International Tables for Crystallography (2006). Vol. F, Chapter 18.1, pp. 369–374.

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),

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

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

position of the first atom and two variables to locate the point on the sphere – a total of five variables instead of six. Every non-redundant constraint reduces the number of degrees of freedom in the model by one. If the second atom in this example were replaced by a group of atoms with known geometry (*e.g.* a phenyl group containing six atoms), the number of positional parameters could be reduced from 21 to eight. Constrained refinement is discussed extensively in *IT* C Chapter 8.3.

The second procedure is to treat the additional information as additional observations. A bond length is assumed to be an observation, based on other crystal structures, which has a mean value and a variance. This observation is added to the data instead of being used to reduce the number of parameters in the model.

The two approaches have different consequences on the ratio of observations to parameters. If we have N_o observations, N_p parameters and N_r non-redundant geometric features to add to the problem, we have either $C = N_o/(N_p - N_r)$ and $(\mathrm{d}C/\mathrm{d}N_r) = C/(N_p - N_r)$ or $C = (N_o + N_r)/N_p$ and $(\mathrm{d}C/\mathrm{d}N_r) = 1/N_p$, where C is the ratio of observations to parameters. The former are parameter *constraints* and the latter are parameter *restraints*. Constraints are more effective at increasing the ratio of observations to parameters, but since these features are built into the model, it is difficult to evaluate how appropriate they actually are for the problem at hand. Restraints provide an automatic evaluation of the appropriateness of the assumed geometry to the current data, because the deviations from the assumed values can be tested for statistical significance.

The most common constraints and restraints applied to macromolecular crystal structures are those which preserve or reinforce the molecular geometry of the amino acid or nucleotide residues (Chapter 18.3). Expected values for the geometry of these structural fragments are available from the small-molecule crystallographic literature and databases. A further step, which reduces the parameter count substantially, is to treat parts of the molecule as a set of linked rigid groups. This is particularly appropriate for aromatic fragments such as the side chains of phenylalanine, tyrosine, tryptophan and histidine, but can also be appropriate for small groups like valine and threonine. The extreme form of this approach is torsion-angle dynamics (Rice & Brünger, 1994), in which the only variables are torsion angles about bonds, and the position and orientation of the whole molecule. This description of the model works well with the right kind of optimization procedure.

Positional restraints can be parameterized in a variety of ways. For example, the geometry of three atoms can be treated as the three distances involved or as two distances and the angle between them. Several of the more popular restrained refinement programs treat the parameters for bond distances, bond angles and planarity as distances with a set of standard deviations. Others treat them as bond distances, bond angles and torsion angles weighted by the energy terms derived from experimental conditions. Different methods of parameterization and weighing have different effects on the refinement process, but to date these differences are not well characterized. The primary effects should be on the approach to convergence, as all of these formulations are normally satisfied by correct structures.

Additional criteria can be added to the model besides simple geometry. Preservation of bond lengths is usually done by adding terms

$$\sum_{\substack{\text{bonded}\\ \text{atoms}}} \left(1/\sigma_{ij}^2\right)\left(d_{ij} - d_{ij}^o\right)^2$$

to the objective function, where d_{ij} is the distance between atoms i and j, d_{ij}^o is the ideal bond length, and σ_{ij} is the weight applied to the bond. This is formally equivalent to treating bond stretching as a spring. Additional energy parameters can be added, such as electrostatic energy terms. Whatever vision of reality is applied to the objective function becomes part of the model.

The atomic displacement factors (B factors) present a different set of problems from the coordinates. The behaviour of these parameters is strongly affected by coordinate errors, and in fact large atomic displacement parameters are frequently used to determine which parts of a structure are likely to contain errors. The B factors are strongly related to the rate at which the diffraction pattern diminishes with resolution and thus cannot be accurately determined unless the diffraction pattern has been measured over a sufficiently wide range of resolution to determine this rate. As a practical matter, it is not feasible to refine individual atomic displacement parameters at resolution less than about 2 Å, and they frequently present problems even in atomic resolution small-molecule structures. In high-resolution small-molecule structures, B factors are frequently represented as anisotropic ellipsoids described by six parameters per atom. In spite of the larger displacements found in macromolecules relative to small molecules, it is rarely possible to support the number of parameters required to refine a structure with independent anisotropic displacement factors. Nevertheless, the B factors of the atoms are essential parts of the crystallographic model. Several methods for reducing the number of independent B factors have been developed. The simplest is group B factors, in which one parameter is refined for all atoms in a particular group of atoms. Another method is to apply a simple model to the change in displacement parameter within a group of atoms. In this treatment, a B factor is refined for one atom, say the C_α atom of an amino-acid residue, and the remainder of the atoms in the residue are assigned displacement parameters that depend on their distance from the C_α atom (Konnert & Hendrickson, 1980). A third method is to enforce similarity of displacement parameters based on the correlation coefficients between pairs of displacement parameters in highly refined high-resolution structures (Tronrud, 1996).

Small-molecule refinement programs also apply restraints to the displacement parameters. The SIMU command of *SHELX* restrains the axes of the anisotropic displacement parameters of bonded atoms to be similar. This approach has been applied to a number of very high resolution macromolecular refinements.

Large B factors do not represent large thermal motions of the atom but rather a distribution of positions occupied by the atom over time or in different unit cells of the crystal. The line between describing atoms with large B factors as distributed about a single point or several points (disordered atoms) is sometimes blurred. At some point, the disorder can become resolved into alternative positions or the atoms disappear from the observable electron density. There are two kinds of disorder that can be easily modelled if data are available to sufficient resolution:

(1) *Static disorder* describes the situation in which portions of the structure have a small number of possible alternative conformations. The atoms in any given unit cell are in only one of the possible conformations, but different cells may have different conformations. Since the diffraction experiment averages the structure over all unit cells in the X-ray beam, the observations correspond to an average structure in which each conformation is weighted according to the fraction of the unit cells containing that conformation. The normal bond-length and angle restraints apply to each conformation, and the fractional occupancy of all conformations should sum to 1.0.

(2) *Dynamic disorder* describes the situation in which portions of the structure are not in fixed positions. This form of disorder is frequently encountered in amino-acid side chains on the molecular surface. The electrons are spread over a sufficiently large volume that the average electron density is very low and the atoms are essentially invisible to X-rays. In such cases, the best model is to simply omit the atoms from the diffraction calculation. They are

commonly placed in the model in plausible positions according to molecular geometry, but this can be misleading to people using the coordinate set. If the atoms are included in the model, the atomic displacement parameters generally become very large, and this may be an acceptable flag for dynamic disorder. The hazard with this procedure is that including these atoms in the model provides additional parameters to conceal any error signal in the data that might relate to problems elsewhere in the model.

At high resolution, it is sometimes possible to model the correlated motion of atoms in rigid groups by a single tensor that describes translation, libration and screw. This is rarely done for macromolecules at present, but may be an extremely accurate way to model the behaviour of the molecules. The recent development of efficient anisotropic refinement methods for macromolecules by Murshudov *et al.* (1999) will undoubtedly produce a great deal more information about the modelling of dynamic disorder and anisotropy in macromolecular structures.

Macromolecular crystals contain between 30 and 70% solvent, mostly amorphous. The diffraction is not accurately modelled unless this solvent is included (Tronrud, 1997). The bulk solvent is generally modelled as a continuum of electron density with a high atomic displacement parameter. The high displacement parameter blurs the edges, so that the contribution of the bulk solvent to the scattering is primarily at low resolution. Nevertheless, it is important to include this in the model for two reasons. First, unless the bulk solvent is modelled, the low-resolution structure factors cannot be used in the refinement. This has the unfortunate effect of rendering the refinement of *all* of the atomic displacement parameters ill-determined. Second, omission or inaccurate phasing of the low-resolution reflections tends to produce long-wavelength variations in the electron-density maps, rendering them more difficult to interpret. In some regions, the maps can become overconnected, and in others they can become fragmented.

18.1.8. Optimization methods

Optimization methods for small molecules are straightforward, but macromolecules present special problems due to their sheer size. The large number of parameters vastly increases the volume of the parameter space that must be searched for feasible solutions and also increases the storage requirements for the optimization process. The combination of a large number of parameters and a large number of observations means that the computations at each cycle of the optimization process are expensive.

Optimization methods can be roughly classified according to the order of derivative information used in the algorithm. Methods that use no derivatives find an optimum through a search strategy; examples are Monte Carlo methods and some forms of simulated annealing. First-order methods compute gradients, and hence can always move in a direction that should reduce the objective function. Second-order methods compute curvature, which allows them to predict not only which direction will reduce the objective function, but how that direction will change as the optimization proceeds. The zero-order methods are generally very slow in high-dimensional spaces because the volume that must be searched becomes huge. First-order methods can be fast and compact, but cannot determine whether or not the solution is a true minimum. Second-order methods can detect null subspaces and singularities in the solution, but the computational cost grows as the cube of the number of parameters (or worse), and the storage requirements grow as the square of the number of parameters – undesirable properties where the number of parameters is of the order of 10^4.

Historically, the most successful optimization methods for macromolecular structures have been first-order methods. This is beginning to change as multi-gigabyte memories are becoming more common on computers and processor speeds are in the gigahertz range. At this time, there are no widely used refinement programs that run effectively on multiprocessor systems, although there are no theoretical barriers to writing such a program.

18.1.8.1. *Solving the refinement equations*

Methods for solving the refinement equations are described in *IT* C Chapters 8.1 to 8.5 and in many texts. Prince (1994) provides an excellent starting point. There are two commonly used approaches to finding the set of parameters that minimizes equation (18.1.4.1). The first is to treat each observation separately and rewrite each term of (18.1.4.1) as

$$w_i[y_i - f_i(\mathbf{x})] = w_i \sum_{j=1}^{N} \left(\frac{\partial f_i(\mathbf{x})}{\partial x_j}\right)(x_j^0 - x_j), \qquad (18.1.8.1)$$

where the summation is over the N parameters of the model. This is simply the first-order expansion of $f_i(\mathbf{x})$ and expresses the hypothesis that the calculated values should match the observed values. The system of simultaneous *observational equations* can be solved for the parameter shifts provided that there are at least as many observations as there are parameters to be determined. When the number of observational equations exceeds the number of parameters, the least-squares solution is that which minimizes (18.1.4.1). This is the method generally used for refining small-molecule crystal structures, and increasingly for macromolecular structures at atomic resolution.

18.1.8.2. *Normal equations*

In matrix form, the observational equations are written as

$$\mathbf{A}\Delta = \mathbf{r},$$

where $\mathbf{A}$ is the M by N matrix of derivatives, Δ is the parameter shifts and $\mathbf{r}$ is the vector of residuals given on the left-hand sides of equation (18.1.8.1). The *normal equations* are formed by multiplying both sides of the equation by $\mathbf{A}^T$. This produces an N by N square system, the solution to which is the desired least-squares solution for the parameter shifts.

$$\mathbf{A}^T\mathbf{A}\Delta = \mathbf{A}^T\mathbf{r} \text{ or } \mathbf{M}\Delta = \mathbf{b},$$

$$m_{ij} = \sum_{k=1}^{M} w_k \left(\frac{\partial f_k(\mathbf{x})}{\partial x_i}\right)\left(\frac{\partial f_k(\mathbf{x})}{\partial x_j}\right),$$

$$b_i = \sum_{k=1}^{M} w_k[y_k - f_k(\mathbf{x})]\left(\frac{\partial f_k(\mathbf{x})}{\partial x_i}\right).$$

Similar equations are obtained by expanding (18.1.4.1) as a second-order Taylor series about the minimum $\mathbf{x}_0$ and differentiating.

$$\Phi(\mathbf{x}-\mathbf{x}_0) \approx \Phi(\mathbf{x}_0) + \left\langle \left(\frac{\partial\Phi}{\partial x_i}\right)_{\mathbf{x}_0} \middle| (\mathbf{x}-\mathbf{x}_0) \right\rangle + \frac{1}{2}\left\langle (\mathbf{x}-\mathbf{x}_0) \middle| \left(\frac{\partial^2\Phi}{\partial x_i \partial x_j}\right)_{\mathbf{x}_0} \middle| (\mathbf{x}-\mathbf{x}_0) \right\rangle,$$

$$\left|\left(\frac{\partial\Phi}{\partial\mathbf{x}}\right)\right\rangle \approx \left|\left(\frac{\partial^2\Phi}{\partial x_i \partial x_j}\right)_{\mathbf{x}_0}\right| (\mathbf{x}-\mathbf{x}_0)\right\rangle.$$

The second-order approximation is equivalent to assuming that the matrix of second derivatives does not change and hence can be computed at $\mathbf{x}$ instead of at $\mathbf{x}_0$.

18.1.8.3. *Choice of optimization method*

First-order methods are generally the most economical for macromolecular problems. The most general approach is to treat the problem as a non-linear optimization problem from the beginning. This strategy is used by *TNT* (Tronrud *et al.*, 1987; Tronrud, 1997) and by *X-PLOR* (Kuriyan *et al.*, 1989), although by very different methods.

TNT uses a preconditioned conjugate gradient procedure (Tronrud, 1992), where the preconditioning function is the second derivatives of the objective function with respect to each parameter. In other words, at each step the parameters are normalized by the curvature with respect to that parameter, and a normal conjugate gradient step is taken. This has the effect that stiff parameters, which have steep derivatives, are scaled down, while soft parameters (such as *B* factors), which have small derivatives, are scaled up. This greatly increases both the rate and radius of convergence of the method.

X-PLOR (and its intellectual descendent, *CNS*) (Chapter 18.2 and Section 25.2.3) uses a simulated annealing procedure that derives sampling points by molecular dynamics. Simulated annealing is a process by which the objective function is sampled at a new point in parameter space. If the value of the objective function at the new point is less than that at the current point, the new point becomes the current point. If the value of the objective function is greater at the new point than at the current point, the Boltzmann probability $\exp(-\Delta E/kT)$ of the difference in function values ΔE is compared to a random number. If it is less than the random number, the new point is accepted as the current point; otherwise it is rejected. This process continues until a sufficiently deep minimum is found that the sampling process never leaves that region of parameter space. At this point the 'temperature' in the Boltzmann factor is reduced, which lowers the probability that the current point will move out of the region. This produces a finer search of the local region. The cooling process is continued until the solution has been restricted to a sufficiently small region. There are many variations of the strategy that affect the rate of convergence and the completeness of sampling. The primary virtue of simulated annealing is that it does not become trapped in shallow local minima. Simulated annealing can be either a zero-order method or a first-order method, depending on the strategy used to generate new sampling points. *X-PLOR* treats the fit to the diffraction data as an additional energy term, and the gradient of that 'energy' is treated as a force. This makes it a first-order method.

The first widely available macromolecular refinement program, *PROLSQ* (Konnert, 1976), uses an approximation to the second-order problem in which the matrix is greatly simplified. The parameters for each atom are treated as a small block on the diagonal of the matrix, and the off-diagonal blocks for pairs of atoms related by geometric constraints are also filled in. The sparse set of linear equations is then solved by an adaptation of the method of conjugate gradients.

The most comprehensive refinement program available for macromolecules is the same as the most comprehensive program available for small molecules – *SHELXL*98 (Sheldrick, 1993; see also Section 25.2.10). The primary adaptations to macromolecular problems have been the addition of conjugate gradients as an optimization method for cases in which the full matrix will not fit in the available memory and facilities to process the polymeric models required for macromolecules.

18.1.8.4. *Singularity in refinement*

Unless there are more linearly independent observations than there are parameters to fit them, the system of normal equations has no solution. The inverse of the matrix does not exist. Second-order methods fail in these circumstances by doing the matrix equivalent of dividing by zero. However, the objective function is still defined and has a defined gradient at all points. First-order methods will find a point at which the gradient is close to zero, and zero-order methods will still find a minimum value for the objective function. The difficulty is that the points so found are not unique. If one computes the eigenvalues and eigenvectors of the matrix of normal equations, one will find in this case that there are some eigenvalues that are very small or zero. The eigenvectors corresponding to these eigenvalues define sets of directions in which the parameters can be moved without affecting the value of the objective function. This region of the parameter space simply cannot be determined by the available data. The only recourses are to modify the model so that it has fewer parameters, add additional restraints to the problem, or collect more data. The real hazard with this situation is that the commonly used refinement methods do not detect the problem. Careful use of cross validation and keeping careful count of the parameters are the only remedy.

18.1.9. Evaluation of the model

Macromolecular model refinement is a cyclic process. No presently known refinement algorithm can remove all the errors of chain tracing, conformation, or misinterpretation of electron density. Other methods must be interspersed with refinement to help remove model errors. These errors are detected by basic sanity checks and the use of common sense about the model. This topic is discussed comprehensively in Part 21 and in Kleywegt (2000).

18.1.9.1. *Examination of outliers in the model*

Refinement-program output listings will normally provide some information on atoms that are showing non-standard bond lengths, bond angles or *B* factors. In addition, there is other software which can help identify non-standard or unusual geometry, such as *PROCHECK* (Laskowski *et al.*, 1993) and *WHAT IF* (Vriend, 1990). These are very useful in identifying questionable regions of structure but should not be completely relied on to identify errors or how the molecular models may be improved. Overall, the constraints in the model must be satisfied exactly, and the restraints should have a statistically reasonable distribution of deviations from the ideal values.

18.1.9.2. *Examination of model electron density*

Refinement of the model to improve the agreement between the observed and calculated diffraction data and the associated calculated phases should result in improved electron-density and ΔF maps. Unexplained features in the electron-density map or difference map are a clear indication that the model is not yet complete or accurate. Careful examination of the Fourier maps is essential. Interactive graphics programs such as *XtalView* (McRee, 1993) and *O* contain a number of analysis tools to aid in the identification of errors in the models.

There are several different types of Fourier maps that can be useful in the correction of the models. This topic is discussed extensively in Chapter 15.2. Usual maps include F_o maps, ΔF maps and $(nF_o - mF_c)$ maps. The Fourier coefficients used to compute the maps should be weighted by estimates of the degree of bias as described in Chapter 15.2. While ΔF maps are very useful in highlighting areas in the maps that reflect the greatest difference between the F_o's and F_c's in Fourier space, they do not show the electron density of the unit cell. Positive and negative regions of a ΔF map may be the result of positional errors of an atom or group of atoms, *B*-factor errors, completely misplaced atoms or missing atoms. F_o maps show the electron density but are biased by the current model. A $(2F_o - F_c)$ map is a combination of an F_o map

and a ΔF map which results in a map better showing the changes due to errors. Some investigators prefer using further amplified ΔF contributions by using a $(3F_o - 2F_c)$ map or higher-order terms.

The contribution of the disordered solvent continuum has been discussed previously. Macromolecular crystals also contain significant quantities of discrete or partially discrete solvent molecules (*i.e.* water). Care needs to be taken in adding solvent to a model. Errors in models generate peaks in Fourier maps that can be interpreted as solvent peaks. Hence, adding solvent peaks too early in the refinement process may, in fact, lead to model errors. Automatic water-adding programs are becoming more common; examples include *SHELXL*98 and *ARP*/*wARP* (Lamzin & Wilson, 1997). These programs check if the waters are with in reasonable bonding distances of hydrogen-bonding atoms. There is a distribution of solvent molecules ranging from ones with low *B* factors at unit occupancy to ones with very large *B* factors. Various criteria are used to decide on a cutoff in the discrete solvent contribution. A rule of thumb for ambient-temperature data sets is frequently about one solvent molecule per residue in a protein molecule. As more data are being collected at cryogenic temperatures, this ratio is tending to go up. Noise is being fitted if too many peaks in a ΔF map are being assigned as solvent molecules. This can also contribute to reducing *R* factors on incorrect models. Solvent sites may not be fully occupied. Because of the large *B* factors and limited range of the diffraction data, the *B* factors and occupancy are highly correlated. Refinement of occupancy does not usually contribute either to improving a model or to reduction of *R* factors in structures with up to 2.0 Å resolution data. Beyond 1.5 Å data, it may be possible to refine solvent water occupancies and *B* factors. At even higher resolution, some programs, such as *SHELXL*98, provide anisotropic refinement methods which may further improve the solvent model while reducing *R* factors including $R_{\rm free}$.

18.1.9.3. *R and* $R_{\rm free}$

Cross validation is a powerful tool for avoiding over-interpretation of the data by a too elaborate model. The introduction of cross validation to crystallography (Brünger, 1992) has been responsible for significant improvement in the quality of structure determinations. A subset of the reflections, chosen randomly, is segregated and not used in the refinement. If the model is correct and the only errors are statistical, these reflections should have an *R* factor close to that of the reflections used in the refinement. Changes to the model should affect both *R* and $R_{\rm free}$ similarly. Kleywegt & Jones (1997) have pointed out that it is necessary to treat the selection of free reflections very carefully in the presence of noncrystallographic symmetry.

18.1.10. Conclusion

It is always important to bear in mind that macromolecular crystal structures are models intended to explain a particular set of observations. Statistical measures can determine how well the model explains the observations, but cannot say whether the model is true or not. The distinction between precision and accuracy must always be kept in mind. The objective should not be simply to obtain the best fit of a model to the data, but, in addition, to find all of the ways in which a model does *not* fit the data and correct them. Until the day when all crystals diffract to atomic resolution, the primary objective of refinement of the models will be to determine just how well the structures are or are not determined.

$$\sigma_{\mathrm{LS,\,Luzz}}(r) = 1.33(N_i/p)^{1/2}[R(s_m)/s_m], \qquad (18.5.8.3)$$

where $R(s_m)$ is the value of R at some value of $s = s_m$ on the selected Luzzati curve. Equation (18.5.8.3) provides a means of making a very rough statistical estimate of error for an atom with $B = B_{\mathrm{avg}}$ (the average B for fully occupied sites) from a plot of R *versus* $2\sin\theta/\lambda$.

The corresponding equation involving R_{free} is

$$\sigma_{\mathrm{LS,\,Luzz}}(r) = 1.33(N_i/n_{\mathrm{obs}})^{1/2}[R_{\mathrm{free}}(s_m)/s_m]. \qquad (18.5.8.4)$$

18.5.8.3. *Comments on Luzzati plots*

Protein structures always show a great range of B values. The Luzzati theory effectively assumes that all atoms have the same B. Nonetheless, the Luzzati method applied to high-angle data shells does provide an upper limit for $\langle\Delta r\rangle$ for the atoms with low B. It is an upper limit since experimental errors and model imperfections are not allowed for in the theory.

Low-resolution structures can be determined validly by using restraints, even though the number of diffraction observations is less than the number of atomic coordinates. The Luzzati method, based preferably on R_{free}, can be applied to the atoms of low B in such structures. As the number of observations increases, and the resolution improves, the Luzzati $\langle\Delta r\rangle$ increasingly overestimates the true $\sigma(r)$ of the low-B atoms.

In the use of Luzzati plots, the method of refinement, and its degree of convergence, is irrelevant. A Luzzati plot is a statement for the low-B atoms about the maximum errors associated with a given structure, whether converged or not.

References

18.1

Adams, P. D., Pannu, N. S., Read, R. J. & Brunger, A. T. (1999). *Extending the limits of molecular replacement through combined simulated annealing and maximum-likelihood refinement*. *Acta Cryst.* D**55**, 181–190.

Booth, A. D. (1946*a*). *A differential Fourier method for refining atomic parameters in crystal structure analysis*. *Trans. Faraday Soc.* **42**, 444–448.

Booth, A. D. (1946*b*). *The simultaneous differential refinement of co-ordinates and phase angles in X-ray Fourier synthesis*. *Trans. Faraday Soc.* **42**, 617–619.

Bricogne, G. (1997). *Bayesian statistical viewpoint on structure determination: basic concepts and examples*. *Methods Enzymol.* **276**, 361–423.

Brünger, A. T. (1992). *Free R-value – a novel statistical quantity for assessing the accuracy of crystal structures*. *Nature (London)*, **355**, 472–475.

Busing, W. R., Martin, K. O. & Levy, H. A. (1962). *ORFLS.* Report ORNL-TM-305. Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA.

Diamond, R. (1971). *A real-space refinement procedure for proteins*. *Acta Cryst.* A**27**, 436–452.

Hendrickson, W. A. (1985). *Stereochemically restrained refinement of macromolecular structures*. *Methods Enzymol.* **115**, 252–270.

Hughes, E. W. (1941). *The crystal structure of melamine. J Am. Chem. Soc.* **63**, 1737–1752.

International Tables for Crystallography (1999). Vol. C. *Mathematical, physical and chemical tables*, edited by A. J. C. Wilson & E. Prince. Dordrecht: Kluwer Academic Publishers.

International Tables for Crystallography (2001). Vol. B. *Reciprocal space*, edited by U. Shmueli. Dordrecht: Kluwer Academic Publishers.

Kleywegt, G. J. (2000). *Validation of protein crystal structures*. *Acta Cryst.* D**56**, 249–265.

Kleywegt, G. J. & Jones, T. A. (1997). *Model building and refinement practice. Methods Enzymol.* **277**, 208–230.

Konnert, J. H. (1976). *A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units*. *Acta Cryst.* A**32**, 614–617.

Konnert, J. H. & Hendrickson, W. A. (1980). *A restrained-parameter thermal-factor refinement procedure*. *Acta Cryst.* A**36**, 344–350.

Kuriyan, J., Brünger, A. T., Karplus, M. & Hendrickson, W. A. (1989). *X-ray refinement of protein structures by simulated annealing: test of the method on myohemerythrin*. *Acta Cryst.* A**45**, 396–409.

Lamzin, V. S. & Wilson, K. S. (1997). *Automated refinement for protein crystallography*. In *Macromolecular crystallography*, Part B, edited by C. C. & R. Sweet, 269–305. San Diego: Academic Press.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures*. *J. Appl. Cryst.* **26**, 283–291.

McRee, D. E. (1993). *Practical protein crystallography*, p. 386. San Diego: Academic Press.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method*. *Acta Cryst.* D**53**, 240–253.

Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). *Efficient anisotropic refinement of macromolecular structures using FFT*. *Acta Cryst.* D**55**, 247–255.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Incorporation of prior phase information strengthens maximum-likelihood structure refinement*. *Acta Cryst.* D**54**, 1285–1294.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood*. *Acta Cryst.* A**52**, 659–668.

Prince, E. (1994). *Mathematical techniques in crystallography and materials science*. 2nd ed. Berlin: Springer-Verlag.

Rice, L. M. & Brünger, A. T. (1994). *Torsion-angle dynamics – reduced variable conformational sampling enhances crystallographic structure refinement*. *Proteins Struct. Funct. Genet.* **19**, 277–290.

Sheldrick, G. M. (1993). *SHELXL93. Program for the refinement of crystal structures*. University of Göttingen, Germany.

Tronrud, D. E. (1992). *Conjugate-direction minimization: an improved method for the refinement of macromolecules*. *Acta Cryst.* **48**, 912–916.

Tronrud, D. E. (1996). *Knowledge-based B-factor restraints for the refinement of proteins*. *J. Appl. Cryst.* **29**, 100–104.

Tronrud, D. E. (1997). *The TNT refinement package*. *Methods Enzymol.* **277**, 306–318.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *An efficient general-purpose least-squares refinement program for macromolecular structures*. *Acta Cryst.* A**43**, 489–501.

Vriend, G. (1990). *WHAT IF: a molecular modeling and drug design program*. *J. Mol. Graphics*, **8**, 52–56.

Watenpaugh, K. D., Sieker, L. C., Herriott, J. R. & Jensen, L. H. (1972). *The structure of a non-heme iron protein: rubredoxin at 1.5 Å resolution*. *Cold Spring Harbor Symp. Quant. Biol.* **36**, 359–367.

Watenpaugh, K. D., Sieker, L. C., Herriott, J. R. & Jensen, L. H. (1973). *Refinement of the model of a protein: rubredoxin at 1.5 Å resolution*. *Acta Cryst.* B**29**, 943–956.

18.2

Abramowitz, M. & Stegun, I. (1968). *Handbook of mathematical functions. Applied mathematics series*, Vol. 55, p. 896. New York: Dover Publications.