International Tables for Crystallography (2006). Vol. F, Section 19.5.7, pp. 447–449.

diffraction pattern (Makowski, 1978) and for entire data sets (Yamashita *et al.*, 1995; Ivanova & Makowski, 1998).

19.5.6.4. *Integration of crystalline fibre data*

The variation of reflection shape in detector space can be determined using a few sharp reflections and taking into account parameters related to crystallite size and disorientation in the specimen (Millane & Arnott, 1986). This allows the integration boundary of a reflection to be determined. Sometimes, the boundary encompasses two or more reflections too close to separate; such reflections are considered to constitute a composite reflection.

19.5.6.5. *Integration of continuous data*

In diffraction from noncrystalline fibres, intensity is a function of R on each layer line. Angular deconvolution (Makowski, 1978; Namba & Stubbs, 1985; Yamashita *et al.*, 1995) or profile fitting (Millane & Arnott, 1986) corrects for disorientation and overlap between adjacent layer lines and may also incorporate background subtraction. The intensity determined in this way should be corrected for geometric and other effects if this has not been done previously (Section 19.5.6.2; Namba & Stubbs, 1985; Millane & Arnott, 1986).

19.5.7. Determination of structures

If the amplitude and phase of each diffracted wave are known, structure determination is, in principle, straightforward (Section 19.5.3.4). In practice, however, the phase problem for fibres is more acute than for single crystals because of the limited resolution of the data, and because the diffracted intensities overlap as a result of disorientation and cylindrical averaging. Patterson methods (MacGillavry & Bruins, 1948; Stubbs, 1987) have sometimes been useful, but the cylindrically averaged Patterson function is usually too complicated for detailed interpretation. Phasing by heavy-atom methods is not practical for polymers with small unit cells because of the difficulties in incorporating heavy atoms into the structures. Structures having small unit cells are instead determined by constructing initial models based on chemical information and the observed helical parameters. Extensions of the isomorphous-replacement method (Namba & Stubbs, 1985) have been useful in determining structures, such as those of helical viruses, in which the unit cells are much larger. In all cases, refinement and evaluation of the model structures are essential. A flow chart of the sequential steps in the determination and refinement of fibre structures with small unit cells is shown in Fig. 19.5.7.1.

19.5.7.1. *Initial models: small unit cells*

For many biopolymers, especially polypeptides, polynucleotides and polysaccharides, the repeating unit is a monomer or a small oligomer and the unit-cell dimensions are in the range 10 to 50 Å. Such unit cells can accommodate one or more polymer helices, packed in an organized fashion.

An initial model is constructed from the primary structure of the repeating unit, using bond lengths, bond angles and some conformation angles derived from surveys of accurate single-crystal analyses. The model must satisfy the observed helical parameters and have reasonable intra- and inter-chain non-bonded, hydrogen-bonded and polar interactions.

This preliminary model provides an approximate solution to the phase problem and a starting point for refinement. Since there is no assurance that the refined model represents the true structure, however, stereochemically plausible alternatives must be carefully considered, refined and objectively adjudicated. Alternatives can include both right- and left-handed helices, single helices, and multistranded helices with parallel and antiparallel strands. The next stage involves the packing arrangement in the unit cell. If two or more helices are present, their positions, orientations and relative polarities must be varied in refinement.

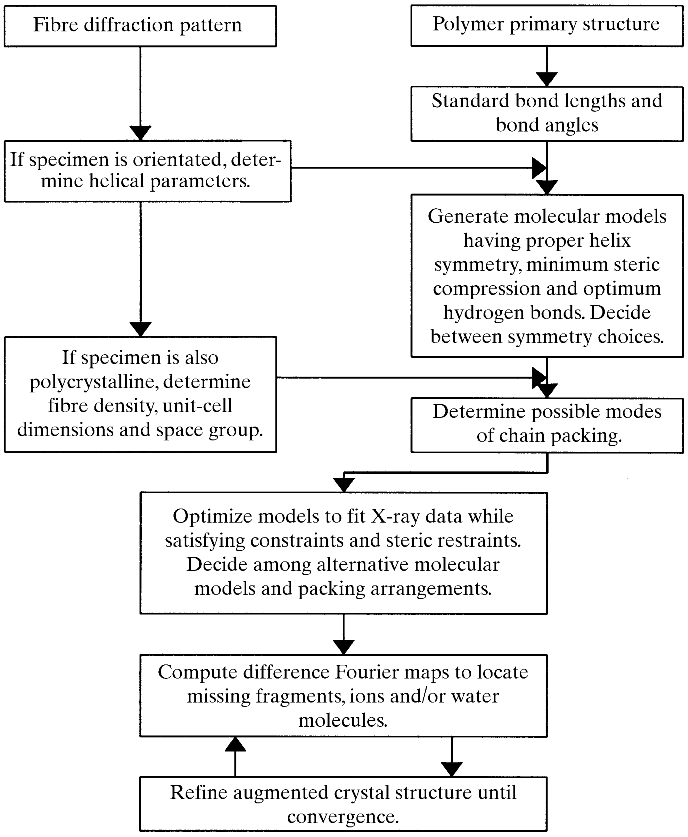
Fibre diffraction pattern
Polymer primary structure
Standard bond lengths and bond angles
If specimen is orientated, determine helical parameters.
Generate molecular models having proper helix symmetry, minimum steric compression and optimum hydrogen bonds. Decide between symmetry choices.
If specimen is also polycrystalline, determine fibre density, unit-cell dimensions and space group.
Determine possible modes of chain packing.
Optimize models to fit X-ray data while satisfying constraints and steric restraints. Decide among alternative molecular models and packing arrangements.
Compute difference Fourier maps to locate missing fragments, ions and/or water molecules.
Refine augmented crystal structure until convergence.

Fig. 19.5.7.1. Flow chart of the principal steps in the determination and refinement of fibre structures with small unit cells.

19.5.7.2. *Refinement: small unit cells*

The widely used linked-atom least-squares (*LALS*) technique (Arnott & Wonacott, 1966; Smith & Arnott, 1978) and the variable virtual bond (*PS*79) method (Zugenmaier & Sarko, 1980) were developed for fibre structures. They are similar in principle to the least-squares refinement procedure for crystalline proteins (Hendrickson, 1985), although bond lengths and bond angles are usually kept fixed in the fibre refinements. The function minimized by the *LALS* program is of the form

$$\Omega = \sum_m w_m \Delta F_m^2 + \sum_i e_i \Delta\theta_i^2 + \sum_j k_j \Delta c_j^2 + \sum_n \lambda_n G_n. \qquad (19.5.7.1)$$

The first term on the right-hand side is the weighted sum of the squares of the differences, ΔF_m, between observed and calculated X-ray structure amplitudes of Bragg reflections or continuous diffraction. Either or both types of data can be used as necessary. The weights, w_m, are inversely proportional to the estimated variance of the data. The second term minimizes the differences, $\Delta\theta_i$, between the expected (standard) values of conformation and bond angles and those in the model; the weights, e_i, are based on empirically determined variances. The third term is designed to take care of non-bonded interactions and thus keep the model free from steric compression. It includes the deviations from target values of both intra- and inter-chain hydrogen bonds and the differences between acceptable and calculated non-bonded distances for those contacts that are smaller than the acceptable limiting values. The

Copyright © 2006 International Union of Crystallography

weights, k_j, are based on the Buckingham energy function for non-bonded contacts and empirical variances for hydrogen bonds. Finally, the fourth term imposes constraints (G_h, with Lagrange multipliers λ_h) for helix connectivity and ring closure, as in a furanose or pyranose, and it vanishes when all such constraints are satisfied. During the refinement, the structure factors are calculated with either the conventional atomic scattering factor f or with a solvent-corrected atomic scattering factor f_w (Fraser *et al.*, 1978; Chandrasekaran & Radha, 1992) given by the function

$$f_w(D) = f(D) - v\sigma_s \exp(-\pi v^{2/3} D^2), \qquad (19.5.7.2)$$

where $D = (R^2 + Z^2)^{1/2}$, σ_s is the electron density of the solvent and v is the excluded volume of the atom. If the van der Waals radius of water is taken as 2 Å, σ_s for water is 0.2984 e Å^{-3}. Equation (19.5.7.2) allows for the solvent contribution to the diffracted intensity and is particularly useful in studying hydrated fibres in which structured and amorphous water can account for up to 50% of the total mass.

19.5.7.3. *Data-to-parameter ratio*

The total number of data used in this refinement process is $M + I + J$, where M, I and J are, respectively, the number of observations in the first three terms of equation (19.5.7.1). If P is the number of parameters refined and H is the number of independent constraints in the last term, then the number of degrees of freedom of the system is $P - H$. The effective number of data is given by $D = (M + I + J) - (P - H)$. The data-to-parameter ratio (*D/P*), a measure of the dependability of the final results, must be greater than one for meaningful refinement. *D/P* is typically in the range 3 to 11 in the analysis of polynucleotide and polysaccharide structures. This ratio is comparable to those commonly reported for single-crystal structures, confirming that fibre-diffraction analysis of polymers, despite the limited number of X-ray data, can yield reliable results.

19.5.7.4. *Initial models: large unit cells*

For large macromolecular aggregates, such as viruses and cytoskeletal filaments, initial models cannot usually be devised using the primary structure of the molecule alone. The largely α-helical filamentous bacteriophages form a rare class of exceptions (Makowski *et al.*, 1980). Molecular-replacement methods, in which initial models are constructed from single-crystal structure determinations of the separated components of the aggregate or from known related structures, can be useful, but because of the limited number of data in a fibre pattern such models can sometimes be difficult to refine.

Multi-dimensional isomorphous replacement (MDIR), an extension of the isomorphous-replacement method of protein crystallography, has been useful in studying helical viruses (Stubbs & Diamond, 1975; Namba & Stubbs, 1985). The dimensions are the real and imaginary parts of the various overlapping structure factors at a given point in the diffraction pattern. Information about both the phases of the structure factors and the relative magnitudes of the overlapping structure factors is obtained from heavy-atom derivatives of the virus; at least twice as many heavy-atom derivatives as the number of significant **G** terms in equation (19.5.3.7) are required. If the structure of a related aggregate is known, MDIR can be combined with molecular replacement (Namba & Stubbs, 1987*a*; Wang & Stubbs, 1994); in this case, fewer derivatives are required.

Layer-line splitting (Franklin & Klug, 1955) arises when the helical symmetry of the scattering particles is close to, but not exactly, integral. For example, tobacco mosaic virus (TMV) has 49.02 subunits in three turns of the viral helix. In this case, the **G** terms in each layer line do not fall at exactly the same Z values in the diffraction pattern. The resulting shifts in the positions of the layer lines can be measured for the native aggregate and, in favourable cases, for heavy-atom derivatives, and used to provide additional phase information (Stubbs & Makowski, 1982). Information from electron microscopy (Beese *et al.*, 1987) and neutron scattering (Nambudripad *et al.*, 1991) has also been used.

19.5.7.5. *Refinement: large unit cells*

Refinement of fibre structures having large unit cells has many parallels to refinement in protein crystallography. Refinement in real space, especially the solvent-flattening approach, has been widely used to improve electron-density maps and is particularly valuable in structure determination of noncrystalline fibres. Since helical aggregates have finite radii, **g** terms [equation (19.5.3.6)] can be set to zero outside a maximum radius and back-transformed to obtain refined estimates of the phases of the **G** terms. More detailed solvent-flattening algorithms can also be used (Namba & Stubbs, 1985).

Molecular models can be refined by methods conceptually related to those of *LALS*. The principal difference is that bond lengths and angles are not kept fixed, but are restrained to remain close to standard values. The restrained least-squares method (Hendrickson, 1985), widely used in protein crystallography, has been adapted (Stubbs *et al.*, 1986) for fibre diffraction and used to refine a number of filamentous virus structures (Namba *et al.*, 1989; Nambudripad *et al.*, 1991). Although effective, the radius of convergence of this method is less than desired, probably because of the limited number of data available from fibre diffraction (Wang & Stubbs, 1993).

Molecular-dynamics methods have been used to increase the radius of convergence of refinement (Wang & Stubbs, 1993). The program *X-PLOR* (Brünger *et al.*, 1987) has been adapted for fibre diffraction and can handle data from both crystalline and noncrystalline fibres. A potential-energy function of the form

$$\Omega = E + S \sum_l \sum_i w_{li} \{[I_o(R_i)]^{1/2} - [I_c(R_i)]^{1/2}\}^2 \qquad (19.5.7.3)$$

is minimized. The first term, E, is an empirical energy function that accounts for distortions in bond lengths, bond angles and conformation angles, and for non-bonded, electrostatic and hydrogen-bonding interactions. The second term accounts for the differences between the observed and calculated X-ray intensities at specific values of R_i on every layer line l; w_{li} is the weight for each observation and S is a normalizing factor. In the most effective use of this method, simulated annealing, the process of heating the structure to a temperature of 3000 to 4000 K is simulated, then the structure is cooled ('annealed') in small increments. At high temperatures, energy barriers between the starting model and structures of lower potential can be overcome; in this way, the radius of convergence of the refinement is increased.

19.5.7.6. *Difference Fourier methods*

As in crystallography, difference maps are used during refinement to correct errors and to identify missing fragments of the model and, in the final stages of refinement, to identify solvent molecules and associated ions.

In crystalline fibre diffraction, the most common difference maps use calculated phases with amplitudes of either $F_o - F_c$ or $2F_o - F_c$. In both cases, weighting the coefficients on the basis of the observed and calculated structure amplitudes has been used to minimize the root-mean-square error in the electron-density maps. Reflections superposed by cylindrical averaging do, however, present problems. One solution is to divide the observed intensity equally among the superposed reflections. This is a reasonable approach in the initial stages of structure analysis, when the reliability of the model is uncertain, and has the advantage of minimizing bias toward the model. Alternatively, the observed intensity may be split in the same ratio as the calculated intensity. This approach, although biased, is more effective for locating solvent molecules and ions in an otherwise well determined structure. Difference Fourier maps have played a significant role in determining the molecular structures and packing arrangements in unit cells mediated by water molecules and cations of several polynucleotide (Chandrasekaran *et al.*, 1995, 1997) and polysaccharide helices (Winter *et al.*, 1975; Chandrasekaran *et al.*, 1988, 1998; Chandrasekaran, Radha & Lee, 1994).

In noncrystalline fibre diffraction, the superposition of intensities due to cylindrical averaging is more serious and must be taken into account. Namba & Stubbs (1987*b*) have shown that the coefficients yielding the most accurate electron-density maps of the full structure have amplitudes of $NG_o - (N-1)G_c$, where N is the number of significant terms in equation (19.5.3.7) (the number of superposed intensities), and the observed intensity is divided in the ratio of the calculated intensity. For filamentous viruses at moderate resolution, N is typically in the range four to six. As in crystallography and crystalline fibre diffraction, maps calculated from amplitudes of $F_o - F_c$ have low noise levels and are most useful for checking the accuracy of final models and for locating solvent molecules.

19.5.7.7. *Evaluation*

As in crystallography, fibre structures are evaluated by statistical measures, such as R values, and by the examination of difference maps. Fibre-diffraction R values are inherently lower than those expected in crystallography, particularly when large numbers of intensities have been superposed by cylindrical averaging (Stubbs, 1989). The largest likely R value for noncrystalline TMV at 3 Å resolution is about 0.31 and for polycrystalline DNA at 3 Å resolution it is about 0.41, both significantly less than the value of 0.59 to be expected from noncentric single-crystal analyses (Millane, 1989).

Comparison of R values alone is not necessarily a reliable way to discriminate between competing models. Such discrimination is often required for structures with small unit cells, for which alternative models are routinely refined (Sections 19.5.7.1 and 19.5.7.2). The relative merits of any pair of competing models can be assessed on the basis of several types of statistics (Arnott, 1980) using Hamilton's significance test (Hamilton, 1965), which considers not only residuals but also numbers of degrees of freedom (Section 19.5.7.3). Such a test is essential. There are many examples in the literature where R values have been lowered by the simple process of increasing the number of degrees of freedom; a decreased R value obtained in this way may or may not have any significance.

Difference Fourier maps have been used to evaluate crystalline fibre diffraction analyses for many years, for example, to reject the controversial Hoogsteen base pairing in double-stranded DNA (Arnott *et al.*, 1965), and later to discriminate between 10- and 11-fold double helices of RNA (Arnott *et al.*, 1967). Difference maps have been essential in the refinement of fibre structures with large unit cells (Namba *et al.*, 1989; Wang & Stubbs, 1994), both to identify errors in early models and to confirm that the final structures contained no major errors or omissions.

19.5.8. Structures determined by X-ray fibre diffraction

The α-helix of several synthetic polypeptides (Pauling & Corey, 1951), the double helix of DNA (Watson & Crick, 1953), the ribbon structure of cellulose (Meyer & Misch, 1937) and the low-resolution structure of tobacco mosaic virus (Barrett *et al.*, 1971) were early examples of structures determined by fibre diffraction. Early workers also examined a number of fibrous proteins (Bailey *et al.*, 1943). In the past 50 years, developments in theory and practice and the availability of fast computers have made it possible to determine and refine about 200 biological polymer structures of varying complexities. The largest repeating units in polypeptides, polynucleotides and polysaccharides solved to date correspond to a tripeptide, a tetranucleotide and a hexasaccharide, respectively.

19.5.8.1. *Polypeptides*

The structural details of the α-helix and β-sheet, the principal secondary-structure elements of proteins, have emerged from the analysis of synthetic polypeptides (Pauling & Corey, 1951, 1953). Analysis of noncrystalline fibre-diffraction patterns led to the triple-helical coiled-coil model of collagen (Ramachandran & Kartha, 1955; Rich & Crick, 1955). Recent studies on the organization of β-sheets in peptides of up to about 45 residues are providing an understanding of the molecular details of amyloid fibrils, related to Alzheimer's disease (Inouye *et al.*, 1993; Malinchik *et al.*, 1998).

19.5.8.2. *Polynucleotides*

The molecular structures of a series of DNA and RNA helices have been determined and refined using data from polycrystalline fibres (Arnott *et al.*, 1969; Chandrasekaran & Arnott, 1989). These include the canonical A, B and C forms of DNA, corresponding, respectively, to 11-, 10- and 9.3-fold right-handed antiparallel Watson–Crick base-paired helices. Structural differences between the three have been attributed to changes in furanose puckerings and helical parameters: the A form has C3-*endo*, but B and C have C2-*endo* or analogous C3-*exo* puckers. All RNA duplexes are members of the A family. Later important structures included the sixfold single helix of poly (C) (Arnott *et al.*, 1976), a compact eightfold double helix for poly d(AT) and poly d(IC) (Arnott *et al.*, 1983), and the left-handed Z-DNA for poly d(GC) (Arnott *et al.*, 1980). Difference Fourier syntheses were instrumental in locating a spine of water molecules in the minor groove and a series of sodium ions and water molecules that bridge the phosphate groups of adjacent DNA molecules in the tenfold helices of poly (dA)·poly (dT) (Chandrasekaran *et al.*, 1995), poly (dA)·poly (dU) and poly d(AI)·poly d(CT) (Chandrasekaran *et al.*, 1997). Data from noncrystalline fibres have been used to determine, among others, the structures of DNA·RNA hybrid duplexes (Arnott *et al.*, 1986), a DNA triple-stranded helix (Chandrasekaran *et al.*, 2000*a*) and two RNA triple-stranded helices (Chandrasekaran *et al.*, 2000*b*,*c*). In each case mentioned, the best model was clearly preferred statistically (Hamilton, 1965) and had an R value between 0.2 and 0.3 to about 3 Å resolution.

19.5.8.3. *Polysaccharides*

Among the three-dimensional structures determined for industrially useful and biologically important polysaccharides are the gel-forming calcium *i*-carrageenan (Arnott, Scott *et al.*, 1974), sodium pectate (Walkinshaw & Arnott, 1981), gellan (Chandrasekaran *et al.*, 1988) and welan (Chandrasekaran, Radha & Lee, 1994), and a series of distinct helical forms of the glycosaminoglycan hyaluronan (Arnott & Mitra, 1984). The conformations of these molecules are delicately controlled by ions, such as sodium, potassium and calcium. The repeating units range from a simple monosaccharide to a branched pentasaccharide.

Specific interactions among the polysaccharides and their associated small molecules can be correlated with their observed properties. A number of neutral polysaccharides, such as cellulose, chitin and mannan, are twofold ribbon-like helices, which aggregate

references