4.5.2.6.3. Patterson functions

In fibre diffraction, the conventional Patterson function cannot be calculated since the individual structure-factor intensities are not available. However, MacGillavry & Bruins (1948) showed that the *cylindrically averaged Patterson function* can be calculated from fibre diffraction data. Consider the function $\hat{Q}(r, z)$ defined by

$$\hat{Q}(r, z) = \sum_{l=0}^{\infty} \int_0^{\infty} \varepsilon_l I_l(R) J_0(2\pi Rr) \cos(2\pi lz/c) 2\pi R \, \mathrm{d}R, \quad (4.5.2.58)$$

where $\varepsilon_l = 1$ for $l = 0$ and 2 for $l > 0$, which can be calculated from the intensity distribution on a continuous fibre diffraction pattern. Using equations (4.5.2.7), (4.5.2.10), (4.5.2.17) and (4.5.2.58) shows that $\hat{Q}(r, z)$ is the cylindrical average of the Patterson function, $\hat{P}(r, \varphi, z)$, of one molecule, *i.e.*

$$\hat{Q}(r, z) = (1/2\pi) \int_0^{2\pi} \hat{P}(r, \varphi, z) \, \mathrm{d}\varphi. \quad (4.5.2.59)$$

The $\hat{}$ symbols on $\hat{P}(r, \varphi, z)$ and $\hat{Q}(r, z)$ indicate that these are Patterson functions of a single molecule, as distinct from the usual Patterson function of a crystal, which contains intermolecular interatomic vectors and is periodic with the same periodicity as the crystal. $\hat{P}(r, \varphi, z)$ is periodic only along z and is therefore, strictly, a Patterson function along z and an autocorrelation function along x and y (Millane, 1990*b*). The cylindrically averaged Patterson contains information on interatomic separations along the axial direction and in the lateral plane, but no information on orientations of the vectors in the lateral plane.

For a polycrystalline system; consider the function $Q(r, z)$ given by

$$Q(r, z) = \sum_l \sum_{h,k} R_{hk} I_l(R_{hk}) J_0(2\pi R_{hk} r) \cos(2\pi lz/c), \quad (4.5.2.60)$$

where the sums are over all the overlapped reflections $I_l(R_{hk})$ on the diffraction pattern, given by equation (4.5.2.24). It is easily shown that $Q(r, z)$ is related to the Patterson function $P(r, \varphi, z)$ by

$$Q(r, z) = (1/2\pi) \int_0^{2\pi} P(r, \varphi, z) \, \mathrm{d}\varphi, \quad (4.5.2.61)$$

where, in this case, $P(r, \varphi, z)$ is the usual Patterson function (expressed in cylindrical polar coordinates), *i.e.* it contains all intermolecular (both intra- and inter-unit cell) interatomic vectors and has the same translational symmetry as the unit cell. The cylindrically averaged Patterson function for polycrystalline fibres therefore contains the same information as it does for noncrystalline fibres (*i.e.* no angular information in the lateral plane), except that it also contains information on intermolecular separations.

Low resolution and cylindrical averaging, in addition to the usual difficulties with interpretation of Patterson functions, has resulted in the cylindrically averaged Patterson function not playing a major role in structure determination by fibre diffraction. However, information provided by the cylindrically averaged Patterson function has, in a number of instances, been a useful component in fibre diffraction analyses. A good review of the application of Patterson functions in fibre diffraction is given by Stubbs (1987). Removing data from the low-resolution part (or all) of the equator when calculating the cylindrically averaged Patterson function removes the strong vectors related to axially invariant (or cylindrically symmetric) parts of the map, and can aid interpretation (Namba *et al.*, 1980; Stubbs, 1987). It is also important when calculating cylindrically averaged Patterson functions to use data only at a resolution that is appropriate to the size and spacings of features one is looking for (Stubbs, 1987).

Cylindrically averaged Patterson functions were used in early applications of fibre diffraction analysis (Franklin & Gosling, 1953; Franklin & Klug, 1955). The intermolecular peaks that usually dominate in a cylindrically averaged Patterson function can help to define the locations of multiple molecules in the unit cell. Depending on the space-group symmetry, it is sometimes possible to calculate the complete three-dimensional Patterson function (or certain projections of it). This comes about because of the equivalence of the amplitudes of overlapping reflections in some high-symmetry space groups. The intensity of each reflection can then be determined and a full three-dimensional Patterson map calculated (Alexeev *et al.*, 1992). The only difficulty is that nonsystematic overlaps are often present, although these are usually relatively few in number and the intensity can be apportioned equally amongst them, the resulting errors usually being small relative to the level of detail present in the Patterson map. For lower space-group symmetries, it may not be possible to calculate a three-dimensional Patterson map, but it may be possible to calculate certain projections of the map. For example, if the overlapped $hk0$ reflections have the same intensities, a projection of the Patterson map down the c axis can be calculated. Since such a projection is along the polymer axes, it gives the relative positions of the molecules in the ab plane. If the combined helix and space-group symmetry is high, an estimate of the electron density can be obtained by averaging appropriate copies of the three-dimensional Patterson function (Alexeev *et al.*, 1992).

4.5.2.6.4. Molecular model building

The majority of the structures determined by X-ray fibre diffraction analysis have been determined by molecular model building (Campbell Smith & Arnott, 1978; Arnott, 1980; Millane, 1988). Most applications of molecular model building have been to polycrystalline systems, although there have been a number of applications to noncrystalline systems (Park *et al.*, 1987; Millane *et al.*, 1988). The approach is to use spacings and symmetry information derived directly from the diffraction pattern, coupled with the primary structure and stereochemical information on the molecule under study, to construct models of all *kinds* of possible molecular or crystal structure. These models are each refined (optimized) against the diffraction data, as well as stereochemical restraints, to produce the best model of each kind. The optimized models can be compared using various figures of merit, and in favourable cases one model will be sufficiently superior to the remainder for it to represent unequivocally the correct structure. The principle of this approach is that by making use of stereochemical constraints, the molecular and crystal structure have few enough degrees of freedom that the parameter space has a sufficiently small number of local minima for these to be identified and individually examined to find the global minimum. The X-ray phases are therefore not determined explicitly.

There are three steps involved in structure determination by molecular model building: (1) construction of all possible molecular and crystal structure models, (2) refinement of each model against the X-ray data and stereochemical restraints, and (3) adjudication among the refined models. The overall procedure for determining polymer structures using molecular model building is summarized by the flow chart in Fig. 4.5.2.2, and is described below.

The helix symmetry of the molecule, or one of a few helix symmetries, can be determined as described in Section 4.5.2.6.2. Different kinds of molecular model may correspond to one of a few different helix symmetries, usually corresponding to different

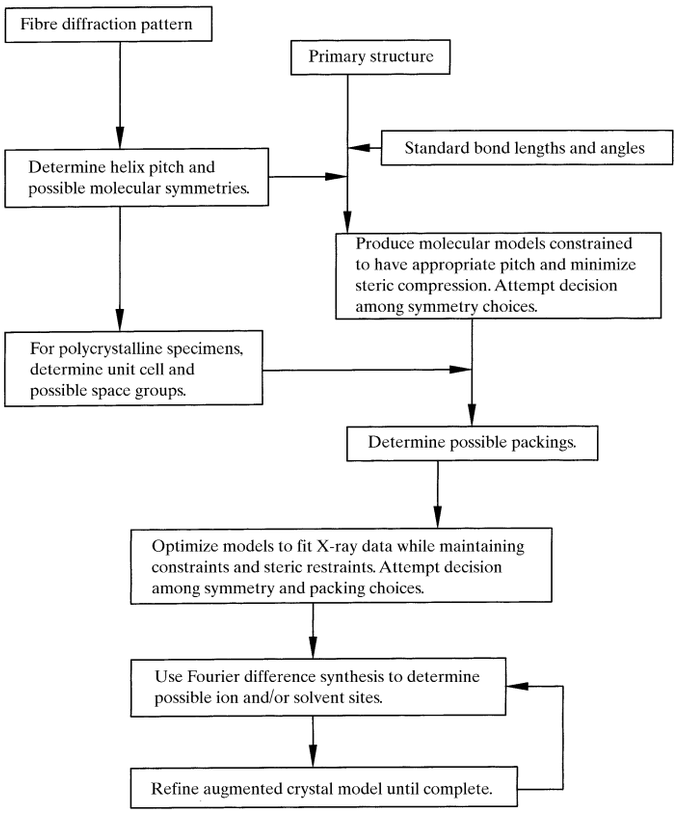

Fig. 4.5.2.2. Flow chart of the molecular-model-building approach to structure determination (Arnott, 1980).

values of v. For example, helix symmetries u_v and u_{u-v}, which correspond to the left- and right-handed helices, cannot be distinguished on the basis of the overall intensity distribution alone. Other examples of different kinds of molecular model may include single, double or multiple helices, parallel or antiparallel double helices, different juxtapositions of chains within multiple helices and different conformational domains within the molecule. For polycrystalline systems, in addition to different kinds of molecular structures, there are often different kinds of possible packing arrangements within the unit cell. There may be a number of possible packings which correspond to different arrangements within the crystallographic asymmetric unit, and there may be more than one space group that needs to be considered.

Despite the apparent large number of potential starting models implied by the above discussion, in practice the number of feasible models is usually quite small, and many of these are often eliminated at an early stage. Definition and refinement of helical polymers [steps (1) and (2) above] are carried out using computer programs, the most popular and versatile being the *linked-atom least-squares* (*LALS*) system (Campbell Smith & Arnott, 1978; Millane *et al.*, 1985), originally developed by Arnott and co-workers in the early 1960s (Arnott & Wonacott, 1966). This system has been used to determine the structures of a wide variety of polynucleotides, polysaccharides, polyesters and polypeptides (Arnott, 1980; Arnott & Mitra, 1984; Chandrasekaran & Arnott, 1989; Millane, 1990*c*). Other refinement systems exist (Zugenmaier & Sarko, 1980; Iannelli, 1994), but the principles are essentially the same and the following discussion is in terms of the *LALS* system. The atomic coordinates are defined, using a linked-atom description, in terms of bond lengths, bond angles and conformation (torsion) angles (Campbell Smith & Arnott, 1978). Stereochemical constraints are imposed, and the number of parameters reduced, by fixing the bond lengths, often (but not always) the bond angles, and possibly some of the conformation angles. The molecular conformation is then defined by the remaining parameters. For polycrystalline systems, there are usually additional variable parameters that define the packing of the molecule(s) in the unit cell. A further source of stereochemical data is the requirement that a model exhibit no overshort nonbonded interatomic distances. These are incorporated by a quadratic nonbonded potential that is matched to a Buckingham potential (Campbell Smith & Arnott, 1978). A variety of other restraints can also be incorporated.

In the *LALS* system, the quantity Ω given by

$$\Omega = \sum_m \omega_m \Delta F_m^2 + \sum_m k_m \Delta d_m^2 + \sum_m \lambda_m G_m = X + C + L \qquad (4.5.2.62)$$

is minimized by varying a set of chosen parameters consisting of conformation angles, possibly bond angles, and packing parameters. The term X involves the differences ΔF_m between the model and experimental X-ray amplitudes – Bragg and/or continuous. The term C involves restraints to ensure that overshort nonbonded interatomic distances are driven beyond acceptable minimum values, that conformations are within desired domains, that hydrogen-bond and coordination geometries are close to the expected configurations, and a variety of other relationships are satisfied (Campbell Smith & Arnott, 1978). The ω_m and k_m are weights that are inversely proportional to the estimated variances of the data. The term L involves constraints which are relationships that are to be satisfied exactly ($G_m = 0$) and the λ_m are Lagrange multipliers. Constraints are used, for example, to ensure connectivity from one helix pitch to the next and to ensure that chemical ring systems are closed. The cost function Ω is minimized using full-matrix nonlinear least squares and singular value decomposition (Campbell Smith & Arnott, 1978).

Structure determination usually involves first using equation (4.5.2.62) with the terms C and L only, to establish the stereochemical viability of each kind of possible molecular model and packing arrangement. It is worth emphasizing that it is usually advantageous if the specimen is polycrystalline, even though the continuous diffraction contains, in principle, more information than the Bragg reflections (since the latter are sampled). This is because the molecule in a noncrystalline specimen must be refined in steric isolation, whereas for a polycrystalline specimen it is refined while packed in the crystal lattice. The extra information provided by the intermolecular contacts can often help to eliminate incorrect models. This can be particularly significant if the molecule has flexible sidechains. The initial models that survive the steric optimization are then optimized also against the X-ray data, by further refinement with X included in equation (4.5.2.62). The ratios $(\Omega_P/\Omega_Q)^{1/2}$ and $(X_P/X_Q)^{1/2}$ can be used in Hamilton's test (Hamilton, 1965) to evaluate the differences between models P and Q. On the basis of these statistical tests, one can decide if one model is superior to the others at an acceptable confidence level. In the final stages of refinement, bond angles may be varied in a 'stiffly elastic' fashion from their mean values if there are sufficient data to justify the increase in the number of degrees of freedom.

If sufficient X-ray data are available, it is sometimes possible to locate additional ordered molecules such as counterions or solvent molecules by difference Fourier synthesis as described in Section 4.5.2.6.5. Their positions can then be co-refined with the polymer structure while hydrogen bonds and coordination geometries are optimized. The resulting structure can then be used to compute improved phases to search for additional molecules. Since the signal-to-noise ratio in fibre difference syntheses is usually low, difference maps must be interpreted with caution. The assignment of counterions or solvent molecules to peaks in the difference synthesis must be supported by plausible

interactions with the rest of the structure and, following refinement of the structure, by elimination of the peak in the difference map and by a significant improvement in the agreement between the calculated and measured X-ray amplitudes.

4.5.2.6.5. Difference Fourier synthesis

Difference Fourier syntheses are widely used in both protein and small-molecule crystallography to detect structural errors or to complete partial structures (Drenth, 1994). The difficulty in applying difference Fourier techniques in fibre diffraction is that the individual observed amplitudes $|F_o|$ are not available. However, difference syntheses have found wide use in fibre diffraction analysis, one of the earliest applications being to polycrystalline fibres of polynucleotides (*e.g.* Arnott *et al.*, 1967). Calculation of a three-dimensional difference map (for the unit cell) from Bragg fibre diffraction data requires that the observed intensity $I_l(R_{hk}) = I_o$ be apportioned among the contributing intensities $|F_{hkl}|^2 = |F_o|^2$. There are two ways of doing this. The intensities may be divided equally among the contributing $(m/2)$ reflections [*i.e.* $|F_o| = (2I_o/m)^{1/2}$], or they may be divided in the same proportions as those in the model, *i.e.*

$$|F_o| = \left(\frac{I_o}{\sum|F_c|^2}\right)^{1/2}|F_c|. \qquad (4.5.2.63)$$

The advantage of the former is that it is unbiased, and the advantage of the latter is that it may be more accurate but is biased towards the model. Equal division of the intensities is often (but not always) used to minimize model bias. Once the observed amplitudes have been apportioned, an $|F_o| - |F_c|$ map can be calculated as in conventional crystallography, although noise levels will be higher owing to errors in apportioning the amplitudes. As a result of overlapping of the reflections, a synthesis based on coefficients $m|F_o| - (m-1)|F_c|$ gives a more accurate estimate of the true density than does one based on $2|F_o| - |F_c|$, as is described below. Difference syntheses for polycrystalline specimens calculated in this way have been used, for example, to locate cations and water molecules in polynucleotide and polysaccharide structures (*e.g.* Cael *et al.*, 1978), to help position molecules in the unit cell (*e.g.* Chandrasekaran *et al.*, 1994) and to help position side chains, and have also been applied in neutron fibre diffraction studies of polynucleotides (Forsyth *et al.*, 1989).

Sim (1960) has shown that the mean-squared error in difference syntheses can be minimized by weighting the coefficients based on the agreement between the calculated and observed structure amplitudes. Such an analysis has recently been conducted for fibre diffraction, and shows that the optimum difference synthesis is obtained by using coefficients (Millane & Baskaran, 1997; Baskaran & Millane, 1999*a*)

$$\left[w_m\frac{|F_c|(I_o)^{1/2}}{(\sum|F_c|^2)^{1/2}} - |F_c|\right]\exp(i\alpha_c), \qquad (4.5.2.64)$$

where m is the number of degrees of freedom as defined in Section 4.5.2.6.1. If the reflections contributing to I_o are either all centric or all acentric, then the weights are given by

$$w_m = \frac{I_{m/2}(X)}{I_{m/2-1}(X)}, \qquad (4.5.2.65)$$

where $I_m(\cdot)$ denotes the modified Bessel function of the first kind of order m, and X is given by

$$X = \frac{\kappa(I_o)^{1/2}(\sum|F_c|^2)^{1/2}}{\sum_j f_j^2}, \qquad (4.5.2.66)$$

where $\kappa = 1$ for centric reflections and 2 for acentric reflections. The form of the weighting function is more complicated if both centric and acentric reflections contribute, but it can be approximated as w' given by

$$w' = (w_{2N_a} + w_{N_c})/2, \qquad (4.5.2.67)$$

where N_a and N_c are the number of acentric and centric reflections, respectively, contributing. Use of the weighted maps reduces bias towards the model (Baskaran & Millane, 1999*b*).

For continuous diffraction data from noncrystalline specimens, the situation is essentially identical except that one works in cylindrical coordinates. Referring to equations (4.5.2.7) and (4.5.2.10), the desired difference synthesis, $\Delta g(r, \varphi, z)$, is the Fourier–Bessel transform of $G_o - G_c$ where G_o and G_c denote the observed and calculated, respectively, Fourier–Bessel structure factors $G_{nl}(R)$. Since G_o is not known, the synthesis is based on the Fourier–Bessel transform of $(|G_o| - |G_c|)\exp(i\alpha_c)$, where α_c is the phase of G_c. As in the polycrystalline case, the individual $|G_o|$ need to be estimated from the data $I_o^{1/2}$ given by equation (4.5.2.17), and can be based on either equal division of the data, or division in the same proportion as the amplitudes from the model.

Namba & Stubbs (1987*a*) have shown that the peak heights in a difference synthesis are $1/m$ times their true value, as opposed to half their true value in a conventional difference synthesis. The best estimate of the true map is therefore provided by a synthesis based on the coefficients $[m|F_o| - (m-1)|F_c|]\exp(i\alpha_c)$, rather than on $(2|F_o| - |F_c|)\exp(i\alpha_c)$. Test examples showed that the noise in the synthesis can be reduced by using a value for m that is fixed over the diffraction pattern and approximately equal to the average value of m over the pattern (Namba & Stubbs, 1987*a*). Difference Fourier maps for noncrystalline systems have been used in studies of helical viruses to locate heavy atoms, to correct errors in atomic models and to locate water molecules (Mandelkow *et al.*, 1981; Lobert *et al.*, 1987; Namba, Pattanayek & Stubbs, 1989; Wang & Stubbs, 1994).

4.5.2.6.6. Multidimensional isomorphous replacement

At low enough resolution, only one Fourier–Bessel structure factor contributes on each layer line of a fibre diffraction pattern, so that only the phase needs to be determined and the situation is no different to that in protein crystallography. If heavy-atom-derivative specimens can be prepared, the usual method of multiple isomorphous replacement (MIR) (Drenth, 1994) can be applied, which in principle requires only two heavy-atom derivatives. At higher resolution, however, more than one Fourier–Bessel structure factor contributes on each layer line. A generalized form of isomorphous replacement which involves using diffraction data from several heavy-atom derivatives to determine the real and imaginary components of each contributing $G_{nl}(R)$ is referred to as *multidimensional isomorphous replacement* (MDIR) (Namba & Stubbs, 1985). MDIR was first described and used to determine the structure of TMV at 6.7 Å resolution (Stubbs & Diamond, 1975; Holmes *et al.*, 1975), and has since been used to extend the resolution to 2.9 Å (Namba, Pattanayek & Stubbs, 1989). A consequence of cylindrical averaging is that large numbers of heavy-atom derivatives are required: at least two for each Bessel term to be separated. The theory of MDIR is outlined here.

The first step in MDIR is location of the heavy atoms in the derivative structures. The radial coordinate of a heavy atom can be determined by analysis of the intensity distribution in the low-