International Tables for Crystallography (2006). Vol. F, Section 18.2.3, pp. 375–377.

18.2. Enhanced macromolecular refinement by simulated annealing

BY A. T. BRUNGER, P. D. ADAMS AND L. M. RICE

18.2.1. Introduction

The analysis of X-ray diffraction data generally requires sophisticated computational procedures that culminate in refinement and structure validation. The refinement procedure can be formulated as the chemically constrained or restrained nonlinear optimization of a target function, which usually measures the agreement between observed diffraction data and data computed from an atomic model. The ultimate goal of refinement is to optimize simultaneously the agreement of an atomic model with observed diffraction data and with *a priori* chemical information.

The target function used for this optimization normally depends on several atomic parameters and, most importantly, on atomic coordinates. The large number of adjustable parameters (typically at least three times the number of atoms in the model) gives rise to a very complicated target function. This, in turn, produces what is known as the multiple minima problem: the target function contains many local minima in addition to the global minimum, and this tends to defeat gradient-descent optimization techniques such as conjugate gradient or least-squares methods (Press *et al.*, 1986). These methods are unable to sample molecular conformations thoroughly enough to find the optimal model if the starting one is far from the correct structure.

The challenges of crystallographic refinement arise not only from the high dimensionality of the parameter space, but also from the phase problem. For new crystal structures, initial electron-density maps must be computed from a combination of observed diffraction amplitudes and experimental phases, where the latter are typically of poorer quality and/or at a lower resolution than the former. A different problem arises when structures are solved by molecular replacement (Hoppe, 1957; Rossmann & Blow, 1962), which uses a similar structure as a search model to calculate initial phases. In this case, the resulting electron-density maps can be severely 'model-biased', that is, they sometimes seem to confirm the existence of the search model without providing clear evidence of actual differences between it and the true crystal structure. In both cases, initial atomic models usually contain significant errors and require extensive refinement.

Simulated annealing (Kirkpatrick *et al.*, 1983) is an optimization technique particularly well suited to overcoming the multiple minima problem. Unlike gradient-descent methods, simulated annealing can cross barriers between minima and, thus, can explore a greater volume of the parameter space to find better models (deeper minima). Following its introduction to crystallographic refinement (Brünger *et al.*, 1987), there have been major improvements of the original method in four principal areas: the measure of model quality, the search of the parameter space, the target function and the modelling of conformational variability.

For crystallographic refinement, the introduction of cross validation and the free R value (Brünger, 1992) has significantly reduced the danger of overfitting the diffraction data during refinement. Cross validation also produces more realistic coordinate-error estimates based on the Luzzati or σ_A methods (Kleywegt & Brünger, 1996). The complexity of the conformational space has been reduced by the introduction of torsion-angle refinement methods (Diamond, 1971; Rice & Brünger, 1994), which decrease the number of adjustable parameters that describe a model approximately tenfold. The target function has been improved by using a maximum-likelihood approach which takes into account model error, model incompleteness and errors in the experimental data (Bricogne, 1991; Pannu & Read, 1996). Cross validation of parameters for the maximum-likelihood target function was essential in order to obtain better results than with conventional target functions (Pannu & Read, 1996; Adams *et al.*, 1997; Read, 1997). Finally, the sampling power of simulated annealing has been used for exploring the molecule's conformational space in cases where the molecule undergoes dynamic motion or exhibits static disorder (Kuriyan *et al.*, 1991; Burling & Brünger, 1994; Burling *et al.*, 1996).

18.2.2. Cross validation

Cross validation (Brünger, 1992) plays a fundamental role in the maximum-likelihood target functions described below. A few remarks about this method are therefore warranted (for reviews see Kleywegt & Brünger, 1996; Brünger, 1997). For cross validation, the diffraction data are divided into two sets: a large *working* set (usually comprising 90% of the data) and a complementary *test* set (comprising the remaining 10%). The diffraction data in the working set are used in the normal crystallographic refinement process, whereas the test data are not. The cross-validated (or 'free') R value computed with the test-set data is a more faithful indicator of model quality. It provides a more objective guide during the model building and refinement process than the conventional R value. It also ensures that introduction of additional parameters (*e.g.* water molecules, relaxation of non-crystallographic symmetry restraints, or multi-conformer models) improves the quality of the model, rather than increasing overfitting.

Since the conventional R value shows little correlation with the accuracy of a model, coordinate-error estimates derived from the Luzzati (1952) or σ_A (Read, 1986) methods are unrealistically low. Kleywegt & Brünger (1996) showed that more reliable coordinate errors can be obtained by cross validation of the Luzzati or σ_A coordinate-error estimates. An example is shown in Fig. 18.2.2.1 using the crystal structure and diffraction data of penicillopepsin (Hsu *et al.*, 1977). At 1.8 Å resolution, the model has an estimated coordinate error of ∼0.2 Å as assessed by multiple independent refinements. As the resolution of the diffraction data is artificially truncated and the model re-refined, the coordinate error (assessed by the atomic root-mean-square difference to the refined model at 1.8 Å resolution) increases monotonically. The conventional R value improves as the resolution decreases and the quality of the model worsens. Consequently, coordinate-error estimates do not display the correct behaviour either: the error estimates are approximately constant, regardless of the resolution and actual coordinate error of the models. However, when cross validation is used (*i.e.*, the test reflections are used to compute the estimated coordinate errors), the results are much better: the cross-validated errors are close to the actual coordinate error, and they show the correct trend as a function of resolution (Fig. 18.2.2.1).

18.2.3. The target function

Crystallographic refinement is a search for the global minimum of the target

$$E = E_{\text{chem}} + w_{\text{X-ray}} E_{\text{X-ray}} \qquad (18.2.3.1)$$

as a function of the parameters of an atomic model, in particular, atomic coordinates. E_{chem} comprises empirical information about chemical interactions; it is a function of all atomic positions, describing covalent (bond lengths, bond angles, torsion angles, chiral centres and planarity of aromatic rings) and non-bonded (intramolecular as well as intermolecular and symmetry-related)

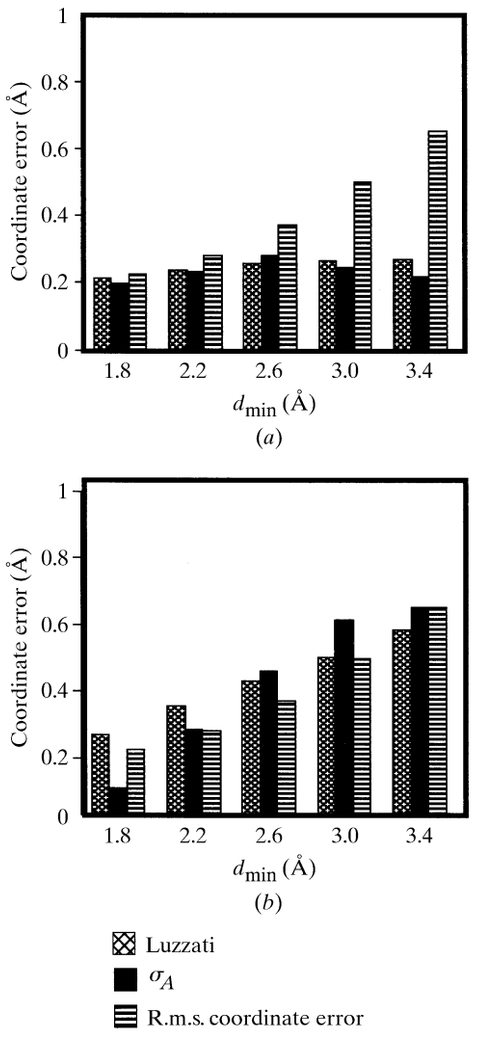

Fig. 18.2.2.1. Effect of resolution on coordinate-error estimates: accuracy as a function of resolution. Refinements were begun with the crystal structure of penicillopepsin (Hsu *et al.*, 1977) with water molecules omitted and with uniform temperature factors. The low-resolution limit was set to 6 Å. Inclusion of all low-resolution diffraction data does not change the conclusions (Adams *et al.*, 1997). The penicillopepsin diffraction data were artificially truncated to the specified high-resolution limit. Each refinement consisted of simulated annealing using a Cartesian-space slow-cooling protocol starting at 2000 K, overall *B*-factor refinement and individual restrained *B*-factor refinement. All refinements were carried out with 10% of the diffraction data randomly omitted for cross validation. (*a*) Coordinate-error estimates of the refined structures using the methods of Luzzati (1952) and Read (1986). All observed diffraction data were used, *i.e.* no cross validation was performed. The actual coordinate errors (r.m.s. differences to the original crystal structure) are shown for comparison. (*b*) Cross-validated coordinate-error estimates. The test set was used to compute the coordinate-error estimates (Kleywegt & Brünger, 1996).

interactions (Hendrickson, 1985). $E_{\text{X-ray}}$ is related to the difference between observed and calculated data, and $w_{\text{X-ray}}$ is a weight appropriately chosen to balance the gradients (with respect to atomic parameters) arising from the two terms.

18.2.3.1. *X-ray diffraction data versus model*

The traditional form of $E_{\text{X-ray}}$ consists of the crystallographic residual, E^{LSQ}, defined as the sum over the squared differences between the observed ($|\mathbf{F}_o|$) and calculated ($|\mathbf{F}_c|$) structure-factor amplitudes for a particular atomic model:

$$E_{\text{X-ray}} = E^{\text{LSQ}} = \sum_{hkl \in \text{working set}} (|\mathbf{F}_o| - k|\mathbf{F}_c|)^2, \qquad (18.2.3.2)$$

where *hkl* are the indices of the reciprocal-lattice points of the crystal and *k* is a relative scale factor.

Minimization of E^{LSQ} can produce improvement in the atomic model, but it can also accumulate systematic errors in the model by fitting noise in the diffraction data (Silva & Rossmann, 1985). The least-squares residual is a limiting case of the more general maximum-likelihood theory and is only justified if the model is nearly complete and error-free. These assumptions may be violated during the initial stages of refinement. Improved targets for macromolecular refinement have been obtained using the more general maximum-likelihood formulation (Bricogne, 1991; Pannu & Read, 1996; Adams *et al.*, 1997; Murshudov *et al.*, 1997). The goal of the maximum-likelihood method is to determine the likelihood of the model, given estimates of the model's errors and those of the measured intensities.

A starting point for the maximum-likelihood formulation of crystallographic refinement is the Sim (1959) distribution, *i.e.*, the Gaussian conditional probability distribution of the 'true' structure factors, $\mathbf{F}$, given a partial model with structure factors $\mathbf{F}_c$ and the model's error (Fig. 18.2.3.1) (Srinivasan, 1966; Read, 1986, 1990) (for simplicity we will only discuss the case of acentric reflections),

$$P_a(\mathbf{F};\, \mathbf{F}_c) = (1/\pi\varepsilon\sigma_\Delta^2)\exp[-(\mathbf{F} - D\mathbf{F}_c)^2/\varepsilon\sigma_\Delta^2], \qquad (18.2.3.3)$$

where σ_Δ is a parameter that incorporates the effect of the fraction of the asymmetric unit that is missing from the model and errors in the partial structure. Assuming a Wilson distribution of intensities, it can be shown that (Read, 1990)

$$\sigma_\Delta^2 = \langle|\mathbf{F}_o|^2\rangle - D^2\langle|\mathbf{F}_c|^2\rangle, \qquad (18.2.3.4)$$

where *D* is a factor that takes into account model error: it is unity in the limiting case of an error-free model and it is zero if no model is available (Luzzati, 1952; Read, 1986). For a complete and error-free model, σ_Δ therefore becomes zero, and the probability distribution, $P_a(\mathbf{F};\, \mathbf{F}_c)$, is infinitely sharp.

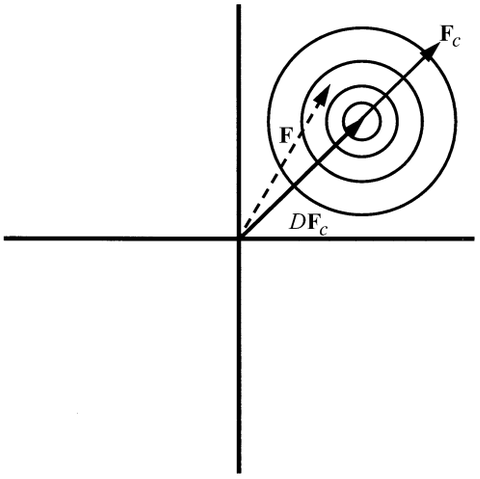

Fig. 18.2.3.1. The Gaussian probability distribution forms the basis of maximum-likelihood targets in crystallographic refinement. The conditional probability of the true structure factor, $\mathbf{F}$, given model structure factors, is a Gaussian in the complex plane [equation (18.2.3.3)]. The expected value of the probability distribution is $D\mathbf{F}_c$ with variance σ_Δ, where *D* and σ_Δ account for missing or incorrectly placed atoms in the model.

Taking measurement errors into account requires multiplication of equation (18.2.3.3) with an appropriate probability distribution (usually a conditional Gaussian distribution with standard deviation σ_o) of the observed structure-factor amplitudes ($|\mathbf{F}_o|$) around the 'true' structure-factor amplitudes ($|\mathbf{F}|$),

$$P_{\text{meas}}(|\mathbf{F}_o|;\,|\mathbf{F}|). \tag{18.2.3.5}$$

Prior knowledge of the phases of the structure factors can be incorporated by multiplying equation (18.2.3.3) with a phase probability distribution

$$P_{\text{phase}}(\varphi) \tag{18.2.3.6}$$

and rewriting equation (18.2.3.3) in terms of the structure-factor moduli and amplitudes of $\mathbf{F} = |\mathbf{F}|\exp(i\varphi)$.

The unknown variables $|\mathbf{F}|$ and φ in equations (18.2.3.3)–(18.2.3.5) have to be eliminated by integration in order to obtain the conditional probability distribution of the observed structure-factor amplitudes, given a partial model with errors, the amplitude measurement errors and prior phase information:

$$P_a(|\mathbf{F}_o|;\,\mathbf{F}_c) = (1/\pi\varepsilon\sigma_\Delta^2)\int \mathrm{d}\varphi\;\mathrm{d}|\mathbf{F}|\;|\mathbf{F}|P_{\text{meas}}(|\mathbf{F}_o|;\,|\mathbf{F}|) \times P_{\text{phase}}(\varphi)\exp\left\{-[|\mathbf{F}|\exp(i\varphi) - D\mathbf{F}_o]^2/\varepsilon\sigma_\Delta^2\right\}. \tag{18.2.3.7}$$

The likelihood, $\mathcal{L}$, of the model is defined as the joint probability distribution of the structure factors of all reflections in the working set. Assuming independent and uncorrelated structure factors, $\mathcal{L}$ is simply the product of the distributions in equation (18.2.3.7) for all reflections. Instead of maximizing the likelihood, it is more common to minimize the negative logarithm of the likelihood,

$$E_{\text{X-ray}} = \mathcal{L} = -\sum_{hkl\in\text{working set}} \log[P_a(|\mathbf{F}_o|;\,\mathbf{F}_c)]. \tag{18.2.3.8}$$

Empirical estimates of σ_Δ [and D through equation (18.2.3.4)] can be obtained by minimizing $\mathcal{L}$ for a particular atomic model. It is generally assumed that σ_Δ and D show relatively little variation among neighbouring reflections. Accepting this assumption, σ_Δ and D can be estimated by considering narrow resolution shells of reflections and assuming that the two parameters are constant in these shells. Minimization of $\mathcal{L}$ can then be performed as a function of these constant shell parameters while keeping the atomic model fixed (Read, 1986, 1997). Alternatively, one can assume a two-term Gaussian model for σ_Δ (Murshudov *et al.*, 1997) and minimize $\mathcal{L}$ as a function of the Gaussian parameters. Note that individual atomic B factors are taken into account by the calculated model structure factors ($\mathbf{F}_c$).

This empirical approach to estimate σ_Δ and D requires occasional recomputation of these values as the model improves. Refinement methods that improve the model structure factors, $\mathbf{F}_c$, will therefore have a beneficial effect on σ_Δ and D. Better estimates of these values will then enhance the next refinement cycle. Thus, powerful optimization methods and maximum-likelihood targets are expected to interact in a synergistic fashion (*cf.* Fig. 18.2.5.1). Structure-factor averaging of multi-start refinement models can provide another layer of improvement by producing a better description of $\mathbf{F}_c$ if the model shows significant variability due to errors or intrinsic flexibility (see below).

In order to achieve an improvement over the least-squares residual [equation (18.2.3.2)], cross validation was found to be essential (Pannu & Read, 1996; Adams *et al.*, 1997) for the estimation of model incompleteness and errors (σ_Δ and D). Since the test set typically contains only 10% of the diffraction data, these cross-validated quantities can show significant statistical fluctuations as a function of resolution. In order to reduce these fluctuations, Read (1997) devised a smoothing method by applying restraints to σ_A values between neighbouring resolution shells where

$$\sigma_A = \left[1 - (\sigma_\Delta/\langle|\mathbf{F}_o|^2\rangle)\right]^{1/2}. \tag{18.2.3.9}$$

Pannu & Read (1996) have developed an efficient Gaussian approximation of equation (18.2.3.7) in cases of no prior phase information, termed the 'MLF' target function. In the limit of a perfect model (*i.e.* $\sigma_\Delta = 0$ and $D = 1$), MLF reduces to the traditional least-squares residual [equation (18.2.3.2)] with $1/\sigma_o^2$ weighting. In the case of prior phase information, the integration over the phase angles has been carried out numerically in equation (18.2.3.7), termed the 'MLHL' target (Pannu *et al.*, 1998). A maximum-likelihood function which expresses equation (18.2.3.7) in terms of observed intensities has also been developed, termed 'MLI' (Pannu & Read, 1996).

18.2.3.2. *A priori chemical information*

The parameters for the covalent terms in E_{chem} [equation (18.2.3.1)] can be derived from the average geometry and (r.m.s.) deviations observed in a small-molecule database. Extensive statistical analyses were undertaken for the chemical moieties of proteins (Engh & Huber, 1991) and polynucleotides (Parkinson *et al.*, 1996) using the Cambridge Structural Database (Allen *et al.*, 1983). Analysis of the ever-increasing number of atomic resolution macromolecular crystal structures will no doubt cause some modifications of these parameters in the future.

It is common to use a purely repulsive quartic function ($E_{\text{repulsive}}$) for the non-bonded interactions that are included in E_{chem} (Hendrickson, 1985):

$$E_{\text{repulsive}} = \sum_{ij}[(cR_{ij}^{\min})^n - R_{ij}^n]^m, \tag{18.2.3.10}$$

where R_{ij} is the distance between two atoms i and j, $R_{ij}^{\min}$ is the van der Waals radius for a particular atom pair ij, $c \leq 1$ is a constant that is sometimes used to reduce the radii, and $n = 2$, $m = 2$ or $n = 1$, $m = 4$. van der Waals attraction and electrostatic interactions are usually not included in crystallographic refinement. These simplifications are valid since the diffraction data contain information that is able to produce atomic conformations consistent with actual non-bonded interactions. In fact, atomic resolution crystal structures can be used to derive parameters for electrostatic charge distributions (Pearlman & Kim, 1990).

18.2.4. Searching conformational space

Annealing denotes a physical process wherein a solid is heated until all particles randomly arrange themselves in a liquid phase and is then cooled slowly so that all particles arrange themselves in the lowest energy state. By formally defining the target, E [equation (18.2.3.1)], to be the equivalent of the potential energy of the system, one can simulate such an annealing process (Kirkpatrick *et al.*, 1983). There is no guarantee that simulated annealing will find the global minimum (Laarhoven & Aarts, 1987). However, compared to conjugate-gradient minimization, where search directions must follow the gradient, simulated annealing achieves more optimal solutions by allowing motion against the gradient (Kirkpatrick *et al.*, 1983). The likelihood of *uphill* motion is determined by a control parameter referred to as *temperature*. The higher the temperature, the more likely it is that simulated annealing will overcome barriers (Fig. 18.2.4.1). It should be noted that the simulated-annealing temperature normally has no physical meaning and merely determines the likelihood of overcoming barriers of the target function in equation (18.2.3.1).

references