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),

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

references