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)

references