Taking measurement errors into account requires multiplication of equation (18.2.3.3) with an appropriate probability distribution (usually a conditional Gaussian distribution with standard deviation σ_o) of the observed structure-factor amplitudes ($|\mathbf{F}_o|$) around the 'true' structure-factor amplitudes ($|\mathbf{F}|$),

$$P_{\text{meas}}(|\mathbf{F}_o|;\, |\mathbf{F}|). \tag{18.2.3.5}$$

Prior knowledge of the phases of the structure factors can be incorporated by multiplying equation (18.2.3.3) with a phase probability distribution

$$P_{\text{phase}}(\varphi) \tag{18.2.3.6}$$

and rewriting equation (18.2.3.3) in terms of the structure-factor moduli and amplitudes of $\mathbf{F} = |\mathbf{F}| \exp(i\varphi)$.

The unknown variables $|\mathbf{F}|$ and φ in equations (18.2.3.3)–(18.2.3.5) have to be eliminated by integration in order to obtain the conditional probability distribution of the observed structure-factor amplitudes, given a partial model with errors, the amplitude measurement errors and prior phase information:

$$P_a(|\mathbf{F}_o|;\, \mathbf{F}_c) = (1/\pi\varepsilon\sigma_\Delta^2)\int \mathrm{d}\varphi\, \mathrm{d}|\mathbf{F}|\, |\mathbf{F}| P_{\text{meas}}(|\mathbf{F}_o|;\, |\mathbf{F}|) \times P_{\text{phase}}(\varphi) \exp\left\{-[|\mathbf{F}| \exp(i\varphi) - D\mathbf{F}_o]^2/\varepsilon\sigma_\Delta^2\right\}. \tag{18.2.3.7}$$

The likelihood, $\mathcal{L}$, of the model is defined as the joint probability distribution of the structure factors of all reflections in the working set. Assuming independent and uncorrelated structure factors, $\mathcal{L}$ is simply the product of the distributions in equation (18.2.3.7) for all reflections. Instead of maximizing the likelihood, it is more common to minimize the negative logarithm of the likelihood,

$$E_{\text{X-ray}} = \mathcal{L} = -\sum_{hkl \in \text{working set}} \log[P_a(|\mathbf{F}_o|;\, \mathbf{F}_c)]. \tag{18.2.3.8}$$

Empirical estimates of σ_Δ [and D through equation (18.2.3.4)] can be obtained by minimizing $\mathcal{L}$ for a particular atomic model. It is generally assumed that σ_Δ and D show relatively little variation among neighbouring reflections. Accepting this assumption, σ_Δ and D can be estimated by considering narrow resolution shells of reflections and assuming that the two parameters are constant in these shells. Minimization of $\mathcal{L}$ can then be performed as a function of these constant shell parameters while keeping the atomic model fixed (Read, 1986, 1997). Alternatively, one can assume a two-term Gaussian model for σ_Δ (Murshudov *et al.*, 1997) and minimize $\mathcal{L}$ as a function of the Gaussian parameters. Note that individual atomic B factors are taken into account by the calculated model structure factors ($\mathbf{F}_c$).

This empirical approach to estimate σ_Δ and D requires occasional recomputation of these values as the model improves. Refinement methods that improve the model structure factors, $\mathbf{F}_c$, will therefore have a beneficial effect on σ_Δ and D. Better estimates of these values will then enhance the next refinement cycle. Thus, powerful optimization methods and maximum-likelihood targets are expected to interact in a synergistic fashion (*cf.* Fig. 18.2.5.1). Structure-factor averaging of multi-start refinement models can provide another layer of improvement by producing a better description of $\mathbf{F}_c$ if the model shows significant variability due to errors or intrinsic flexibility (see below).

In order to achieve an improvement over the least-squares residual [equation (18.2.3.2)], cross validation was found to be essential (Pannu & Read, 1996; Adams *et al.*, 1997) for the estimation of model incompleteness and errors (σ_Δ and D). Since the test set typically contains only 10% of the diffraction data, these cross-validated quantities can show significant statistical fluctuations as a function of resolution. In order to reduce these fluctuations, Read (1997) devised a smoothing method by applying restraints to σ_A values between neighbouring resolution shells where

$$\sigma_A = \left[1 - (\sigma_\Delta/\langle|\mathbf{F}_o|^2\rangle)\right]^{1/2}. \tag{18.2.3.9}$$

Pannu & Read (1996) have developed an efficient Gaussian approximation of equation (18.2.3.7) in cases of no prior phase information, termed the 'MLF' target function. In the limit of a perfect model (*i.e.* $\sigma_\Delta = 0$ and $D = 1$), MLF reduces to the traditional least-squares residual [equation (18.2.3.2)] with $1/\sigma_o^2$ weighting. In the case of prior phase information, the integration over the phase angles has been carried out numerically in equation (18.2.3.7), termed the 'MLHL' target (Pannu *et al.*, 1998). A maximum-likelihood function which expresses equation (18.2.3.7) in terms of observed intensities has also been developed, termed 'MLI' (Pannu & Read, 1996).

18.2.3.2. *A priori chemical information*

The parameters for the covalent terms in E_{chem} [equation (18.2.3.1)] can be derived from the average geometry and (r.m.s.) deviations observed in a small-molecule database. Extensive statistical analyses were undertaken for the chemical moieties of proteins (Engh & Huber, 1991) and polynucleotides (Parkinson *et al.*, 1996) using the Cambridge Structural Database (Allen *et al.*, 1983). Analysis of the ever-increasing number of atomic resolution macromolecular crystal structures will no doubt cause some modifications of these parameters in the future.

It is common to use a purely repulsive quartic function ($E_{\text{repulsive}}$) for the non-bonded interactions that are included in E_{chem} (Hendrickson, 1985):

$$E_{\text{repulsive}} = \sum_{ij}[(cR_{ij}^{\min})^n - R_{ij}^n]^m, \tag{18.2.3.10}$$

where R_{ij} is the distance between two atoms i and j, $R_{ij}^{\min}$ is the van der Waals radius for a particular atom pair ij, $c \leq 1$ is a constant that is sometimes used to reduce the radii, and $n = 2$, $m = 2$ or $n = 1$, $m = 4$. van der Waals attraction and electrostatic interactions are usually not included in crystallographic refinement. These simplifications are valid since the diffraction data contain information that is able to produce atomic conformations consistent with actual non-bonded interactions. In fact, atomic resolution crystal structures can be used to derive parameters for electrostatic charge distributions (Pearlman & Kim, 1990).

18.2.4. Searching conformational space

Annealing denotes a physical process wherein a solid is heated until all particles randomly arrange themselves in a liquid phase and is then cooled slowly so that all particles arrange themselves in the lowest energy state. By formally defining the target, E [equation (18.2.3.1)], to be the equivalent of the potential energy of the system, one can simulate such an annealing process (Kirkpatrick *et al.*, 1983). There is no guarantee that simulated annealing will find the global minimum (Laarhoven & Aarts, 1987). However, compared to conjugate-gradient minimization, where search directions must follow the gradient, simulated annealing achieves more optimal solutions by allowing motion against the gradient (Kirkpatrick *et al.*, 1983). The likelihood of *uphill* motion is determined by a control parameter referred to as *temperature*. The higher the temperature, the more likely it is that simulated annealing will overcome barriers (Fig. 18.2.4.1). It should be noted that the simulated-annealing temperature normally has no physical meaning and merely determines the likelihood of overcoming barriers of the target function in equation (18.2.3.1).

references