Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Introduction to macromolecular refinement

Dale. E. Tronrud

Howard Hughes Medical Institute and Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA

Correspondence e-mail: dale@uoxray.uoregon.edu

Received 5 April 2004
Accepted 21 September 2004

The process of refinement is such a large problem in function minimization that even the computers of today cannot perform the calculations to properly fit X-ray diffraction data. Each of the refinement packages currently under development reduces the difficulty of this problem by utilizing a unique combination of targets, assumptions and optimization methods. This review summarizes the basic methods and underlying assumptions in the commonly used refinement packages. This information can guide the selection of a refinement package that is best suited for a particular refinement project.

1. Introduction

Refinement is the optimization of a function of a set of observations by changing the parameters of a model.

This is the definition of macromolecular refinement at its most basic level. To understand refinement, we need to understand the definitions of its various parts. The four parts are 'optimization, 'a function', 'observations' and 'the parameters of a model'.

While formally different topics, these concepts are tightly connected. One cannot choose an optimization method without considering the nature of the dependence of the function on the parameters and observations. In some cases, one's confidence in an observation is so great that the parameters are devised to make an inconsistent model impossible. These observations are then referred to as constraints.

This paper will discuss each of these topics in detail. An understanding of each topic and their implementation in current programs will enable the selection of the most appropriate program for a particular project.

2. Observations

The 'observations' include everything known about the crystal prior to refinement. This set includes commonly noted observations, such as unit-cell parameters, structure-factor amplitudes, standardized stereochemistry and experimentally determined phase information. In addition, other types of knowledge about the crystal, which are usually not thought about in the same way, include the primary structure of the macromolecules and the mean electron density of the mother liquor.

For a particular observation to be used in refinement, it must be possible to gauge the consistency of the model with this observation. Current refinement programs require that this measure be continuous. If a property is discrete, some

mathematical elaboration must be created to transform the measure of the model's agreement into a continuous function.

As an example consider chirality; the C^{α} atom of an amino acid is in either the D or the L configuration. It cannot be 80% L and 20% D. Since the agreement of a model to this piece of knowledge is discrete, the derivative of the agreement function is not informative. To allow the correction of this sort of error, most programs use some function in which the chirality is expressed as its sign (*e.g.* a chiral volume or improper dihedral angle). Since the additional information in these residual functions is simply the ideal bond lengths and angles, restraining chirality in this fashion causes geometrical restraints to be included in the refinement *via* two different routes. This duplication makes it difficult to assign proper weights to this information.

This problem is not encountered with most types of observations. Diffraction amplitudes, bond lengths and angles calculated from the model can be varied by small changes in the model's parameters.

Each observation should be accompanied by an indication of its confidence. If the errors in an observation follow a normal distribution then the confidence in the observation is indicated by the standard deviation ('σ') of that distribution. In more complicated situations, a complete description of the probability distribution will be required. Experimental phases are examples of this difficult class of observations. Their uncertainties can be quite large and multiple maxima are possible, as is the case with SIR phases.

2.1. Stereochemical restraints

When a diffraction data set is missing high-resolution reflections, the details of the molecule cannot be visualized. Fortunately, the molecule can be viewed as a set of bond lengths, bond angles and torsion angles, instead of the conventional view of a set of atoms floating in space (see Fig. 1). The advantage derived from this geometrical view of a structure is that the values of the bond lengths and angles and their standard uncertainties are known from high-resolution small-molecule structures (Allen, 2002).

To be honest, the most interesting aspects of a molecule are the angles of rotation about its single bonds. If the φ and ψ angles of the backbone of the polypeptide chain and the χ angles of the side chains were known, most of the questions about the structure could be addressed. The scatter of bond angles and planarity seen in accurate structures is large enough, however, that one cannot constrain a model to 'ideal' values. For example, if a peptide ω angle (the angle of rotation about the C and N atoms in the peptide bond) differs from the value that results in a planar peptide bond [as it can; see König *et al.* (2003) as one example of many] but is forced into a plane, the protein's backbone will be distorted over many residues to compensate for the error. Refinement with constrained bond lengths and angles was implemented in the early 1970s in Diamond's (1971) real-space refinement program, but was eventually abandoned, in part because of this problem.

Even though stereochemical constraints on bond lengths and angles do not work, this knowledge can still be applied as restraints. Constraints simplify the refinement problem by reducing the number of parameters. Restraints instead work by increasing the number of observations; a penalty is imposed for deviations from ideal stereochemistry, just as a penalty is imposed for deviations from observed structure-factor amplitudes.

The practice in the field is to track the ratio of the number of observations to the number of parameters; the larger the ratio, the better the quality of the result of the refinement. This useful metric is only valid when the observations are independent of each other and when the parameters are related to the different types of observations in roughly comparable ways.

As a trivial example of this issue, consider two data-collection strategies: (i) 10% of the reflections are each measured ten times and (ii) 100% of the reflections are each measured once. Clearly, the latter data set contains more information. While a reflection measured ten times is recorded with greater precision than one only measured once, it does not provide as much information as ten unique reflections. In acknowledgment of this difference, we average together the multiple measurements of each reflection and count 'merged' reflections instead of the total measurements.

With stereochemical restraints (*i.e.* observations), the situation is not as clear. The knowledge that a particular pair of atoms is separated by 1.3 Å is redundant with some of the diffraction data of 1.3 Å resolution and higher but independent from diffraction data of resolution much lower than 1.3 Å. In a refinement problem with 2 Å resolution diffraction data, this stereochemical restraint would count as one independent observation. In a problem with 0.9 Å data, it would be redundant to the diffraction data and probably would not add any useful additional information to the refinement. This information redundancy requires that the geometrical restraints be weighted relative to the diffraction restraints in an essentially empirical fashion. When the resolution of the diffraction data set is low, we want to give the geometrical restraints a high weight. When we have a high-resolution data set, we want to use a low weight, dropping to zero when the limit of diffraction is so high that the stereochemical restraints are providing no additional information. The probabilistic formulation used for maximum likelihood should adjust this weighting automatically, but in practice it is usually necessary to choose an empirical weighting factor to maintain reasonable stereochemistry for all parts of the model.

3. The parameters of the model

The general perception of the parameters of a molecular model are dominated by the PDB file format (Bernstein *et al.*, 1977). In this format a molecule is a collection of atoms, each defined by its location, a '*B*' factor and an occupancy. (Each atom also has an atomic type, but since these are not continuously variable they are not considered as parameters here.)

The *B* factor provides an estimate of an atom's vibration about its central position. The usual form is either to define the *B* factor as isotropic, meaning that the atom vibrates equally in all directions and can be visualized as lying within a sphere, or to define an anisotropic *B* factor, which describes vibration of the atom within an ellipsoid centered at the atomic coordinate. Six parameters are required to define such an ellipsoid (Stout & Jensen, 1989). The *B* factor is isotropic when the off-diagonal elements of this matrix are equal to zero and the diagonal elements are all equal to each other. Therefore, only one number is required to define an isotropic *B* factor.

In the traditional formulation, each atom is defined by (i) three numbers that give its location in the unit cell, (ii) either one number for an isotropic *B* factor or six numbers for an anisotropic *B* factor and (iii) one number for its occupancy. These numbers form the principle set of parameters for the model.

In a medium-sized protein there are about 2500 atoms. With five parameters for each atom there would be 12 500 parameters and with ten parameters per atom there would be 25 000 parameters. For such a protein, a diffraction data set to 2 Å resolution would contain about 22 000 reflections. Since the mathematical relationship between the structure factors and the model is nonlinear, macromolecular refinement will not produce useful results unless there are many times more reflections in the data set than parameters in the model. Clearly, a refinement of a model with anisotropic *B* factors at 2 Å resolution will be problematic and one with isotropic *B* factors is borderline.

(This difficulty is not restricted to molecules of a particular size. The larger the molecule the greater the number of parameters, but the unit cell will also increase in size, which increases the number of reflections at a given resolution. The ratio of observation to parameters essentially depends only on resolution, for all sizes of molecules. There is some effect arising from the solvent content of the cell, with large solvent-content cells resulting in relatively larger sets of reflections at the same resolution.)

At such resolutions something must be done to simplify the parameterization of the model (or increase the number of observations). Simplification can be achieved by imposing constraints on the parameters, forcing the model to be exactly consistent with the prior knowledge or recasting the parameters into some form where the lack of compliance is impossible. The creation of program code to implement these solutions can be very difficult. Some of the traditional solutions were devised because of limitations of time and computer technology and are not easily justified today.

The first parameter to go is the occupancy. Because the difference-map feature that results from an error in occupancy is very similar to that resulting from an error in an isotropic *B* factor, only quite high-resolution diffraction data can generate difference maps that have sufficient clarity to distinguish the two. Since the two parameters are linked in this fashion, it would be advantageous to eliminate one of them. Fortunately, most of the atoms in the crystal are chemically bonded together and have the same occupancy, which is very close to 1.0. Applying this knowledge as a constraint allows the model to be refined with one fewer parameter per atom.

While this simplification is appropriate for the atoms in the macromolecule, it is not for individual water molecules. It is quite likely that particular water molecules are not present with full occupancy, but the problem of discriminating between a low occupancy and a high *B* factor remains. The traditional solution is to again hold the occupancy fixed at 1.0. While this solution is not credible, it does allow the atom to refine to flatten the difference map and it lowers the *R* value almost as much as refining both parameters would. Because of this constraint, the *B* factor must be redefined to be a combination of motion and occupancy. This generalization in interpretation of the *B* factor is implicit in most macromolecular models, but is not clearly stated in the deposited models.

Unless the resolution of the diffraction data is very high, refinement of a model containing anisotropic *B* factors results in structures that are physically unreasonable. To avoid this absurdity, refinement is performed with isotropic *B* factors. This choice is not made because the motions of the atoms are

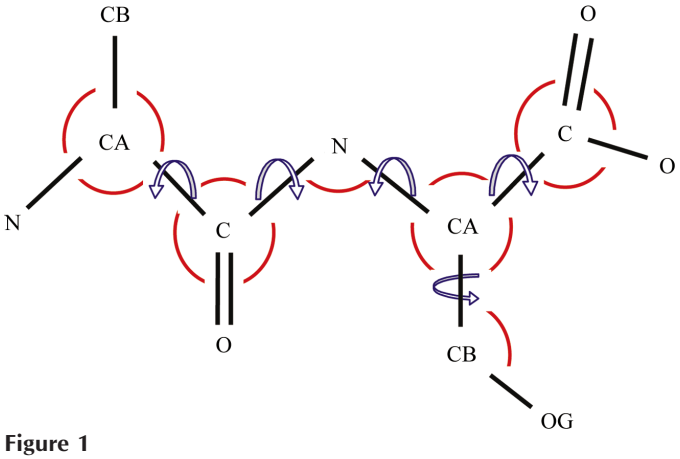

Figure 1
Stereochemical restraints in a dipeptide. This figure shows the bonds, bond angles and torsion angles for the dipeptide Ala-Ser. Black lines indicate bonds, red arcs indicate bond angles and blue arcs indicate torsion angles. The values of the bond lengths and bond angles are, to the precision required for most macromolecular-refinement problems, independent of the environment of the molecule and can be estimated reliably from small-molecule crystal structures. The values of most torsion angles are influenced by their environment and, although small-molecule structures can provide limits on the values of these angles, they cannot be determined uniquely without information specific to this crystal. It is instructive to note that this example molecule contains 12 atoms and requires 36 degrees of freedom to define their positions (12 atoms times three coordinates for each atom). The molecule contains 11 bonds, 14 bond angles and five torsion angles, which together define 30 degrees of freedom. The unaccounted-for degrees of freedom are the six parameters that define the location and orientation of the entire dipeptide. This result is general; the sum of the number of bonds, the number of bond angles, the number of torsion angles and six will always be three times the number of atoms. Other stereochemical restraints, such as chiral volume and planarity, are redundant. For example, the statement that the carbonyl C atom and the atoms that bond to it form a planar group is equivalent to saying that the three bond angles around the carbonyl C atom sum to 360°. These types of restraints are added to refinement packages to compensate for their (incorrect) assumption that deviations from ideality for bond angles are independent of each other.

actually believed to be isotropic, but simply to limit the number of parameters. The result is the paradox that the crystals that probably have the largest anisotropic motions are modeled with isotropic B factors.

3.1. Rigid-body parameterization

One common restructuring of the standard set of parameters is that performed in rigid-body refinement. When there is an expectation that the model consists of a molecule whose structure is essentially known but whose location and orientation in the crystal are unknown, the parameters of the model are refactored. The new parameters consist of a set of atomic positions specified relative to an arbitrary coordinate system and up to six parameters to specify how this coordinate system maps onto the crystal: up to three to describe a translation of the molecule and three to define a rotation. The traditional set of coordinates is calculated from this alternative factorization with the equation

$$\mathbf{x_t} = \mathbf{R}(\theta_1, \theta_2, \theta_3)\mathbf{x}_r + \mathbf{t},$$

where $\mathbf{x_t}$ is the positions of the atoms in the traditional crystallographic coordinate system, $\mathbf{R}(\theta_1, \theta_2, \theta_3)$ is the rotation matrix, which rotates the molecule into the correct orientation, and $\mathbf{t}$ is the translation required to place the properly orientated molecule into the unit cell.

In principle, all of these parameters could be refined at the same time, but refinement is usually performed separately for each parameter class, because of their differing properties. The values of the orientation and location parameters are defined by diffraction data of quite low resolution and the radius of convergence of the optimization can be increased by ignoring the high-resolution data. In addition, in those cases where rigid-body refinement is used, one usually knows the internal structure of the molecule quite well, while the location and orientation are more of a mystery.

For this reason, molecular replacement can be considered to be a special case of macromolecular refinement. Since the internal structure of the molecule is known with reasonable certainty, one creates a model parameterized as the rigid-body model described above. One then 'refines' the orientation and location parameters. Since this is a small number of parameters and no good estimate for starting values exists, one uses search methods to locate an approximate solution and gradient descent optimization to fine-tune to orientation parameters.

The principal drawback of the rigid-body parameterization is that macromolecules are not rigid bodies. If the external forces of crystal packing differ between the crystal where the model originated and the crystal where the model is being placed, then the molecule will be deformed. Optimizing the rigid-body parameters alone cannot result in a final model for the molecule.

3.2. NCS-constrained parameterization

When the asymmetric unit of a crystal contains multiple copies of the same type of molecule and the diffraction data are not of sufficient quantity or quality to define the differences between the copies, it is useful to constrain the non-crystallographic symmetry (NCS) to perfection. In such a refinement the parameterization of the model is very similar to that of rigid-body refinement. There is a single set of atomic parameters [positions, B factors and occupancies (usually constrained equal to unity)] for each type of molecule and an orientation and location (six parameters) for each copy.

As with rigid-body refinement, the orientation and location parameters are refined separately from the internal structure parameters. Firstly, the orientation and location parameters are refined at low (typically 4 Å) resolution while the atomic parameters are held fixed. The atomic parameters are then refined against all the data while the external parameters are held fixed.

Both rigid-body refinement and constrained NCS refinement have a problem with parameter counts. When the location and orientation parameters are added to create a rigid-body model, the total number of parameters in the model increases by six, but the new parameters are redundant. For example, the entire molecule can be moved up the y axis by changing the rigid-body y coordinate or by adding a constant to all the y coordinates of the individual atoms. This type of redundancy does not create a problem when one class of parameters are held fixed. If all the parameters are refined at once, however, it is at best confusing and at worst (when the optimization method uses second derivatives) it will cause numerical instabilities.

The constrained NCS parameterization has the same shortcoming as rigid-body parameterization. Each copy of the macromolecule experiences a different set of external forces as a result of their differing crystal contacts and it is expected that the copies will respond by deforming in differing ways. The constraint that their internal structures be identical precludes the model from reflecting these differences. If the diffraction data are of sufficient resolution to indicate that the copies differ but are not high enough to allow refinement of unconstrained parameters (without explicit consideration of NCS), then the model will develop spurious differences between the copies (Kleywegt & Jones, 1995).

Relaxing the constraints and implementing NCS restraints is the usual solution chosen to overcome this problem. Most implementations of NCS restraints continue to assume that the molecules are related by a rigid-body rotation and translation, except for the random uncorrelated displacements of individual atoms. If two molecules differ by an overall bending, the NCS restraints will impede the models from matching that shape. The program *SHELXL* (Sheldrick & Schneider, 1997) contains an option for restraining NCS by suggesting that the torsion angles of the related molecules be similar, instead of the positions of the atoms being similar after rotation and translation. By removing the rigid-body assumption from its NCS restraints, this program allows deformations that are suppressed by other programs.

3.3. Torsion-angle parameterization

The replacement of atomic coordinates by torsion angles dramatically reduces the total number of parameters (see Fig. 1). This decrease is advantageous when the resolution of the diffraction data is quite low (lower than 3 Å). At these resolutions there are many fewer reflections to define the values of parameters in the traditional model. Even with the addition of bond-length and angle information as restraints, these models tend to get stuck in local minima or overfit the data.

Increasing the weight on the stereochemical restraints to compensate for the lack of diffraction data does not work well because the sizes of the off-diagonal elements of the normal matrix also increase in significance (see §5.2.3), which causes optimization methods that ignore these elements to become ineffective.

Simulated annealing also has difficulty accommodating high weights on bond lengths and angles (Rice & Brünger, 1994). When the 'force constant' of a bond is large, the bond's vibrational frequency increases. The highest frequency motion determines the size of the time step required in the slow-cooling molecular-dynamics calculation, so increasing the weight on stereochemistry greatly increases the amount of time taken by the slow-cooling calculation.

The programs commonly used to refine models at these low resolutions [*X-PLOR* (Brünger *et al.*, 1987) and *CNS* (Brünger *et al.*, 1998)] use simulated-annealing and gradient-descent methods of optimization. Optimization methods that use the off-diagonal elements of the normal matrix are not used in these circumstances, because their radii of convergence are not large enough to correct the errors that typically are found in low-resolution models.

One solution to the problem of large stereochemistry weights is to choose a parameterization of the model where the bond lengths and angles simply cannot change. If the parameters of the model are the angles of rotation about the single bonds, the number of parameters drops considerably and there is no need for a stereochemical weight (it is effectively infinite). There are on average about five torsion angles and about eight atoms per amino acid. Changing from an atomic model to a torsion-angle model will replace 24 positional parameters with five angular parameters. This nearly fivefold reduction in parameters greatly improves the observation-to-parameter ratio, in addition to improving the power of simulated-annealing and gradient-descent optimization. The nature of torsion-angle parameters makes the implementation of their refinement much more difficult than that of the other parameters described here. When working with atomic positions, for example, one can estimate the shifts to be applied by considering the improvement in the residual function by moving each atom in turn, holding the other atoms fixed in space. This form of calculation cannot be performed with torsion-angle parameters. If the first main-chain torsion angle is varied, the majority of the molecule is moved out of density and any amount of shift is rejected. The variation of a torsion angle can only lead to improvement if other torsion angles are simultaneously varied in compensatory fashion. The most flexible solution to this problem to date is described by Rice & Brünger (1994).

This parameterization is the same as that of Diamond (1971) (although the actual method of optimization is quite different) and suffers the same limitations. If there are real deviations from ideal bond angles, a model that ignores that possibility will be distorted. The modern implementation in *CNS* (Rice & Brünger, 1994) is not seriously affected by this problem for two reasons. Firstly, these refinements are performed at rather low resolution and the distortions are not as significant and secondly, the torsion-angle refinement is followed by conventional refinement after the large errors have been corrected.

3.4. TLS *B*-factor parameterization

Probably the most significant inappropriate constraint applied generally to protein models is the isotropic *B* factor. It is quite certain that atoms in crystals that diffract to resolutions lower than 2 Å move anisotropically and yet they are routinely modeled as isotropic. While the excuse for this choice is the undeniable need to reduce the number of parameters in the model, this clearly is not a choice likely to improve the fit of the model to the data.

Schomaker & Trueblood (1968) described a parameterization that allows the description of anisotropic motion with many fewer parameters than an independent anisotropic *B* factor for each atom. This parameterization is called TLS (translation, libration and screw). In this system the motion of a group of atoms is described by three matrices, one for a purely translational vibration of the group, a second for libration (or wobbling) of the group about a fixed point and a third for a translation and libration that occurs in concert. The explicit assumption of TLS *B* factors is that the group of atoms move as a rigid unit. More complicated motions can be modeled by nesting several TLS groups within a larger group, creating a tree-like data structure.

TLS *B* factors are difficult to implement as parameters in a refinement program. The programs *RESTRAIN* (Haneef *et al.*, 1985) and, more recently, *REFMAC* (Murshudov *et al.*, 1997; Winn *et al.*, 2001) include the option of refining TLS *B* factors.

In the TLS formalism, 20 parameters are used to describe the motion of the entire group of atoms. Since the anisotropic *B* of one atom requires six parameters, any TLS group composed of more than three atoms results in a decrease in the total number of parameters. Of course, a large number of small TLS groups will not reduce the parameter count very much and will only be refinable with higher resolution data than a TLS model containing large groups. Then again, a TLS model composed of large groups might not be able to mimic the set of anisotropic *B* factors required to fit the data.

In the absence of a related structure refined with anisotropic *B* factors at atomic resolution, it is difficult to objectively define rigid groups larger than side chains with aromatic rings.

4. The function

In crystallographic refinement, three functions are commonly used. They are the empirical energy function, the least-squares residual and maximum likelihood.

4.1. Empirical energy

The idea that the best model of a protein would be that with the lowest energy has been used since the early 1970s (for an example, see Levitt, 1974). To a person with a background in biochemistry, such a measure is quite intuitive. The program will give the difference between two conformations or two models in kcal mol^{-1}, which is a familiar unit.

There are two principal problems with this function as a refinement residual. The first problem is that it has been impossible so far to devise an empirical energy function that is accurate enough to reproduce experimental results. If the function is not reliable, the models generated using it cannot be trusted either. The second problem is that there is no statistical theory underlying this function. None of the vast array of mathematical tools developed in other fields can be applied to an analysis of the quality of the model or the nature of its remaining errors.

While the refinement packages *X-PLOR* (Brünger *et al.*, 1987) and *CNS* (Brünger *et al.*, 1998) use the language of energy in their operation, the actual function used is closer to one of the other two functions. It is important to remember that these programs are not even attempting to calculate 'energies' that relate to binding energies and stability.

4.2. Least squares

Least squares is the simplest statistical method used in macromolecular refinement. Like empirical energy, the history of least squares in macromolecular structure determination extends back to the 1970s (Konnert, 1976) and the approach continues to be used today.

The least-squares residual function is

$$f(\mathbf{p}) = \sum_{i}^{\text{all data}} [Q_o(i) - Q_c(i, \mathbf{p})]^2 / \sigma_o(i)^2, \qquad (1)$$

where $Q_o(i)$ and $\sigma_o(i)$ are the value and standard deviation for observation number i. $Q_c(i, \mathbf{p})$ is the model's prediction for observation i using the set of model parameters $\mathbf{p}$. The larger the difference between the observation and the model's prediction, the worse the model. The more precisely we know an observation, the more important that observation becomes in the overall sum. One varies the parameters of the model to find a set that gives the lowest sum of deviants.

The values of the parameters found by minimizing this function are those that have the smallest individual standard deviation or the smallest probable error (Mandel, 1984). This statement is only true, however, if the assumptions of the method are correct. The assumptions of least squares are that the errors in the observations obey a normal distribution with completely known ('observed') variances and that, given perfect observations and the best parameters, the model would predict the observations perfectly.

In recent years it has been shown (Bricogne, 1988, 1993; Read, 1990) that these assumptions are incorrect in many refinement problems. The simplest example occurs when the model is incomplete, say missing a domain. With an imperfect model of this type it is impossible for any set of parameters to reproduce all the observations. The refinement function must account for the unknown contribution of the unmodeled part of the molecule and least squares cannot do that.

4.3. Maximum likelihood

To construct a refinement function that does not make the assumptions of least squares, one must generalize the method. Such a generalization is called maximum likelihood. Currently, maximum-likelihood options are available in the programs *CNS* (Brünger *et al.*, 1998), *REFMAC* (Murshudov *et al.*, 1997) and *BUSTER*/*TNT* (Bricogne & Irwin, 1996; Tronrud *et al.*, 1987). These programs are listed in order of increasing sophistication of their implementation of maximum likelihood.

Maximum likelihood is a generalized statistical framework for estimating the parameters of a model on the basis of observations (Bricogne, 1997; Sivia, 1996, p. 64). This approach differs from least squares in that maximum likelihood can accommodate observations with uncertainties of arbitrary character and model parameters whose values are also expected to have such uncertainties.

While the maximum-likelihood method is completely general, macromolecular refinement is such a difficult problem that no computer can perform a likelihood refinement in complete generality. The authors of each computer program must make particular assumptions about the nature of the uncertainties in the observations and the parameters of the final model in order to produce a program that will produce a result in a reasonable amount of time.

While least squares is rather simple and is usually implemented similarly in all programs, maximum likelihood depends critically on a detailed model of how errors are distributed and the consequences of these errors. Each implementation of maximum likelihood makes it own set of assumptions and one may work better than another in any particular problem.

4.3.1. Overview of Bayesian inference. Maximum likelihood itself is an approximation of the general Bayesian inference procedure (Sivia, 1996). Bayesian inference is a means of combining all information known about a problem in a completely general fashion.

One starts by calculating, for every combination of values of the parameters of the model, how probable that set of parameters is when all of the information known prior to the current experiment is considered. In crystallographic refinement, this information would include basic properties (*e.g.* that anisotropic B factors must be positive definite and isotropic B factors must be positive), stereochemical information (*e.g.* atom CA of a particular residue is about 1.52 Å from atom C and its isotropic B factor is about 4 $Å^2$ smaller) and various conventions (*e.g.* that at least one atom of each

molecule should lie in the conventional asymmetric unit of the unit cell). This probability distribution is named the prior distribution.

The second probability distribution is called the likelihood distribution. This distribution contains, for every combination of values for the parameters of the model, the probability that the experiment would have turned out as it did, assuming that set of values was correct. If the predicted outcome of the experiment for a particular set of values differs from the actual experimental results by much more than the expected uncertainty in both the measurements and the ability of the model to predict, then the probability is quite low.

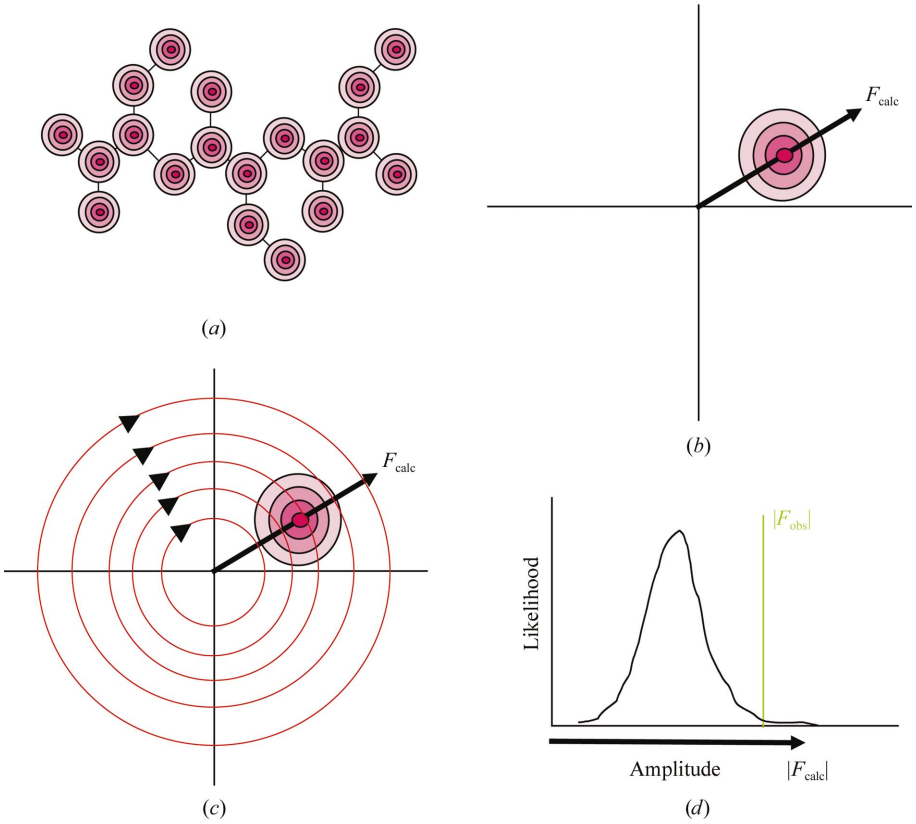

Figure 2
Probability distributions for one reflection in the maximum-likelihood worldview. (*a*) The maximum-likelihood method begins with the assumption that the current structural model itself contains errors. This figure represents the probability distributions of the atoms in the model. Instead of a single location, as assumed by the least-squares method, there is a cloud of locations that each atom could occupy. While not required by maximum likelihood, the computer programs available today assume that the distributions of positions are normal and have equal standard deviations [the value of which is defined to be that value which optimizes the fit of the model to the test set of diffraction data (Pannu & Read, 1996; Brünger, 1992)]. (*b*) The distribution of structures shown in (*a*) results in a distribution of values for the complex structure factors calculated from that model. An example of one of the distributions is shown. The value of the structure factor calculated from the most probable model is labeled F_{calc}. The nonlinear relationship between real and reciprocal space causes this value not to be the most probable value for the structure-factor distribution. As shown by Read (1986), the most probable value has the same phase as F_{calc} but has an amplitude that is only a fraction of that of F_{calc}. This fraction, conventionally named D, is equal to unity when the model is infinitely precise and is zero when the model is infinitely uncertain. The width of the distribution, named σ_{calc}, also arises from the coordinate uncertainty and is large when D is small and zero when D is unity. The recognition that the structure factor calculated from the most probable model is not the most probable value for the structure factor is the key difference between least squares and the current implementations of maximum likelihood. (*c*) In refinement without experimental phase information, the probability distribution of the calculated value of the structure factor must be converted to a probability distribution of the amplitude of this structure factor. This transformation is accomplished by mathematically integrating the two-dimensional distribution over all phase angles at each amplitude. This integral is represented by a series of concentric circles. (*d*) The probability distribution for the amplitude of the structure factor. The bold arrow below the horizontal axis represents the amplitude of F_{calc}, calculated from the most probable model. As expected, the most probable amplitude is smaller than $|F_{calc}|$. With this distribution the likelihood of any value for $|F_{obs}|$ can be evaluated, but more importantly one can calculate how to modify the model to increase the likelihood of $|F_{obs}|$. In this example, the likelihood of $|F_{obs}|$ is improved by either increasing $|F_{calc}|$ or increasing the precision of the model. This action is the opposite of the action implied by the least-squares analysis of Fig. 3.

Any set of values are only worth considering if they have high probabilities in both distributions. Therefore, the two distributions are multiplied to generate a new probability distribution, called the posterior probability, which includes all of the information about the values of the parameters. If the posterior distribution contains a single well defined peak, that peak is the solution. The width of the peak would indicate how precisely these values are known. If there are multiple peaks of about the same height or if there is a single peak that is diffuse, then the experiment has not produced information sufficient to distinguish between the various possible sets. In this case, one can study the posterior probability distribution to help design the next experiment.

Unfortunately, calculating the prior and likelihood distributions for all combinations of values for the parameters of a macromolecular model is well beyond the capability of current computers.

As described here, the posterior probability is not normalized. To normalize it, one must divide it by the probability of the experimental data given what was known about such data prior to the experiment. In the case of diffraction data this information would include Wilson (1942, 1949) statistics and the non-negativity of structure-factor amplitudes. Since we have one set of experimental data this normalization factor is simply one number and can be ignored without affecting the shape of the posterior probability distribution.

4.3.2. The maximum-likelihood approximation. The maximum-likelihood method depends on the assumption that the likelihood distribution has a single peak whose location is approximately known. This assumption allows one to ignore nearly all of

the volume of the distribution and concentrate on the small region near the starting model. Finding the values for the parameters that result in the greatest likelihood reduces to a function-optimization operation very similar in structure to that used by the least-squares refinement programs of the past. To increase this similarity, the negative logarithm of the likelihood function is minimized in place of maximizing the likelihood itself.

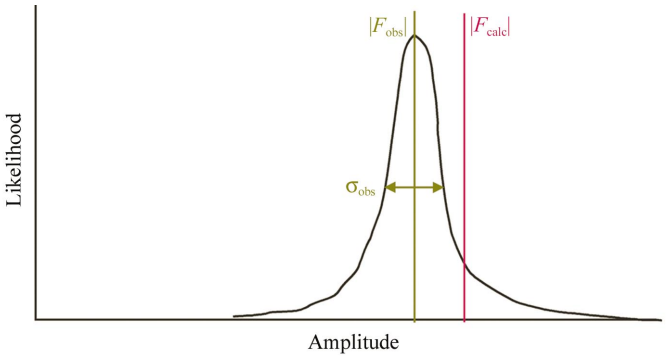

Figure 3
Probability distribution for one reflection in the least-squares worldview. In least-squares analysis it is assumed that the observed and calculated structure factors have exactly the same phase, so the only error to consider is in the magnitude of the observation. The true value of $|F_{obs}|$ is assumed to be represented by a one-dimensional Gaussian centered at its measured value and with a spread related to its estimated standard uncertainty, σ_{obs}. The calculated amplitude is assumed to have no spread at all. In this example, the parameters of the model should be modified to cause $|F_{calc}|$ to decrease.

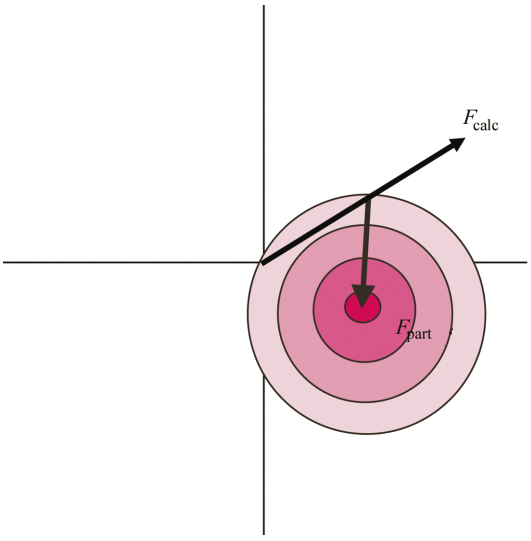

Figure 4
Probability distribution for maximum likelihood in the presence of unbuilt structure. This figure shows the probability distribution in the complex plane for the case where, in addition to the modeled parts of the crystal, there is a component present in the crystal for which an explicit model has not been built. This distribution is an elaboration of that shown in Fig. 2(*b*). That distribution is convoluted with the probability distribution of the structure factor calculated from the envelope where the additional atoms are believed to lie and weighted by the number of atoms in this substructure (which can be represented as a distribution centered on the vector F_{part}). The resulting distribution has a center that is offset by F_{part} and a width that is inflated relative to that of Fig. 2(*b*) by the additional uncertainty inherent to the unbuilt model.

The basic maximum-likelihood residual is

$$f(\mathbf{p}) = \sum_{i}^{\text{all data}} [Q_o(i) - \langle Q_c(i, \mathbf{p})\rangle]^2/[\sigma_o(i)^2 + \sigma_c(i, \mathbf{p})^2], \qquad (2)$$

where the symbols are very similar to those in (1). In this case, however, the quantity subtracted from $Q_o(i)$ is not simply the equivalent quantity calculated from the parameters of the model but the expectation value of this quantity calculated from all the plausible models similar to **p**. $\sigma_c(i, \mathbf{p})$ is the width of the distribution of values for $Q_c(i, \mathbf{p})$ over the plausible values for **p**. For diffraction data, the 'quantities' are the structure-factor amplitudes. The expectation value of the amplitude of a structure factor ($\langle|F_{calc}|\rangle$) calculated from a structural model, which itself contains uncertainties, is calculated by integrating over all values for the phase, as in Fig. 2(*c*). The mathematics of this integral are difficult and beyond the scope of this overview. The calculation of $\langle|F_{calc}|\rangle$ is discussed by Pannu & Read (1996) and Murshudov *et al.* (1997).

The maximum-likelihood method also depends on the assumption that the prior probability distribution contains no information. This assumption is certainly not valid in macromolecular refinement, where there is a wealth of information

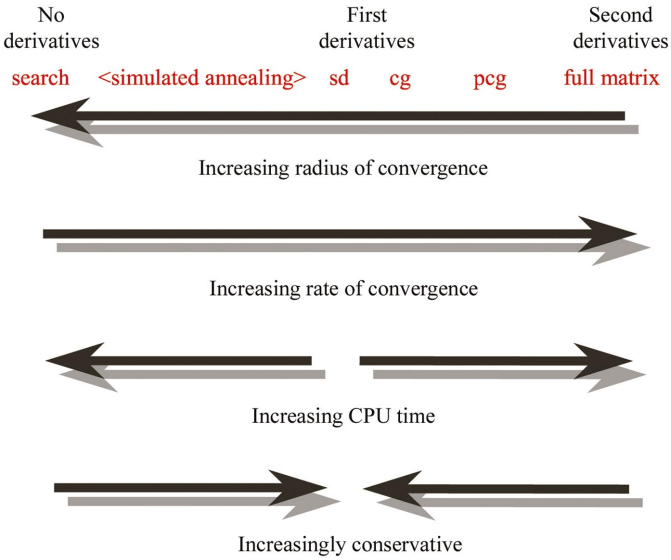

Figure 5
The principal properties of optimization methods considered here are the 'rate of convergence', 'radius of convergence', 'CPU time' and 'conservativity'. The rate of convergence is the number of iterations of the method required to reach an optimum solution. The radius of convergence is a measure of the accuracy required of the starting model. The CPU time represents the amount of time required to reach the optimum. The conservativity is a measure of the tendency of a method of optimization to preserve the values of parameters when changes would not affect the fit of the model to the data. The locations of several optimization methods on these continuums are indicated by the placement of their names. The search method uses no derivatives and is located furthest to the left. The simulated-annealing method occupies a range of positions, which is controlled by the temperature of the slow-cooling protocol. Steepest descent (sd) uses only first derivatives, while the conjugate-gradient (cg), preconditioned conjugate-gradient (pcg) and full-matrix methods use progressively more second derivatives.

about macromolecules. Somehow, maximum likelihood must be modified to preserve this knowledge. This problem is overcome by the authors of the current refinement programs by including the stereochemical information in the likelihood calculation as though it were the results of the 'experiment', essentially the same approach as that taken in least-squares programs.

Perhaps a simpler way of viewing this solution is to call the procedure 'maximum posterior probability' and optimize the product of the likelihood and prior distributions by varying the values of the parameters in the neighborhood of a starting model.

4.3.3. Comparing maximum likelihood and least squares. Fig. 3 shows the mathematical world that crystallographic least-squares refinement inhabits. There are two key features of least squares that are important when a comparison to maximum likelihood is made: (i) the identification of the measurement of the observation as the only source of error and (ii) the absence of any consideration of the uncertainty of the phase of the reflection. Figs. 2 and 4 show probability distributions used in maximum-likelihood equivalent to Fig. 3.

A fundamental difference between the least-squares worldview and that of maximum likelihood is that least squares presumes that small random changes in the values of the parameters will cause small random changes in the predicted observations. While atomic positions are recorded to three places beyond the decimal point in a PDB file, this degree of precision was never intended to be taken seriously. Usually somewhere in the paper a statement similar to 'the coordinates in this model are accurate to 0.15 Å' is made. When calculating structure factors to be compared with the observed structure-factor amplitudes, the structure factor of the particular model listed in the deposition is not the value desired. Instead, the central (or best) structure factor of the population of structures that exist within the error bounds quoted by the author is needed. When there is a linear relationship between the parameters of the model and the observations, this distinction is not a problem. The center of the distribution of parameter values transforms to the center of the distribution of observations.

When the relationship is not linear this simple result is no longer valid. One must be careful to calculate the correct expectation value for the predicted observation with consideration of the uncertainties of the model. This complication was anticipated by Srinivansan & Parthasarathy (1976) and Read (1986), but was not incorporated into refinement programs until the 1990s.

The mathematical relation that transforms a coordinate model of a macromolecule into structure factors is shown in Fig. 2. The uncertainty in the positions and B factors of the model causes the expectation value of the structure factor to have a smaller amplitude than the raw calculated structure factor but the same phase. The greater the uncertainty, the smaller the amplitude of the expectation value, with the limit of complete uncertainty being an amplitude of zero. As expected, when the uncertainty of the values of the parameters increases the uncertainty of the prediction of the structure factor also increases.

Fig. 4 shows the Argand diagram for the case where one also has atoms in the crystal which have not been placed in the model. If one has no knowledge of the location of these atoms then the vector $F_{\rm part}$ has an amplitude of zero and the phase of the center of the distribution is the same as that calculated from the structural model (as was the case in Fig. 2). If, however, one has a vague idea where the unbuilt atoms lie, their contribution ($F_{\rm part}$) will have a non-zero amplitude and the center of the probability distribution for this reflection will have a phase different from that calculated from the current model. The ability to alter the probability distribution by adding this additional information reduces the bias of the distribution toward the model already built. Such models can only be refined with *BUSTER*/*TNT* (Roversi *et al.*, 2000) at this time.

5. The optimization method

Function-minimization methods fall on a continuum (see Fig. 5). The distinguishing characteristic is the amount of information about the function that must be explicitly calculated and supplied to the algorithm. All methods require the ability to calculate the value of the function given a particular set of values for the parameters of the model. Where the methods differ is that some require only the function values (simulated annealing is such a method; it uses the gradient of the function only incidentally in generating new sets of parameters), while others require the gradient of the function as well. The latter class of methods are called gradient-descent methods.

The method of minimization that uses the gradient and all of the second derivative (*i.e.* curvature) information is called the 'full-matrix' method. The full-matrix method is quite powerful, but the requirements of memory and computations for its implementation are beyond current computer technology except for small molecules and smaller proteins. Also, for reasons to be discussed, this algorithm can only be used when the model is very close to the minimum – closer than most 'completely' refined protein models. For proteins, it has only been applied to small molecules (<2000 atoms) that diffract to high resolution and have previously been exhaustively refined with gradient-descent methods.

The distance from the minimum at which a particular method breaks down is called the 'radius of convergence'. It is clear that the full-matrix method is much more restrictive than the gradient-descent methods and that gradient-descent methods are more restrictive than simulated annealing. Basically, the less information about the function calculated at a particular point, the larger the radius of convergence will be.

5.1. Search methods

Of the many methods of minimizing functions, the simplest methods to describe are the search methods. Pure search methods are not used in macromolecular refinement

because of the huge amount of computer time that would be required, but are routinely used in molecular replacement. To determine the best orientation of the trial model in a crystal, one simply calculates the fit of the model to the observations for an exhaustive set of trials. Once the entire set of calculations have been completed, the best one is simple to identify.

The common motif of search methods is that they each have some means of selecting which combination of parameters to test and simply keep track of the best one found so far. One can systematically sample all combinations or randomly pick values for the parameters. If the function being minimized has some property that restricts the possible solutions, this information can be used to guide the search (such as packing restrictions in molecular replacement).

The more combinations tested, the greater the chance that the absolute best solution will be stumbled upon and the greater the precision of the answer. It is rare for a search method to find the best parameters exactly. Usually, the answer from a search method is used as the starting point for a gradient-descent minimization, which will fine-tune the result.

Simulated annealing (Kirkpatrick *et al.*, 1983; Otten & van Ginneken, 1989) is a search method; a random set of models are compared with the observations. Because it is known that the correct model must have good bond lengths and angles, the random model generator is chosen to ensure that all its output has reasonable geometry. The random generator used is a molecular-dynamics simulation program. Since the parameters of the model have 'momentum' they can 'move' through flat regions in the function and even over small ridges and into different local minima. The 'annealing' part of the method is to start with high 'velocities' ('temperature'), to allow the model great freedom, and slowly reduce the momentum until eventually the model is trapped in a minimum that is hoped to be the global minimum.

The explanation of simulated annealing involves a lot of quotes. These words (*e.g.* momentum and temperature) are analogies and should not be taken too seriously.

The principal advantage of simulated annealing is that it is not limited by local minima and can therefore correct errors that are quite large, thus saving time by reducing the effort required for manual rebuilding of the model.

The principal disadvantage is the large amount of computer time required. Since so much time is required to complete a proper slow-cool protocol, the protocols used in crystallographic refinement are abbreviated versions of what is recommended in the wider literature. Because of this compromise, the model can get trapped with poor conformations. It also becomes possible that some regions of the model that were correct at the start will be degraded by the process. To reduce the chance of this degradation occurring, the slow-cool runs should be very slow and the starting temperature should be lowered when the starting model is better (*e.g.* when the starting model is derived from the crystal structure of a molecule with very similar sequence and/or the addition of a relatively small adduct).

5.2. Gradient-descent methods

An analysis of the full-matrix method and all gradient-descent methods begins with the Taylor series expansion of the function being minimized. For a generic function $[f(\mathbf{p})]$ the Taylor series expansion is

$$f(\mathbf{p}) = f(\mathbf{p}_0) + \left|\frac{\mathrm{d}f(\mathbf{p})}{\mathrm{d}\mathbf{p}}\right|^t_{\mathbf{p}=\mathbf{p}_0} (\mathbf{p}-\mathbf{p}_0) + \frac{1}{2}(\mathbf{p}-\mathbf{p}_0)^t \left|\frac{\mathrm{d}^2 f(\mathbf{p})}{\mathrm{d}\mathbf{p}^2}\right|_{\mathbf{p}=\mathbf{p}_0} (\mathbf{p}-\mathbf{p}_0) + \cdots, \quad (3)$$

where $\mathbf{p}_0$ is the current set of parameters of the model and $\mathbf{p}$ is the parameters of any similar model. In all refinement programs the higher order terms (represented by '...') are ignored. This assumption has considerable consequences, which will be discussed later.

We can change the nomenclature used in (3) to more closely match those in refinement by defining $\mathbf{p}_0$ to be the parameters of the current model and $\mathbf{s}$ to be a 'shift vector' that we want to add to $\mathbf{p}_0$. $\mathbf{s}$ is equal to $\mathbf{p}-\mathbf{p}_0$. The new version of (3) is

$$f(\mathbf{p}_0+\mathbf{s}) = f(\mathbf{p}_0) + \left|\frac{\mathrm{d}f(\mathbf{p})}{\mathrm{d}\mathbf{p}}\right|^t_{\mathbf{p}=\mathbf{p}_0} \mathbf{s} + \frac{1}{2}\mathbf{s}^t \left|\frac{\mathrm{d}^2 f(\mathbf{p})}{\mathrm{d}\mathbf{p}^2}\right|_{\mathbf{p}=\mathbf{p}_0} \mathbf{s}, \quad (4)$$

and its derivative is

$$\left|\frac{\mathrm{d}f(\mathbf{p})}{\mathrm{d}\mathbf{p}}\right|_{(\mathbf{p}=\mathbf{p}_0+\mathbf{s})} = \left|\frac{\mathrm{d}f(\mathbf{p})}{\mathrm{d}\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_0} + \left|\frac{\mathrm{d}^2 f(\mathbf{p})}{\mathrm{d}\mathbf{p}^2}\right|^t_{\mathbf{p}=\mathbf{p}_0} \mathbf{s}. \quad (5)$$

Since the first and second derivatives can be calculated given any particular value for $\mathbf{p}_0$, this equation allows the gradient of the function to be calculated given any shift vector. In addition, the equation can be inverted to allow the shift vector to be calculated given the gradient of the function and the second-derivative matrix.

At the minimum (or maximum) of a function, all components of the gradient are zero. Therefore, we should be able to calculate the shift vector between the current model ($\mathbf{p}_0$) and the minimum. The equation for this is simple,

$$\mathbf{s} = -\left|\frac{\mathrm{d}^2 f(\mathbf{p})}{\mathrm{d}\mathbf{p}^2}\right|^{-1}_{\mathbf{p}=\mathbf{p}_0} \left|\frac{\mathrm{d}f(\mathbf{p})}{\mathrm{d}\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_0}. \quad (6)$$

The full-matrix method uses this equation, evaluated with the current parameters, to calculate $\mathbf{s}$. [The matrix does not have to be inverted to solve this equation for $\mathbf{s}$ (Golub & van Loan, 1989; Konnert, 1976).] $\mathbf{s}$ is then added to $\mathbf{p}_0$ to give the set of parameters that cause the function to be minimal and, in the case of refinement, the best fit to the observations.

In the classic example of fitting a line to a set of points, one evaluates this single expression and the minimum is discovered. The truncated Taylor series is exact and the shift vector is also exact. In refinement something is obviously different. In macromolecular refinement the higher-order terms of (4) are not equal to zero, resulting in the shift vector giving only the approximate location of the minimum.

The quality of the estimate is limited by the size of the terms that are ignored. The terms of the Taylor series have increasing powers of $\mathbf{s}$. If $\mathbf{s}$ is small, these higher-order terms

also become small. Therefore, as $\mathbf{p}_0$ becomes closer to the minimum, $\mathbf{s}$ becomes more accurate. Dropping the higher order terms of the series creates the limited radius of convergence of these methods.

The full-matrix method and all the gradient-descent methods that are derived from it become a series of successive approximations. An initial guess for the parameters of the model ($\mathbf{p}_0$) is manufactured in some way. For the shift vector to actually give an improved set of parameters, the guess must be sufficiently close to the minimum. The 'sufficiently close' criterion is difficult to formulate exactly.

The property of the full-matrix method that compensates for its restricted radius of convergence is its 'power of convergence'. If the starting model is within the radius of the full-matrix method, that method will be able to bring the model to the minimum quicker than any other method.

5.2.1. The normal matrix. The aspect of the full-matrix minimization method that prevents it being used in common refinement is the difficulty in calculating the term

$$\left|\frac{\mathrm{d}^2 f(\mathbf{p})}{\mathrm{d}\mathbf{p}^2}\right|^{-1}_{\mathbf{p}=\mathbf{p}_0}. \tag{7}$$

This matrix written out in full is

$$\begin{pmatrix} \frac{\partial^2 f(\mathbf{p})}{\partial p_1^2} & \frac{\partial^2 f(\mathbf{p})}{\partial p_2\,\partial p_1} & \cdots & \frac{\partial^2 f(\mathbf{p})}{\partial p_n\,\partial p_1} \\ \frac{\partial^2 f(\mathbf{p})}{\partial p_1\,\partial p_2} & \frac{\partial^2 f(\mathbf{p})}{\partial p_2^2} & \cdots & \frac{\partial^2 f(\mathbf{p})}{\partial p_n\,\partial p_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{p})}{\partial p_1\,\partial p_n} & \frac{\partial^2 f(\mathbf{p})}{\partial p_2\,\partial p_n} & \cdots & \frac{\partial^2 f(\mathbf{p})}{\partial p_n^2} \end{pmatrix}^{-1}. \tag{8}$$

This matrix contains $n \times n$ elements, where n is the number of parameters in the model. In a typical case n will be of the order of 10 000. The number of elements in the second-derivative matrix, often called the normal matrix, would be 100 000 000. It takes a lot of computer time to calculate it (Tronrud, 1999), a lot of memory to store it and a lot more computer time to solve for the shifts. The gradient-descent methods make various assumptions about the importance of different parts of the normal matrix in order to reduce these requirements.

To understand the relative importance of the different elements of the normal matrix, one needs to understand the meanings of each part. The most important classification of the elements is the distinction between the elements on the diagonal and those off it. The elements on the diagonal are affected by a single parameter and are therefore easier to analyze. The off-diagonal elements are affected jointly by two parameters.

The information contained in the off-diagonal elements describes how the effect on the function of changing one parameter is affected by changes in a second. In essence, it is related to the correlation of the two parameters. It is instructive to consider the simple case where each parameter is varied in turn. Parameter a is varied to minimize the function. Parameter b is then changed. If the off-diagonal element for a and b has a non-zero value then parameter a will have to be readjusted and the larger that value the greater the adjustment required.

The diagonal elements contain information about the effect of a parameter's value on its own effect on the function. This, of course, will always be large. (If the diagonal element is zero then any value for that parameter will be equivalent, a property that is usually undesirable in a parameter.)

5.2.2. Sparse-matrix method. One can examine the relationship between the parameters in the model to determine which pairs will have significant off-diagonal elements in the normal matrix. The pairs whose off-diagonal elements are predicted to be small can then be ignored. Such selective attention only pays off when the vast majority of elements can be discarded.

With some functions all the off-diagonal elements may be ignored, while other functions do not allow any to be ignored. One must treat functions on a case-by-case basis to determine which elements to use. An analysis of the residual function for X-ray diffraction shows that the size of the off-diagonal elements is related to the extent of electron-density overlap of the two atoms (Agarwal, 1978). Since atoms are fairly compact, all off-diagonal terms between parameters in atoms are negligible, except for atoms bonded to one another, and the terms for those pairs are small. Since an atom has a large overlap with its own electrons, the diagonal elements are very large compared with any off-diagonal ones.

The stereochemical restraints commonly used in protein refinement have a different pattern. Here, the parameters of atoms connected by a bond distance or angle have a strong correlation. Atoms not restrained to one another have no correlation at all. The off-diagonal terms that are non-zero are as significant as the diagonal ones.

This knowledge allows one to calculate the normal matrix as a sparse matrix, *i.e.* the vast majority of the off-diagonal elements are never calculated and do not even have computer memory allocated for their storage. The only elements calculated are the diagonal ones (including contributions from both the crystallographic and stereochemical restraints) and the off-diagonal elements for parameters from atoms directly connected by geometric restraints.

Even with the simplification of the normal matrix introduced by the sparse approximation, the problem of solving for the parameter shifts is difficult. There are an enormous number of numerical methods available for solving problems like this and these methods depend strongly on the nature of the approximation to the normal matrix being used. It is important to note, however, that each method includes assumptions and approximations that should be understood before the method is used.

The refinement programs *PROLSQ* (Hendrickson & Konnert, 1980), *REFMAC* (Murshudov *et al.*, 1997) and *SHELXL* (Sheldrick & Schneider, 1997) use the sparse-matrix approximation to the normal matrix. They calculate the shifts to apply to the model using a method called 'conjugate gradient' (Konnert, 1976), which is unrelated to the conjugate-gradient method used to minimize functions (Fletcher & Reeves, 1964). It is a sign of confusion to state that *X-PLOR*

(Brünger *et al.*, 1987) and *CNS* (Brünger *et al.*, 1998) use the same method as these programs.

5.2.3. Diagonal matrix. A further step in simplifying the normal matrix is made by ignoring all off-diagonal elements. The normal matrix becomes a diagonal matrix, which is inverted by simply inverting each diagonal element in turn. In essence, working with the matrix becomes a one-dimensional problem. Since any correlation between parameters has been assumed to be zero, the shift for a particular parameter can be calculated in isolation from the shifts of all other parameters. With this approximation, the full-matrix of equation (6) becomes

$$s_i = -\left|\frac{\partial f(\mathbf{p})}{\partial p_i}\right|_{\mathbf{p}=\mathbf{p}_0} \Big/ \left|\frac{\partial^2 f(\mathbf{p})}{\partial p_i^2}\right|_{\mathbf{p}=\mathbf{p}_0}. \quad (9)$$

5.2.4. Steepest descent. A further simplification can be made if all the diagonal elements of the normal matrix have the same value. In this case, none of the elements need be calculated. The average value can be estimated from the behavior of the function value as the parameters are shifted. The shift for a particular parameter is simply

$$s_i = -\left|\frac{\partial f(\mathbf{p})}{\partial p_i}\right|_{\mathbf{p}=\mathbf{p}_0}. \quad (10)$$

The steepest descent method has the advantage of a large radius of convergence. Since the gradient of a function points in the steepest direction uphill, the steepest descent method simply shifts the parameters in the steepest direction downhill. This method is guaranteed to reach the local minimum, given enough time.

Any method that actually divides by the second derivative is subject to problems if the curvature is negative or, worse yet, zero. Near a minimum, all second derivatives must be positive. Near a maximum, they are all negative. As one moves away from the minimum, the normal matrix elements tend toward zero. The curvature becomes zero at the inflection point that surrounds each local minimum.

The full-matrix method becomes unstable somewhere between the minimum and the inflection point. The diagonal approximation method has a similar radius of convergence, although larger than that of the full-matrix method. The steepest descent method, however, simply moves the parameters to decrease the function value. The method will move toward the minimum when the starting point is anywhere within the ridge of hills surrounding the minimum.

The steepest descent method is very robust. It will smoothly converge to the local minimum whatever the starting parameters are. However, it will require a great number of iterations and therefore a great deal of time to do so.

The problem with steepest descent is that no information about the normal matrix is used to calculate the shifts to the parameters. Whenever the assumptions break down (the parameters have correlation and have different diagonal elements) the shifts will be inefficient.

5.2.5. Conjugate gradient. Just as one can calculate an estimate for the slope of a function by looking at the function value at two nearby points, one can estimate the curvature of a function by looking at the change in the function's gradient at two nearby points. These gradients are routinely calculated in steepest descent refinement. The gradient is calculated, the parameters are shifted a little and the gradient is calculated again. In steepest descent the two gradients are never compared, but if they were some information about the normal matrix could be deduced.

The conjugate-gradient method (Fletcher & Reeves, 1964) does just this. The analysis of Fletcher and Reeves showed that the steepest descent shift vector can be improved by adding a well defined fraction of the shift vector of the previous cycle. Each cycle essentially 'learns' about one dimension of curvature in the n-dimensional refinement space (where n is the number of parameters in the model.) Therefore, after n cycles everything is known about the normal matrix and the minimum is found.

The shift vector for cycle $k + 1$ using the conjugate gradient is

$$\mathbf{s}_{k+1} = -\left|\frac{\mathrm{d}f(\mathbf{p})}{\mathrm{d}\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_k} + \beta_{k+1}\mathbf{s}_k, \quad (11)$$

where β_{k+1} is the ratio of the length of the function's present gradient to that of the previous cycle. During the first cycle there is no previous cycle and therefore the first cycle must be steepest descent.

The fundamental limitation of the conjugate-gradient method is that it is guaranteed to reach the minimum in n cycles only if the Taylor series does indeed terminate, as assumed in (4). If there are higher order terms, as there are in crystallographic refinement, then n cycles will only get the model nearer to the minimum. One should start over with a new run of n cycles to get the model even closer.

Even n cycles is a lot in crystallography. No one runs thousands of cycles of conjugate-gradient refinement, nor can so many cycles be run with current software, because the shifts become too small to be represented with the precision of current computers. Small shifts are not necessarily unimportant ones. These small shifts can add up to significant changes in the model, but they cannot be calculated.

The conjugate-gradient method was elaborated by Powell (1977). This paper included a discussion of an alternative equation for the calculation of β_k, which was equivalent for a quadratic function but gave superior results for some non-quadratic functions. In addition, a strategy was described for restarting the conjugate-gradient search more often than once every n cycles, which avoids, to a limited extent, cycles with very small shifts.

X-PLOR and *CNS* use the conjugate-gradient method as modified by Powell (1977), subsequent to simulated annealing.

5.2.6. Preconditioned conjugate gradient. The conjugate-gradient method is better than the steepest descent method because the former uses some information about the normal matrix to improve the quality of the shift vector. It would seem

Table 1
Properties of a selection of refinement programs.

This table lists a summary of the properties of six commonly used refinement programs. The meanings of the various codes are as follows. Parameters: *xyzb*, position, isotropic *B* factor and occupancy; aniso, anisotropic *B* factor; TLS, group TLS *B* factors used to generate approximate anisotropic *B* factors; torsion, only allow variation of angles of rotation about single bonds; free, generalized parameters, which can be used to model ambiguity in twining, chirality or static conformation. Function: EE, empirical energy; LS, least squares; ML, maximum likelihood using amplitude data; MLφ:, maximum likelihood using experimentally measured phases; ML?, maximum likelihood using envelopes of known composition but unknown structure. Method: SA, simulated annealing; CG, Powell variant conjugate gradient; PCG, preconditioned conjugate gradient; Sparse, sparse-matrix approximation to the normal matrix; FM, full matrix calculated for normal matrix.

Program	Parameters	Function	Method
BUSTER/TNT	*xyzb*	ML, MLφ, ML?	PCG
CNS	*xyzb*, torsion	EE, LS, ML, MLφ	SA, CG
REFMAC	*xyzb*, TLS, aniso	LS, ML, MLφ	Sparse, FM
SHELXL	*xyzb*, aniso, free	LS	Sparse, FM
TNT	*xyzb*	LS	PCG
X-PLOR	*xyzb*, torsion	EE, LS, ML, MLφ	SA, CG

reasonable to believe that the shift vector could be improved further if additional information were added. For instance, the diagonal elements of the normal matrix can be calculated directly and quickly.

All this information is combined together in the preconditioned conjugate-gradient method (Axelsson & Barker, 1984; Tronrud, 1992). This method operates like the conjugate-gradient method except that the preconditioned method uses the shifts from the diagonal matrix method for its first cycle instead of those from the steepest descent method. The shift vector for the preconditioned conjugate gradient is

$$\mathbf{s}_{k+1} = -\left|\frac{\mathrm{d}f(\mathbf{p})}{\mathrm{d}\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_k} \Bigg/ \left|\frac{\mathrm{d}^2 f(\mathbf{p})}{\mathrm{d}\mathbf{p}_i^2}\right|_{\mathbf{p}=\mathbf{p}_k} + \beta'_{k+1}\mathbf{s}_k, \qquad (12)$$

where the trick is calculating β'_{k+1} correctly. This matter is discussed in detail by Tronrud (1992).

6. Summary

Table 1 summarizes the properties of the refinement programs discussed in this review. The field of macromolecular refinement is blessed with a variety of programs that can be used to improve our structural models. With a firm understanding of the differences between these programs, one should be able to choose the one that best fits the needs of any project.

This work was supported in part by NIH grant GM20066 to B. W. Matthews.

References

Agarwal, R. C. (1978). *Acta Cryst.* A**34**, 791–809.
Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Axelsson, O. & Barker, V. (1984). *Finite Element Solution of Boundary Value Problems*, ch. 1, pp. 1–63. Orlando, FL, USA: Academic Press.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Bricogne, G. (1988). *Acta Cryst.* A**44**, 517–545.
Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.
Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.
Bricogne, G. & Irwin, J. J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Brünger, A. T., Adams, P. D., Clore, G. M., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Brünger, A. T., Kuriyan, K. & Karplus, M. (1987). *Science*, **235**, 458–460.
Diamond, R. (1971). *Acta Cryst.* A**27**, 436–452.
Fletcher, R. & Reeves, C. (1964). *Comput. J.* **7**, 81–84.
Golub, G. H. & van Loan, C. F. (1989). *Matrix Computations*, 2nd ed. Baltimore: John Hopkins University Press.
Haneef, I., Moss, D. S., Stanford, M. J. & Borkakoti, N. (1985). *Acta Cryst.* A**41**, 426–433.
Hendrickson, W. A. & Konnert, J. H. (1980). *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, ch. 13, pp. 13.01–13.26. Bangalore: Indian Academy of Sciences.
Kleywegt, G. J. & Jones, T. A. (1995). *Structure*, **3**, 535–540.
König, V., Vértesy, L. & Schneider, T. R. (2003). *Acta Cryst.* D**59**, 1737–1743.
Konnert, J. H. (1976). *Acta Cryst.* A**32**, 614–617.
Kirkpatrick, S. C. D., Gelatt, J. & Vecchi, M. P. (1983). *Science*, **220**, 671–680.
Levitt, M. (1974). *J. Mol. Biol.* **82**, 393–420.
Mandel, J. (1984). *The Statistical Analysis of Experimental Data.* New York: Dover.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Otten, R. H. J. M. & van Ginneken, L. P. P. P. (1989). *The Annealing Algorithm.* Boston: Kluwer Academic Publishers.
Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–669.
Powell, M. J. D. (1977). *Math. Program.* **12**, 241–254.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.
Rice, L. M. & Brünger, A. (1994). *Proteins*, **19**, 277–290.
Roversi, P., Blanc, E., Vonrhein, C., Evans, G. & Bricogne, G. (2000). *Acta Cryst.* D**56**, 1316–1323.
Schomaker, V. & Trueblood, K. N. (1968). *Acta Cryst.* B**24**, 63–76.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.
Sivia, D. S. (1996). *Data Analysis: A Bayesian Tutorial.* Oxford University Press.
Srinivansan, R. & Parthasarathy, S. (1976). *Some Statistical Application in X-ray Crystallography.* Oxford: Pergamon Press.
Stout, G. H. & Jensen, L. H. (1989). *X-ray Structure Determination: A Practical Guide*, 2nd ed, pp. 424–426. New York: John Wiley & Sons.
Tronrud, D. E. (1992). *Acta Cryst.* A**48**, 912–916.
Tronrud, D. E. (1999). *Acta Cryst.* A**55**, 700–703.
Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.
Wilson, A. J. C. (1942). *Nature* (*London*), **150**, 151–152.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Acta Cryst.* D**57**, 122–133.