p)$ is represented by a series in powers of $(p-1)^{1/2}$:

$$\mu(p) = (p-1)^{1/2}[a_0 + a_1(p-1) + a_2(p-1)^2 + \ldots], \tag{63}$$

where a_0, a_1, ... are some coefficients. Substituting this expression for $\mu(p)$ in (58) and collecting the terms with equal powers in the Taylor expansion at the right side of the equation, we obtain linear equations for the coefficients $a_0, a_1, \ldots$. This gives an asymptotic formula

$$\mu(p) = [6(p-1)]^{1/2}\left[1 - \frac{11}{20}(p-1) + \frac{3889}{5600}(p-1)^2 + \ldots\right], \quad \text{for } p \to 1+. \tag{64}$$

A similar approach results in an asymptotic formula

$$\mu(p) = p\{1 - 2\exp(-2p^2) - (8p^2 - 2)[\exp(-2p^2)]^2 - (48p^4 - 24p^2 + 2)[\exp(-2p^2)]^3 + \ldots\}, \quad \text{for } p \to \infty. \tag{65}$$

Numerical tests with these formulae show that the approximation (64) may be applied for $1 < p < 1.3$, while for $p \ge 1.3$ the formula (65) is more suitable.

One more way to obtain asymptotic formulae for $\mu(p)$ is to derive the differential equation for this function and to look for its solutions in the form of (63) or (65). Differentiating the identity

$$\mu(p) \equiv p\tanh[p\mu(p)], \tag{66}$$

we obtain an equation for $\mu(p)$ in the form

$$\frac{d\mu}{dp} = -\frac{\mu}{p}\left(1 + \frac{2}{p^2 - \mu^2 - 1}\right). \tag{67}$$

The conditions

$$\mu(1) = 0, \quad \mu(p) > 0, \quad \text{for} \quad p > 1, \tag{68}$$

must be attached to identify the unique solution.

B4. Acentric reflections

For acentric reflections,

$$\psi(x; p) = x^2 - \ln[I_0(2px)], \tag{69}$$

$$\psi'(x; p) = 2[x - pM(2px)] \tag{70}$$

and

$$\psi''(x; p) = 2\left\{1 - 2p^2\left[1 - \frac{1}{2px}M(2px) - M(2px)^2\right]\right\}. \tag{71}$$

Here,

$$M(z) = I_1(z)/I_0(z) \tag{72}$$

and the equation

$$\frac{dM(z)}{dz} = 1 - \frac{1}{z}M(z) - M(z)^2 \tag{73}$$

was used, which follows immediately from the main properties of the modified Bessel functions.

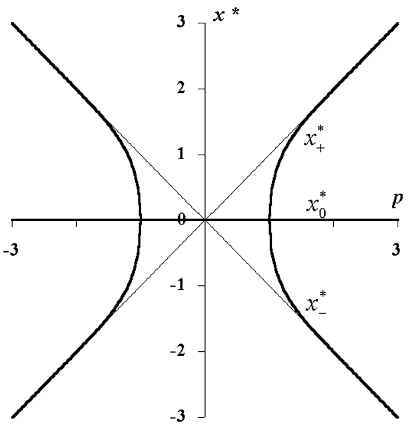

Figure 9
Bifurcation diagram for the solutions of the equation $x = p\tanh(px)$. The solutions x^* are shown as a function of the parameter p.

Numerical tests had shown similar behaviour of the function $\psi(x; p)$ for centric and acentric cases, while the authors failed to find rigorous mathematical proofs for the uniqueness of the minimum of $\psi(x; p)$ for positive x in the acentric case.

For $p < 1$, the quadratic approximation of the function (69) becomes

$$\psi(x; p) \simeq (1 - p^2)x^2 \tag{74}$$

and for $p > 1$ it is

$$\psi(x; p) \simeq \text{constant} + 2[1 - p^2 + \mu^2(p)][x - \mu(p)]^2. \tag{75}$$

Similarly to the centric case, asymptotic formulae may be derived for the function $\mu(p)$ defined now as the positive solution of the equation

$$x = p\frac{I_1(2px)}{I_0(2px)}. \tag{76}$$

For p close to 1, the function is approximated by

$$\mu(p) = (p-1)^{1/2}\left[2 - \frac{5}{6}(p-1) + \frac{199}{144}(p-1)^2 - \frac{3547}{2880}(p-1)^3 + \frac{93451}{82944}(p-1)^4 + \dots\right], \quad \text{for } p \to 1+. \tag{77}$$

For large p, an asymptotic formula is

$$\mu(p) = p\left[1 - \frac{1}{4}\left(\frac{1}{p}\right)^2 - \frac{3}{32}\left(\frac{1}{p}\right)^4 - \frac{9}{128}\left(\frac{1}{p}\right)^6 - \frac{141}{2048}\left(\frac{1}{p}\right)^8 - \dots\right], \quad \text{for } p \to \infty. \tag{78}$$

Similarly to the centric case, the first and second formulae may be applied for $1 < p < 1.3$ and for $p \geq 1.3$, respectively.

The differential equation for the function $\mu(p)$ is now

$$\frac{\mathrm{d}\mu}{\mathrm{d}p} = -\frac{\mu}{p}\left(1 + \frac{1}{p^2 - \mu^2 - 1}\right). \tag{79}$$

This work was supported by RFBR grant 00-04-48175 and by CNRS through the 'post rouge' position for VL. The authors thank C. Lecomte for his support of the project.

References

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

Afonine, P., Lunin, V. Y. & Urzhumtsev, A. G. (2001). *CCP4 Newslett. Protein Crystallogr.* **39**, 52–56.

Afonine, P., Lunin, V. Y. & Urzhumtsev, A. G. (2002). *CCP4 Newslett. Protein Crystallogr.* **40**, http://www.ccp4.ac.uk/newsletters.html.

Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–455.

Bricogne, G. & Irwin, J. (1996). *Proceedings of the CCP4 Study Weekend*, pp. 85–92. Daresbury Laboratory, Warrington, England.

Brünger, A. (1997). *Methods Enzymol.* **277**, 366–396.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLabo, 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.

Fokine, A. V., Afonine, P. V., Mikhailova, I. Yu., Tsygannik, I. N., Mareeva, T. Yu., Nesmeyanov, V. A., Pangborn, W., Li, N., Duax, W., Siszak, E. & Pletnev, V. Z. (2000). *Russ. J. Bioorg. Chem.* **26**, 512–519.

Lunin, V. Y. (1997). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, *NATO ASI Series C*, Vol. 507, pp. 451–454.

Lunin, V. Y., Lunina, N. L., Petrova, T. E., Urzhumtsev, A. G. & Podjarny, A. D. (1998). *Acta Cryst.* D**54**, 726–734.

Lunin, V. Y. & Skovoroda, T. P. (1995). *Acta Cryst.* A**51**, 880–887.

Lunin, V. Y. & Urzhumtsev, A. (1984) *Acta Cryst.* A**40**, 269–277.

Lunin, V. Y. & Urzhumtsev, A. (1999). *CCP4 Newslett. Protein Crystallogr.* **37**, 14–28.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

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

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). D**54**, 1285–1294.

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

Petrova, T. E., Lunin, V. Y., Lunina, N. L. & Skovoroda, T. P. (1999). *Biophysics*, **44**, 18–22.

Petrova, T. E., Lunin, V. Y. & Podjarny, A. D. (1999). *Acta Cryst.* A**55**, 739–745.

Petrova, T. E., Lunin, V. Y. & Podjarny, A. D. (2000). *Acta Cryst.* D**56**, 1245–1252.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

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

Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.

Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.

Skovoroda, T. P. & Lunin, V. Y. (2000). *Crystallogr. Rep.* **45**, 195–198.

Srinivasan, R. & Parthasarathy, S. (1976). *Some Statistical Applications in X-ray Crystallography*. Oxford: Pergamon Press.

Urzhumtsev, A. G., Skovoroda, T. P. & Lunin, V. Y. (1996). *J. Appl. Cryst.* **29**, 741–744.