International Tables for Crystallography (2006). Vol. F, Section 18.1.4, pp. 369–370.

18. REFINEMENT

18.1. Introduction to refinement

BY L. F. TEN EYCK AND K. D. WATENPAUGH

18.1.1. Overview

Methods of improving and assessing the accuracy of the positions of atoms in crystals rely on the agreement between the observed and calculated diffraction data. Calculation of diffraction data from an atomic model depends on the theoretical model of scattering of X-rays by crystals discussed in *IT* B (2001) Chapter 1.2. The properties of the measured data are discussed in *IT* C (1999) Chapters 2.2 and 7.1–7.5, and the mathematical basis of refinement of structural parameters is discussed in *IT* C Chapters 8.1–8.5. This chapter concentrates on the special features of macromolecular crystallography.

18.1.2. Background

Macromolecular crystallography is not fundamentally different from small-molecule crystallography, but is complicated by the sheer size of the problems. Typical macromolecules contain thousands of atoms and crystallize in unit cells of around a million cubic ångstroms. The large size of the problems has meant that the techniques applied to small molecules require too many computational resources to be directly applied to macromolecules. This has produced a lag between macromolecular and small-molecule practice beyond the limitations introduced by generally poorer resolution. In essence, macromolecular refinement has followed small-molecule crystallography. Additional complexity arises from the book-keeping required to describe the macromolecular structure, which is usually beyond the capabilities of programs designed for small molecules.

Fitting the atom positions to the calculated electron-density maps (Fourier maps) was a standard method until the introduction of least-squares refinement technique in reciprocal space by Hughes (1941). A less computationally intense method of calculating shifts using difference Fourier maps (ΔF methods) was introduced by Booth (1946*a*,*b*). By the early 1960s, digital computers were becoming generally available and least-squares refinement methods became the method of choice in refining small molecules. The program *ORFLS* developed by Busing *et al.* (1962) was perhaps the most extensively used. In the late 1960s, as protein structures were being determined by multiple isomorphous replacement (MIR) methods (see Part 12 and Chapter 2.4 in *IT* B), methods of improving the structural models derived from the electron-density maps were being studied. Diamond (1971) introduced the use of a constrained chemical model in the fitting of a calculated electron-density model to an MIR-derived electron-density map in a 'real-space refinement' procedure. Diamond commented that phases derived from a previous cycle of real-space fitting could be used to calculate the next electron-density map, but this was not done. Watenpaugh *et al.* (1972) first showed in 1971 that ΔF refinement methods could be applied to both improve the model and extend the phases from initial MIR or SIR (single isomorphous replacement) experimental phases. Watenpaugh *et al.* (1973) also applied least-squares techniques to the refinement of a protein structure for the first time using a 1.54 Å resolution data set. Improvement of the phases, clarification of the electron-density maps and interpretation of unknown sequences in the structure were clearly evident, although chemical restraints were not applied. The adaptation by Hendrickson & Konnert of the restrained least-squares refinement program developed by Konnert (Konnert, 1976; Hendrickson, 1985) became the first extensively used macromolecular refinement program. At this time, refinement of protein models became practical and nearly universal. Model refinement improved models derived from structures determined by isomorphous replacement methods and also provided the means to improve structural models of related protein structures determined by molecular replacement methods (see Part 13 and *IT* B Chapter 2.3).

By the 1980s, it became clear that additional statistical rigour in macromolecular refinement was required. The first and most obvious problem was that macromolecular structures were often solved with fewer observations than there were parameters in the model, which leads to overfitting. Recent advances include cross validation for detection of overfitting of data (Brünger, 1992); maximum-likelihood refinement for improved robustness (Pannu & Read, 1996; Murshudov *et al.*, 1997; Bricogne, 1997; Adams *et al.*, 1999); improved methods for describing the model with fewer parameters (Rice & Brünger, 1994; Murshudov *et al.*, 1999); and incorporation of phase information from multiple sources (Pannu *et al.*, 1998). These improvements in the theory and practice of macromolecular refinement will undoubtedly not be the last word on the subject.

18.1.3. Objectives

A variety of methods are employed to improve the agreement between observed and calculated macromolecular diffraction patterns. Some of the more popular methods are discussed in the different sections of this chapter. In part, the different methods arise from focusing on different goals during different stages of model refinement. Bias generated by incomplete models, and radius of convergence, are important considerations at early stages of refinement, because the models are usually incomplete, contain significant errors in atom parameters and may carry errors from misinterpretation of poorly phased electron-density maps. During this stage of the process, the primary concern is to determine how the model of the chain tracing and conformation of the residues should be described. In later stages, after the description of the model has been determined, the objective is to determine accurate estimates of the values of the parameters which best explain the observed data. These two stages of the problem have different properties and should be treated differently.

18.1.4. Least squares and maximum likelihood

'Improving the agreement' between the observed and calculated data can only be done if one first decides the criteria to be used to measure the agreement. The most commonly used measure is the L_2 norm of the residuals, which is simply the sum of the squares of the differences between the observed and calculated data (*IT* C Chapter 8.1),

$$L_2(\mathbf{x}) = \|w_i[y_i - f_i(\mathbf{x})]\| = \sum_i w_i[y_i - f_i(\mathbf{x})]^2, \qquad (18.1.4.1)$$

Copyright © 2006 International Union of Crystallography

where w_i is the weight of observation y_i and $f_i(\mathbf{x})$ is the calculated value of observation i given the parameters $\mathbf{x}$. In essence, least-squares refinement poses the problem as 'Given these data, what are the parameters of the model that give the minimum variance of the observations?'. The L_2 norm is strongly affected by the largest deviations, which is not a desirable property in the early stages of refinement where the model may be seriously incomplete. In the early stages, it may be better to refine against the L_1 norm,

$$L_1 = \sum_i w_i |[y_i - f_i(\mathbf{x})]|,$$

the sum of the absolute value of the residuals. At present, this technique is not used in macromolecular crystallography.

The observable quantity in crystallography is the diffracted intensity of radiation. Fourier inversion of the model gives us a complex structure factor. The phase information is normally lost in the formulation of $f_i(\mathbf{x})$. This is a root cause of some of the problems of least-squares refinement from poor starting models. Many of the problems of least-squares refinement can be addressed by changing the measure of agreement from least squares to maximum likelihood, which evaluates to the likelihood of the observations given the model. In this formulation, the problem is posed as 'Given this model, what is the probability that the given set of data would be observed?'. The model is adjusted to maximize the probability of the given observations. This procedure is subtly different from least squares in that it is reasonably straightforward to account for incomplete models and errors in the model in computing the probability of the observations. Maximum-likelihood refinement is particularly useful for incomplete models because it produces residuals that are less biased by the current model than those produced by least squares. Maximum likelihood also provides a rigorous formulation for all forms of error in both the model and the observations, and allows incorporation of additional forms of prior knowledge (like additional phase information) into the probability distributions.

The likelihood of a model given a set of observations is the product of the probabilities of all of the observations given the model. If $P_a(\mathbf{F}_i; \mathbf{F}_{i,c})$ is the conditional probability distribution of the structure factor $\mathbf{F}_i$ given the model structure factor $\mathbf{F}_{i,c}$, then the likelihood of the model is

$$L = \prod_i P_a(\mathbf{F}_i; \mathbf{F}_{i,c}).$$

This is usually transformed into a more tractable form by taking the logarithm,

$$\log L = \sum_i \log P_a(\mathbf{F}_i; \mathbf{F}_{i,c}).$$

Since the logarithm increases monotonically with its argument, the two versions of the equation have maxima at the same values of the parameters of the model. This formulation is described in more detail in Chapter 18.2, in *IT* C Section 8.2.1 and by Bricogne (1997), Pannu & Read (1996), and Murshudov *et al.* (1997).

18.1.5. Optimization

Once the choice of criteria for agreement has been made, the next step is to adjust the parameters of the model to minimize the disagreement (or maximize the agreement) between the model and the data. The literature on optimization in numerical analysis and operations research, discussed in *IT* C Chapters 8.1–8.5, is very rich. The methods can be characterized by their use of gradient information (no gradients, first derivatives, or second derivatives), by their search strategy (none, downhill, random, annealed, or a combination of these), and by various performance measures on different classes of problems. These will be discussed more fully in Section 18.1.8.

18.1.6. Data

Resolution, accuracy, completeness and weighting of data all have an impact on the refinement process. Small-molecule crystals usually, but not always, diffract to well beyond atomic resolution. Macromolecular crystals do not generally diffract to atomic resolution. Macromolecular structures are by definition large, which in turn means that the unit cells are large and the number of diffracting unit cells per crystal is small when compared to small-molecule crystals of similar size. Fortunately, the situation can be partially offset with the use of the much more intense radiation generated by synchrotrons (Part 8) and by improved data-collection methods (Parts 7–11). Synchrotron-radiation sources designed to produce intense beams of X-rays for the study of materials are becoming much more readily available. As a consequence, both higher resolution and statistically better data can be obtained. Improvements in area-detector technology, protein purification, cryocrystallography and data-integration software beneficially influence the refinement process.

Refinement of crystal structures is a statistical process. There is no substitute for adequate amounts of accurate, correctly weighted data. Lower accuracy can be accommodated by increased amounts of data and correct weighting. Unfortunately, determining the correct weighting for macromolecular diffraction data is difficult. Maximum-likelihood methods are more robust than least-squares methods against improperly weighted data.

It has been clearly demonstrated that the best procedure for refining small molecules is to include all of the observations as integrated intensities, properly weighted, without preliminary symmetry averaging. Inclusion of weak data and refinement on diffracted intensity does not change the results very much, but has a strong effect on the precision of the parameter estimates derived from the refinement.

The long-standing debate as to whether refinement should be against structure-factor amplitudes or diffracted intensity has been resolved for small-molecule crystallography. Refinement against intensity is preferred because it is closer to the experimentally observed quantity, and the statistical weighting of the data is superior to that obtained for structure-factor amplitudes. If the model is correct and the data are reasonably good, the primary distinction between the two approaches is in the standard uncertainties of the derived parameters, which are usually somewhat better if the refinement is against diffracted intensity.

18.1.7. Models

Atomic resolution models are generally straightforward. A reasonably well phased diffraction pattern at atomic resolution shows the location of each atom. The primary problem (which can be substantial) is deciding how to model any disorder that may be present. Structural chemistry is derived from the model. Macromolecular models generally have most of the structural chemistry built in as part of the model. This approach is required as a direct consequence of having too little data at too limited a resolution to determine the positions of all of the atoms without using this additional information.

There are two procedures for building structural chemistry into a model. The first is to use known molecular geometry to reduce the number of variables. For example, if the distance between two atoms is held constant, the locus of possible positions for the second atom is the surface of a sphere centred on the first atom. This means that the position of the second atom can be specified given the

references