

Biophysical Electron Microscopy

Basic Concepts and Modern Techniques

Edited by

P. W. HAWKES

CNRS, Toulouse, France

U. VALDRÈ

University of Bologna, Bologna, Italy



ACADEMIC PRESS
Harcourt Brace Jovanovich, Publishers
London San Diego New York
Boston Sydney Tokyo Toronto

7 Image Analysis of Electron Micrographs

M. F. MOODY

The School of Pharmacy, London University, 29/39 Brunswick Square,
London WC1N 1AX, UK

7.1 INTRODUCTION

7.1.1 Need for image processing

If electron microscope techniques were perfect, this chapter would be unnecessary. But even the best conceivable instruments and techniques yield images with deficiencies, some of which can be reduced by subsequent processing of the micrographs. In increasing order of severity, the three most serious corrigible deficiencies are projection, loss of phase information and noise. Projection (i.e. the contribution of the entire thickness of the specimen to the transmitted wave) and loss of phase information (which contributes to the intensity of the image only when that is defocused enough to limit its resolution) can both, in principle, be corrected completely, provided there is no noise. But, at high magnification, noise is unavoidable; we are allowed only a choice between its two main sources. We can use an adequate exposure for the image, and the electrons will destroy all high-resolution detail in the specimen*; or we can limit the exposure to preserve the specimen, and the "image" will consist of relatively few electrons with a somewhat random distribution.

If we take the second option, the higher potential resolution of an undamaged specimen can be realized only by summing very many identical images. This is possible only if the structure has a large number of identical repeating units—i.e. an accurate, undistorted, symmetry. If, however, the symmetry is low, we are forced to accept the first option, and the resolution is then limited by specimen damage. The type of symmetry present in the specimen thus determines both the optimal level of exposure (and type of preparation), and also the extent to which the micrographs can be corrected. Thus the most useful image-processing techniques exploit symmetries with the largest number of repeating units: two-dimensional crystals and helices. At the lowest level, however, where the particles are asymmetrical and disoriented, it is questionable whether image correction will ever be able to proceed very far.

* This destruction is the consequence of two features of the electron: its low mass (giving a long wavelength unless its momentum is high, when there is sufficient energy to cause serious specimen damage); and, even more, its charge (needed for focusing, but capable of ionizing the specimen). (Damage would be very much lower with, for example, neutrons of thermal energies, but their use in a high-resolution microscope faces almost insuperable technical difficulties.) In addition to the two basic, unavoidable sources of noise, electron micrographs generally suffer from an additional (and even larger) noise resulting from irregularities in the supporting film or in the negative stain.

The processing of micrographs thus follows a logical sequence. First the symmetry (if any) is determined, then it is exploited for image enhancement by averaging, and finally the effects of projection are removed by combining the data from several different images. The arrangement of sections in this chapter (after the first introductory one) follows this order.

7.1.2 Previous accounts of the subject

The most detailed general references include the books by Misell (1978) and by Saxton (1978); Baker's (1981) bibliography lists the literature up to the end of 1979. More recently, several excellent accounts of specific techniques have appeared; these will be referred to in the appropriate sections of this chapter. By referring the reader to such publications, it has been possible to prevent this chapter from growing to an even more inordinate length.

7.1.3 Purposes of this chapter

This chapter has two somewhat distinct purposes. First, it aims to introduce microscopists, whose training has been biological rather than physical, to a series of techniques that use relatively sophisticated mathematical methods. There are several reviews that cover primarily the applications of image analysis. These include Crowther and Klug (1975), Aebi *et al.* (1982), Crowther (1982), and Aebi *et al.* (1984). Reference to these, or to original papers, permits microscopists to judge the relevance of these techniques to their own problems. If the techniques are relevant, and the microscopist wishes to understand how they work, a clear and (preferably) non-mathematical introduction is then needed. This chapter therefore starts with a long section outlining the techniques' two main theoretical pillars, symmetry-theory and Fourier transforms.

The second purpose is to survey all the image-analysis techniques, including the recent ones, but concentrating particularly on those that have had the greatest number of successful applications. An attempt has been made, both to survey them from a coherent viewpoint (to make apparent their interconnections), and to make the main sections of this survey as non-mathematical as possible. Although there are places where mathematics is indispensable, these have been concentrated in a few parenthetical sections that can be omitted without rendering the account incomprehensible. (These sections are 7.2.3(i)–7.2.4, 7.4.5(a), (b), 7.4.6(d), 7.4.7(b), 7.5.3(b), 7.5.3(d) and 7.6.4(c)–(d). The first of these groups gathers together the mathematics of helical diffraction and of diffraction from rotationally symmetric objects, topics for which there is no convenient summary.)

7.1.4 Starting image processing

Two very different classes of research worker, with very different needs, might embark on using image-processing techniques. One class, the biological electron microscopist, is probably best advised to collaborate, at least initially, with an established group, though an understanding of the basis of the techniques would probably make the results

of the collaboration more useful and relevant. The other class is a research group already equipped in some area of molecular biophysics (such as crystallography or NMR) which wishes to start processing electron micrographs. For hardware, they will need access to an optical diffractometer, a microdensitometer giving accurate computer-readable data, and a computer that is preferably above the 16-bit minicomputer level. (An output device for displaying optical density arrays as pictures would also be very convenient.) As for software, many packages are available; see the descriptions by Smith (1978), Saxton *et al.* (1979), Frank *et al.* (1981), Trus and Steven (1981), Van Heel and Keegstra (1981) and Hegerl and Altbauer (1982). In addition to these, the original software written at Cambridge is now also installed and adapted at Brandeis University and at the EMBL (Heidelberg).

7.2 BASIC PRINCIPLES

7.2.1 Outline of biological symmetry

In symmetry theory, as in any area of applied mathematics, one studies the mathematical implications of a hypothesis that could be true in systems of interest; then experimental tests of these implications can reveal the validity of the hypothesis. In symmetry theory, the hypothesis concerns the arrangement of identical subunits (which may be single macromolecules) in a larger aggregate. The hypothesis is that, being identical, these subunits have no reason to become arranged in any way that differentiates between them. All the subunits in the aggregate are thus supposed to be indistinguishable—not only in their internal structure, but also in the ways in which they associate with their neighbours. This indistinguishability is called *equivalence*. Identical proteins do not invariably associate in this way; indeed, given the constant thermal movements within protein molecules, it is most unlikely that any two "identical" proteins could have the same structure at any given moment. Nevertheless, many protein aggregates do show the structures to be expected if their subunits were exactly equivalent, presumably because the time-averaged properties are identical. This is most useful, since some sort of symmetry is essential for the picture averaging required for high-resolution electron microscopy.

We now look at the mathematical implications of the hypothesis of equivalence. An aggregate of equivalent subunits will have *symmetry*: the aggregate will look exactly the same after it has been moved in certain ways. For example, rotation about a certain axis and by an appropriate angle may bring each subunit into exactly the position that was previously occupied by another subunit. Movements that have this effect (symmetry operations) cannot include reflections or inversions if the subunits contain only L-amino acids or D-sugars. Then the only permitted symmetry operations are rotations, translations (i.e. uniform movements with no rotational component), and combinations of them (screw displacements).

Only certain combinations of symmetry operations have the required self-consistency to be realized in symmetrical aggregates. These permissible combinations are called *symmetry groups*. Many symmetry groups contain, as subgroups, smaller permissible combinations; for example, an object with six-fold rotational symmetry must also have

two-fold rotational symmetry. The symmetry groups are best classified according to the number of independent translations.

(a) *Point-groups*

When this number is zero, so that rotations are the only symmetry operations, we have the point-groups. A single N -fold rotation axis constitutes the cyclic point-group C_N , the symmetry of a ring of subunits. If there is a double ring, which is the same when turned upside-down (requiring two-fold axes perpendicular to the N -fold axis), we have the dihedral point-group D_N . (Dihedral symmetry is often found in enzyme complexes, and it is also the highest point-group symmetry that is consistent with helical symmetry.) The remaining point-groups are the tetrahedral, octahedral and icosahedral groups, named after the regular polyhedra which possess these point-group symmetries. The icosahedral point-group, with 60 equivalent positions, describes the symmetry of many spherical viruses, and special techniques have been developed to exploit this symmetry for three-dimensional reconstruction from micrographs (Section 7.6).

(b) *Line-groups*

One independent translation generates, by itself, a line of subunits. (If such a line possesses a two-fold axis, it is non-polar.) Aggregates of fibrous proteins (collagen, tropomyosin, etc.) often give band patterns with this sort of one-dimensional symmetry. They are one-dimensional in projection, because the molecules are so flexible that order is easily lost in the direction perpendicular to the fibre's length. A different symmetry type is given if the single independent translation is combined with a rotation to yield a screw displacement. Repeated application of this gives helical symmetry. (The great biological importance of this justifies a special section (7.2.3) for helical structures and their Fourier transforms.) Certain types of helical symmetry apply to lattices, as well as to isolated fibres. Of these, the most important is the two-fold screw axis. Successive subunits are rotated by 180° about the helix axis, so that they point alternately to the left and right. If such a structure is viewed in projection, the left- and right-pointing subunits appear as mirror images. Such an arrangement is called a *glide-line* (see Fig. 7.52). (Note that, although glide- or mirror-lines cannot apply to the three-dimensional structures of biological macromolecules, they can apply to the projections of these structures.)

(c) *Plane-groups*

Two independent translations, without rotational symmetry, generate a sheet with the symmetry of the plane-group $p1$. The sheet can also have an overall rotational symmetry, but only with two-, three-, four- or six-fold axes (giving the plane-groups $p2$, $p3$, $p4$ or $p6$; International Tables for Crystallography, 1983). Aggregates of globular proteins can give thin crystals whose images are based on these plane-groups. The most favourable cases for image analysis occur when the molecules form monolayers; these arise naturally in some cell membranes and walls, especially in bacteria. Because of the distinction between the interior and exterior of the cell, the two surfaces of such a monolayer display quite different parts of the molecules, consistent with the plane-groups $p1$, $p2$, $p3$, $p4$ or $p6$. But synthetic monolayer crystals that form in solution often show no such distinction

between the two surfaces. This implies that there must be two-fold rotation- or screw-axes relating the two surfaces, and the symmetry must be one of the two-sided plane groups listed by Holser (1958).

(d) *Space-groups*

Finally we come to symmetries with three independent translations. These are the classical crystallographic space-groups (International Tables for Crystallography, 1983). As before, operations (such as mirror- or glide-planes, or centres of symmetry) that change chirality cannot be present if all the amino acids are "L" and all the sugars are "D". Then, instead of the full 230, only 65 space-groups are permitted. Although of fundamental importance in protein crystallography, these are much less useful in electron microscopy; structures thick enough to have the full symmetry of any space-group are unlikely to give good images.

Sometimes molecules form symmetrical groupings before they associate to form the crystal. The rotational symmetry of the groupings may, or may not, apply to the crystal as a whole. It can do so only if certain conditions are satisfied: the symmetry axis must be two-, three-, four- or six-fold, and it must be positioned correctly in the crystal (e.g. perpendicular to the sheet in a one-sided crystal). If these conditions are not met, the symmetry axis is local (or non-crystallographic), and applies only within a small region of the crystal.

7.2.2 Introduction to Fourier transforms

Fourier transforms are of fundamental importance in the theory of image formation and also in most of the methods which have been developed for image analysis and processing. The accounts of these topics given here use, in the main, a set of results from Fourier transform theory which are applied in simple intuitive or geometrical form. Because of its simplicity, this form of the theory is very widely used in initial investigations. It is often possible to see the main outlines of a problem in this way, and to postpone application of the detailed mathematical theory to the point where precise numerical values are required. (Readable mathematical accounts are given by Lipson and Taylor, 1958, and by Bracewell, 1986; see also Chapter 3.)

(a) *Simple one-dimensional pictures and Fourier analysis*

We start by considering the basic principles of Fourier analysis. Suppose that we wish to analyse a periodic curve—a curve that repeats exactly after a certain distance a . (This curve—Fig. 7.1b—could be thought of as the microdensitometer trace along an electron micrograph—Fig. 7.1a—of a fibrous protein aggregate showing periodic banding; we shall refer to it as the "picture".) We are interested in determining its component periodicities, each of which will be called a Fourier component (or simply a component). The type of component used by Fourier analysis is a cosine wave (Fig. 7.1c,d), completely defined by three variable parameters (Fig. 7.1c). These are *amplitude* (half the difference in height between a peak and a trough); *period* (the distance between two successive

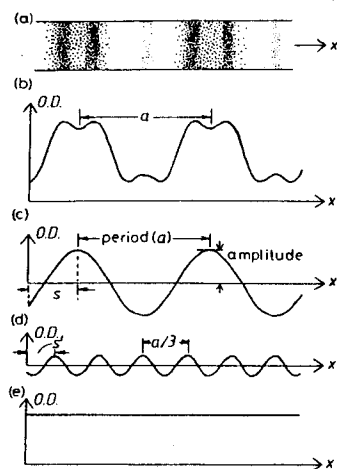


Fig. 7.1 An illustration of the principles of one-dimensional Fourier analysis. (a) An electron micrograph shows periodic banding. (b) Its optical density trace plots optical density (O.D.) as a function of the x -coordinate. This periodic function is the sum of the three Fourier components shown in (c), (d) and (e). (c) The three defining parameters of the basic Fourier component are indicated. s is the shift of the first maximum; $s = \text{period} \times \text{phase}/360^\circ$. (d) This Fourier component has a different shift s' , and a period which is a submultiple of the basic period a . (e) This Fourier component is a constant term that causes the minimum of the sum (curve (b)) to be greater than zero.

peaks); and the *shifts* (relative positions) of different cosine waves, measured by their *phases*.

How should we measure phases? Should we simply measure the distances by which all the cosine waves must be shifted, relative to some arbitrary line? This plan has the disadvantage that the same shift has a far greater effect on a wave of short period (Fig. 7.1d) than on one of long period (Fig. 7.1c). So it would seem best to use the ratio shift/period. This ratio could have any value from zero to infinity. However, the cosine waves repeat exactly when shifted by the length of their period. Consequently a shift by any whole number of periods makes no difference. Therefore we need to represent the ratio so that it repeats after every whole number. This can be done by using an angle, given by 360° times the ratio of shift/period, to represent phase.

If the band pattern were symmetrical about the origin (Fig. 7.2a), then all its Fourier components would be cosine waves with one of two possible phases. A cosine wave (Fig. 7.2b) is symmetrical about the origin, where it has either its maximum or its minimum. If it has its maximum there, its phase is (+) or 0° ; if it has its minimum there, its phase is (-) or 180° .

By means of Fourier analysis, we can find a set of such cosine waves whose sum equals the original curve (or picture). Thus the periodic curve in Fig. 7.1b equals the sum of the cosine waves in Figs 7.1c and d, plus the uniform density in Fig. 7.1e. The set of component cosine waves obtained from a picture is unique; there is only one set to analyse, or to compare with the set obtained from another picture.

Once we have found the Fourier components of a picture, we need to represent them in a diagram. Since each component of Fig. 7.1 has three numbers (period, amplitude

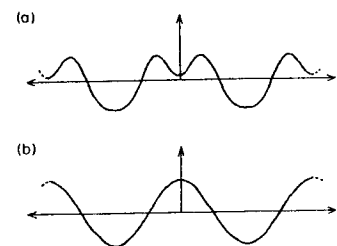


Fig. 7.2 (a) This function is symmetrical about the origin (i.e. it is the same at x and at $-x$). (b) One of its Fourier components, a cosine wave with no shift. (The other type of component is a negative cosine wave, equivalent to a shift of half its period.)

and phase), it might seem that three-dimensional space would be appropriate. But, if we used this for one-dimensional pictures, how could we represent two-dimensional pictures? It turns out that the most suitable representation is to use one number as the coordinate along an axis (so a one-dimensional picture has a one-dimensional transform). Then, at each point along the axis, we have two numbers, that is, a vector.

Which of the three quantities (period, amplitude, phase) should we represent along the axis? When plotting the picture, the axis represents distance. So it seems that we should use period, the only number that measures a distance. However, the periods of the possible Fourier components vary from some minimum (a cosine wave with the small oscillations needed to represent the finest detail in the picture), up to infinity (an infinite period is needed to provide the uniform background level in the picture, as in Fig. 7.1e). A more convenient range is obtained if, instead of using the period, we use its reciprocal (called *spatial frequency*). Instead of ranging from a minimum to infinity, spatial frequency ranges from zero to some maximum. The remaining two numbers (amplitude and phase) are represented by a vector. Since phase is measured as an angle, we use it to set the angle of the vector. The vector's length therefore represents the amplitude, which ranges from zero (for a Fourier component that is missing in the picture), up to some maximum (corresponding to the strongest Fourier component).

So we are led to a representation like that in Fig. 7.3. Each vector represents a Fourier component, as we have just explained. The position of the vector (the position of its widest part in Fig. 7.3) gives the spatial frequency ($1/\text{period}$) of the component. The vector at the origin, with zero spatial frequency and hence infinite period, gives the background density of the picture. If (as we assume in this figure) the picture is periodic with a repeat a , then all the component cosine waves must also repeat after a distance

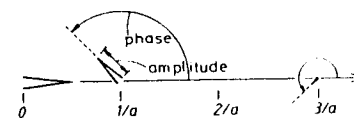


Fig. 7.3 Representation of the three Fourier components of Fig. 7.1 in a single diagram. Each component is a vector, drawn as a V-shape whose point is the tip of the vector. The wide part of the "V" is the vector's base, and lies on the X-axis. The vector's angle is the Fourier component's phase, and its length is the component's amplitude. The origin of each vector is positioned at the reciprocal of the component's period, along a spatial-frequency or X-axis.

a. Therefore they must have periods $a, a/2, a/3, \dots$ or spatial frequencies $1/a, 2/a, 3/a, \dots$. Since the spatial frequencies are the reciprocals of real distances, the spatial frequency coordinate axis is considered to lie in reciprocal space. To remind us of the connection between the two axes, the real axis is labelled x , and the reciprocal (or spatial frequency) axis is labelled X .

(b) *General pictures and the Fourier transform*

Non-periodic pictures. The representation shown in Fig. 7.3 is satisfactory for the one-dimensional periodic density distribution shown in Fig. 7.1a, but it will need three extensions to represent other types of picture. First, pictures often fail to repeat exactly. This causes no problem with their representation. Vectors are now required with spatial frequencies that are not exact multiples of $1/a$ —indeed, we need an infinity of vectors, continuously distributed as a function of the spatial frequency X . Such a distribution is called a Fourier transform, or F.T. (which, for brevity, we shall often refer to simply as a transform, when there is no ambiguity). The other two extensions will be discussed below.

Complex pictures. These are pictures that show variations not merely of amplitude—darkness—but also of phase. The electron waves that have passed through the specimen are like this, and their phase distribution is important if we are discussing imaging processes (Chapter 4). The need for complex pictures also arises when we discuss the Fourier transform of a Fourier transform, for then the “picture” is itself a transform, and its pixels have both amplitude and phase. The problem is that such a picture has twice as much information as the conventional kind, so there must also be twice as much information in its transform. But how are we to extend our transform representation to encompass twice as much information? We are already using vectors (each point carries two independent numbers), and there is no convenient way to pack twice as much information (four numbers) into each point. So we must double the length of the transform: the spatial frequency axis becomes extended in the negative direction, and we add components with negative spatial frequencies. Because of its generality, this extended representation will be used from now on. But, while using it, we shall often be discussing the transform of an ordinary picture, in which only the amplitude (optical density) varies. For such pictures, both sides of the axis are unnecessary: the positive side suffices, as in Fig. 7.3. In this case the negative side contains no new information, and must mirror in some way the distribution on the positive side. What is the form of this mirroring? It turns out that the amplitudes on both sides are mirrored exactly. However, the phase on the negative side of the axis, instead of being identical to the phase on the positive side, has the opposite sign (Fig. 7.4). This special type of symmetry is found only in the transforms of pure amplitude pictures, and is called Friedel symmetry.

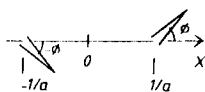


Fig. 7.4 Spatial frequencies (X) are now allowed to be negative, so as to represent the transforms of complex functions. This diagram represents the transform of a real function; so the transform vectors (described in Fig. 7.3) at $1/a$ and $-1/a$ must have the relationship (Friedel symmetry) shown here.

The phases are even simpler if the picture also has two-fold symmetry (i.e. looks the same when turned upside-down). Then only pure cosine waves are needed as its Fourier components, as with the one-dimensional case shown in Fig. 7.2. The phases are then only 0° or 180° .

Two-dimensional pictures. To represent these, we need Fourier components that cover the two-dimensional x, y -plane. Three such components are sketched in Fig. 7.5c, e and g. When added together, they yield the periodic picture (a). To represent these two-dimensional components, we need two spatial frequency axes (X and Y), so our reciprocal

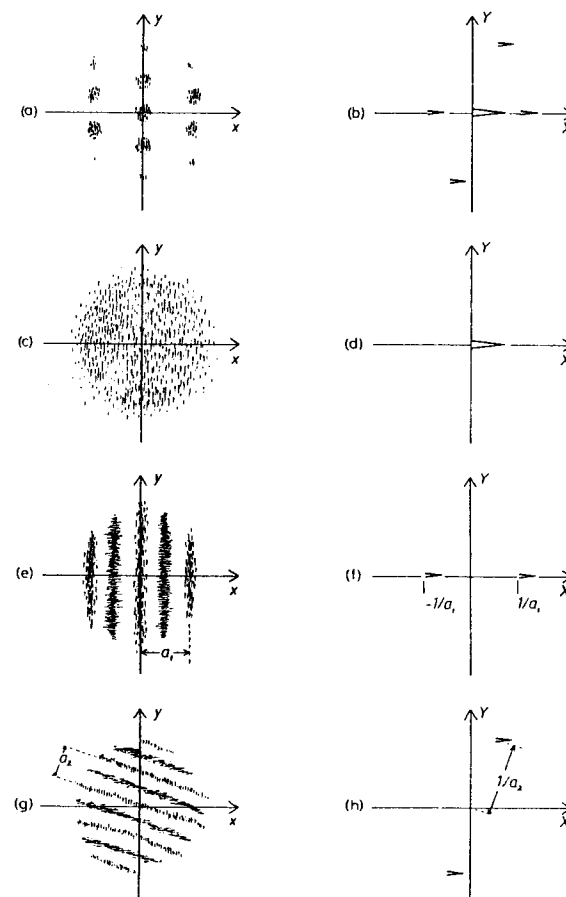


Fig. 7.5 The Fourier transform (F.T.) of a two-dimensional picture. (a) A periodic (repeating) two-dimensional picture. (b) F.T. of (a), which is the sum of (d), (f) and (h). (c) Simplest Fourier component of (a): uniform density. (The vertical shading indicates that it is positive.) (d) F.T. of (c). (e) Simplest non-uniform Fourier component of (a): cosine waves parallel to the y -axis, with period a_1 . (f) F.T. of (e). (g) Third Fourier component of (a): oblique cosine waves with period a_2 . (h) F.T. of (g).

space becomes two-dimensional. The simplest component is the uniform density (c). This is a cosine wave of infinite period, i.e. of zero spatial frequency, so its corresponding vector (d) lies at the origin. The next simplest (e) has bands parallel to the y -axis and varies only along the x -axis. Consequently this is really a one-dimensional component, and its representation is confined to the corresponding spatial frequency axis (X -axis). Since the bands in (e) are real (the phase variations are discontinuous), we have two vectors with Friedel symmetry, i.e. of equal length, and with equal and opposite phases (here zero). Finally, we consider the tilted set of bands (g). Where shall we put the vectors for this? Our choice must be consistent with what we have already used for the bands (e). There the two vectors lie on a line (X -axis) that is perpendicular to the bands; and the distance of each vector from the origin is the reciprocal of the period a_1 of the bands. Maintaining this rule with the bands (g), we have the vectors shown in (h). The picture (a), being the sum of the three components (c), (e) and (g), has a transform (b) which is the sum of the vectors in (d), (f) and (h).

Suppose we rotate this two-dimensional picture; what will happen to the transform? Rotation of the picture will rotate all the sets of density bands. And rotation of these will rotate the positions (not the phases) of the vectors that represent them in the transform. (For their position is fixed by the condition that the line joining them to the origin must be perpendicular to the bands.) Consequently, rotation of the picture causes the same rotation of the transform.

A picture consisting of a single point. The picture in Fig. 7.5a was periodic, needing only a very few components, represented by the vectors in Fig. 7.5b. We next consider a non-periodic picture—the very simplest, a point. To begin with, we place the point at the origin. A point (or, in a one-dimensional plot, a peak) can be represented by superposing a whole series of pure cosine waves (Fig. 7.6). All the waves have their maxima coinciding at the origin, where they add together to give a peak. Elsewhere, the waves interfere, and their sum gives nothing. Now all these waves (the Fourier components of the point) have the same amplitude, the same phase (0° , since they are all in phase at the origin), and a continuous spread of spatial frequencies. The transform of the peak (Fig. 7.7a) is thus given by a uniform distribution of vectors (Fig. 7.7b).

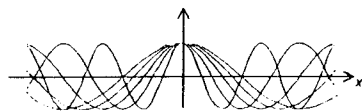


Fig. 7.6 Illustration of how a peak at the origin can be constructed by summing cosine waves of all possible periods. A small selection of those waves is shown here; they reinforce near the origin. At larger positive or negative values of x , some of the cosine waves become negative and others positive, so that the sum there approximates to zero.

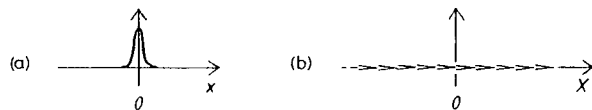


Fig. 7.7 A Fourier transform pair representing the fact illustrated in Fig. 7.6. (a) A peak at the origin. As in Fig. 7.6, this can be represented as the sum of cosine waves of all different periods. (b) Each of these component cosine waves is a vector of zero phase.

Next suppose that the peak was not at the origin, but displaced from it (Fig. 7.8a). We still need the same set of cosine waves, but now their maxima must add together at the new position of the point. To move the maxima to the new position, the waves need a phase shift. But a constant phase shift would move all the waves by the same fraction of their period, i.e. by different distances, and there would be no one point at which they would all get into phase. So the phase shift must vary with spatial frequency, and it turns out that the required variation is a uniform rotation, as shown in Fig. 7.8b.

Finally, suppose that the peak is a point in two-dimensional space. If this "point-peak" were at the origin of real space, we would apply the same argument as for a one-dimensional peak at the origin. So we should need a continuous distribution of two-dimensional cosine waves, all with 0° phase and with every possible spatial frequency and direction. (Reciprocal space would be completely filled with identical arrows, all pointing to the right.) If the required point-peak were displaced from the origin (Fig. 7.9a), the phases of the vectors would change, as before. Along the line joining the point-peak to the origin of real space, we should have a one-dimensional picture, as in Fig. 7.8a, with the transform in Fig. 7.8b. That one-dimensional diagram contains all the features of the two-dimensional picture. The perpendicular direction is redundant; in the transform, it contains only exact copies of the vectors in Fig. 7.8b. So the actual distribution that we need for a point-peak in two-dimensional space is shown in Fig. 7.9b. The amplitude is everywhere constant, but the phase rotates uniformly. This

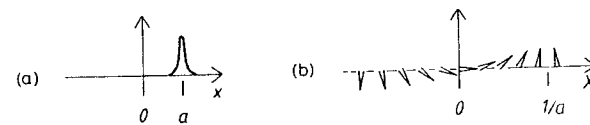


Fig. 7.8 (a) The peak at the origin in Fig. 7.7 is shifted to $x = a$. (b) In order to bring all the cosine waves into coherence at $x = a$, their phases must rotate uniformly, as shown here.

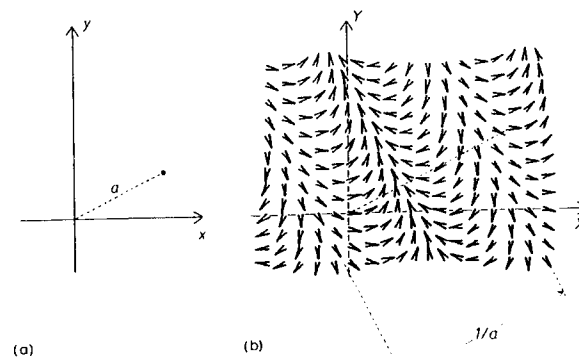


Fig. 7.9 (a) Two-dimensional picture consisting of a small point a distance a from the origin. (b) F.T. of (a), consisting of vectors of the same length, and showing the rotation seen in Fig. 7.8. The field of vectors has the appearance of plane complex waves. The wavefronts are perpendicular to the vector in (a), and their separation is the reciprocal of its length.

“phase wave” has its wave fronts perpendicular to the line joining the point-peak to the origin of real space. The period is the reciprocal of the length of this line.

If the point-peak were in the same position on the other side of the real space origin, Fig. 7.9a would be turned through 180° . What would be the corresponding phase-wave? The rotation of the point-peak is equivalent to rotating the coordinate axes while keeping the point-peak fixed. Consider only the phases on the x -axis. The new phase at a distance $+x$ from the origin would be the same as the old phase at $-x$. Two consequences follow. The phase at the origin, where $+x = -x$, is unchanged. And the phase changes now found when moving from the origin in a positive direction must be the same as those previously encountered when moving in a negative direction. Figures 7.8 and 7.9 show that, when moving in a negative direction, the phase rotated clockwise; so this is the sense in which it must now rotate when moving in a positive direction.

We can obtain a pair of points (as in Fig. 7.11a) by adding together a point in the upper right quadrant (Fig. 7.9a) and a corresponding point in the lower left quadrant. The transform of this pair of points is then the sum of Fig. 7.9b, and of this same Figure when the vectors have a reversed direction of rotation. Because of the close connection between the phase waves of these two transforms, we get a simple result when we add them together. This is shown, for just the X -axis of reciprocal space, in Fig. 7.10b. The vectors from Figs 7.8 or 7.9 are shown by continuous lines, and those from the rotated picture are shown by broken lines. The superposition of the two sets of vectors at each point is shown in Fig. 7.10c. It is clear that all the vectors resulting from this superposition have either 0° or 180° phase. This means that they are either positive (0°) or negative (180°). In two dimensions, we obtain the result shown in Fig. 7.11: a picture (a) consisting of just two point-peaks (or pixels, or vectors), of equal amplitude and zero phase, gives a transform (b) consisting of a single density wave, of the sort used as components in Fig. 7.5.

But suppose the phases of the two vectors on the left were not zero; what connection between their phases would ensure that the transform was a single density wave? At the origin of the transform, we sum all the vectors of the picture. If these vectors have Friedel symmetry, as in Fig. 7.4, they can be divided into pairs with the same amplitude but opposite phase. When added together, the members of such a pair must (like the

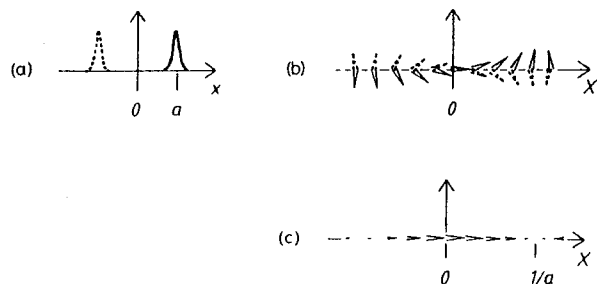


Fig. 7.10 (a) A pair of peaks, the right-hand peak (continuous line) being the same as that in Fig. 7.8. (b) The F.T. of each peak in (a) (the vectors with a continuous line refer to the right-hand peak). Note that the vectors from each peak rotate in opposite directions, and are consequently related by mirror symmetry about the X -axis. (c) The sum of the two sets of vectors in (b), i.e. the overall F.T. of (a).

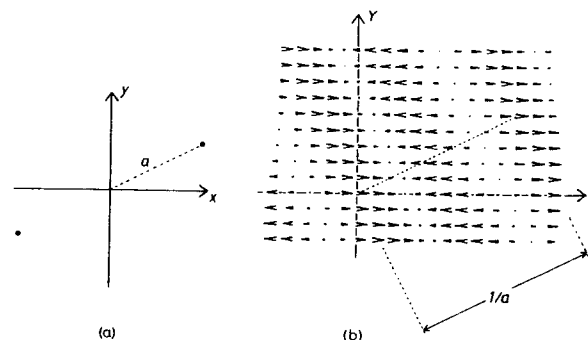


Fig. 7.11 (a) Two-dimensional picture consisting of a pair of peaks, each a distance a from the origin. (b) F.T. of (a); the vectors are horizontal (i.e. real), and arranged in bands that repeat with a spatial frequency $1/a$. (The intensity, however, repeats with a spatial frequency of $1/2a$.)

vectors of Fig. 7.10b), give a resultant that is parallel to the X -axis. Elsewhere in the transform, the contributions of each member of a pair will rotate by equal angles in opposite directions (as in Fig. 7.10b). So the resultants must always be parallel to the X -axis, i.e. we have a single density wave.

(c) *The Fourier transform of a Fourier transform*

We now have two alternative ways of viewing the Fourier transform. The first is more appropriate to an ordinary picture, without phase variations. We find the picture's component periodic density bands, as in Fig. 7.1 or Fig. 7.5. Each set of bands gives rise to a pair of vectors in the transform. For a picture without phase variations, the vectors have Friedel symmetry.

In the second way of viewing the transform, we start by dividing the picture into pixels. If the “picture” is a Fourier transform, the “pixels” are vectors. Each pixel or vector gives rise, in the transform of the picture, to a phase wave like that in Fig. 7.9. The wavefronts are perpendicular to the line joining the pixel to the origin of real space, and the phase of the wave, at the origin of reciprocal space, is given by the phase of the pixel. A pair of equal pixels, symmetrically arranged about the origin, and with Friedel symmetry, gives rise to a density wave in which the peaks alternate in sign.

Bearing in mind these two ways of viewing Fourier transformation, we consider the effect of two successive Fourier transformations, i.e. of Fourier-transforming a Fourier transform. We view the first transformation in the first way, i.e. as a decomposition of an ordinary picture into density waves, each of which gives rise to a pair of vectors in reciprocal space (as in Fig. 7.5). But we view the second transformation in the second way, i.e. as the conversion of each pair of vectors, related by Friedel symmetry, into a corresponding density wave. So, by double Fourier transformation, density waves become converted into vector pairs, and then back again into density waves. It would seem that we end with the same picture with which we began. This is very nearly true—it is identical (for a two-dimensional picture) except for being rotated through 180° .

(d) Six basic rules for Fourier transforms

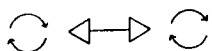
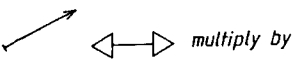
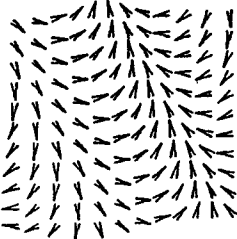
Intuitive prediction of the effects of Fourier transformation is much facilitated by means of six rules or theorems. These can be grouped into three pairs. (See Fig. 7.12 for a summary.) Although we still shall refer to the object which is Fourier-transformed as the "picture", implying that it is two-dimensional, nevertheless these rules apply equally to three-dimensional objects.

Algebraic: linearity. The first two rules concern the effect of algebraic operations. First we consider addition or subtraction. Addition, as applied to pictures, means superposition. Subtraction, defined as the inverse of addition, means the removal of one picture from another, if necessary by making the density negative. The first rule simply states that the addition or subtraction of two pictures causes the addition or subtraction of their Fourier transforms. Transformation is thus described as linear. In a linear system, scale factors are reproduced faithfully: a picture can be multiplied by, say, 3 through two superpositions with itself; the addition rule implies that the transform undergoes the same process, and is also multiplied by 3.

ALGEBRAIC

- (a) Linearity $\pm \longleftrightarrow \pm$
- (b) Convolution $\times \longleftrightarrow *$

ISOMETRIC MOVEMENT

- (c) Rotation 
 - (d) Translation 
- 

DISTORTION

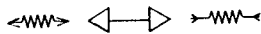
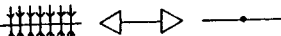
- (e) Scale 
- (f) Projection 

Fig. 7.12 The three pairs of rules concerning F.T.s, explained in the text. (F.T. pairs are indicated by the double-headed arrow.)

Algebraic: multiplication. The simplicity of the rule for the addition or subtraction of two pictures does not, however, extend to their multiplication. The multiplication of two pictures is defined in the following way. The pictures are superposed, and the densities of the corresponding pixels are multiplied together to give the density of the pixel in the product. In such a multiplication, zero always wins (i.e. the product is zero if either of the factors is). So the multiplication of two pictures gives a picture with the combined blank spaces of both. We can make use of this to impose a screen on a picture. When a photograph is converted into dots or thin lines ready for half-tone reproduction, the process can be viewed as one of multiplication by an array of dots or lines representing the screen. This process is referred to as "sampling" a picture.

How is the transform of a "product picture", resulting from the multiplication of two pictures, related to their individual transforms? It is not simply the two transforms multiplied together; instead, they are *convoluted* with each other. Convolution is an important operation that we shall encounter frequently. If two images (pictures or transforms) are convoluted together, each pixel of the first picture is replaced by an entire copy of the second. The density (or intensity or strength) of this copy is proportional to that of the pixel. Also, the final result is the same irrespective of which image is put first in the convolution. It is most easily grasped when one of the pictures consists of a few points, like the stars in a constellation. After convolution with the other picture, each star is replaced by a copy of it, the density of the copy being proportional to that of the corresponding star (Fig. 7.13). Multiplication and convolution arise commonly when analysing micrographs. If the micrograph consists of many identical images repeated on a lattice, then it can be viewed as one image convoluted with the lattice. Imaging defects, which apply to all parts of the picture, can also be viewed as a convolution. On the other hand, masking out unwanted parts of a picture can be viewed as multiplication with a masking function that is zero under the mask, and unity within the aperture.

Isometric movement: rotation. Rotating the picture produces exactly the same rotation on the transform (see Section 7.2.2(b)). This has the effect that any rotational symmetry present in the object is also present in the transform.

Isometric movement: translation. Translation is uniform movement without rotation. Translation of a picture produces multiplication of the transform by a complex wave. This is because the translated picture can be regarded as the original picture convoluted with a displaced vector (such as that in Fig. 7.9a). So its transform is the original transform multiplied by the transform of the vector; and that (shown in Fig. 7.9b) is a complex wave. We have only to enlarge our definition of multiplication to encompass pictures with phase as well as amplitude, i.e. complex vectors. When multiplying these, the amplitudes are multiplied together (as with real vectors), and the phases are added together.



Fig. 7.13 A series of points arranged in the pattern of a familiar stellar constellation is convoluted (*) with a circle. This replaces each point by a copy of the circle.

Distortion: scale rule. This states that uniform compression of the picture, in some direction, produces uniform stretching of the transform, in the same direction and by the same ratio (Fig. 7.14a, b). It follows that uniform stretching of the picture must also produce uniform compression of the transform; and also that an overall enlargement or reduction of the size of the picture produces an overall reduction or enlargement (respectively) of the size of the transform. The scale rule is true because uniform stretching of the picture enlarges periods in the direction of stretch, i.e. reduces spatial frequencies in this direction, and hence moves the corresponding vectors in the transform closer to the origin of reciprocal space.

Distortion: projection rule. Suppose we compress a picture in some chosen direction. If we continue compressing it in this way, we shall eventually squeeze the entire picture onto a line that is perpendicular to the direction of compression (left of Fig. 7.14). Onto this line we shall have projected the picture. During this process, the transform will be stretched in the direction along which the picture was compressed (right of Fig. 7.14). Only a small part of the transform will remain unaltered: the part along a line, perpendicular to the stretch direction, and passing through the origin (right of Fig. 7.14b). Continued stretching will eventually remove all other parts of the transform into outer reciprocal space. So we find that projection of the picture, onto some line, produces a transform that derives exclusively from the central line with the same orientation (Fig. 7.14c). Consequently the one-dimensional transform of the projection is the central section of the transform, taken in a parallel direction.

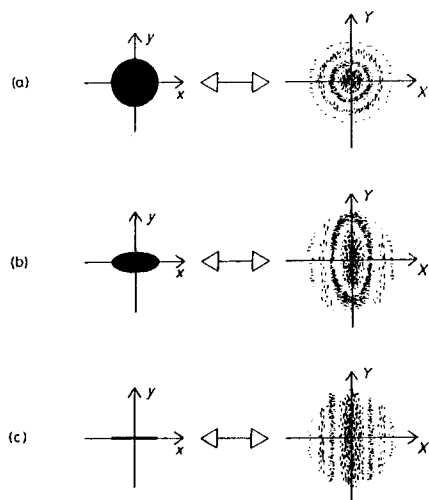


Fig. 7.14 (a) A picture (consisting of a uniform circle) and its F.T. Since the picture has circular symmetry, so also has the F.T. (b) The circle has become squashed along the y -axis, so that it is now an ellipse. Correspondingly, the F.T. has become stretched along the Y -axis. (c) The squashing of the circle has been carried to the limit where it is just a line of density on the x -axis. Correspondingly, the stretching of the F.T. has been carried to the limit where it consists only of the part previously on the X -axis, but now extending uniformly along the Y -direction.

Suppose we project the picture onto a line, and then project everything along this line onto a point, which now registers the total density of the picture. The corresponding part of the picture's transform is now its central point. So the value of the transform there equals the total density of a picture.

(e) Simple one-dimensional transforms

Some idealized types of picture recur so frequently in our discussions that we present their Fourier transforms now. We start with one-dimensional pictures, for which the density can be plotted as a function of the single spatial dimension (x).

"Shah"-function. This is named after the Cyrillic character ω (which apparently derives originally from an Egyptian hieroglyph for "garden"). It consists of equally-spaced, identical peaks (left of Fig. 7.15). We have seen (Section 7.2.2(b)) that a single peak can be generated by the superposition of cosine waves of identical amplitude, zero phase, and with all possible periods. However, the peaks of the shah-function have the constant spacing a , so all the component density waves must have their periods restricted to submultiples of a . It is therefore reasonable that the transform of the shah-function is as shown in the right of Fig. 7.15, i.e. another shah-function (or "comb").

Rectangle. This is not a two-dimensional rectangle, but a one-dimensional picture consisting of a uniform line, of length b . The plot of its density, shown on the left of Fig. 7.16a, is rectangular—hence its name. Its transform is shown on the right of Fig. 7.16a. This has a peak at the origin, since the average density of the "rectangle" is positive. Elsewhere the function (called a "sinc" function) oscillates, passing through zero at nodes with spatial frequencies $(1/b, 2/b, 3/b, \dots)$ (b is the length of the "rectangle"). We can see that the nodes must be positioned at just these spatial frequencies by the following argument. If we take the "rectangle" and convolute it with a shah-function of spacing b , all the copies of the "rectangle" will fit exactly together without gap or overlap, giving a uniform density (Fig. 7.16b). This uniform density line can be regarded as a "rectangle" of infinite length, i.e. infinitely stretched out. Applying the scale rule, its transform will be the transform of the "rectangle", but squashed to zero width, i.e. to a single peak. The transform of the shah-function in Fig. 7.16c is also known, and the transform of convolution is multiplication. So only one term remains unknown in the new equation (Figs 7.16b, c); this is the transform of Fig. 7.16a. We see that the transform of the original "rectangle" (i.e. the "sinc" function) is such that, when multiplied by a comb with peaks at $0, \pm b, \pm 2b, \pm 3b, \dots$, it yields a single peak at the origin. The sinc function must therefore annihilate the shah's peaks at $\pm b, \pm 2b, \pm 3b, \dots$. It can only do this by having nodes at just these points.

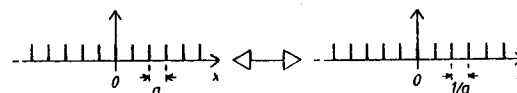


Fig. 7.15 On the left, a "shah-function" consists of equidistant peaks spaced a distance a apart. On the right, its F.T. consists of the same function with peaks spaced $1/a$ apart.

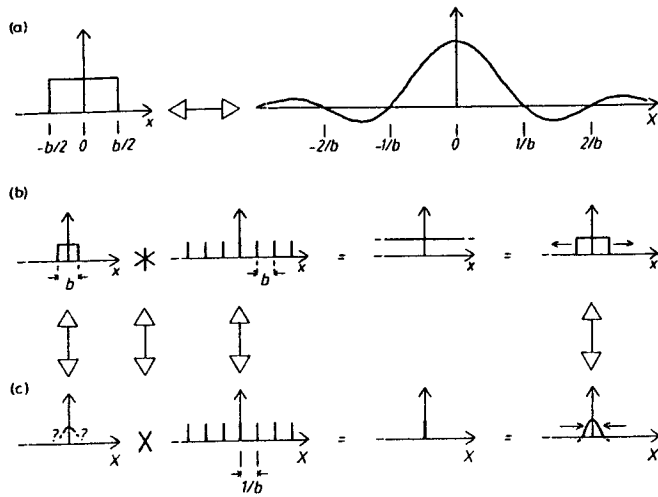


Fig. 7.16 (a) A rectangular density-function of width b (on the left) has the “sinc-function” F.T. shown on the right. (b) If the rectangular density-function (left) is convoluted ($*$) with a shah-function of the same spacing, it gives a uniform density, which equals a rectangular density-function stretched out to infinity. (c) The F.T. of (b). The F.T. of convolution ($*$) is multiplication; of the shah-function, another shah-function; and of the rectangular density-function stretched to infinity, a peak. As explained in the text, this result demonstrates that the zeroes of the sinc-function must lie at multiples of $1/b$.

Triangle. Like the “rectangle”, the “triangle” is also the name of a density-function. It describes a line whose density rises uniformly up to a maximum, from which it then uniformly declines. This density-function (left of Fig. 7.17b) is given when many copies of the rectangle function are added together along their length, as in Fig. 7.17a. But this process is just convolution with a line density function. So the “triangle” density of Fig. 7.17b can be regarded as the convolution of two “rectangles” with each other. Writing this as an equation (right of Fig. 7.17b), we take transforms of each term on the right-hand side (Fig. 7.17c). It is apparent that the required transform is the square of the transform of the “rectangle” (i.e. the square of the “sinc” function). So the oscillations of that transform become both smaller and positive (left of Fig. 7.17c).

It is as if the pronounced oscillations in the transform of the “rectangle” were caused by its sharp edges, whose removal in the “triangle” damped the oscillations. On this argument, we should suppose that further smoothing of the edges of the density function would yield an even smoother transform. The limit would be reached with a very smoothly declining function whose equally smoothly declining transform was indistinguishable from it in shape. Such a function exists: it is the bell-shaped Gaussian distribution curve so important in statistics (Fig. 7.18). (But, although having the same shape, a Gaussian and its transform have reciprocal widths.)

(f) Simple two-dimensional transforms

Square lattice. We now progress to two-dimensional density functions or pictures, but we shall relate these to the one-dimensional functions considered above. We start with

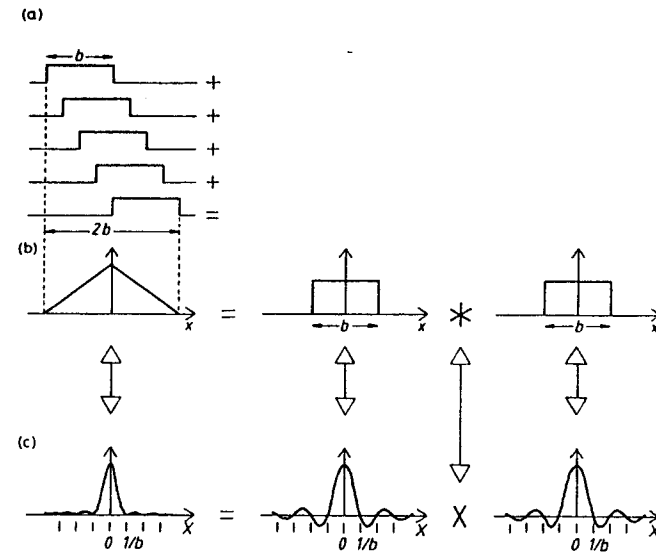


Fig. 7.17 (a) Many superpositions of a rectangular density-function, of width b , yield a triangular density-function of width $2b$. (b) These multiple superpositions of the rectangular density-function (left) are equivalent (right) to the convolution of this function with itself. (c) F.T. of (b). The rectangular density-function gives the sinc-function (Fig. 7.16), and convolution ($*$) gives multiplication. Hence the F.T. of the triangular density-function is the square of the sinc-function; this is shown on the left.

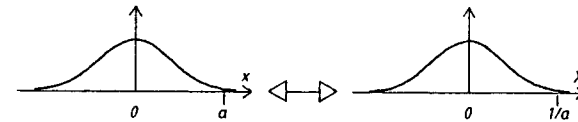


Fig. 7.18 The Gaussian curve with variance a^2/π (left) has the F.T. (right), which is the same function, but with variance $1/(\pi a^2)$. Thus the curve widths are reciprocal; stretching the left curve compresses the right one.

the square lattice (left of Fig. 7.20a), but we have to approach this in stages. The first stage is a set of equidistant parallel lines (Fig. 7.19). We obtain these lines from the one-dimensional comb function by indefinitely stretching it perpendicular to its length, so that each point becomes an infinite line. By the scale rule, the transform is infinitely compressed in the same direction. So we obtain the row of peaks shown on the right of Fig. 7.19, and we are ready for the next stage. Let two such sets of equidistant parallel lines, oriented perpendicular to each other, be multiplied together (right of Fig. 7.20a). Since zero always wins in multiplication, we are left with density only at those points where neither picture was zero, i.e. at a set of points on a square lattice. We have now expressed the square lattice, in an equation, as two sets of parallel lines multiplied together. Next we take the transforms of each term on the right-hand side of the equation (Fig. 7.20b), replacing multiplication by its transform, convolution. To convolute the two rows of peaks, we replace each peak of one row by the whole of the other row. We obtain another square lattice, but the lattice spacing is the reciprocal of that in the first.

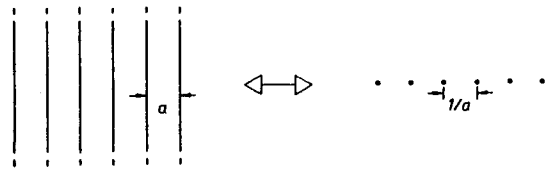


Fig. 7.19 A two-dimensional picture consisting of infinite, equidistant, parallel straight lines (left) has a F.T. (right) consisting of a line of equidistant points.

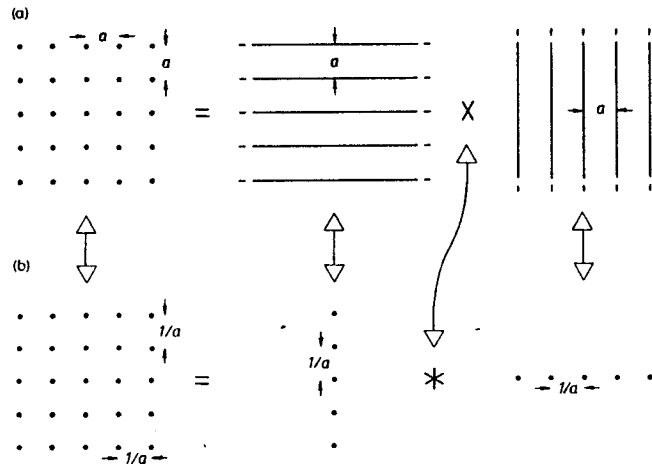


Fig. 7.20 (a) A square lattice (left) can be represented (right) as the product of two sets of parallel lines. (b) F.T. of (a). On the right, the parallel lines give (as in Fig. 7.19) lines of points, and multiplication gives convolution (*). Hence (left) the F.T. of the square lattice is another square lattice.

Arbitrary two-dimensional lattice. Any two-dimensional lattice can be obtained from a square lattice by subjecting it to stretch and compression. The directions of stretch and compression, which must be oriented appropriately with respect to the lattice, deform it by shear. So we can write an equation with the arbitrary lattice on the left-hand side (Fig. 7.21a). When we take its transform, we use the scale rule to turn the stretch into a compression, and the compression into a stretch. This effectively turns the stretch-compression process (i.e. the shear) through 90° (Fig. 7.21b). So, when applied to the square lattice, it results in a (reciprocal) lattice with the same shape as the real lattice on the left-hand side, but rotated by 90° . This simple relationship between real and reciprocal lattices applies only in two dimensions, but it is useful in the analysis of optical diffraction patterns.

Square and parallelogram. The square can be constructed by the convolution of two perpendicular lines: a copy of the second line is placed at every point of the first (Fig. 7.22). As usual, we complete the transform equation (below), using the transform of a one-dimensional line. Multiplying these two transforms together, we obtain the transform of the square.

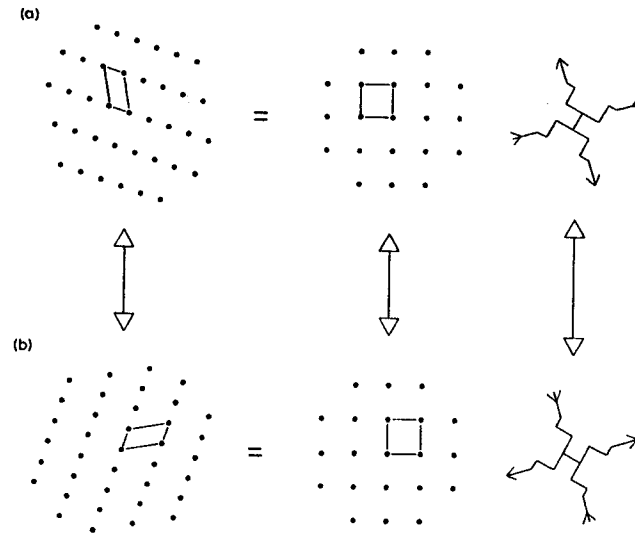


Fig. 7.21 (a) Points arranged in a lattice of arbitrary shape (left) can be represented (right) as a square lattice subject to suitable compression and stretching. (These distortions are indicated by the shear-symbol on the extreme right; this is composed of a compression-symbol, with arrows pointing inward, combined with a stretching symbol, with arrows pointing outward.) (b) F.T. of (a). On the right, the square lattice gives another square lattice, as in Fig. 7.20. The compression gives stretching, and the stretching gives compression (as in rule (e) of Fig. 7.12); this is equivalent (right) to rotating the distortion through a right-angle. (So the shear-symbol on the right has turned through a right-angle.) Hence (left) the F.T. of the lattice is the same lattice rotated through a right-angle.

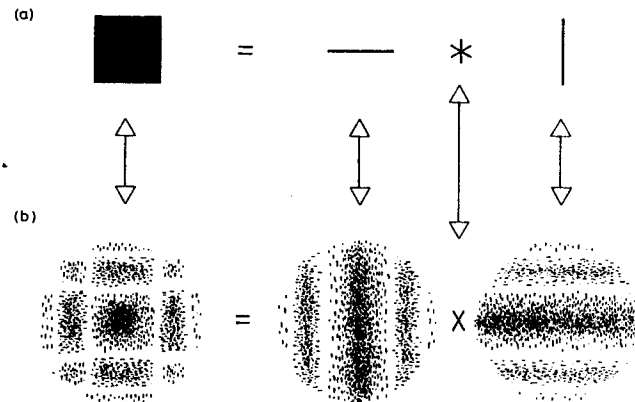


Fig. 7.22 (a) A square of uniform density (left) can be represented as the convolution of two line segments (right). (b) F.T. of (a). Each line segment gives (right) a sinc-function (see Fig. 7.16) stretched along the direction perpendicular to that segment. They are multiplied together (right), giving the transform of the original square.

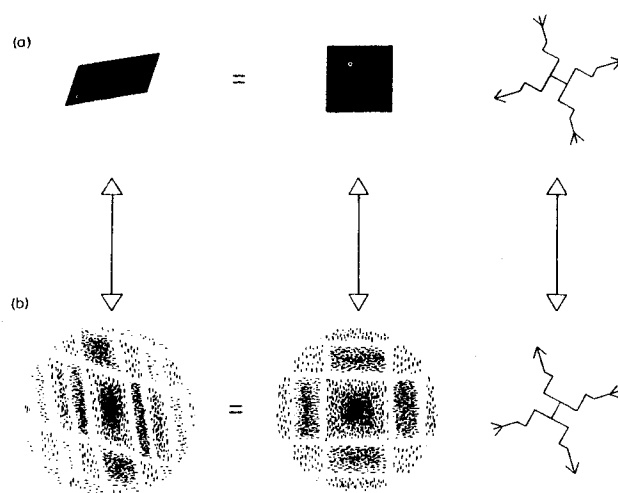


Fig. 7.23 (a) A parallelogram of uniform density (left) can be represented as a distorted square (right). (See legend to Fig. 7.21 for an explanation of the shear-symbol.) (b) F.T. of (a). As in Fig. 7.21, the distortion becomes rotated through a right-angle. Consequently the F.T. of the square (middle) becomes distorted (left) to resemble the original parallelogram rotated through a right-angle.

The transform of the parallelogram can be obtained from that of the square by stretching and compression (Fig. 7.23), rather as in the case of the arbitrary lattice.

Pictures lacking two-fold symmetry. The transforms of the above shapes consist only of positive or negative regions ("real" transforms). How does this simplification occur? Positive transform points have phase 0° , and negative points have phase 180° (so that the vector points in the negative direction). If the picture should happen to be the same when turned upside-down (two-fold axis), then so also must be the transform (by the rotation rule). But a two-fold axis applied to the transform replaces each point by its Friedel mate. We recall that, for any transform of a picture that has no phase variations, the Friedel mate has the same amplitude, but its phase has the negative value (Fig. 7.4). If the Friedel mate is to be identical, the negative of the phase must be the same as the phase. This is true only if the phase is 0° and 180° . That is the reason why the transforms of the square and rectangle are "real".

We conclude these examples of two-dimensional transforms with one which is not "real", because the picture does not have two-fold symmetry. The transform of such a picture has continuously varying phase as well as amplitude. To represent it, we have to show both amplitude and phase at each point. We do this (as in Figs 7.3, 7.5 or 7.9) by placing, on lattice points, small vectors with pointed tips and wide bases, like conifers. Their lengths represent the transform's amplitude, and their orientation its phase.

We consider the transform of a simple picture: three points at the vertices of an equilateral triangle (Fig. 7.24a). This transform is shown in Fig. 7.24b. Since the picture has a three-fold symmetry axis, so also does the transform, in which the X -axis values reappear on the symmetrically related X' and X'' lines. The picture has no-phase

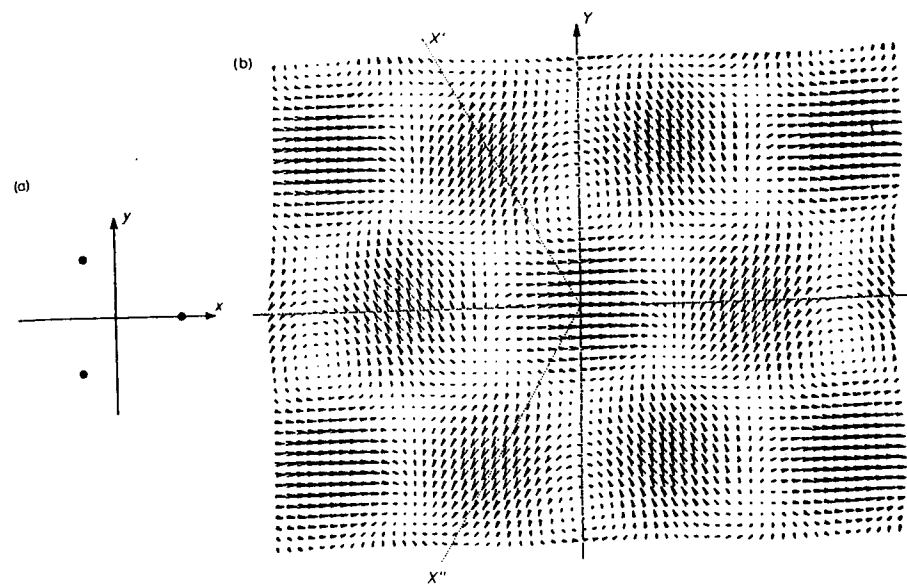


Fig. 7.24 (a) Three points at the vertices of an equilateral triangle. (b) F.T. of (a). It is represented, as in Fig. 7.3, by vectors. X' and X'' are lines where the F.T. has the same values as along X (a consequence of the three-fold symmetry of (a)).

variations, so the transform has Friedel symmetry, and its amplitude consequently has two-fold symmetry. Thus the amplitude has both two-fold and three-fold symmetry, i.e. six-fold symmetry. The transform can be seen to consist of circular "blobs" of amplitude of about the same size, this size being related to the size of the triangle in the picture. (That feature is explained in Section 7.2.2(h).) Where the amplitude of each "blob" is substantial, the phase is nearly constant. Virtually all the changes in phase are confined to the boundaries between adjacent "blobs". The amplitude is zero only at a few isolated points ("nodes"). This situation differs from that in "real" transforms, where the "blobs", which are either positive or negative, are separated by lines where the transform is zero.

(g) *Transforms of pictures of crystals*

As explained under Section 7.2.1, crystals can have translational symmetry in one, two or three dimensions (the first two being the most useful in electron micrographs). In each case, the translational symmetry is compatible with certain rotational symmetries.

A crystal's translational symmetry has simple consequences for the Fourier transform. For the crystal can be represented as the convolution of its lattice with the molecule (or group of molecules) that constitute the repeating structure. The image or picture of the crystal can likewise be represented by a convolution which puts a copy of some motif (the image of the repeating structure) at each lattice point (Fig. 7.25a). Applying the convolution theorem, the transform of this picture is the transform of the motif, sampled at the points of the reciprocal lattice (Fig. 7.25b).

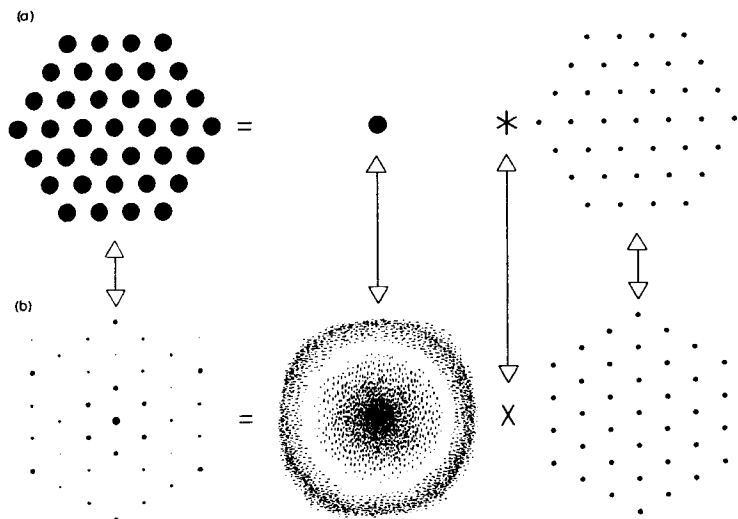


Fig. 7.25 (a) Uniform circles arranged on a hexagonal lattice (left). This equals (right) the convolution of one such circle with a hexagonal lattice. (b) F.T. of (a). The F.T. of the lattice (right) is the same lattice rotated through 90° (as in Fig. 7.21). The F.T. of the circle (centre) also has circular symmetry (rule (c) of Fig. 7.12). The original pattern's F.T. (left) is the product of the two F.T.s on the right; i.e., it is the middle F.T. "sampled" at the points of the F.T. on the right.

If the crystal also has an axis of rotational symmetry, then the transform will possess the same axis of symmetry. The intensities of the spots in the transform, the only visible features of an optical diffraction pattern, must show Friedel symmetry, since the original picture had no phase variations. For a two-dimensional transform, this means that the intensities have two-fold symmetry, even if the crystal has no rotational symmetry. This two-fold Friedel symmetry of the intensities becomes combined with any other rotational symmetry the transform may possess. Thus a crystal with a three-fold axis ($p3$, the symmetry of bacteriorhodopsin) gives a transform with three-fold symmetry, but the intensities show six-fold symmetry. If the transform intensities have two-, four- or six-fold symmetry, it is therefore necessary to look at the phases to see if the true symmetry is only half of this.

It is also possible that the repeating structure (motif) has rotational symmetry that is not present in the crystal as a whole. Then the transform of the motif (first term on the right of Fig. 7.25b) has the rotational symmetry of the motif (the circle directly above it), but the reciprocal lattice (second term in Fig. 7.5b) lacks this symmetry. (It is a case of the lowest symmetry winning.) So the rotational symmetry of the motif is no longer clearly apparent on the left of Fig. 7.25b. However, there exist special techniques for finding it.

(h) Sampling theorem

Pictures of finite extent give Fourier transforms that are smooth functions. As with the smooth functions in mathematical tables, it might be thought that such transforms can

be conveyed adequately only by many columns of figures, giving the values sampled at very close intervals. However, this is not the case: the sampling interval can be quite coarse if the correct procedure is used to reconstruct the transform.

The proof is sketched in Fig. 7.26. It is given for one-dimensional distributions (still called "pictures") and their transforms, but the theorem applies in two- or three-dimensional space. On the first line (Fig. 7.26a), we write down an identity: the picture to be transformed (left) equals itself repeated indefinitely, provided this is then multiplied by a "rectangle" function that annihilates all the copies except one. On the second line (b), we represent the indefinitely repeated copies of the picture as a convolution of the picture with a "shah" function. We now have an extended identity, in real space, and on the third line (c) we take Fourier transforms of each term. The transforms of the "shah" and the "rectangle" were treated in Section 7.2.2(e). The operations of multiplication and convolution are transforms of each other. The Fourier transform of the picture is represented, for brevity, by "F.T.(picture)". In the fourth line (d) this transform is represented explicitly, so that the transform equals itself, first multiplied by a comb, and then convoluted with a "sinc" function (see the right of Fig. 7.16a). On the last line (e) we show the consequence of multiplying the transform by the "shah": we get that transform, sampled at points corresponding to the peaks of the "shah".

We have now obtained the sampling theorem itself (Fig. 7.26e). If we have a picture of width D , its transform may be sampled to $0, \pm 1/D, \pm 2/D, \dots$ and then reconstructed exactly from these sampled values. To do so, one replaces each of the sampling peaks by an appropriate sinc function, whose central maximum has the same height as the sampling peak. Adding together all the different sinc functions gives the convolution, i.e. reconstructs the transform.

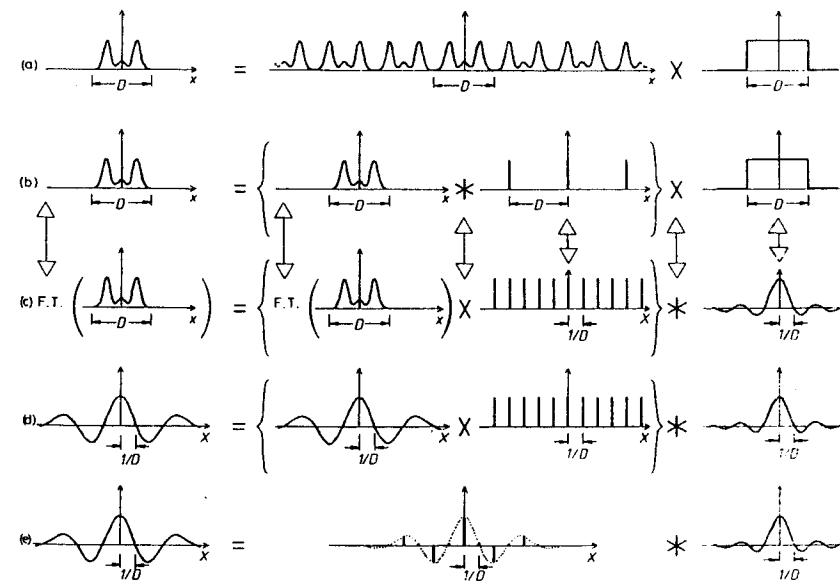


Fig. 7.26 A pictorial "proof" of the Whittaker-Shannon sampling theorem (see text). The original identity (a) gives (b), whose F.T. (c) gives (d), from which the sampling theorem follows in (e).

The sampling theorem defines in a precise way the intuitive idea that a transform is "chunky", being made of "blobs" of roughly uniform amplitude and phase, whose size depends inversely on the size of the picture. On the other hand, its actual application involves some approximation; for any picture of finite width will have a transform that extends to infinity, so the sampling theorem reconstruction is—strictly—an infinite series. But the transforms of finite pictures with limited resolution drop rapidly to very low levels, so that the reconstruction converges quickly. Only a few terms are needed to approximate the picture to within experimental error.

7.2.3 Fourier transforms of helices

To understand the methods used for analysing and reconstructing helical structures, we need some background knowledge of helical Fourier transforms. The mathematical equations—necessary for writing or using computer programs—are summarized in Section 7.2.3(i). But, as with the Fourier transforms of linear and planar periodic structures, a qualitative and intuitive understanding is the tool used most frequently. Consequently, we here derive the essential features of helical Fourier transforms in a non-mathematical fashion. First, however, we must explain the geometry of helical structures.

(a) Helical symmetry

Helical structures are common among biological macromolecules for a very simple geometrical reason. Suppose two identical molecules (subunits) bind to each other. The nature of the binding defines the spatial relationship of the second subunit to the first. If a third subunit binds to the second, and if the binding is the same as before, we should expect this spatial relationship to be repeated. Suppose this is so, and that more and more subunits are added, all with exactly the same spatial relationship to their neighbours. What shape will the aggregate assume? In general, we shall have generated a helix. As more and more subunits are added, the helix extends further and further in a certain direction, the helix axis. (For convenience, we shall suppose that this always runs vertically, i.e. along the z -axis.) The spatial relationship of neighbouring subunits can be described as a screw operation, consisting of a rotation of the subunit about the helix axis, coupled with a translation (uniform movement) along it. Any helical symmetry is completely defined by two numbers, the rotation Ω and the translation h (Fig. 7.27). These parameters can be chosen in various ways, depending on which particular set of helical lines is chosen (see below).

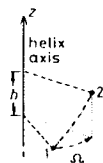


Fig. 7.27 Unit 1 of a helix is moved to the position of unit 2 by a rotation (the twist angle Ω) and a translation (the rise distance h).

7. Image Analysis of Electron Micrographs

An isolated helix can also possess overall rotational symmetry. If so, there will be some rotation axis about which the helix can be turned until it coincides with itself. If the angle through which it must be turned is $360^\circ/N$, this is called an N -fold rotation axis. In general, of course, a rotation would also move the helix axis; but it is essential that this should always coincide with itself. This poses no problem if the rotation axis is identical with the helix axis. Any other orientation of the rotation axis causes problems (for $N > 1$), with one exception. It is possible for the rotation axis to lie perpendicular to the helix axis, bringing that axis into coincidence with itself by a rotation of 180° ($N = 2$). These conclusions can be summarized by saying that the helix can possess only point-group rotational symmetry (cyclic C_N or dihedral D_N), with the N -fold axis along the helix axis. (As examples, DNA has the point-group D_1 , and the extended T4 bacteriophage sheath (Fig. 7.28) has the point-group C_6 .)

Because helical symmetry needs only two parameters, it can be represented as a two-dimensional diagram on paper. This is done as follows. Suppose we have a three-dimensional model of a helical structure such as a virus. Choose some point of each subunit and mark it. We now have a set of points that lie on the surface of a cylinder. If we were to cut that cylinder along a line parallel to the helix axis, and then to open it flat, we should have a two-dimensional diagram (called the "radial projection") of the helical lattice. This diagram can also represent the point-group symmetry of the helix. An N -fold axis along the helix axis causes each lattice point to become converted into N such points, all with the same position along the helix axis (same z -coordinate). In the diagram, therefore, each lattice point becomes replaced by N such points with

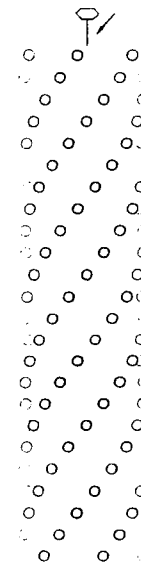


Fig. 7.28 Positions of the 144 subunits in the extended sheath of bacteriophage T4 (Moody, 1973). Each of the 24 annuli contains six subunits. This gives the helix a six-fold axis, symbolized by the hexagon at the top. The helices arrowed at the top are the same as those indicated in Fig. 7.29.

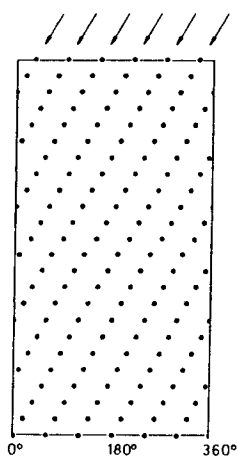


Fig. 7.29 Radial projection for the helix of Fig. 7.28, showing its outside surface. The arrows at the top indicate the helices arrowed in Fig. 7.28.

the same z -coordinate (Fig. 7.29). A two-fold axis perpendicular to the helix axis generates no new lattice points, so this must be represented by two-fold crystallographic symbols on the lattice points.

We shall refer to this two-dimensional diagram as the helical projection lattice. If we choose any two points in this lattice and join them with a straight line, we can also join the corresponding pairs of points so as to produce a set of parallel lines. These represent helical lines (called simply helices) in the original three-dimensional helix (Fig. 7.28). By this procedure, we could find an infinite number of sets of helical lines, so we need some way of choosing the most useful set. The different sets of helices differ in their number (n), and in their pitch (P , which is the vertical distance that must be travelled along a helix before we are vertically above the starting point). n must be a multiple of the rotation axis N , but sets of lines can always be chosen that contain the minimum number, N , of helical lines. Of these, the lines with the largest pitch constitute the *basic helices*. The choice of this set serves to fix the values of h and Ω that we use to define the helix. With the basic helices, h and Ω will have their smallest possible (absolute) value, which is why these helices are the most convenient choice. Instead of using Ω , we can use the pitch $P = h(360^\circ/\Omega)$, or the number of units/turn $= P/h = 360^\circ/\Omega$, denoted by $1/m$.

(b) *Fourier components of a helix*

In considering the Fourier transforms of helical structures, we apply the same general approach used earlier (Sections 7.2.2(a) and 7.2.2(b)) with linear and planar periodic ones. We imagine the helical structure to be composed of a number of standard, simplified density-waves, each of which gives a reasonably simple diffraction pattern. Three general types of density-wave are required.

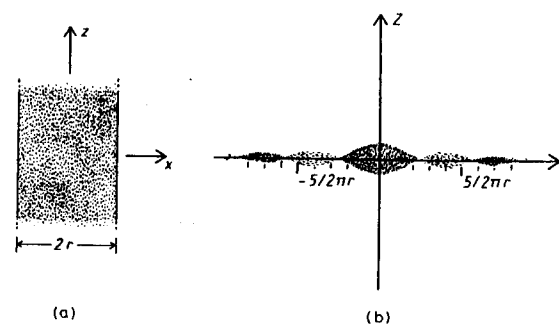


Fig. 7.30 (a) The simplest component of any helical structure which is confined to a thin cylinder. This is a thin cylinder of uniform density (positive, indicated by the vertical shading). (b) The F.T. of (a) (confined to the XY -plane) is sectioned in the XZ -plane. (Vertical shading indicates positive density, horizontal shading indicates negative density: the width of the shading indicates the amplitude of the transform.)

Uniform density wave. The first type of density-wave we need is a thin cylinder of uniform density (Fig. 7.30a). Such a cylinder, coaxial with the z -axis, is unaffected by any amount of stretching parallel to the z -axis, so its transform is unaffected by any amount of compression parallel to the Z -axis. Therefore the transform can exist only on the XY -plane. On this plane it has a large peak at the origin (since the cylinder has no negative density regions), surrounded by concentric circular rings (since the transform, like the cylinder, must be circularly symmetric). These rings have alternating signs (Fig. 7.30b).

Cylindrical density-waves. Next we need a cylindrical Fourier component consisting of ring-shaped sinusoidal density-waves of period h/m (Fig. 7.31a). This component could be obtained by multiplying a thin uniform cylinder (see above) with a set of parallel planar sheets of sinusoidally varying density, parallel to the xy -plane and with

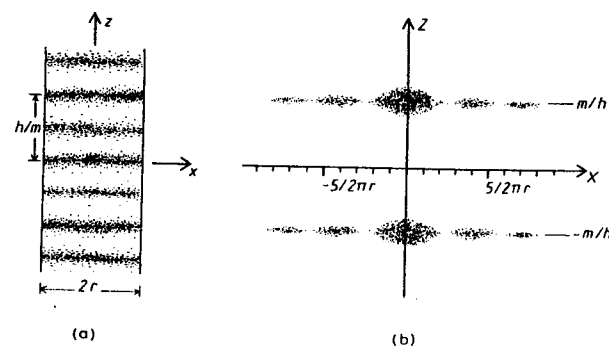


Fig. 7.31 (a) The second type of helical Fourier component: circular bands of alternating sign and of period h/m . (Vertical shading: positive; horizontal shading: negative.) (b) The F.T. of (a) (shown sectioned in the XZ -plane) is confined to the two layer-planes $Z = \pm m/h$.

a repeat distance h/m . The transform of the latter is two points at $Z = \pm m/h$ (see Section 7.2.2(b)). The transform of the former is shown in Fig. 7.30b. By the convolution theorem, the transform of our set of ring-shaped density-waves is the convolution of those two transforms, i.e. it is two identical copies of the cylinder transform (see above), on the planes $Z = \pm m/h$ (Fig. 7.31b).

Helical density-waves. The final type of cylindrical Fourier component (Fig. 7.32a) consists of a set of helical density-waves, of alternating sign, on a thin cylindrical surface of radius r . This can be regarded as a set of two-dimensional sinusoidal density-waves (as used in Section 7.2.2(b)) on a flat surface which is then rolled up to form a cylinder. There are n helices (i.e. n positive, and n negative, peaks of density on any equatorial line), each of pitch P , so the repeat in the z direction is P/n .

Layer-planes. A section through the transform (amplitude only) of this set of helical density-waves is shown in Fig. 7.32b. Most of the features can be deduced by intuitive arguments. First, the set of density-waves in Fig. 7.32a is unaltered if it is convoluted with a line of equidistant points, spaced P/n apart, and oriented parallel to the z -axis. (It is unaltered, since the result of this convolution merely superposes the density-waves successively on themselves, in perfect register.) By the convolution theorem, the transform is therefore unaffected if multiplied by the transform of the line of points, i.e. by a set of equidistant parallel planes perpendicular to the z -axis, and spaced n/P apart. Consequently the transform cannot exist except on this set of "layer-planes" at $\pm n/P, \pm 2n/P, \pm 3n/P, \dots$

If the set of convoluting points had been spaced apart by some fraction of P/n ($P/2n, P/3n, \dots$), then the density-waves would have been superposed on themselves out of register, and would have cancelled each other out. The convolution theorem therefore implies that multiplication of the transform by planes spaced with any multiple of n/P ($2n/P, 3n/P, \dots$) destroys the transform—i.e. there is no transform to be sampled at $\pm 2n/P, \pm 3n/P, \dots$. The transform is thus confined to $Z = \pm n/P$. (Fig. 7.32 shows the case where $n = 1$.)

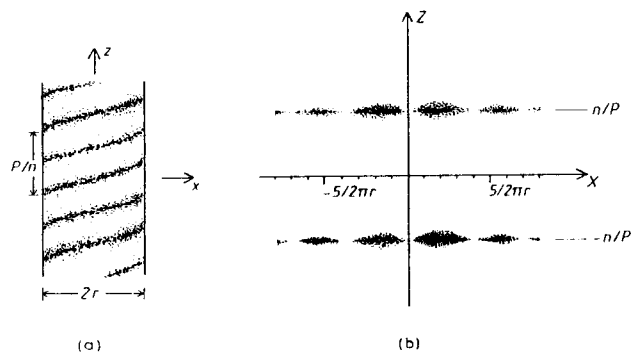


Fig. 7.32 (a) The third type of helical Fourier component: n helical density waves of alternating sign and of pitch P . (Vertical shading: positive; horizontal shading: negative.) (b) The F.T. of (a) (shown sectioned in the XZ -plane) is confined to the two layer-planes $Z = \pm n/P$. (See Fig. 7.34 for a three-dimensional representation of this.)

What is the distribution of the transform's amplitude on these layer-planes? First we note that rotation of the transform about the Z -axis is equivalent to rotation of the density-waves about the z -axis. Now this is equivalent to translating (shifting) the waves along the z -axis, which merely changes the phase of the transform (see rule (d) in Fig. 7.12). So the transform's amplitude must be rotationally symmetric about the Z -axis (Fig. 7.34). We therefore need consider only one section of this amplitude distribution, as was shown in Fig. 7.32.

Amplitude on layer-planes. Some idea of the transform amplitude on this section can be obtained by considering what would happen to the transform if the cylinder in Fig. 7.32a had a very large diameter, and if the number of helices increased proportionately to the diameter. In the limit, the cylinder's curvature could be ignored, so it could be represented approximately by a group of flat sheets joined together along lines parallel to the z -axis—a sort of polygonal cylinder. On each sheet, the density-waves would be flat and straight, but spaced as they were in the original cylinder. We can see their appearance by cutting that cylinder (along a line parallel to its axis) and opening it out. We should then see a pattern of which a small part is shown in Fig. 7.33a. As in Fig. 7.5g and h, its transform consists of a pair of spots, one of which is shown in Fig. 7.33b. By the same argument used when discussing Fig. 7.32, the Z -coordinate of this spot is n/P . This is the reciprocal of the spacing, in the z -direction, of the lines in Fig. 7.33a. Those lines are straight, so the X -coordinate of the spot in Fig. 7.33a will be given by the reciprocal of the lines' spacing in the x -direction—i.e. by $n/(2\pi r)$. So, if the X -coordinate (= R -coordinate) of the peak is denoted by R_M , we have $2\pi r R_M = 2\pi r(n/2\pi r) = n$ when n is large.

When n is small, the curvature of the cylindrical surface can no longer be ignored, so $2\pi R_M$ is not simply n . A better approximation is $2\pi R_M = 0.9 + 1.1n$, except for the first few orders. (More accurate values are shown by the open circles in Fig. 7.43; the amplitude distribution will be considered in more detail later (Section 7.2.3(g)) in connection with the half-helix.)

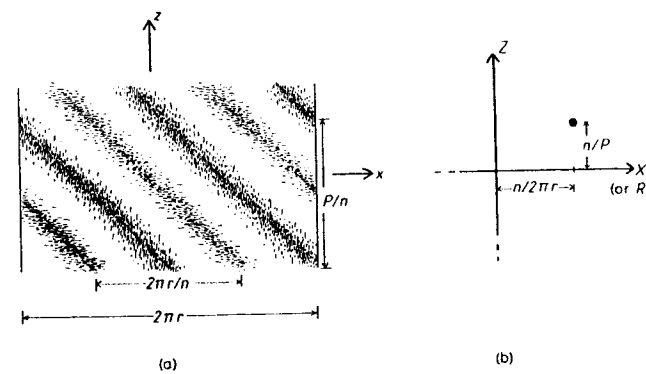


Fig. 7.33 (a) A thin cylinder with n helical density-waves, like that in Fig. 7.32a, has been cut parallel to the z -axis and opened out flat. (b) The F.T. of this sheet (if n were large) would consist of just two spots (of which only one is shown here). Because of the circular symmetry of the amplitude in a layer-plane (Figs 7.34 and 7.36), the X -coordinate is equivalent to the radial R -coordinate (see Fig. 7.48).

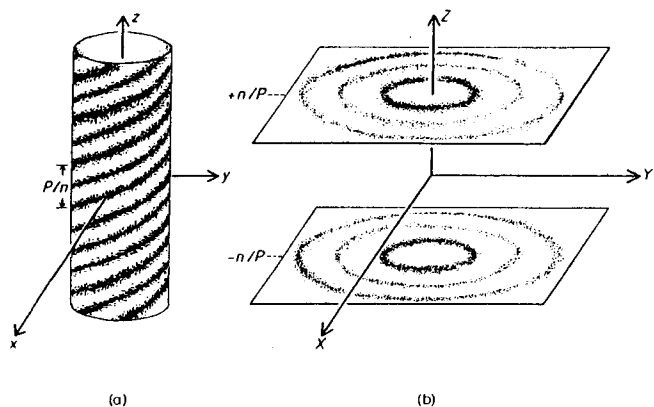


Fig. 7.34 (a) A set of n helical density-waves, as in Fig. 7.32a. (b) The two layer-planes in the F.T. of (a).

Before considering the phase of the transform, let us summarize our conclusions about its amplitude. A set of cylindrical helical density-waves of pitch P/n is shown in Fig. 7.34a. Its F.T. (Fig. 7.34b) exists only on the two planes $Z = \pm n/P$. On each plane, the transform amplitude consists of concentric circular rings, the innermost being the strongest.

Phase on layer-planes. Now we consider the phase of these rings. Figures 7.30b and 7.32b show sections of the transform in a vertical plane. Consider what happens if that section plane is rotated about the Z -axis. Each transform section corresponds to the projection of the helical density-waves onto a vertical plane that rotates about the z -axis (Fig. 7.35). Rotating this projection plane (Fig. 7.35) is equivalent to rotating the helix, which is equivalent to shifting the helix along the z -axis. Shifting (translating) an object causes a phase shift in the transform (Section 7.2.2(d)), which is constant along lines perpendicular to the translation. Since the translation is along the z -axis, the phase shift is constant in directions perpendicular to the Z -axis, i.e. on layer-lines. All parts of a layer-line therefore receive the same phase shift: this is a phase rotation proportional to the translation of the helix, and hence proportional to the

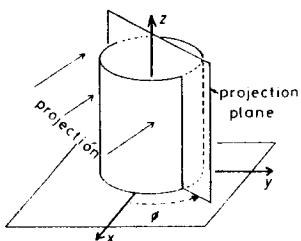


Fig. 7.35 The cylinder in Fig. 7.34a is being projected onto a projection plane that rotates through a variable angle ϕ . (The projection of the top half onto the xz -plane is shown in Fig. 7.32a.)

rotation of the projection plane, and to the identical rotation of the transform section plane. As this rotates, the transform phase also rotates. Since there are n helical density-waves, the projection is exactly repeated every $360^\circ/n$, and so also is the phase of the corresponding section of the transform. This is most easily seen by looking down onto one of the layer-planes (Fig. 7.36). The phase will be zero along a radius line corresponding to a real section of the transform, as in the XZ -plane of Fig. 7.32b. The phase rotates uniformly with the radius line, and next becomes zero when the line has rotated by $360^\circ/n$ ($360^\circ/3 = 120^\circ$ in Fig. 7.36). Furthermore, it can be shown (by considering the effect of inverting the helical waves and their transform by 180° rotation about the x - and X -axes) that the phases on the two layer-planes of Fig. 7.34b rotate in opposite directions.

This n -fold rotational symmetry of the phase on a layer-plane has consequences affecting the parity of n , the number of helices in any set. If n is even, the phase is the same after a 180° rotation. For any section through the transform, the portion on the left of the Z -axis is related to that on the right by a 180° rotation. So, if n is even, the phases (as well as the amplitudes) on a layer-line have mirror symmetry. However, if n is odd, a 180° rotation of the transform section plane rotates the phase also by 180° , i.e. makes the phase vectors point in the opposite direction. This is shown on the left side of the 0° line in Fig. 7.36.

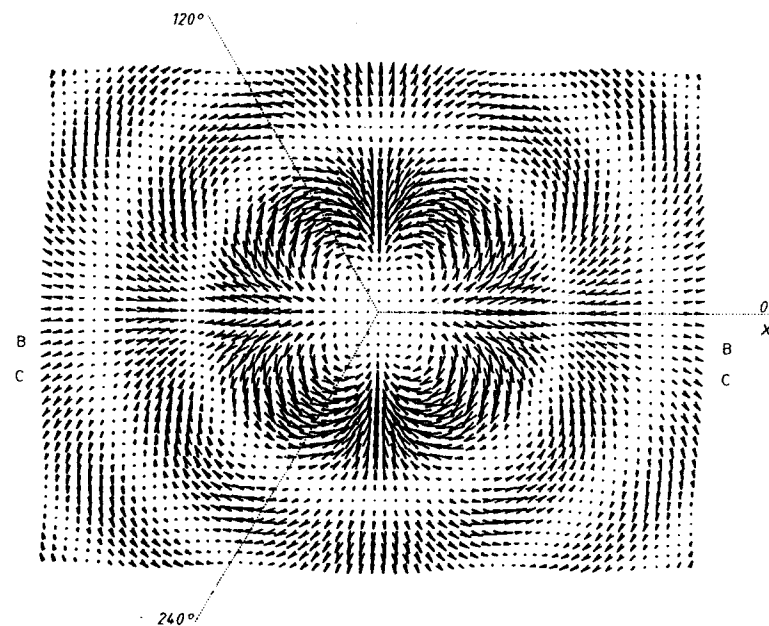


Fig. 7.36 The upper layer-plane from the F.T. of a set of three right-handed helical density-waves. Since $n = 3$, the circular rings of transform amplitude have phases that rotate three times in a complete revolution. Rows B and C are the lines along which the layer-plane is intersected by tilted section-planes (see Fig. 7.64).

(c) The (n, Z) plot

Now we have described the three basic types of helical component and their transforms. By adding appropriate quantities of the density-waves in Figs 7.30a and 7.32a, we could obtain a series of n helices with no negative density. If the correct amount of the density-waves in Fig. 7.31a were then added as well, we should obtain a structure composed of annuli, each containing n dark blobs, and with successive annuli related by the screw in Fig. 7.32a. Therefore, by taking enough of these three types of component (with all necessary values of the variable parameters, such as P , n , h and the cylinder radius), and then adding them all together, any helical structure can be constructed. Actually these components would suffice to construct any kind of structure whatever. If the structure is to have strict helical symmetry, therefore, there must be restrictions as to which components are allowed to contribute to it. These restrictions can be put into a particularly clear form if we use a device called the (n, Z) plot.

Consider the transform of a density-wave when the thin cylinder to which it is confined is cut parallel to the z -axis and opened out flat. When this is done with the sinusoidal component in Fig. 7.32a or 7.34a, we obtain Fig. 7.33, whose transform (ignoring effects due to the finite width of the sheet) is a pair of points in the XZ -plane. (One is shown in Fig. 7.33.) When this operation is performed on any of the other density-waves, the corresponding transform (again ignoring the effects of finite width) also consists of sharp peaks. Note that the positions of the peaks from the flattened surfaces approximately correspond to those from the unflattened, cylindrical density waves.

Now apply the same procedure to a complete helical lattice (still supposing, for simplicity, that it is confined to a thin cylindrical surface). Open out the helix in Fig. 7.37a (here we show its inside surface), and obtain the two-dimensional lattice ("radial projection") in Fig. 7.37b. Ignoring the effects of finite width, its transform will consist of sharp peaks on a lattice (Fig. 7.37c). Next, divide the X -axis (lower scale) of this

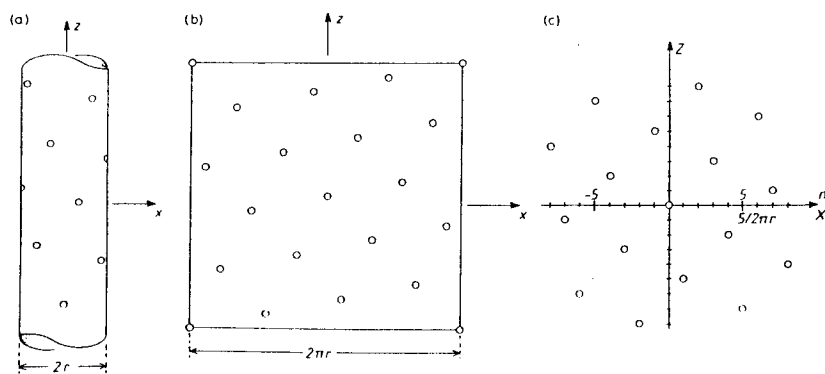


Fig. 7.37 Relation between a helical lattice and its (n, Z) plot. (a) Helical lattice drawn on a thin cylinder. (b) The cylinder has been cut parallel to its axis and opened out so as to show its inside surface. (This gives a radial projection, similar to that in Fig. 7.29, but showing the other surface.) (c) When the lattice in (b) is rotated through 90° , and the axes have been scaled and relabelled (see text), we obtain the (n, Z) plot.

transform into units of $\Delta X = 1/(2\pi r)$ and then relabel it as the n -axis (upper scale). The transform lattice is thereby converted into an (n, Z) plot.

The (n, Z) plot corresponding to any helical lattice contains a list of all the helical density-waves that are compatible with it. This list can be read off the (n, Z) plot as follows. The point at the (n, Z) origin refers to a uniform density (Fig. 7.30a). All other points occur as pairs, with the origin bisecting the line that joins them. Each pair refers to a set of helical density-waves (Fig. 7.32b). From a pair of points, choose that with a positive n -coordinate. Then the number n of helical density-waves in this set is simply the n -coordinate of the point. The axial spacing P/n is given by the reciprocal of the Z -coordinate of the point. (If this is negative, it means that the helices are left-handed.) If $n = 0$, i.e. if the points are on the Z -axis, then they refer to a set of annular density-waves (Fig. 7.31a), with a spacing $h = 1/Z$.

The (n, Z) plot corresponding to any helical lattice is not only very useful, but also easily found from the "radial projection". Open this out so as to show the view from inside the helix (i.e. not as in Figs 7.28 and 7.29). Rotate this view of the radial projection lattice (Fig. 7.37b) through 90° , so that the z -axis is now horizontal. The z -coordinate of each point will be found to be a multiple of some finite z -value. (This is because all points in the original helical lattice had z -coordinates that were multiples of the rise distance h .) Find this z -value (which will be called Δz), and put marks where z is $0, \Delta z, 2\Delta z$, etc. Relabel the z -axis as the n -axis, relabelling the marks as $n = 0, N, 2N, \dots$ (for a helical rotation axis of order N). Now we have marked the n -axis of our (n, Z) plot. Next, relabel the vertical axis as Z , and find the appropriate scale. The easiest way to do this is to find the rise distance h of the original helix, and to mark the first point above the origin that lies on the Z -axis as $1/h$. This completes our (n, Z) plot.

(d) The helical "selection rule"

The helical lattice is defined by the three numbers N (the order of the rotation axis), h (the rise distance) and P (the pitch of the basic helices). If we know their values, the (n, Z) plot can be calculated without first plotting out the helical lattice. We look for the minimal set of cylindrical density-waves which, when added together, will give a density blob at each lattice point (Fig. 7.38a). To begin with, we must represent the basic helices. Their pitch is P and their number n is N , the order of the rotation axis and the minimum possible number for any set of helices. To represent them, the set of helical density-waves (like those in Fig. 7.32a or 7.34a) must run along these basic helices, so they will need N positive (and N negative) helices of pitch P . They will give rise to two points in the (n, Z) plot, with coordinates $(N, N/P)$ and $(-N, -N/P)$ (the first is shown as point A in Fig. 7.38b). We shall also need a set of annular density-waves (like those in Fig. 7.31a) with the positive peaks separated by h , so their Z -coordinate in the (n, Z) plot is $1/h$ (or $-1/h$). Since their n -coordinate is zero, we obtain the point B in Fig. 7.38b. Finally, we add the point O at the origin of the (n, Z) plot (corresponding to the uniform cylindrical density-wave in Fig. 7.30a).

We now have two shortest vectors \vec{OA} and \vec{OB} of the (n, Z) plot. We shall, of course, need other helical Fourier components to represent all the detail of the helix. All these components correspond to points in a lattice on the (n, Z) plot, and that lattice can be generated from the two vectors \vec{OA} and \vec{OB} . So the general lattice point $j(\vec{OA}) + m(\vec{OB})$

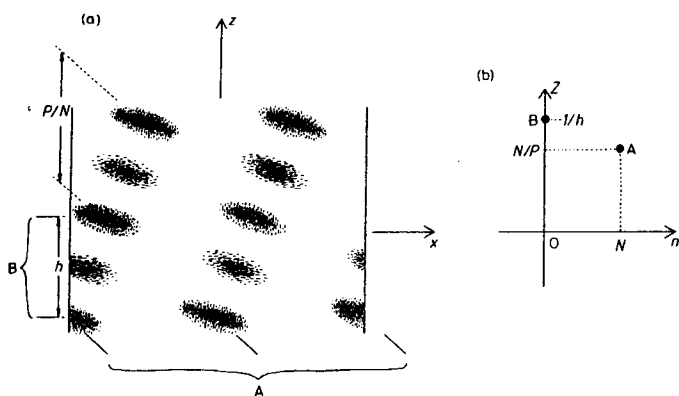


Fig. 7.38 Diagram to illustrate how a helical lattice, as represented in its simplest form by the positive (vertically shaded) spots, can be used to construct the lattice of an (n, Z) plot. (a) $1/N$ th of the radial projection of the helical lattice, viewed from the inside of the helix. (N is the order of the rotation axis of the helix; in this diagram, $N = 2$.) A indicates density-waves along the basic helices (N in number), and B indicates density-waves along the annuli. The intersection of these two sets of waves generates the spots shown here. (b) Part of the corresponding (n, Z) diagram. Point A has derived from the helices 'A' in (a); point B is derived from the annuli 'B' in (a).

will have (n, Z) coordinates given by $n = jN$ and $Z = jN/P + m/h$, where j and m are any integer (positive or negative). This gives

$$Z = n/P + m/h \quad (1)$$

(n is a multiple of N), called the *helical selection rule*. If, instead of using P , we define the basic screw operation with the twist angle Ω° , then the selection rule assumes the form

$$Z = n(\Omega/360^\circ) + m/h \quad (2)$$

(e) Helix with an exact repeat

Let us suppose that the helix repeats exactly when translated a distance c along its axis (the z -axis). Then it will not be changed if it is convoluted with a set of points arranged along the z -axis and separated by c . Consequently, its Fourier transform will not be changed if multiplied by a set of planes, parallel to the XY -plane, and separated by $1/c$. This means that the transform exists only on the planes $Z = 0, 1/c, 2/c, 3/c, \dots$. We therefore put $Z = L/c$ (L is any integer), and the helical selection rule becomes $L/c = n/P + m/h$, or $L = (c/P)n + (c/h)m$. Now (c/P) is the number of turns of the basic helix per repeat ($= t$), and (c/h) is the number of units on each basic helix per repeat ($= u$) (note that both u and t are integers). The helical selection rule now becomes

$$L = tn + um \quad (3)$$

(As before, L , n and m are any \pm integer, and n must be a multiple of N , the rotation axis.) This equation defines a lattice of points on an (n, L) plot, the form taken by the (n, Z) plot when $Z = L/c$.

Certain differences follow when the helix repeats exactly, so that the transform is confined to the layer-planes $Z = L/c$. First, we can use the integer L to denote the layer-plane containing a particular reflection. Second, there must be points in the (n, Z) plot with the same value of Z but different values of n . (If n_1, m_1 satisfy the integral helical selection rule, so also do $(n_1 + u), (m_1 - t)$.) In the transform, the two or more different patterns like those in Fig. 7.36 will be superposed in the same layer-plane and will therefore interfere. The resultant amplitude, with contributions from two different n -values, cannot have circular symmetry. This could greatly alter a diffraction pattern and complicate interpretation.

The limited resolution of the transforms of most electron micrographs so restricts the number of layer-lines that such interference is fairly rare. However, the possibility is increased if the helical particle is short, since this increases the width of the layer-lines. (Layer-line interference is not, strictly, an all-or-none affair; the layer-line profile is a sinc function.)

The original papers on helical diffraction (Cochran *et al.*, 1952; Klug *et al.*, 1958) used exclusively the notation appropriate for helices with an exact repeat. This usage, which seems to have been a vestige of the crystallographic approach to helices*, is no longer relevant when analysing the patterns of particles under no constraint to have an exact repeat. Nevertheless, it is still often used, perhaps because it is enshrined in certain computer programs. It can be made to work by choosing a repeat that is long enough to fit the data to within the required accuracy. But there is always the temptation to simplify the numbering of the layer-lines by approximating the structure to one with a smaller repeat. Then accuracy is lost unnecessarily in stating the helical parameters. Moreover, if any later improvement is made in measuring these parameters (or if the helix should change its exact structure under certain conditions), a complete change is necessary in the assignments of L to the layer-lines. So it is extremely difficult to calculate the error in the twist angle when the approximation of a short repeat is used. All this is quite unnecessary; the (n, Z) plot, which makes no assumptions about the repeat of the helix, and which consequently allows all measurable quantities to be determined in the usual ways (e.g. least-squares), is just as easy to use.

(f) Structure not confined to a single radius

Now we drop the requirement that the helical structure should be confined to a thin cylinder. Since the helical structure now has thickness, we represent it as the sum of a series of concentric cylinders with successively incremented radii. On each cylinder there is a density pattern. Although the patterns usually differ, each fits the same helical lattice. Each cylinder is thus a thin helix of the type considered in Section 7.2.3(b), and its transform will be confined to layer-planes defined by the selection rule. Since the selection

* If exact helical symmetry applies to an entire three-dimensional lattice, there must be an integral number of units (actually 1, 2, 3, 4 or 6) in the repeat. This fact had been so thoroughly taught to crystallographers that there was a small revolution in structural thinking when it was pointed out (Crane, 1950; Pauling and Corey, 1951) that it need not apply to single molecules. However, the number of units per repeat then moved only to a (small) rational number, not to a real number. The mathematics of helical diffraction were recast by Ramachandran (1960) in a form appropriate to a real number of units per repeat, but this involved a substantial change in notation. Here we make the smallest possible change, so that papers using helical diffraction can be followed easily.

rule is independent of the cylinder radius, the layer-planes of each of the component transforms (from different cylinders) will have identical Z-coordinates. Consequently they can interfere. We now consider what effects this will produce.

From the transform of the thick helix, select one pair of layer-planes. Suppose that $n = n_1$ and $Z = Z_1$ on them. What features of the thick helix do the planes relate to? They relate to the sum, from all the concentric cylinders, of sets of sinusoidal helical density-waves. These waves will differ in size, corresponding to the different radii of the corresponding cylinders. They also differ in amplitude and phase (corresponding to the different density-patterns). However, all the waves will have the same number (n_1) and pitch (n_1/Z_1). (We are supposing that the helix has no exact repeat, so that layer-planes with different n 's will have different Z 's.) Since $n = n_1$ for the waves on each cylinder, the contributions of each cylinder to the transform will have an amplitude with circular symmetry, and a phase that rotates n_1 times along a circular path with its centre on the Z-axis. So their superposition will also have these two features.

Suppose that, for a helix with effectively no repeat, we knew all the values of the transform along one radial line on a layer-plane. Then the circular symmetry of the amplitude would mean that we knew the amplitude over the entire layer-plane. Also, the n_1 -fold rotation of the phase would allow us to predict the phase over the entire layer-plane. This applies to every layer-plane; even though n is usually different on different planes, the phase can be predicted provided n is known. So an axial transform section is all that we need before we can reconstruct the entire transform of any helix, provided the repeat is sufficiently long. With this proviso, the three-dimensional transform of a helix can be adequately represented by a two-dimensional one composed of layer-lines. Such a two-dimensional transform is simply the transform of the projection of the helix. This result is of obvious importance in the three-dimensional reconstruction of helical structures from projection data (Section 7.5.4).

(g) *Transforms of flattened or unequally contrasted helices*

During specimen preparation, helices are often imperfectly preserved or stained. This will change the transform, and it is important to recognize and, if possible, correct for these changes. Exact correction may be difficult, since the distortions are not simple; the most common distortion, flattening, produces neither a flat sheet, nor even an elliptical cylinder, but a shape resembling an unrisen pastry (Seymour and DeRosier, 1987). Moreover, the effects of distortion can be even more complicated unless they are uniform along the length of the particle. To give some qualitative appreciation of these effects, we therefore assume uniformity, and consider only a few simplified distortion shapes.

If distortion is uniform, the changes are independent of z . So they affect only the distribution of amplitude and phase within each (X, Y) plane of reciprocal space. The positions of the layer-planes (or -lines) are therefore unaffected, and the previous selection rule will still apply. However, the connection between the order n of a layer-line, and the distribution of amplitude on it, will alter.

Although this complicates interpretation of the transform, such interpretation is still reasonably simple if the main structural features lie on a thin (although distorted) cylinder. The easiest case is that where the thin cylinder (Fig. 7.39a) has been squashed flat (Fig. 7.39b). (This is not the situation in Fig. 7.37b, since a flattened cylinder has two surfaces.) So the transform contains two reciprocal lattices related by a mirror plane

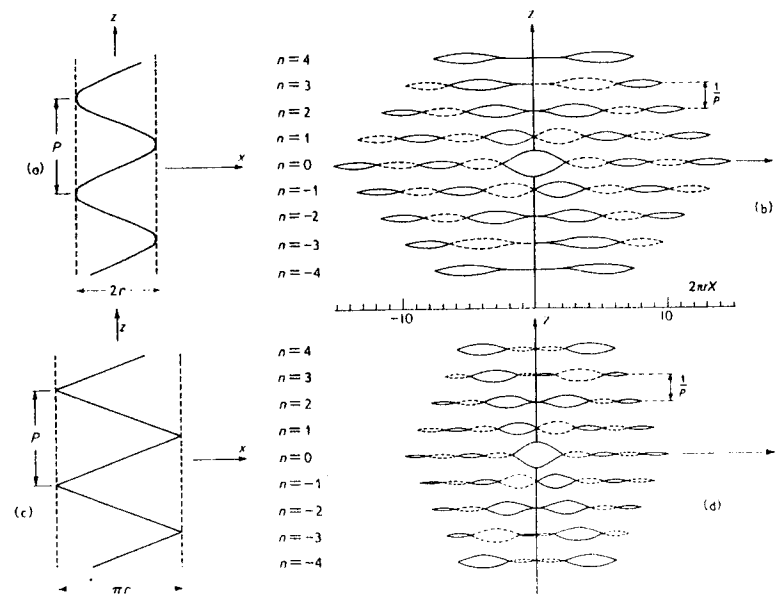


Fig. 7.39 Comparison of the F.T. of a circular, and of a flattened, thin helix (Moody, 1967a). (a) Projection of a continuous circular helix (radius r) onto the xz -plane. (b) F.T. of (a). (Amplitude is indicated by the separation of the paired lines; negative regions of the transform are shown by broken lines.) The principal maxima on each layer-line lie closest to the Z -axis. (c) The circular helix of (a) has been completely flattened into the xz -plane. (d) F.T. of (c). This is quite similar to the F.T. in (b). The principal maxima have been displaced and the subsidiary maxima diminished.

(Fig. 7.53c). Therefore each layer-line contains not just one peak (as in an (n, Z) plot), but two. These will be broadened because of the finite width of the flattened cylinder, so they will interfere. This displaces the peaks from their lattice positions, the displacement being greatest for $n = 1$ and diminishing as n increases (elliptical points in Fig. 7.43). The effect is to produce a pattern qualitatively similar to the transform of a helix.

Another case that can be treated exactly is the half-helix. Since this reveals the origin of the peaks in the undistorted helix, we treat the two together. Figure 7.40 shows a left-handed helical wire, with the upper part continuous and the lower part broken. If both parts are diffracting, the diffraction pattern will be that shown in Fig. 7.41. These curves represent the amplitude distributions in the F.T. of a helical structure. On a layer-plane of order n , the distribution is a Bessel function $J_n(X)$. These are characterized by a central gap, of width roughly proportional to n (see Section 7.2.3(b)), and flanked by two broad peaks.

The origin of the peaks in the helix's F.T. (Fig. 7.41) is clearer if we consider the F.T. of just half the helix (Fig. 7.42). This half-helix is shown as the continuous line in Fig. 7.40 (i.e. the part represented by the broken line has been removed). It will be seen that its F.T. (Fig. 7.42) preserves just the peaks on the right-hand side of the whole-helix's F.T. (Fig. 7.41). (When comparing these Figures, the Bessel functions $J_0(X)$, $J_1(X)$, etc. in Fig. 7.41 correspond with layer-planes, 0, 1, etc. in Fig. 7.42.) So the peaks on the



Fig. 7.40 The projection of a thin helix, as in Fig. 7.39a. It is left-handed, and its upper surface (whose projection gives the F.T. shown in Fig. 7.42) is drawn as a continuous line. The diameter is $1/\pi$ so that Fig. 7.41 shows, with the correct scale, the F.T. of both surfaces.

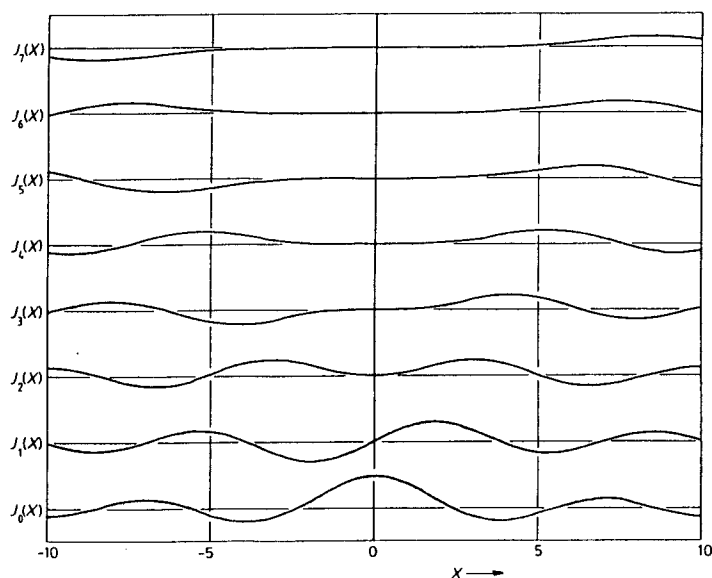


Fig. 7.41 A plot of the Bessel functions of orders 0 to 7. These correspond to the F.T. of both the continuous and broken lines of Fig. 7.40. The different curves corresponding to different layer-lines, the zero-order layer-line (bottom curve) passing through the origin. If the radius of the original helix were r , the abscissa scale shown here could be taken to measure $2\pi Rr$.

The even orders are symmetrical about the origin, but the odd orders are antisymmetrical, i.e. their amplitudes are equal at X and $-X$, but their phases differ by 180° . Moreover, whereas the zero-order Bessel function has a peak at the origin, the higher-order functions are zero there, and do not become large until X reaches a value that is approximately proportional to the order.

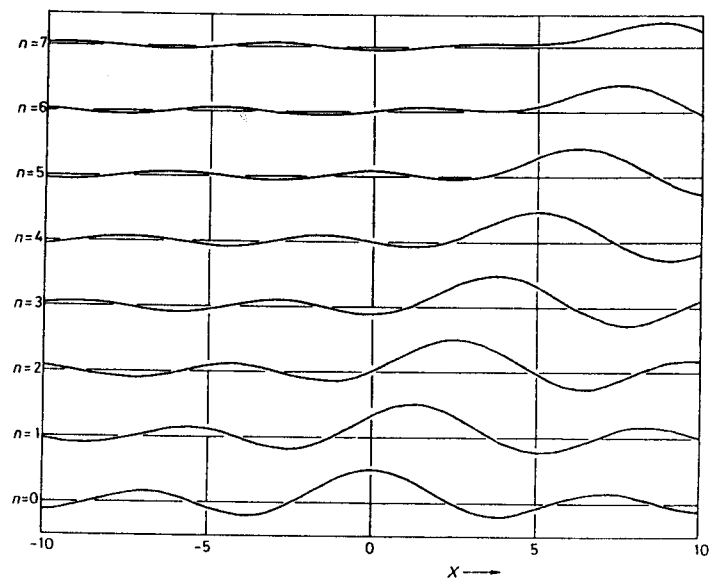


Fig. 7.42 F.T. of the continuous line of Fig. 7.40, i.e. of a half-helix (Moody, 1967a). Consequently only half the peaks of Fig. 7.41 are present on each layer-line. (However, the peaks are double the height of those in Fig. 7.41 since the vertical scale was chosen to make the zero-order layer-lines identical in the two figures.) All the highest peaks of the different layer-lines are positive, and they fall approximately on a line through the origin.

right-hand side of Fig. 7.41 derive from the top surface of the helix (i.e., from the continuous line in Fig. 7.40). The bottom surface (broken line in Fig. 7.40) gives rise to the peaks on the left-hand side of Fig. 7.41. Note also that the string of peaks in Fig. 7.42 lies nearly along a straight line (unlike those in Fig. 7.41). Because the two peaks on a layer-line can no longer interfere, their positions are not distorted, as in Fig. 7.39.

(h) Representing calculated helical structures

Suppose that we have calculated the three-dimensional structure of a helix, and now wish to represent it on a sheet of paper. Like any other three-dimensional structure, it can be represented as the image of a two-dimensional surface corresponding to some chosen density level. Thus the original helical reconstructions were shown as photographs of balsa models. Nowadays, these pictures can be generated in a computer. However, such representations of a contour surface cannot show the density function. This can be presented by a suitable two-dimensional section or projection. We start by considering sections of helices.

Sections. It is easiest to calculate Fourier transforms on surfaces where one of the coordinates is constant. So, if the transform is stored as a function of the Cartesian coordinates (X, Y, Z) , the sections of the corresponding density distribution that are

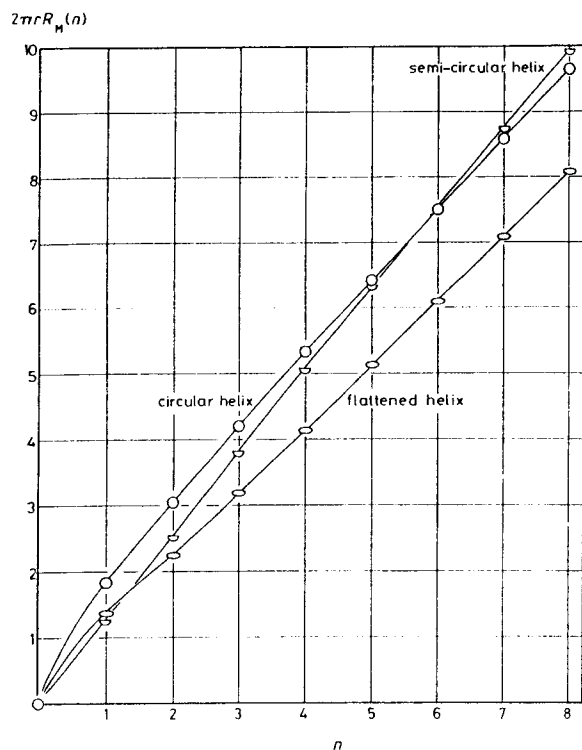


Fig. 7.43 Positions of the first maxima in the F.T. of helices whose supporting cylinders have various shapes. The circular and semicircular helices correspond to the F.T.s in Figs 7.41 and 7.42, respectively. (Adapted from Table 2 of Moody, 1967a.)

easiest to calculate lie on planes perpendicular to x , y or z . However, a helical transform calculated by Equation (19) (see below) will be stored as a function of the cylindrical coordinates (R, Φ, Z) , so the most easily calculated sections will have r , ϕ or z constant. (See Fig. 7.48 for the definition of these coordinates.) When z is constant, we have sections perpendicular to the z -axis (Fig. 7.44), which are used for building a balsa model of the helix. When ϕ is constant, we obtain sections on radial planes joined at the z -axis; these are useful for displaying circularly symmetric features of the structure. When r is a constant one calculates sections on concentric cylinders. Such sections can be plotted on paper, as if the cylinder had been cut and opened out flat. They can conveniently represent the structural features at the most important radii.

Radial projections. The calculation of any *section* of a particle requires the entire transform, but *projections* can be calculated using only parts of it. We are familiar with projections along parallel straight lines. But projections can also be made along lines that are curved (though they must not intersect). The structure density is integrated along each of the curved lines; then a surface is chosen that intersects the lines and, at

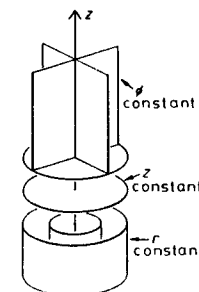


Fig. 7.44 Three principal surfaces for representing sections of a helical structure.

each intersection point, the integrated density is marked. Such non-rectilinear projections are often appropriate for a helical structure, and some of them can be calculated easily by using the mathematics of helical transforms. The two most useful helical projections are of this type. Their utility depends on the particle not having an exact repeat, so that every layer-plane of its transform must have only one order n_j .

Consider first the projection of the particle along the set of concentric circles centred on the particle's axis, and parallel to the xy -plane (Fig. 7.45). This projection could be obtained if the particle's density were first averaged by rotation about its axis, and the resulting rotationally symmetric structure were then represented by a section on the xz -plane. Instead of calculating the complete density and then averaging it, we can obtain the rotational average directly from the helix's transform. We have just to calculate a back-transform, from only those layer-planes that are rotationally symmetric. Where shall we find such layer-planes? Since the particle is supposed to have no exact repeat, no pairs of layer-planes can interfere; on each layer-plane, therefore, the amplitude is rotationally symmetric. However, the phase rotates n times, and cannot be rotationally symmetric unless $n = 0$. Consequently, the rotational average is found by taking the transform only of those layer-planes for which $n = 0$. These occur at $Z = m/h$, where h is the rise distance (Section 7.2.3(a) and Fig. 7.27).

Helical projections. Next we consider the helical projection. It has been pointed out that, in a helical structure, there are many different sets of helices. In a helical projection we choose some set (of pitch P_s and number n_s), and project the structure down them (Fig. 7.46). There is another way to represent this projection. Suppose that the helix is convoluted with a set of points lined up along the z -axis, and spaced (P_s/n_s) apart. (P_s/n_s) equals the spacing between successive turns of our chosen helices. The convolution

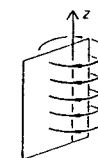


Fig. 7.45 Projection of a helical structure along circles, centred about the helix axis z (rotationally-averaged projection).

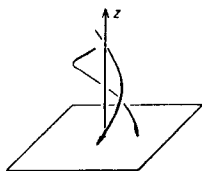


Fig. 7.46 Projection of a helical structure along a given set of helices (helical projection).

will thus superpose, on any given turn of these helices, subunits from all the remaining turns, preserving their angular orientation. If (as we are supposing) the helix has no exact repeat, none of these subunits will superpose exactly on each other; each will be displaced by short distances along the turn of the helix (Fig. 7.47). If the displacements are short enough, the resulting distribution will approximate that of the particle helically averaged—that is, averaged by being continuously screwed along the chosen set of helices.

How can we calculate this projection directly from the transform of the helix? The convolution described above is equivalent to multiplying the transform by a set of parallel planes at $Z = \pm(n_s/P_s), \pm 2(n_s/P_s), \dots$. Note the effect of this multiplication. Since the helix has no exact repeat, every layer-plane has a different Z -coordinate, equal to n/P , and referring to n helical density-waves of pitch P . The layer-planes will be grouped into sets, where the i th set has Z -coordinates $n_i/P_i, 2n_i/P_i, 3n_i/P_i, \dots$. The lowest member of each set refers to a series of density-waves whose positive crests run along an actual set of helices in the structure. (The higher members contain 2, 3, etc. times as many density-waves as there are corresponding helices in the structure; the layer-line with $n = 1$ in Fig. 7.39 is in the first set, and the layer-lines with $n = 2, 3$, etc. belong to higher sets.) Multiplying the transform by planes at $Z = \pm(n_s/P_s), \pm 2(n_s/P_s), \dots$ therefore means sampling the transform at a particular set of layer-planes. If those planes, and no others, are back-transformed, the helically averaged density is calculated. This may be adequately represented by a section in the axial (yz -) plane, or the basal (xy -) plane.

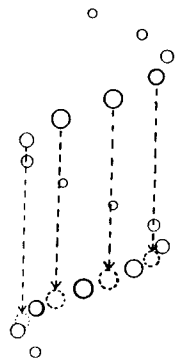


Fig. 7.47 The helical projection is equivalent to convolution of the helical structure with a set of equidistant points parallel to its axis, having a spacing equal to the pitch of the desired helices. This has the effect of superposing the subunits from different turns of the helix.

(i) Review of Fourier transforms in Cartesian coordinates

For reference, and as a basis for the corresponding formulae for helical or rotationally-symmetric structures, we list the standard equations in Cartesian coordinates (see, for example, Bracewell, 1986).

A density function $f(x) = f(x, y, z)$ has the Fourier transform $F(\mathbf{X}) = F(X, Y, Z)$, where

$$F(\mathbf{X}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\mathbf{x}) \exp(2\pi i \mathbf{X} \cdot \mathbf{x}) d^3 \mathbf{x} \quad (4)$$

This has the inversion

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(\mathbf{X}) \exp(-2\pi i \mathbf{x} \cdot \mathbf{X}) d^3 \mathbf{X} \quad (5)$$

Two density functions f and g have a convolution (\star) defined by the function

$$f(\mathbf{x}) \star g(\mathbf{x}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\mathbf{u}) g(\mathbf{x} - \mathbf{u}) d^3 \mathbf{u} \quad (6)$$

The two functions f and g also have a cross-correlation function defined by

$$f(\mathbf{x}) \star \star g(\mathbf{x}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\mathbf{u}) g(\mathbf{x} + \mathbf{u}) d^3 \mathbf{u} \quad (7)$$

Whereas the convolution operation is independent of the order of f and g , the cross-correlation operation is not. If f and g are the same function, then the cross-correlation function is called an "auto-correlation function". By substituting the convolution of Equation (6) in place of $f(\mathbf{x})$ in Equation (4), and changing the order of integration, we obtain the convolution theorem:

$$F(\mathbf{X})G(\mathbf{X}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{f(\mathbf{u}) \star g(\mathbf{x})\} \exp(2\pi i \mathbf{X} \cdot \mathbf{x}) d^3 \mathbf{x} \quad (8)$$

It can be similarly proved that:

$$F^*(\mathbf{X})G(\mathbf{X}) = F(-\mathbf{X})G(\mathbf{X}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{f(\mathbf{x}) \star \star g(\mathbf{x})\} \exp(2\pi i \mathbf{X} \cdot \mathbf{x}) d^3 \mathbf{x} \quad (9)$$

If f and g are the same function, then Equation (9) gives the transform of the auto-correlation function:

$$|F(\mathbf{X})|^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{f(\mathbf{x}) \star \star f(\mathbf{x})\} \exp(2\pi i \mathbf{X} \cdot \mathbf{x}) d^3 \mathbf{x} \quad (10)$$

By taking the Fourier inversion of this equation, we express the auto-correlation function as the transform of $|F(\mathbf{X})|^2$. (Hence the auto-correlation function, like its transform, must have a centre of symmetry.) We then equate to zero the variable common to both sides, and obtain the three-dimensional Cartesian form of the Parseval relation:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f^2(\mathbf{u}) d^3 \mathbf{u} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |F(\mathbf{X})|^2 d^3 \mathbf{X} \quad (11)$$

("Parseval relations" show how the "length" of a "vector" changes after transformation of the functional space; such equations are useful in relating the density scale factors of a structure or picture and its transform.)

(j) *Fourier transforms of helical structures*

A helical structure has a unique axis (identified with z) along which the translational component of the screw acts, and which is the axis of the rotational component. The natural coordinate system for discussing the transforms of helices is therefore cylindrical polar coordinates.

Fourier transform in cylindrical polar coordinates. Cylindrical polar coordinates (Fig. 7.48) are defined by the following equations:

$$\begin{aligned} x &= r \cos \phi & X &= R \cos \Phi \\ y &= r \sin \phi & Y &= R \sin \Phi \\ z &= z & Z &= Z \end{aligned} \quad (12)$$

Making these substitutions in the Cartesian Fourier transform Equation (4), we obtain:

$$F(R, \Phi, Z) = \int_{-\infty}^{\infty} \int_0^{2\pi} \int_0^{\infty} f(r, \phi, z) \exp \{2\pi i [Rr \cos(\Phi - \phi) + Zz]\} r dr d\phi dz \quad (13)$$

We wish to make the best use of the fact that the helix is periodic in both ϕ and z . We note that a two-dimensional repeating structure which is periodic in x and y gives a particularly simple transform—it exists only at the points of the reciprocal lattice in the XY -plane. Now the transform of that two-dimensional repeating structure contains the term $\exp[2\pi i(Xx + Yy)]$; so we expect a similar simplification if a similar form in Φ and Z could somehow be introduced into Equation (13). Z and z are already in the correct combination, and we need only introduce the required change into Φ . This is possible through use of the generating function for Bessel functions (Watson, 1958):

$$\exp \left\{ \frac{1}{2} u(t - t^{-1}) \right\} = \sum_{n=-\infty}^{\infty} t^n J_n(u) \quad (14)$$

from which the substitution $t = i \exp(iv)$ yields:

$$\exp(iu \cos v) = \sum_{n=-\infty}^{\infty} \exp\{in(v + \pi/2)\} \cdot J_n(u) \quad (15)$$

In this last equation, v is no longer the argument of the cosine, but occurs as the combination nv in the exponential. Setting $u = 2\pi Rr$ and $v = (\Phi - \phi)$ in Equation (15), then Equation (13) gives us:

$$F(R, \Phi, Z) = \sum_{n=-\infty}^{\infty} \exp\{in(\Phi + \pi/2)\} \int_0^{\infty} J_n(2\pi Rr) 2\pi r \left[\frac{1}{2\pi} \int_0^{2\pi} \int_{-\infty}^{\infty} f(r, \phi, z) \exp\{i(2\pi Zz - n\phi)\} dz d\phi \right] dr \quad (16)$$

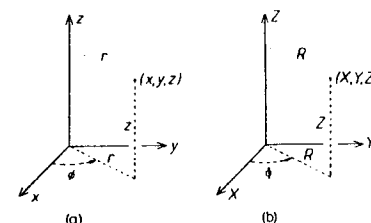


Fig. 7.48 Cylindrical coordinate systems in (a) real and (b) reciprocal space, used for the equations of helical diffraction theory.

In this equation we note that z and ϕ occur in the same sort of expressions as do x and y in an ordinary Cartesian transform*. However, since n is not a continuous variable but an integer, there is a summation over n instead of an integral. So Equation (16) allows us to obtain simple expressions for the case where $f(r, \phi, z)$ is periodic in ϕ and z . However, r and its reciprocal coordinate R , though they occur in the combination rR , are separated from Z and ϕ in a quite different part of the transform.

" n -Transforms". Certain stages on the route to the complete helical transform have been defined and given symbols. Reference to the Fourier transform formula in Equation (16) shows that transformation proceeds by a different route for each of the three cylindrical polar coordinates.

The z and Z coordinates are not changed from their Cartesian definition, and transformation uses the same expression $\exp(2\pi i Zz)$ that occurs in the Cartesian Fourier transform, Equation (4).

The transformation with respect to r proceeds by a separate route. It is no longer strictly a Fourier transform, for which the kernel is a complex exponential, but rather a Hankel (or Fourier-Bessel) transform, for which the kernel is a Bessel function. (This will become clearer later; see Equations (18) and (21).)

The transformation with respect to ϕ proceeds by a curious double route. First, the density function $f(r, \phi, z)$ is Fourier transformed, within the double integral at the right of Equation (16), so that ϕ is replaced by the integer variable n . Finally, this " n -coordinate" is transformed, by means of an infinite Fourier series, into the required reciprocal coordinate Φ . Thus, in between the original density function $f(r, \phi, z)$ and its Fourier transform $F(R, \Phi, Z)$, we have "transforms" in which ϕ is replaced by n . Because of the simplicity of the z - Z transform, this is always performed first, so all the useful " n -transforms" are functions of Z . However, there are two states for the radial variables (r or R), so there are two of these intermediate " n -transforms". The first is the expression in square brackets in Equation (16):

$$g_n(r, Z) = \frac{1}{2\pi} \int_0^{2\pi} \int_{-\infty}^{\infty} f(r, \phi, z) \exp\{i(2\pi Zz - n\phi)\} dz d\phi \quad (17)$$

* We could exchange the order of Φ and ϕ , obtaining an alternative form of Equation (16) with the combination $(2\pi Zz + n\phi)$ inside the z, ϕ integral. Though more symmetrical, this requires a wholesale redefinition of many quantities in the F.T.

The complete r -integral of Equation (16) is our second "n-transform":

$$G_n(R, Z) = \int_0^\infty g_n(r, Z) J_n(2\pi Rr) 2\pi r dr \quad (18)$$

Using $G_n(R, Z)$, the polar coordinate F.T. of Equation (16) may be written more simply as:

$$F(R, \Phi, Z) = \sum_{n=-\infty}^{\infty} \exp\{in(\Phi + \pi/2)\} G_n(R, Z) \quad (19)$$

Using the orthogonality of complex exponentials, this equation may be solved for $G_n(R, Z)$, giving:

$$G_n(R, Z) = \frac{1}{2\pi i^n} \int_0^{2\pi} \exp(-in\Phi) F(R, \Phi, Z) d\Phi \quad (20)$$

Equation (18) represents $G_n(R, Z)$ as a Hankel transform of $g_n(r, Z)$. Using the inversion theorem for these transforms (see the comments following Equation (27)), we also have

$$g_n(r, Z) = \int_0^\infty G_n(R, Z) J_n(2\pi Rr) 2\pi R dR \quad (21)$$

The connections between the density function, its transform, and the two "n-transforms" are shown in the following table:

$$\begin{array}{c}
 f(r, \phi, z) \xleftrightarrow{[\phi, n]} \\
 \downarrow [z, Z] \\
 g_n(r, Z) \\
 \uparrow [r, R] \\
 G_n(R, Z) \xleftrightarrow{[n, \Phi]} F(R, \Phi, Z)
 \end{array} \quad (22)$$

The inverse transform in polar coordinates. The inversion formula will be needed for the three-dimensional reconstruction of helices (Section 7.6.3). We start with the inversion of Equation (4) in Cartesian coordinates, i.e. Equation (5) above. Then we make the substitutions from Equation (12), and obtain the analogue to Equation (13):

$$f(r, \phi, z) = \int_{-\infty}^{\infty} \int_0^{2\pi} \int_0^\infty F(R, \Phi, Z) \exp\{-2\pi i[Rr \cos(\Phi - \phi) + Zz]\} R dR d\Phi dZ \quad (23)$$

The negative sign of the argument of the exponential function requires us to use the complex conjugate of Equation (15):

$$\exp(-iu \cos v) = \sum_{n=-\infty}^{\infty} \exp\{-in(v + \pi/2)\} J_n(u) \quad (24)$$

Using the previous substitutions $u = 2\pi Rr$ and $v = (\Phi - \phi)$, Equation (24) can be substituted into Equation (23) to give us the inversion formula:

$$f(r, \phi, z) = \sum_{n=-\infty}^{\infty} \exp\{in(\phi - \pi/2)\} \int_0^\infty J_n(2\pi Rr) 2\pi R \times \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} \int_0^{2\pi} F(R, \Phi, Z) \exp\{-i(2\pi Zz + n\Phi)\} d\Phi dZ \right] dR \quad (25)$$

Note that Equation (5) can be obtained from Equation (4) by interchanging the lowercase letters f, x, y, z with the corresponding capital ones, and by taking the complex conjugate of the right-hand side. We might expect that the same operations on Equation (16) would yield Equation (25), but that is not the case. Actually, the operations yield a correct inversion formula, which differs from that in Equation (25) through choosing the negative of the argument of the cosine in Equation (23), i.e. the negative of v in Equation (24). This means that $n\phi$ and $n\Phi$ should both have their signs changed. When this is done, Equation (25) is obtained.

The "n-transforms" are also useful in the inversion formula. If, in Equation (25), we perform the Φ -integration first, we can use the right-hand side of Equation (20) to replace that Φ -integral, and obtain

$$f(r, \phi, z) = \sum_{n=-\infty}^{\infty} \exp(in\phi) \int_{-\infty}^{\infty} \exp(-2\pi i Zz) \left[\int_0^\infty G_n(R, Z) J_n(2\pi Rr) 2\pi R dR \right] dZ \quad (26)$$

Equation (21) now gives us

$$f(r, \phi, z) = \sum_{n=-\infty}^{\infty} \exp(in\phi) \int_{-\infty}^{\infty} g_n(r, Z) \exp(-2\pi i Zz) dZ \quad (27)$$

(The Hankel transform inversion formulae can be obtained directly from the F.T. inversion formulae, e.g. by substituting Equation (25) into Equation (16), though the manipulations are less messy if the two-dimensional equations of Section 7.2.3(k) are used.)

Effects of helical symmetry. If the helical symmetry is defined by a screw with a rise distance h and a twist angle Ω , then $f(r, \phi, z) = f(r, \phi + \Omega, z + h)$. Thus the transform $F(R, \Phi, Z)$ is unchanged if $z \rightarrow z + h$ and $\phi \rightarrow \phi + \Omega$. This will be true if, in Equation (16), the part following i in the exponential function is changed only by some multiple of 2π , say $2\pi m$. So $2\pi Zh - n\Omega = 2\pi m$, or

$$Zh = m + n(\Omega/2\pi) \quad (28)$$

Thus the transform can exist only on layer-planes whose Z -coordinate satisfies Equation (28), which is consequently called the "selection rule" (a term originally applied to Equation (30) by Klug *et al.* (1958)—presumably by analogy with quantum mechanics—but later extended more loosely to Equation (28), etc.) From this basic form, other variants of the selection rule can be obtained. If, instead of the twist angle Ω , we use the pitch P , defined by $P = h(2\pi/\Omega)$, then

$$Z = m/h + n/P \quad (29)$$

If there is an exact z -repeat of length c , i.e. if $f(r, \phi, z) = f(r, \phi, z + c)$, then $2\pi Zc$ must be a multiple of 2π , or $Zc = L$ ($L = \text{any integer}$). Substituting this in Equation (29), we obtain

$$L/c = m/h + n/P$$

so

$$L = m(c/h) + n(c/P)$$

so

$$L = um + tn \quad (30)$$

where $u (= c/h)$ is the number of units per repeat, and $t (= c/P)$ is the number of turns per repeat. Equation (30) is a linear equation with integral coefficients and two variables (m, n). Given one solution (m_1, n_1) , the next is $(m_1 + j, n_1 - ju/t)$, where j is the smallest integer that makes $j(u/t)$ integral. Thus the separation between interfering orders is proportional to the number of units per turn, i.e. to the length of the repeat. The bigger the difference between the interfering orders, the bigger will be the distance between their peaks, provided the interfering orders are both positive. However, this need not be the case, and the most confusing interferences involve orders of opposite sign.

Although infinitely many Bessel orders can interfere in theory, nevertheless there is, in practice, an upper limit n_m to the orders that appear in the transform:

$$n_m < \{2\pi(r/d) - 0.9\}/1.1 \quad (31)$$

where r is the particle radius and d is the resolution of the micrograph.

We have seen that the effect of helical symmetry is to confine the F.T. to "layer-planes", perpendicular to the Z -axis, whose Z -coordinate satisfies Equation (29). This converts Z -integrals [Equations (25)–(27)] into sums. However, the layer-planes are not, of course, delta-functions in the Z -direction. If the particle length is λ , the Z -dependence of the transform of g or G is of the form $\sin\{\pi\lambda(Z - Z_j)\}/\pi\lambda(Z - Z_j)$. (This thickness could be still further increased by some forms of disorder, e.g. periodic perturbations along z .) It is this finite thickness of layer-planes that prevents their interference being an event of extreme rarity.

Effects of point-group symmetry. Helical symmetry is compatible with cyclic or dihedral point-group symmetry, of which the former is a subgroup of the latter.

If the density function $f(r, \phi, z)$ has cyclic symmetry C_N , with the N -fold axis along z , then $f(r, \phi, z) = f(r, \phi + 2\pi/N, z)$. Applying the argument used in the case of helical symmetry, $2\pi n/N$ must be a multiple of 2π . Consequently, n/N must be an integer, i.e. the orders of all Bessel functions must be multiples of the order of the rotation axis.

If the density function $f(r, \phi, z)$ has dihedral symmetry D_N , there are additional two-fold axes perpendicular to the N -fold axis. If the x -axis is chosen to coincide with one of these, $f(r, \phi, z) = f(r, -\phi, -z)$. Then the sine integral of the complex exponential vanishes, and we have:

$$g_n(r, Z) \rightarrow \frac{1}{4\pi} \int_{-\infty}^{\infty} \int_0^{2\pi} f(r, \phi, z) \cos(2\pi Zz - n\phi) d\phi dz \quad (32)$$

(k) *Transforms in planar polar coordinates*

The plane polar coordinate system is the natural one for analysing pictures of particles with rotational symmetry; see Sections 7.4.6 and 7.5.3(c). This coordinate system is easily obtained by merely suppressing the z - (or Z -) coordinate of cylindrical polar coordinates. We start with the n -transforms. Equation (17) gives

$$g_n(r) = \frac{1}{2\pi} \int_0^{2\pi} f(r, \phi) \exp(-in\phi) d\phi \quad (33)$$

and Equation (20) gives

$$G_n(R) = \frac{1}{2\pi i^n} \int_0^{2\pi} F(R, \Phi) \exp(-in\Phi) d\Phi \quad (34)$$

These definitions have more symmetry now that the (z, Z) coordinates have been suppressed. It can be seen that $g_n(r)$ has the same connection with the real distribution $f(r, \phi)$ that $i^n G_n(R)$ does with its transform $F(R, \Phi)$. Next we express the n -transforms as a Hankel transform pair. Equation (21) gives

$$g_n(r) = \int_0^{\infty} G_n(R) J_n(2\pi Rr) 2\pi R dR \quad (35)$$

and Equation (18) gives the symmetrically-related inversion formula

$$G_n(R) = \int_0^{\infty} g_n(r) J_n(2\pi Rr) 2\pi r dr \quad (36)$$

We proceed to employ these in the F.T. and inversion formulae. The F.T. formula of Equation (16) gives, in two dimensions,

$$F(R, \Phi) = \sum_{n=-\infty}^{\infty} \exp\{in(\Phi + \pi/2)\} \int_0^{\infty} J_n(2\pi Rr) 2\pi r \left[\frac{1}{2\pi} \int_0^{2\pi} f(r, \phi) \exp(-in\phi) d\phi \right] dr \quad (37)$$

Similarly the inversion formula of Equation (25) gives

$$f(r, \phi) = \sum_{n=-\infty}^{\infty} \exp\{in(\phi - \pi/2)\} \int_0^{\infty} J_n(2\pi Rr) 2\pi R \left[\frac{1}{2\pi} \int_0^{2\pi} F(R, \Phi) \exp(-in\Phi) d\Phi \right] dR \quad (38)$$

Substituting $g_n(r)$ from Equation (33) for the integral within square brackets in Equation (37), we obtain

$$\begin{aligned} F(R, \Phi) &= \sum_{n=-\infty}^{\infty} \exp\{in(\Phi + \pi/2)\} \int_0^{\infty} g_n(r) J_n(2\pi Rr) 2\pi r dr \\ &= \sum_{n=-\infty}^{\infty} \exp\{in(\Phi + \pi/2)\} G_n(R) = \sum_{n=-\infty}^{\infty} \exp(in\Phi) i^n G_n(R) \end{aligned} \quad (39)$$

where the middle step used Equation (36). The overall Equation (39) expresses the transform as a Fourier series in Φ (in which every function is necessarily periodic). The

coefficients are $i^n G_n(R)$, and are given by the Fourier integral of Equation (34). We can perform the same operations on Equation (38), using Equation (34) to express the integral in square brackets. We obtain:

$$\begin{aligned} f(r, \phi) &= \sum_{n=-\infty}^{\infty} \exp\{in(\phi - \pi/2)\} \int_0^{\infty} \{i^n G_n(R)\} J_n(2\pi Rr) 2\pi R \, dR \\ &= \sum_{n=-\infty}^{\infty} \exp(in\phi) g_n(r) \end{aligned} \quad (40)$$

where the last step used Equation (35). We thereby express the density function as a Fourier series in ϕ , of which the Fourier coefficients are $g_n(r)$ and the appropriate Fourier integral is Equation (33).

These F.T. relations form the basis for analysing rotational symmetry by Fourier methods. They can be connected with the corresponding correlation methods. To start with, we note that Equation (40) expresses the "picture" density, at a given radius r , as a Fourier series in the angle ϕ . Through this Fourier series, the angular convolution of two pictures can be related to the product of the corresponding Fourier coefficients:

$$\frac{1}{2\pi} \int_0^{2\pi} f^{(1)}(r, \psi) f^{(2)}(r, \phi - \psi) \, d\psi = \sum_{n=-\infty}^{\infty} g_n^{(1)}(r) g_n^{(2)}(r) \exp(in\phi) \quad (41)$$

More useful than the convolution is the cross-correlation function. This can be obtained similarly:

$$\begin{aligned} \frac{1}{2\pi} \int_0^{2\pi} f^{(1)}(r, \psi) f^{(2)}(r, \psi + \phi) \, d\psi &= \sum_{n=-\infty}^{\infty} g_n^{(1)}(r) [g_n^{(2)}(r)]^* \exp(-in\phi) \\ &= \sum_{n=-\infty}^{\infty} [g_n^{(1)}(r)]^* g_n^{(2)}(r) \exp(in\phi) \end{aligned} \quad (42)$$

(The two versions of the right-hand side are complex conjugates, equal since the left-hand side is real.) Both of the above equations apply only at a given radius, whereas comparisons of particle symmetries must involve integration over r . Such integrals can be converted into R -integrals involving the n -transforms $G_n(R)$. A simple form of this conversion is seen in the Parseval relation

$$\int_0^{\infty} |g_n(r)|^2 2\pi r \, dr = \int_0^{\infty} |G_n(R)|^2 2\pi R \, dR \quad (43)$$

This is useful since the quantity on the left is the "power spectrum" of n -fold rotational components (Section 7.4.6(a)). When this approach is applied to the cross-correlation function (Equation (42)), it gives:

$$\begin{aligned} \int_0^{\infty} r \, dr \int_0^{2\pi} f^{(1)}(r, \psi) f^{(2)}(r, \psi + \phi) \, d\psi &= \sum_{n=-\infty}^{\infty} \exp(in\phi) \\ &\int_0^{\infty} [G_n^{(1)}(R)]^* G_n^{(2)}(R) 2\pi R \, dR \end{aligned} \quad (44)$$

The corresponding auto-correlation function is

$$\int_0^{\infty} r \, dr \int_0^{2\pi} f(r, \psi) f(r, \psi + \phi) \, d\psi = \sum_{n=-\infty}^{\infty} \exp(in\phi) \int_0^{\infty} |G_n(R)|^2 2\pi R \, dR \quad (45)$$

The point of these equations is that the significant information in a picture $f(r, \phi)$ is contained in a relatively small number of $G_n(R)$, which can be calculated from the F.T. by Equation (34) or (39). From these few $G_n(R)$, the cross-correlation could be calculated with a sum replacing the angular ψ -integral. (The same principle, applied to solid angles, underlies the "fast rotation function" used in crystallography: Crowther, 1972; see also Dodson, 1985.)

7.2.4 Correlation methods

(a) Definition and utility of correlation functions

Linear or rotational periodicities in a picture imply that there are symmetry operations—translations or rotations—that leave it unchanged. So, instead of looking for periodicities, we could search for the symmetry operations. We would therefore translate or rotate the picture relative to itself, and measure how well it matches. Or we may have obtained (by averaging) a clear image of some substructure, and wish to determine the symmetry operation by finding where all copies of the substructure are to be found. Again, we need to translate or rotate the clear image and find where it gives a good match in a more noisy picture. Translating or rotating images presents no problem. But how are we to measure—rapidly—how well they match?

For simplicity, the approach is described for the one-dimensional case. (Extension to two dimensions is simple, involving little more than the replacement of scalars by vectors.) Suppose that we have a substructure $f(x)$, and that we are looking for its matches in a picture $g(x)$. We shift $f(x)$ by a possible translation t ; this changes it to $f(x - t)$. Now we need a criterion to measure how well it matches with the picture $g(x)$. A good criterion is the least-squares one: we take the mismatches at each point, square them (to make them positive) and add them together. The resulting mean-square error will reach its lowest value at the position of best match. So we adjust t to minimize the positive integral:

$$\begin{aligned} \int_{-\infty}^{\infty} \{f(x - t) - g(x)\}^2 \, dx &= \int_{-\infty}^{\infty} \{[f(x - t)]^2 + [g(x)]^2\} \, dx \\ &\quad - 2 \int_{-\infty}^{\infty} f(x - t)g(x) \, dx \end{aligned} \quad (46)$$

The first integral on the right-hand side is positive and independent of t . (The integral of the square of $f(x - t)$ is independent of t because of the infinite range of integration.) To minimize the right-hand side, therefore, we must maximize the last integral. This integral defines the *cross-correlation function* of $f(x)$ and $g(x)$. (It applies not only to cases where x and t are distances, but also when they are angles: rotational cross-correlation).

A peak in the cross-correlation function (X.C.F.) of two pictures indicates a relative position where they match well. The auto-correlation function (A.C.F.) is simply the X.C.F. of two identical pictures. Since a picture matches itself perfectly when there is no displacement, the auto-correlation function (A.C.F.) has a high peak at the origin. From this peak, the A.C.F. will (if the picture is finite) decrease in every direction. It becomes

zero when the displacement equals the picture's width, since the two copies of the picture then no longer overlap. (The X.C.F. similarly vanishes when the displacement equals the sum of the pictures' widths.) These limits will always apply in the case of translational correlations, for every width must be finite. But they do not apply to rotational correlations, as any angle—however large—is equivalent to one below 360° .

There are simple yet powerful rules for Fourier transforms (summarized in Fig. 7.12). Unfortunately, this is not true for correlation functions, because they involve products of functions (i.e. they are not linear like F.T.s). However, there is a useful convolution rule for A.C.F.s.: the A.C.F. of the convolution of two pictures is the convolution of their A.C.F.s.

(b) *A.C.F. of a periodic picture*

What is the value of the A.C.F. for analysing micrographs? First consider the A.C.F. of an infinite lattice of points. When displaced by a multiple of the lattice vector, all the lattice points coincide exactly, and the A.C.F. will have a peak. Any other displacement will make all the lattice points coincide with blank spaces, and the A.C.F. will be zero. So the A.C.F. of a lattice is the same lattice.

Next consider the A.C.F. of a picture with translational symmetry (like a wallpaper pattern), extending infinitely in all directions. Such a picture can be represented as a repeating motif, convoluted with a lattice (as in Fig. 7.25a). Therefore, by the convolution rule at the end of the previous section, its A.C.F. is the A.C.F. of the lattice (= the same lattice), convoluted with the A.C.F. of the motif. This last, like the A.C.F. of any finite object, will have a peak at the origin, from which it will decrease to zero in any direction at a distance equal to the object's width. When convoluted with the lattice, the origin peak will become repeated, generating a kind of blurred image of the lattice. Thus the lattice of the picture will be easily recognizable in the A.C.F., more so than in the original picture (unless the original motif had a single strong peak). This means that the A.C.F. helps to reveal linear periodicities (= translational symmetries).

However, the clarity of the repeat in the A.C.F. will be diminished by the convolution. For the motif's A.C.F., being twice its width in each direction, occupies two lattice repeats in each direction. The origin peak of the motif's A.C.F. thereby becomes superposed with two tails of the A.C.F. Thus the repeated A.C.F. peaks are less recognizable than the sharp peaks in a Fourier transform. The visibility of the peaks is further reduced when the repeating lattice is of finite extent (as in the case of translational, but not rotational, repeats).

(c) *Connections with F.T.s.*

There is a very close connection between the X.C.F. of two functions and their convolution. To obtain the convolution of $f(x)$ and $g(x)$, we would take each point of $g(x)$ and convert the corresponding value into a copy of $f(x)$. At the point $x = x_1$, $g(x)$ has the value $g(x_1)$, so the copy of $f(x)$ that we would place here is $g(x_1) \cdot f(x - x_1)$. Adding all these copies together, we obtain:

$$\int_{-\infty}^{\infty} g(x_1) f(x - x_1) dx_1 \rightarrow \int_{-\infty}^{\infty} f(t - x) g(x) dx \quad (47)$$

making the substitutions $x \rightarrow t$ and $x_1 \rightarrow x$. Referring to Equation (46), we see that this is the X.C.F. of $f(-)$ and g , where $f(-)$ gives, for $+x$, the same value that f gives for $-x$. Thus the X.C.F. of f and g in Equation (46) is the convolution of $f(-)$ and g . Suppose the F.T.s of f and g are F and G , respectively. Then, since F^* is the F.T. of $f(-)$, it follows that the F.T. of the X.C.F. of f and g is F^*G .

This close connection between the X.C.F. (or the A.C.F.) and the product of F.T.s is useful in various ways. It provides a quick method for calculating correlation functions, using the "fast F.T." (see Section 7.4.2(b)). It also provides a simple connection between the results of correlation analysis and those obtained from the picture's F.T.

Summarizing the main formulae of this section,

$$\begin{aligned} \text{X.C.F.}(f, g) &\equiv f(x) \star g(x) \equiv \int_{-\infty}^{\infty} f(u - x) g(u) du \\ &= \int_{-\infty}^{\infty} f(v) g(x + v) dv = \int_{-\infty}^{\infty} \{F(-X)G(X)\} \exp(-2\pi i Xx) dX \quad (48) \end{aligned}$$

7.3 IMAGE ANALYSIS BY INSPECTION

7.3.1 Introductory survey

In the following sections we shall be concerned with analysing images that show different types of symmetry or quasi-symmetry. When the image detail is degraded by noise, its reconstruction requires additional, independent, copies of the same image. Symmetry provides these, and is therefore usually an essential prerequisite for image enhancement. Symmetry can also provide the different views required for three-dimensional reconstruction. Finally, symmetry can be of intrinsic interest by revealing the probable patterns of interaction of macromolecular systems.

In images of biological structures we look for the common types of symmetry, which were briefly surveyed in Section 7.2.1. They involve translations and rotations, either in isolation or in combination. Translational symmetry produces aggregates with very many repetitions of the "signal" in the translation direction(s). Where present, this is the first symmetry to be exploited to give enhanced images which are then used to determine any rotational symmetry that may also be present. This sequence is followed in the image analysis of helical particles and of one- or two-dimensional crystals.

The earliest (and simplest) methods for determining symmetry involved examination of the micrograph, either directly or after transformation by some analogue procedure such as optical diffraction. The first applications of numerical methods often fitted into this general approach, e.g. by providing computed diffraction patterns with phases. Here we refer to all such methods as "image analysis by inspection" since, after processing (to produce the F.T., etc.), further analysis is transferred from the computer to the microscopist, who can contribute experience, judgment and understanding. In this section, we discuss such methods in detail, postponing to the next section any methods which require a more continuous use of the computer.

For determining either translational or rotational symmetries, three general methods have been used. We introduce them briefly here, and then proceed to discuss in detail the two most important methods.

(a) *Superposition methods*

The earliest method, requiring the least equipment, proceeded directly to the enhancement stage. A composite image was formed by the superposition of many copies of the original image, all related by the supposed symmetry (rotational or translational) (McLachlan, 1958; Markham *et al.*, 1963). If the symmetry had been guessed correctly, the composite image showed enhanced contrast and detail; an incorrect guess smeared the image. This method is very slow when searching for a completely unknown symmetry. Its use has been confined to symmetry refinement (i.e., finding exact translational parameters, or a choice among a few possible orders of rotational symmetry). Even here, it has suffered from the subjectivity of the criteria used to judge the correct solution.

(b) *Fourier methods*

Of the other two general methods, the most popular uses Fourier analysis. As explained in Section 7.2.2, this views all pictures in terms of a large set of basic density distributions. For images with translational symmetry, the appropriate shapes are straight bands of alternating density (Fig. 7.5). For helical particles, they are helical density waves (Figs 7.32 and 7.34). For images with only rotational symmetry, they are thin annuli whose density alternates with a regular rotational periodicity; see Section 7.4.6.

The translational density distributions are revealed immediately by the Fourier transform of the picture. A picture that shows symmetry does so by being formed from some special subset of the possible density distributions. (Thus, for a one-dimensional picture with translational symmetry, all the density distributions must have spatial frequencies that are multiples of the picture's.) The particular subset used is characteristic of the symmetry, which is revealed by finding the subset.

(c) *Correlation methods*

The third general method uses the correlation functions introduced in Section 7.2.4. There it was explained how the Fourier and correlation methods, so apparently different, turn out to be closely related. They measure symmetry fits in essentially similar ways, but they differ in convenience. The Fourier method is quicker to apply, especially in its analogue form (optical diffraction). So a preliminary search for picture symmetry is best undertaken by this means. But the Fourier method is far less powerful when symmetry is approximate, i.e. when the structure is distorted or disordered. Then it is necessary to use correlation methods.

However, correlation methods are relatively new, and have not yet been applied so extensively that we can be sure about the set of problems for which they will be most effective. At present they seem better adapted to refining symmetry, rather than to finding it originally. Since they make extensive use of computer processing, they will be discussed mostly in Section 7.4.4.

7.3.2 Analogue image processing methods

Each of the methods just described requires appropriate manipulations of the information contained in a micrograph. These can be achieved either by analogue or by digital

7. Image Analysis of Electron Micrographs

methods. In either case, the micrograph is first illuminated, giving a brightness distribution of the transmitted light. Analogue methods use optical techniques to transform this directly into a pattern appropriate to Fourier or correlation analysis. Though fast and cheap, such methods can achieve very few types of transformation in a straightforward fashion.

Digital techniques use electronic sensors and circuitry to transform the brightness distribution into numbers, which are then processed with a computer. Given sufficient time and cost of processing, any method of analysis can be used. (Numerical methods for finding symmetry will be described in Section 7.4.) The cost and slowness of digital methods have fallen steadily with the development of computer hardware. When image-processing was developed in the 1960s, digital methods were rarely used, but the last 20 years have reversed the situation. However, analogue methods still have a great advantage in speed, and therefore remain useful, even if only to select the best images for digital processing.

(a) *Visual analysis of electron micrographs*

Initial image processing is inevitably done with the eye, to assess specimen preservation, staining (etc.), and imaging conditions. Obvious periodicities should be noted, and less obvious ones should be looked for by viewing the micrograph obliquely (such a view approximates to a projection along that direction). It is important not to be content with the two-dimensional appearance of the micrograph, but to try to interpret the underlying three-dimensional structure of the specimen. Measurements can be made, and compared with the predictions of different classes of model structures. (Such measurements can take account of unusual distortions, such as helical transitions: Moody, 1973.) If the periodicities and other features are clearly distinguishable, visual methods can be quite useful. (An example is given in Section 7.3.4(a).)

(b) *Optical diffractometer*

This was originally developed for use in X-ray crystallography (the history is described by Taylor and Lipson, 1964). Its first application to electron micrographs by Klug and Berger (1964) started the quantitative phase of micrograph image analysis, and it has since remained the most important analogue device.

The underlying principle is simple. Any lens illuminated by a coherent beam of light, parallel to the lens axis, gives a sharp focal spot on that axis. If a diffraction grating is introduced into the parallel beam before the lens, the reinforcement of scattered waves yields diffracted beams, each of parallel light. They diverge at angles that vary reciprocally with the spacing of the grating, i.e. the angles are proportional to its spatial frequency. The lens focuses each scattered beam to a point, displaced from the lens axis by a distance proportional to its angle, if that is small. Consequently, the distribution of light intensity at the focal plane is a plot of the spatial frequencies of the grating. Generalizing, it is a plot of the spatial frequencies in *any* transparent object (even if non-periodic) which is inserted into the parallel beam of light.

This pattern (the *optical diffraction pattern* or optical transform of the transparent object) gives the intensity of the diffracted light; the relative phases of the spots are lost. (It is possible to measure them by interference techniques, but the same effort would be

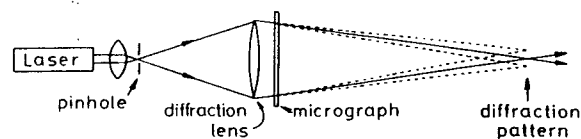


Fig. 7.49 Essential elements of an optical diffractometer. Parallel coherent (usually red) light from a laser is focused onto a pinhole which excludes non-coherent light. The diverging light is brought, by the diffraction lens, to a focus on the right. Each diffracted beam is brought to a different focus at the same plane.

better repaid by using numerical methods.) Essentially the same optical diffraction pattern is obtained if the transparent object is illuminated by light that is slightly divergent or convergent, instead of exactly parallel, or if the distance of the transparent object from the lens is varied somewhat. (The main effect of any of the above changes is not on the intensities, but on the relative phases, of the diffracted spots; see Goodman, 1968.)

For details of the construction and use of an optical diffractometer, see the book by Taylor and Lipson (1964), or the reviews by Markham (1968), Johansen (1975), Gibbs and Rowe (1977) and Erickson *et al.* (1978). The essentials (Fig. 7.49) are the laser light source; the short focal lens and pinhole to remove non-coherent light; and the diffraction lens. To obtain diffraction patterns of reasonable size from ordinary micrographs, it is necessary to use a long path length (even when using the red light of a helium-neon laser). But all the elements of the diffractometer can be made easily accessible by "folding" the path with surface-coated mirrors.

Although electron micrographs on glass plates often give satisfactory diffraction patterns in air, the patterns can sometimes be distorted or misleading. This is because variations in the emulsion thickness, correlated with optical density, cause phase differences in the transmitted light. These contribute at least as much as the amplitude variations to the optical diffraction pattern;* they can be minimized by immersing the electron micrograph in oil of refractive index about 1.53, contained in a cell bounded by optical flats or by the two main diffraction lenses (Berger and Harker, 1967). But such refinements compromise the diffractometer's chief virtues, speed and simplicity. When accuracy is more important, it is preferable to use numerical methods.

(c) Optical correlator

Correlation functions can be obtained by analogue devices, optical correlators. Much of their development was also undertaken for X-ray crystallography (Buerger, 1959; Hosemann and Bagchi, 1962). In its simplest form, the optical correlator consists of two identical copies of the electron micrograph on photographic plates which are held parallel, and in the same orientation (Fig. 7.50). Consider the light that follows a line joining some common feature (A and A') in both plates. After passing through the first plate, the light's intensity becomes multiplied by the transmittance at that point. The same happens when it passes through the second plate, so the final intensity is

* Another peculiarity of optical diffraction is that the relevant optical transmission factor is not the optical density (as it is for densitometry). For optical density affects the intensity ($\psi\psi^*$) of the transmitted light, whereas it is the wave function ψ that is involved in interference phenomena.

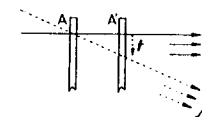


Fig. 7.50 Principle of the optical correlator. Two parallel copies of the micrograph (A, A') are illuminated from the left by incoherent light. (See text for the mechanism of its operation.)

proportional to the square of the transmittance, since it passes through the same feature on each plate. Rays of light parallel to this also pass through identical features on both plates. If we took all the parallel rays and added them together, the relative intensity would measure the sum of all the squared transmittances.

Now consider a ray of light in some other direction. It "sees" different features on the two plates, the features being related by some translation t . Rays parallel to it also see different features, which are related by the same translation t . So, if we add together all these parallel rays, we get the sum over products of transmittances from points related by the same translation t . $|t|$ is proportional to the angle of the rays (for small angles).

We can sum together all the parallel rays in any direction by placing a lens after the second plate. This focuses all parallel sets of rays to point images in the focal plane. The position of each focal point depends on the angle of its parallel rays, and hence on the translation t . (Thus the distribution of intensity around the focus is proportional to the auto-correlation function (Equation (48)), where $f(x)$ and $g(x)$ are both equal to the transmittance of the plate.) Since the lens sorts out all sets of parallel rays, we can now illuminate the first plate with all possible sets, i.e. by incoherent illumination such as from a light-box. The two plates in Fig. 7.50 need not be the same; if they are different, the optical correlator gives the cross-correlation function (Equation (48), where $f(x)$ and $g(x)$ are the transmittances of the two plates). Finally, the lens is not needed if the two plates have different magnifications, the larger being placed closer to the light source.

Although the analogue optical correlator is used occasionally (e.g. by Fiskin and Beer, 1968), its popularity has never approached that of the optical diffractometer. Two plates are needed for the correlator (but only one for the diffractometer), and the peaks from optical correlation are less sharp than those from diffraction (Section 7.2.4(b)). Correlation methods have recently become very popular (Sections 7.4.4, 7.4.6(f)), but they are usually applied by numerical methods. However, it is possible that the optical correlator might be used (as is the optical diffractometer) to scan micrographs rapidly in a search for areas suitable for numerical processing.

7.3.3 Applications to one- and two-dimensional lattices

Most applications of the methods of image analysis by inspection have used the optical diffractometer, followed by computed F.T.s with phases. (Their calculation is described in Section 7.4.2(b).) These applications have exploited the suitability of F.T.s for studying translational symmetry, and concentrated on appropriate specimens. Fortunately, these are fairly common, since translations allow many small subunits to create a structure sufficiently large to create a cellular component. Aggregates based on lattices are therefore common. Usually they are two-dimensional (e.g. sheets of subunits, or cross-sections of

ordered bundles of filaments), but sometimes one-dimensional images are also seen, e.g. with fibrous proteins like collagen. (Such images are, however, projections of more complex arrangements.)

There are several stages in the analysis of symmetry based on lattices (called plane-groups). First we detect and measure the lattice. Next, we find the complete plane-group, of which the lattice is only part. Knowing this, we can clarify the image by one of the techniques of image enhancement described in Section 7.5. In this section we are concerned only with symmetry determination.

a) Detecting the lattice in the presence of noise

The most difficult problem is often detecting the existence of a lattice. It is here that the optical diffractometer finds its principal use. Large areas of micrographs can be surveyed rapidly, and those regions that contain two-dimensional crystals immediately reveal themselves by the presence of a lattice of sharp spots in the optical diffraction pattern. This diffraction (reciprocal) lattice can be quite clear, even when no lattice whatever is visible in the micrograph. (This is the case, for example, with micrographs of unstained specimens obtained at very low electron doses: Unwin and Henderson, 1975.) How is the optical diffractometer so effective?

The diffractometer, as explained above, gives the (squared amplitude of the) F.T. of the transmittance of the micrograph. Our question is thus reformulated as, how can the F.T. of a very noisy crystal image contain clear sharp peaks? First, we note that such a crystal image is really the image of a unit cell, convoluted with a lattice of points. As explained in Section 7.2.2(g) (see especially Fig. 7.25), this gives the F.T. of the unit cell sampled at the reciprocal lattice points. The sampling "points" will be sharp peaks if the picture's lattice is undistorted and reasonably large.*

This explains how the diffraction pattern consists of a lattice of sharp peaks. But how is it able to emerge from an extremely noisy image? To understand that, we must consider the F.T. of noise. This may seem a hopeless task, since no two samples of noise are ever exactly the same. However, we are interested only in the overall, statistical, features of its F.T. The key to deriving these is to note that noise is equivalent to information without any redundancies. If a message has had all repetitions and other predictable features removed, then it will look exactly like noise to someone who is ignorant of the language or code in which it is written. Image noise, in short, is an extremely complex function that can hold the maximum amount of information, given the resolution of the image. Consequently its F.T. must have a similar level of complexity, so it too will resemble noise.

It would therefore seem that the F.T. of noise is simply noise. However, there is one important correction to this statement. Some forms of image noise are always dark; they never make the image brighter. In that case, their projection onto a straight line will be a nearly uniform density distribution, whose F.T. will have a peak at the origin. So, unless the image noise is as often bright as dark, its F.T. will have a sharp peak at the origin, but be roughly uniformly noisy elsewhere.

* A small crystal is just an infinite crystal multiplied by a "mask" function, giving an F.T. in which the sharp peaks of the reciprocal lattice become convoluted with the F.T. of the mask, and are thereby broadened.

Armed with its F.T., we now consider exactly how noise will degrade the image. There are two common ways. First, noise may be added to an otherwise perfect image. For example, the supporting film adds a background of irregular thickness; and, at low exposures, photographic "fog" is serious. Such noise will modify the perfect F.T. by adding the F.T. of noise, i.e. by adding noise to the "perfect" F.T. (A sharp peak is also added at the origin.) The second way for noise to enter an image will be more prevalent in grainy images. These can be crudely represented as perfect images multiplied by a "grain function".* That function would be zero except at the (random) positions of grain particles. This "grain function" is just "dark noise" (all grain is "dark"); its F.T. will be uniform noise with a sharp peak at the origin. But that transform must now be convoluted with the F.T. of the perfect image. Thus, every point of the perfect F.T. must be replaced by a sharp peak (which produces no significant change), plus uniform noise. That second feature adds uniform noise to the overall F.T.

We thus obtain essentially the same result, irrespective of how the noise was added: we get the original F.T., plus uniform noise. But the effect of adding noise to the F.T. is quite different from that of adding it to the original picture. For the F.T. of a crystal has all the information concentrated into a few sharp peaks; so the addition of noise is nothing like so serious as with the original image, where the structural information was also spread uniformly.

(b) Detecting the lattice in the presence of other lattices

The F.T. can also be used to reveal the crystal's symmetry even when other structures, including other crystals, are superposed on it (Finch *et al.*, 1967). The most common situation occurs when the structure of interest is a large sheet that has folded over, superposing its two halves. Unfolded regions may be impossible to find (e.g. because the structure is a wide, thin, hollow tube that has flattened) or undesirable (because the staining or preservation in the folded region is interesting). Then we have to determine separately the symmetries of the upper and lower sheets, as a preliminary to image enhancement by filtering. How will this be easier when using the F.T.?

First consider the connection between the image of the superposed sheets, and the images of each sheet in isolation. Ignoring structural or staining changes following superposition, the transmittance of the pair of sheets will, to a reasonable approximation, equal the product of their transmittances. However, if the scattering is very small, each transmittance will be nearly 1. Representing the transmittance of sheet number j , then, as $\{1 - f_j(x)\}$, the transmittance of the pair is $\{1 - f_1(x)\}\{1 - f_2(x)\} \doteq 1 - \{f_1(x) + f_2(x)\}$, as the last term in the expansion is negligible. When the transmittances are very high, therefore, the combined image is approximately the sum of the individual images (and, consequently, the combined F.T. is approximately the sum of the individual F.T.s). The low contrast of micrographs, so often the subject of regret, has its compensations.

Now return to the problem of disentangling the two lattices. Using F.T.s, this is quite simple, provided the crease of the fold is not in an unlucky direction. The situation is illustrated in Fig. 7.51. The sheet in (a) has folded at a vertical crease, superposing the two areas of hexagonal lattice. In (b), the F.T. of (a) has separated the spots corresponding

* The irregular structure of negative stain can probably be considered a source of "multiplicative noise".

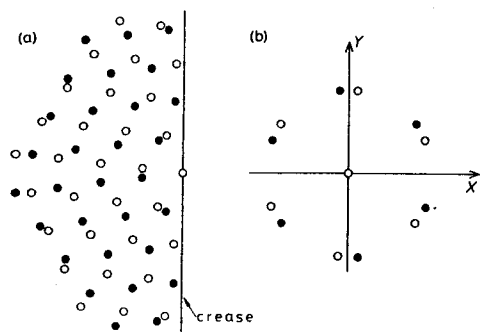


Fig. 7.51 (a) A thin sheet with a hexagonal lattice is folded at a crease. Filled circles: upper sheet; open circles: lower sheet. (b) The diffraction pattern of the folded sheet separates the diffraction patterns of its two parts.

to the F.T.s of the two surfaces, and there is no overlap. However, if the crease had lain perpendicular to a prominent lattice direction (i.e. a short lattice vector), then spots would have been superposed. The separability of the two surfaces by optical diffraction (or filtering) is dependent on a "random" orientation of the crease.

(c) Indexing the reciprocal lattice

Indexing means assigning to each reciprocal lattice point its coordinates in terms of two basic vectors of the lattice (usually the two shortest vectors that are not parallel). If these vectors are \mathbf{a} and \mathbf{b} , then each point in the lattice can be represented by $i\mathbf{a} + j\mathbf{b}$, where (i, j) are its coordinates. The lattice vectors \mathbf{a} and \mathbf{b} must be chosen so that every reciprocal lattice point can have integral coordinates assigned to it (see also Chapter 3, Section 3.3.2).

With the diffraction pattern of a single sheet, the process is very simple, though care should be taken to ensure that the chosen lattice really includes all the significant diffraction spots. Sometimes there are weak spots in the centres of the parallelograms formed by the strong spots. If they were ignored when choosing the basic vectors, these spots would require fractional coordinates, showing that one or both of the basic vectors should be changed for shorter ones.

The process is slightly more difficult when the lattice is part of a folded sheet, as in Fig. 7.51. In this case, half the diffraction spots derive from the upper sheet, and half from the lower. It is necessary to decide which spots are which, before indexing them. (The difference is indicated by the empty and filled circles in Fig. 7.51; we shall now denote the two surfaces by the subscripts 1 and 2.) We proceed as follows. We choose one of the spots closest to the origin, and denote its vector \mathbf{a}_1 . Then \mathbf{a}_2 is the vector from the origin to the corresponding spot, reflected in the Z -axis. (The vectors \mathbf{a}_1 and \mathbf{a}_2 differ only in referring to surfaces 1 and 2.) Next, we find the shortest vector (or one of the shortest vectors) that is not parallel to either \mathbf{a}_1 or \mathbf{a}_2 ; call it \mathbf{b} . The problem is to find if it is \mathbf{b}_1 or \mathbf{b}_2 . This can be decided by trial: first we try \mathbf{b}_1 , and generate the inner part of the lattice from \mathbf{a}_1 and \mathbf{b}_1 . Next we try (in the same way) \mathbf{b}_2 . One of the lattices will fit the diffraction spots, and the other will not, showing which choice was correct.

(d) Finding the plane-group

Once the lattice peaks have been obtained in an optical diffraction pattern or a calculated F.T., the complete space-group should be found.

The first step is simple: measuring the lattice parameters. In Section 7.1.2(f), it was shown that the shape of the real lattice is the same as that of the reciprocal lattice (i.e. of the lattice on which the diffraction spots lie), but just rotated through 90° . So we have only to measure accurately the shape of the diffraction lattice. This shape is completely defined by the lengths of two lattice vectors, and the angle between them. Thus only three parameters need to be extracted (by least-squares procedures) from a considerable amount of data. (These data consist of two coordinates per transform peak, but the relative accuracy of the three parameters depends on the distribution of these peaks.)

Once the lattice is known, the next step is to find which of the 17 plane-groups (International Tables for Crystallography, 1983) applies to the image. Several techniques are useful in finding this.

First, the lattice shape (from the shape of the reciprocal lattice) helps to limit the choice. Five of the groups ($p3$, $p31m$, $p3m1$, $p6$, $p6m$) have an exactly hexagonal lattice; three ($p4$, $p4m$, and $p4g$) have an exactly square lattice; and seven other plane-groups have exactly rectangular lattices.

Second, we look at the pattern of intensities of the diffraction spots. This has the rotational symmetry of the plane-group, with the addition of a two-fold axis (because of the Friedel symmetry of the F.T.). Thus, if the intensities have six-fold symmetry, this shows that the plane-group has either three-fold or six-fold symmetry. Besides this, there is significance in the pattern of intensities along the main lattice-lines that pass through the origin. Three of the groups (pg , pgg and $p4g$) contain glide-lines (see Section 7.2.1). Images containing these lines give, after projection perpendicular to them, a density pattern in which the repeat is halved (Fig. 7.52). Consequently, the spacing of spots along the parallel section through the F.T. must be doubled. This means that, along this line, every second spot is missing. Such "systematic absences" are listed in the description of each plane group in International Tables for Crystallography (1983).

Finally, we look at the phases of the diffraction spots, obtained from a numerical F.T. Whereas the diffraction intensities are constrained, by Friedel symmetry, to have a two-fold axis, the phases are not. So a two-fold axis in the phases means that the original lattice also has one. As a final check on the plane-group, a filtered image of the crystal should be compared with the appearances expected from the different plane-groups.

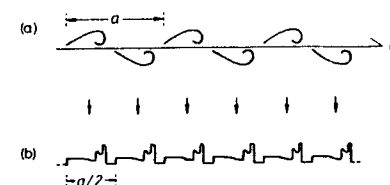


Fig. 7.52 (a) Commas arranged in a two-dimensional pattern with a glide-line (g). The repeat distance of the pattern is a . (b) Its projection, perpendicular to the glide-line. The repeat distance of the projection is $a/2$.

In determining the plane-group, there are two problems peculiar to electron micrographs. First, distortions may be present (Section 7.4.4), so that (for example) a truly hexagonal crystal may have its lattice angle a few degrees away from 60° . Moreover, phases that should be the same can often differ significantly. Thus the apparent symmetry may be lower than the real symmetry. The other problem is that electron micrographs usually have a very limited resolution, so that subunits that really differ may appear indistinguishable. This problem has the opposite effect: the apparent symmetry may be higher than the real one. For these reasons, determining the symmetry of structures from micrographs requires more judgement, and is less certain, than when using X-ray crystallography.

(e) *Using the plane-group*

Knowledge of the plane-group is useful in several ways. First, it will be needed when preparing enhanced images (e.g. by Fourier filtering). Symmetry elements, such as a rotation axis, additional to the lattice symmetry can be exploited to give an even better image.

The second use is in revealing the pattern of interactions of the molecules in the picture. If the specimen is a section of a much thicker structure, then attention should be directed to the possible three-dimensional space-group (these are also listed in International Tables for Crystallography, 1983). However, if the specimen is a sheet of monomolecular thickness, and if the plane-group is found to be $p1$, $p2$, $p3$, $p4$ or $p6$, then it is likely that the image can be interpreted directly in terms of single molecules. (But the exact boundaries of the molecule are uncertain; the plane-group determines only their repeating pattern, and their projected area.) If the picture's symmetry should correspond to one of the other 12 plane-groups, then the picture's plane-group probably represents a projection, in which a mirror-line derives from a two-fold axis in the specimen, and a glide-line from a two-fold screw axis. Thus the true symmetry of the sheet would be one of the two-sided plane groups listed by Holser (1958).

7.3.4 Applications to helical particles

Symmetry determination is not quite so straightforward with helical particles as with two-dimensional crystals. We should first recall exactly what parameters need to be found in the case of strict helical symmetry. First, every helical particle has a *screw displacement*, consisting of a translation (h , the rise distance) parallel to the helix axis, combined with a rotation (Ω , the twist angle) about the axis (Fig. 7.27). (A single screw displacement generates all the different sets of helices that can be seen in the micrograph.) In addition to the screw displacement, the particle may have rotational symmetry: it could have one of the point-groups C_N or D_N (Section 7.2.1), in which the N -fold axis is parallel to the helix axis. (No other point-group is consistent with a screw displacement.) To find the point-group, we need to determine the rotational symmetry about the particle's axis (i.e. the N -fold axis), and also whether or not it is polar (i.e. whether there is a perpendicular two-fold axis, as in D_N).

(a) *Finding the structure of helical particles by direct examination*

Although the structural analysis of helices usually requires Fourier methods, it is occasionally possible to carry it out by visual methods alone. The extended T4 phage sheath (Figs. 7.28, 7.29 and Section 7.3.4(f)) shows clear annuli, which simplify the direct analysis of its images (Moody, 1967b). First, the spacing of the annuli gives the rise distance h . Second, the helical lattice is completely determined by the intersection of the annuli with any set of helices whose number is the minimum ($= N$, the order of the rotation axis). Such helices are clearly visible, allowing the analysis to be completed.

If no annuli are visible, the analysis is more difficult. It is then necessary to see clearly at least two different sets of helices which intersect only at lattice points (which they are likely to do, given the poor resolution of micrographs). For each set, two parameters are needed: the axial repeat (z), which is the distance, measured parallel to the helix axis, between successive helices of the set; and the number (n) of helices in the set. n is positive for right-handed helices, and negative for left-handed ones, whereas z is always positive.

The two sets of helices give us four numbers (z_1, z_2, n_1, n_2). We first calculate N , the order of the rotation axis, which is the highest common factor (H.C.F.) of n_1 and n_2 . Then the minimum rise distance h is given by

$$h = |Nz_1z_2/(n_1z_1 - n_2z_2)| \quad (49)$$

Also, the pitch P of the basic helices is given by

$$P = |Nz_1z_2/(k_1z_1 - k_2z_2)| \quad (50)$$

where k_1 and k_2 satisfy the equation

$$k_1n_2 - k_2n_1 = N \quad (51)$$

(k_1 and k_2 will be obtained when using Euclid's algorithm to find the H.C.F. of n_1 and n_2 .) From P and h , the twist angle Ω of the basic helices can be found from

$$\Omega = 360^\circ(h/P) \quad (52)$$

(b) *Analysis of helical F.T.s: introduction*

Unless the helical structure is exceptionally clear, serious structural analysis will need some form of image processing. Apparently all such analyses have employed Fourier (rather than correlation) methods (see Stewart, 1988, for a review of them). The helical F.T. will probably be seen first by optical diffraction, as part of a rapid survey of the micrographs. This shows only the F.T. amplitudes; but the initial stages of analysis are concerned only with these. Some examples of optical diffraction patterns are presented in Figs 7.58, 7.59 and 7.60. Note that these are relatively sharp in the Z -direction, but spread out in the perpendicular X -direction. (This is because helices are longer in the z -direction, which also contains the translational repeat.) So measurements of Z -coordinates are precise, and those conclusions that can be drawn from them are relatively certain. The X -coordinates, that relate to the angular parts of the symmetry, are much more difficult to interpret.

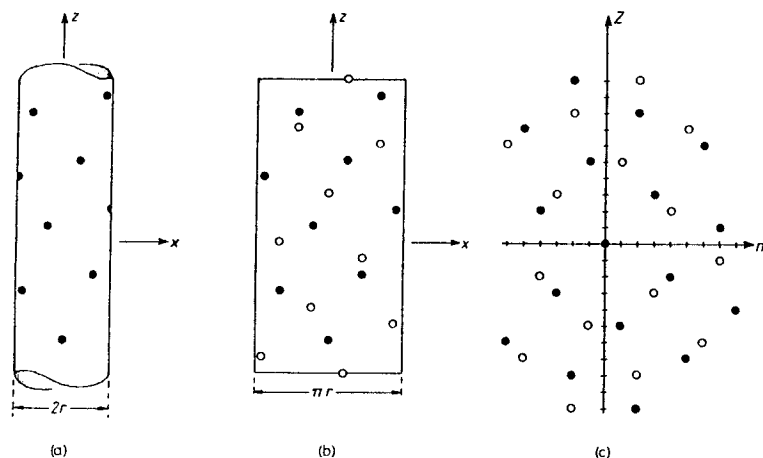


Fig. 7.53 The separation of upper and lower surfaces by optical diffraction, illustrated for a plane sheet in Fig. 7.51, applies also to helices. (a) A helical pattern attached to a thin cylinder. (b) The cylinder has been flattened (filled circles: upper sheet; open circles: lower sheet). (c) (n, Z) plot of the diffraction pattern (this is closely related to the pattern itself; see Fig. 7.37). Each surface contributes its own reciprocal lattice in this plot.

The Z -axis is a line of approximate mirror symmetry. As explained in Section 7.2.3(g), a helical lattice (Fig. 7.53a) is not very different after flattening (Fig. 7.53b). We then obtain a folded sheet with two creases. We have already seen how optical diffraction allows us to disentangle the two parts of a folded sheet (Fig. 7.51). In the same way, the flattened helix would give a diffraction pattern resembling Fig. 7.53c, where the filled circles would derive from the upper surface in Fig. 7.53b. This diffraction pattern, though only a rough representation of the original helical lattice, can be made into a more exact one by converting Fig. 7.53c into an (n, Z) plot. This we do by rescaling the X -axis to give an n -axis (as explained in Section 7.2.3(c)). (However, because of the "dextrist" convention that right-handed helices have a positive n , the open circles must now refer to the upper surface.)

Most of the analysis of the F.T. of a helix is concerned with obtaining this (n, Z) plot. The first step is to disentangle the two lattices (represented by open and filled circles). (The problem is that each layer-line gives us $|n|$ and $|Z|$, from which we get the alternative pairs of lattice points $\pm(|n|, |Z|)$ and $\pm(|n|, -|Z|)$; these differ if $|n|$ is not zero.) The second, and much bigger, step is to extract exact n -coordinates from the observed X -dependence of the transform amplitudes (and other data).

We shall soon look at these steps in more detail. However, it might be useful to clarify first our use of (n, Z) plots, rather than the more common (n, L) plots. As explained in Section 7.2.3(c), the (n, Z) plot is more appropriate when there is no exact repeat. However, it is quite valid to use the (n, Z) plot even when there is a short, exact repeat, just as the (n, L) plot can be used (to within experimental error) for indexing patterns that lack a repeat, provided the "repeat" is made sufficiently large. The existence, or non-existence, of a "repeat" really refers to the existence, or non-existence, of interference

between layer-lines. That is an experimental fact, whereas the use of (n, Z) or (n, L) plots is only a question of convenience and choice.

(c) *The preliminary (n, Z) plot*

The first stage in analysing a single particle's transform is indexing, i.e. relating all the significant reflections to two vectors. This is best illustrated by a concrete example; we shall use a hypothetical structure, and suppose that its diffraction pattern has layer-lines from which we obtain the Z -coordinates of the rows of circles in the unindexed (n, Z) plot of Fig. 7.54.

Approximate n -coordinates. Before we can plot these rows, however, we need at least some estimate of their n -coordinates. Each of these is calculated from two measurements.

First we estimate, from the diffraction pattern, the X -coordinate (= R -coordinate) of the principal maximum for each layer-line (see Figs 7.58, 7.59 and 7.60, for examples). Denote it by $R_M(n)$, where n is the (unknown) order of the layer-line. (We shall assume that there is only one clear principal maximum on each layer-line, and that there is no interference between the layer-lines; this is usually the case, but some exceptions are discussed in Section 7.3.4(j).)

Next, we obtain estimates for the particle radius. Different features of a particle are often present at slightly different radii. Ideally, we want the radius of the helical lines that give rise to the particular layer-line whose n -value we are calculating. This may be different for different helical lines, and anyway it is difficult to determine directly. So we choose the radius where most of the helical structure seems to lie, but make allowance for the range of radii present in the particle. Thus we have a minimum, a most probable, and a maximum radius. For each, we calculate the quantities $2\pi r R_M(n)$ for every layer-line. Then we use Fig. 7.43 to estimate the n -coordinate from each of these quantities. This gives us a range of possible n -coordinates for each layer line.

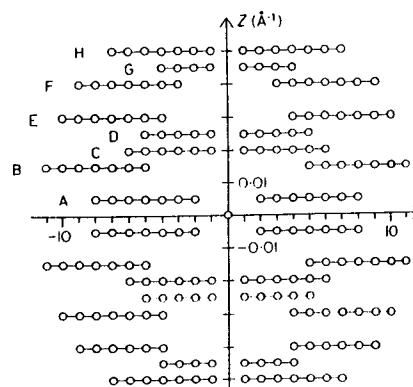


Fig. 7.54 The first of a series of figures to illustrate the analysis of helical diffraction patterns (described in the text). The (hypothetical) diffraction pattern has undergone a preliminary analysis to measure the precise Z -coordinates, and to estimate the range of the (unsigned) n -coordinates.

We use this range to construct an unindexed (n, Z) plot like that of Fig. 7.54. There the range of n is fairly generous, so the outlying values are much less likely than the middle ones. Also, the range has one clear boundary, since there are no circles placed on the Z -axis (i.e. with $n = 0$). This is because there is a *qualitative* difference when $n = 0$: there is only one (central) principal maximum on the layer-line, instead of two.

Initial indexing. Given the unindexed (n, Z) plot of Fig. 7.54, we now seek to index it, i.e. reduce it to a combination of two vectors. Call these $\mathbf{a} = (n_1, Z_1)$ and $\mathbf{b} = (n_2, Z_2)$. Then every row of circles must be expressible as $i\mathbf{a} + j\mathbf{b}$, so that its n -coordinate = $in_1 + jn_2$, and its Z -coordinate = $iZ_1 + jZ_2$. We need to find \mathbf{a} and \mathbf{b} from the data in Fig. 7.54.

Because of the greater accuracy of the Z -coordinates, we start with them. Let us take layer-line A as defining vector \mathbf{a} ; the right side of A gives us $\mathbf{a} = ([2 \rightarrow 8], 0.005)$. (If we chose the left side, we should generate the mirror image of the (n, Z) plot, implying a helix of opposite hand. But determination of the hand is a difficult matter best postponed until later.) Next, we choose layer-line C (instead of B) to define \mathbf{b} . (This is reasonable since, although C has a larger Z -coordinate than B, it has a smaller n -coordinate.) But we can choose either the left or right side of layer-line C; so we have two possibilities, $\mathbf{b}_+ = ([1 \rightarrow 6], 0.02)$ or $\mathbf{b}_- = ([-6 \rightarrow -1], 0.02)$.

Which of the two is correct? We try each possibility, and see which best fits the data in Fig. 7.54. Since the difference lies in vector \mathbf{b} , the best test involves the highest multiple of \mathbf{b} . Consider, therefore, $\mathbf{a} + 2\mathbf{b}$. The first possibility is $\mathbf{a} + 2\mathbf{b}_+ = ([2 \rightarrow 8], 0.005) + 2([1 \rightarrow 6], 0.02) = ([4 \rightarrow 20], 0.045)$. The second possibility is $\mathbf{a} + 2\mathbf{b}_- = ([2 \rightarrow 8], 0.005) + 2([-6 \rightarrow -1], 0.02) = ([-10 \rightarrow 6], 0.045)$. At $Z = 0.045$ we have the layer-line G, with $n = -4 \rightarrow -1$ or $1 \rightarrow 4$. \mathbf{b}_+ gave us $n = 4 \rightarrow 20$, which overlaps with line G at only $n = 4$. \mathbf{b}_- , on the other hand, gave us $n = -10 \rightarrow 6$, which overlaps with line G over its entire range. Therefore \mathbf{b}_- is clearly preferred, since the edges of the n -ranges are the least likely values.

(d) Improving the n -coordinate estimates

Parity. We now have two vectors, \mathbf{a} and \mathbf{b} , with which to index Fig. 7.54. Each vector has a precise Z -coordinate, but a wide range of n -coordinates. The next task is to restrict this range—to a single value, if possible. We have already exploited all the information in the F.T. amplitudes, so we should next make use of the phases. Comparison of the phases at $+Z$ and $-Z$ provides information about the parity of n : if n is even, the phases are equal; if n is odd, the phases differ by 180° (Section 7.2.3(b)). (With an undistorted perfectly aligned particle, the phase difference can be only 0° or 180° ; any other value implies experimental error, distortion, tilt, etc., which must be quite large to make the true phase difference uncertain. However, the particle position and orientation may need to be corrected, as described in Section 7.4.5.)

Let us suppose that this analysis has given us the parity of n for each layer-line of Fig. 7.54. We incorporate this information into the provisional (n, Z) plot of Fig. 7.55. n -Values with the correct parity are indicated by circles, and the excluded ones by dots. Note that these new data confirm our choice of \mathbf{b} , since \mathbf{b} would have used the excluded $n = 4$. We have also removed the less likely half of each layer-line (i.e. the part with either positive or negative n), to take account of our exclusion of \mathbf{b}_+ .

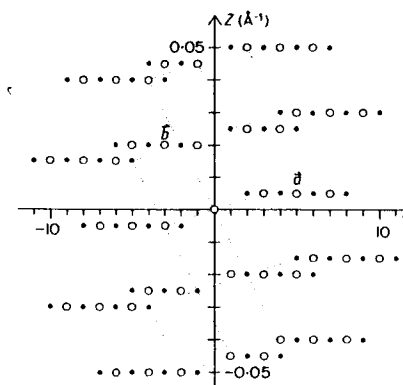


Fig. 7.55 The second stage in the analysis. The layer-lines have been analysed to give two vectors, and the parities of the n -coordinates have been determined. This excludes many otherwise possible (n, Z) lattices, like the one shown.

We now use Fig. 7.55 to refine our estimates of \mathbf{a} and \mathbf{b} . So we try to draw (n, Z) lattices that employ only permitted n -values (open circles). This attempt will exclude many of the alternatives. (For example, the lattice indicated by broken lines in Fig. 7.55 passes through many excluded n -values, indicated by dots.) It will be found that only one lattice fits the data in Fig. 7.55, and this employs the vectors $\mathbf{a} = (5, 0.005)$ and $\mathbf{b} = (-3, 0.02)$, as marked on Fig. 7.55.

Helix counting. This example has shown how knowledge of the parity of n , combined with the need to construct an (n, Z) lattice, can yield a unique solution (Fig. 7.56) from a very unpromising beginning (Fig. 7.54). However, there are cases where a unique solution is not obtainable in this way. Then it is necessary to look for other sources of information: we should need to count the helices in some set. This could be done, for example, if it were possible to find a view of the particle nearly down its long axis; that might allow us to count (i.e. assign n to) some set of helices. If we could do this and (which is not too difficult) find which set of helices we are counting, we should know n for some specific layer-line. This would fix one vector of the (n, Z) lattice.

Suppose, however, that it proves impossible to obtain any different views of the particle. Then there are other ways of improving our estimates of the numbers of helices. Prominent helices can be optically filtered (Section 7.5.2), yielding clearer images that may give unambiguous counts. If the particles are undistorted, a more quantitative method is to obtain images of them tilted through different angles about the particle axis. In the F.T., the phase of a prominent layer-plane can be plotted so as to reveal how rapidly it changes with angle (Finch, 1972b). This is connected with n , since the phase rotates n times in one complete revolution around the layer-plane. (Before this is possible, however, the F.T. must be corrected for the position and orientation of the helix axis; see Section 7.4.5.)

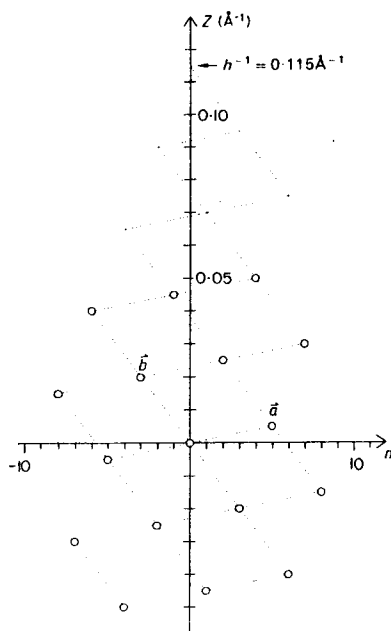


Fig. 7.56 Final (n, Z) plot of the helix. Each open circle represents a layer-line of Fig. 7.54, with an acceptable n -coordinate (Fig. 7.55). In the upper part of the plot, unobserved layer-lines have been added to allow the rise distance h to be measured.

(e) *The final (n, Z) plot and the helix structure*

Using the vectors \mathbf{a} and \mathbf{b} from Fig. 7.55, we can draw the final (n, Z) plot of Fig. 7.56. This has just one lattice point (open circle) corresponding to each layer-line of Fig. 7.54. We can use the lattice in Fig. 7.56 to find the rise distance h , even though this was too small to give a layer-line in the original F.T. If we need accurate values of the helical parameters, then the Z -coordinates of \mathbf{a} and \mathbf{b} can be refined by a least-squares fit to all the layer-line data (Smith and Aebi, 1976).

We can also use the (n, Z) lattice to generate a picture of the original helical structure. To do this, we first re-draw the lattice after reflection in the Z -axis, and then rotate it through 90° . This gives us the lattice in Fig. 7.57, representing the outside of the helix. Besides the lattice, however, Fig. 7.57 includes scales on the ϕ - and z -axes. How did we get these?

The ϕ -axis scale was obtained in the following way. Figure 7.56 shows lattice points with $n = 1, 2, \dots$; that is, their H.C.F. is 1, which is the order of the rotation axis. So the helix has no rotational symmetry, and there can be only one lattice point with a given z -coordinate. Therefore, in Fig. 7.57, the distance between the two closest lattice points on the ϕ -axis is marked as 360° .

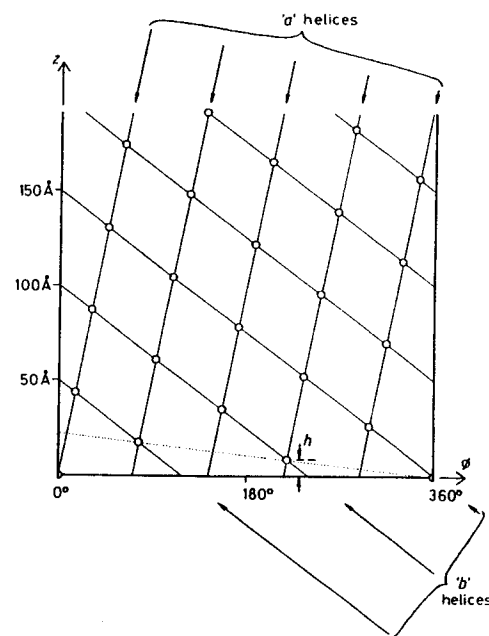


Fig. 7.57 Helical lattice corresponding to the (n, Z) plot of Fig. 7.56. The 'a' and 'b' helices correspond to the basic vectors \mathbf{a} , \mathbf{b} .

The z -axis scale would usually be based on the rise distance h . Figure 7.56 shows how h was found from the (n, Z) plot. In Fig. 7.57, h is the distance of the lowest lattice point above the ϕ -axis (i.e. above the line where $z = 0$; it is marked in Fig. 7.57). However, the particular (n, Z) plot in Fig. 7.56 allows an even simpler method to be used. The vector \mathbf{b} in that figure has $Z = 0.02 \text{ \AA}^{-1}$. This Z -coordinate is the reciprocal of the z -spacing between any two adjacent helices of a set. In Fig. 7.57, therefore, $1/0.02 = 50 \text{ \AA}$ is the z -coordinate at which the first b -helix meets the z -axis above the origin.

Notice how other features of the helical lattice of Fig. 7.57 relate to the (n, Z) plot of Fig. 7.56. Vector \mathbf{b} has n -coordinate -3 , and there are three left-handed b -helices. Vector \mathbf{a} has n -coordinate $+5$, and there are five right-handed a -helices. All the layer-lines of Fig. 7.57 have Z -coordinates that are multiples of 0.005 \AA^{-1} . Therefore the helical lattice will have a repeat at $1/0.005 = 200 \text{ \AA}$. This corresponds to the fact that all the layer-lines in Fig. 7.54 have Z -coordinates at multiples of 0.005 \AA^{-1} .

The hand of the helix has not yet been determined; Fig. 7.57 might represent the inside, rather than the outside. In determining the hand (to which we turn in Section 7.3.4(h)), it is useful to have a model of the helix, represented by Fig. 7.57 folded into a cylinder. This allows us to predict the appearance of, for example, shadowed micrographs for each of the two possible hands, and thus to interpret experimental results immediately.

(f) Some examples of helix structure determination

The previous sections showed most of the general features of a helical lattice determination. However, every particle poses its own special problems. To gain a wider appreciation of these, we now summarize the main features of three structure determinations.

Extended T-even bacteriophage sheath. The optical diffraction pattern of the (negatively-stained) extended sheath was first analysed by Krimm and Anderson (1967). Figure 7.58b shows the diffraction pattern from an extended T4 sheath (Fig. 7.58a) embedded in ice (Lepault and Leonard, 1985). The analysis of this pattern is much facilitated by the clear meridional spots (at the top or bottom) that derive from the annuli. Their n -coordinate is clearly zero, and their Z -coordinate can be measured; so we have established one of the two necessary vectors for the (n, Z) lattice. This is shown on the Z -axis of Fig. 7.58c.

The next clearest layer-line lies close to the equator. Its Z -coordinate can also be measured, but it is not easy to establish its n -coordinate. The problem arises from the fact that the corresponding helices are relatively deep, so that there are contributions from a range of different radii. Instead of calculations based on peak positions in the

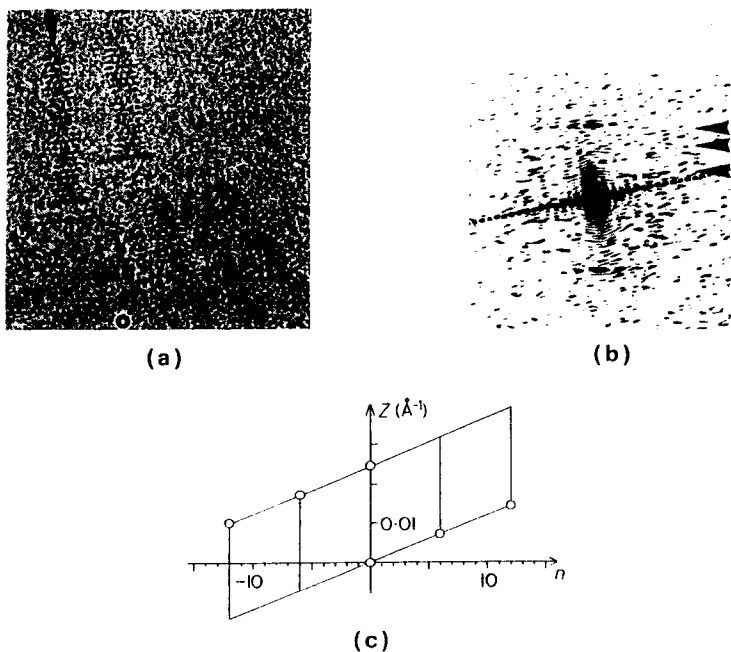


Fig. 7.58 (a) Frozen-hydrated tails of bacteriophage T4 with extended sheaths (Lepault and Leonard, 1985). (b) Optical diffraction patterns from tails like those in (a) (Lepault and Leonard, 1985). (c) The (n, Z) plot of the extended sheath, based on negatively-stained specimens (Krimm and Anderson, 1967), but the same as that found in frozen-hydrated specimens.

diffraction pattern, it is easier to count the helices in micrographs, where it is clear that there are either four or six. Various pieces of evidence show that there are six, so we now have the two vectors defining the (n, Z) -lattice (Fig. 7.58c). Note that our two defining vectors have $n = 0$ or 6, and both are multiples of 6. This must consequently be true for all points on the (n, Z) lattice. Thus the extended sheath has a six-fold axis (Fig. 7.28).

T-even bacteriophage polysheath. Polysheath (Fig. 7.59a) is an aberrant structure with apparently the same helical lattice as the contracted sheath. Presumably because of its length and straightness, polysheath gives beautifully clear optical diffraction patterns (Fig. 7.59b). However, the meridional layer-line is missing; sheath contraction reduces

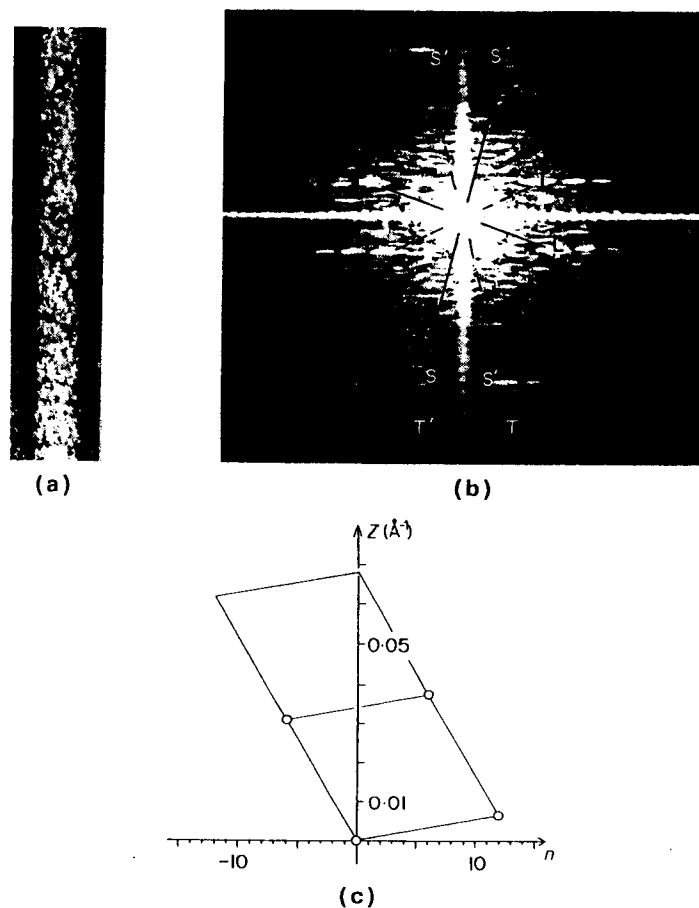


Fig. 7.59 (a) Negatively-stained polysheath of bacteriophage T4 (Moody, 1967a). (b) Optical diffraction pattern from (a) (Moody, 1967a). (c) (n, Z) plot for polysheath (virtually the same as for contracted sheath).

its spacing to beyond the resolution of the micrograph. Analysis of the optical diffraction pattern (Moody, 1967a, 1971) gives three points in the (n, Z) lattice (Fig. 7.59c). As always, their Z -coordinates can be measured, but their n -coordinates pose a greater problem. The diffraction pattern merely shows the range of n , and establishes that the n -coordinates of point S and T are equal. From this, it follows that $n(S) = n(T) = n(L)/2$. Thus the (n, Z) plot can be completed if one set of helices can be counted. Fortunately, this can be done in contracted sheath. These particles often attach to the grid by one end, allowing us to count the set of helices that give rise to point L (Moody, 1967a). There are 12 helices, i.e. $n(L) = 12$, so $n(S) = 6$. Thus all points in the (n, Z) lattice have n 's which are multiples of 6, as in the extended sheath; rotational symmetry is conserved during contraction process.

Microtubules. Electron micrographs of microtubules (Fig. 7.60a) show prominent axial striations (protofilaments). Actually, these are often not quite parallel to the helix axis, but form right-handed helices of very long pitch ("supertwist": Mandelkow and Mandelkow, 1985). However, analysis of the microtubule lattice is easier if we ignore

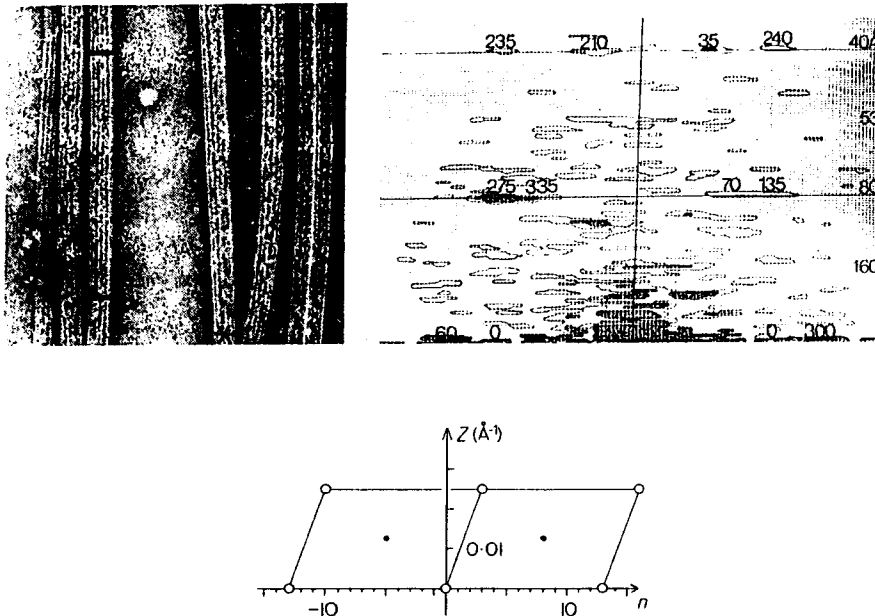


Fig. 7.60 (a) Microtubules from a *Trichonympha* flagellum (Grimstone and Klug, 1966). (b) Computed F.T. of (a) (Amos and Klug, 1974). (c) (n, Z) plot of the microtubule helix. The lattice represented by the open circles is that of the tubulin subunits, ignoring the difference between α - and β -tubulin. The dots indicate the superlattice found in flagellar A-microtubules, reflecting the difference between α - and β -tubulin. (Figures (a) and (b) are reprinted from Amos and Klug (1974), with permission from the Company of Biologists, Ltd.)

this minor complication, and take the protofilaments as exactly parallel to the helix axis. Then the subunits along a protofilament are related by a short translation parallel to the helix axis, i.e. there is a short repeat, which can cause interference between layer-lines.

The structure of microtubules (singlet A-tubules from cilia) was analysed by Amos and Klug (1974). (See also Amos's (1979), (1982) reviews.) The computed diffraction pattern (Fig. 7.60b) shows reflections on the equator and on two layer-lines, at $1/80 \text{ \AA}$ and $1/40 \text{ \AA}$. However, the layer-lines at $1/80 \text{ \AA}$ arise through the difference between α -tubulin and β -tubulin, which shows up to varying extents in different specimens. The equatorial spots relate to the protofilaments, whose number is the n -coordinate of the spots. The peak positions indicated $n = 11-14$, but the phases differ by about 180° , restricting the alternatives to 11 or 13. Direct counting of the protofilaments showed the number to be 13. This gives one vector in the (n, Z) plot: see the circles at the left and right ends of the n -axis in Fig. 7.60c. From the other spots' peak positions and phases (implying parities), only one lattice was possible (Fig. 7.60c), though the determination of hand proved surprisingly troublesome.

A consequence of this (n, Z) lattice is that there is interference between two sets of peaks (with $n = 8$ and $n = -5$) on the $1/80 \text{ \AA}$ layer-line (filled circles in Fig. 7.60c). Since these are of opposite parity, they reinforce on one side and interfere on the other. (This sort of effect is shown more graphically in a view of a layer-plane: see Fig. 7.61.) The interference is less striking with microtubules (where the orders differ by 3) than with bacterial flagella (where they are -5 and 6 ; Finch and Klug, 1972). In either case,

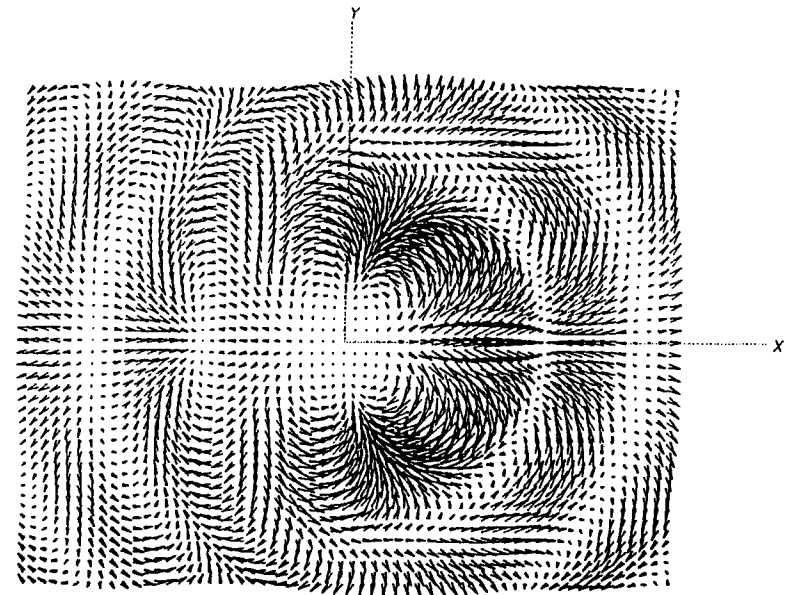


Fig. 7.61 Two superposed (interfering) layer-planes, one with $n = 3$ and the other with $n = 4$. Since these numbers have opposite parities, interference is constructive on one side ($+X$), and destructive on the other ($-X$).

the effect is to reinforce strongly the spot on one side of the Z -axis, at the expense of the other side. This has the effect of generating seemingly "one-sided" images, with a set of oblique striations resembling the truly one-sided images in shadowed preparations. However, the striations in microtubules or bacterial flagella are merely superposition effects, and tilt sometimes one way, sometimes the other. It may not be a coincidence that this effect is seen in helical structures used for swimming. The need for mechanical strength may be the reason, in each case, for the subunits forming long chains nearly parallel to the helix axis (e.g. the protofilaments of microtubules). As we have seen, this causes interference between layer-lines, so that two peaks of similar order and opposite parity will be sure to generate the "one-sided" effect.

(g) *Checking that the (n, Z) lattice is not a superlattice*

It is possible that the particle's helical lattice may actually be larger than the lattice found by the procedure just outlined. This could be because the subunits belong to two classes with slightly different structures (like the α - and β -tubulin monomers), or because other macromolecules attach to the helix, but with several helical subunits per macromolecule. In either case, the larger lattice would have to include several lattice points of the smaller lattice in a regular fashion. Consequently, there would have to be additional lattice points in the (n, Z) plot. When that plot was originally constructed, however, this was not noticed: the corresponding layer-lines were invisible through being lost in the noise. But they might be revealed by averaging the transforms of enough particles.

For example, the particles might form parallel aggregates. Although unsuitable for many purposes (because of sampling of the transform by the aggregate's lattice), these often give strong transforms. They would be likely to reveal any faint spots indicating that the true (n, Z) lattice is smaller than the one we have determined. If this were the case, exact determination of the new lattice would be facilitated by the stringent condition that it must include all the points of the old one. (The layer-lines at $\pm 1/80 \text{ \AA}$ in Fig. 7.60(b), (c), being of variable intensity, can be regarded as a superlattice of the main tubulin lattice, shown as open circles.)

(h) *Hand of the helical lattice*

We must now consider the hand or chirality of the helical lattice. This is connected with our (previously arbitrary) choice of the sign of n associated with a particular sign of Z . Finding chirality in the microscope is quite straightforward in principle. One must remember that most optical systems (electron microscopes, photographic enlargers, or slide- (not overhead-) projectors), having no mirror, cannot yield a mirror image of the object. The image will have the same chirality as a shadow of the object projected by the light (or electron) source. However, care must be taken to note the surface facing illumination, both in the case of the grid in the microscope, and of films or plates in the enlarger. If the emulsion surface always faces the light source, the chirality is unchanged from that of the electron microscope image. Numerical densitometry and display might introduce problems, however, depending on the programs.

There are several reliable electron microscopic methods for determining helix chirality. Perhaps the simplest uses a metal-shadowed preparation of the particle. If this shows

any recognizable helices (and perhaps optical diffraction might be needed to reveal them), it reveals their direction on the top of the particle, and hence their hand. Then we use the model (Section 7.3.4(e)) of the helical structure to identify these helices.

Another reliable method tilts the grid to give a series of micrographs. In principle, the tilt or rotation can be about any axis in the plane of the grid, and still give the information necessary to decide the hand of the helix. In practice, the best direction of this axis depends on the pitch of the most prominent set of helical grooves. If these have a short pitch (as with tobacco mosaic virus), the biggest change in the particle's appearance is produced when the tilt axis is perpendicular to the helix axis. Then a helix develops cusps on one side of the particle while, on the other side, the turns become even smoother. If the helices are very clear, a single image (with all orientations and angles carefully noted) could suffice to establish the hand.

If the helices are not very clear, image-processing may be needed to reveal which side has the cusps. Finch (1972a) used separate optical diffraction of the left and right halves of the image of the tilted particle. When the tilt has made the cusps sharpest, the side of the particle containing them gives the appearance of stacks of annuli. Therefore its diffraction pattern shows meridional spots, in contrast to that of the other side; this distinction can reveal which side has the cusps. (Finch also applied a quantitative technique of DeRosier and Moore, 1970; see Section 7.4.5(b).)

On the other hand, if the most prominent set of helices have a very long pitch, then the clearest change in appearance would be produced by tilting the particle about its long axis. This rotates the helices, and the resulting images are best analysed by reference to their calculated F.T.s; see Finch (1972b).

Less reliable indications of the hand of a helix might be extracted from the tendency of the exposed end of a lightly-stained particle to contract relative to the embedded end ("anisotropic contraction"; Moody, 1967a). However, ordinary unequal staining seems to be of no help, since there is no invariable rule as to which surface will be stained more strongly. Finally, great caution should be used when interpreting the apparently one-sided images given by particles with a very short repeat (Section 7.3.4(f)).

(i) *Polarity of the helix*

Suppose we have now determined the screw and the rotation axis of the helical particle. Its symmetry determination will be complete when we have answered one remaining question: is it polar (i.e. is its point-group dihedral) or not?

Sometimes it is obvious that the particle is polar, as (for example) from the strongly-directional arrow-head shapes of "decorated" actin (Huxley, 1963). Most often, however, the particle image looks non-polar, and—if it is really polar—this is apparent only by functional tests such as whether subunits polymerize equally onto both ends (non-polar) or preferentially onto one (polar). But it would be much easier if polarity could be detected by image processing.

If the helix is non-polar, it must have two-fold axes perpendicular to the particle. These axes will give the F.T. a two-fold symmetry, but no change will be apparent in the amplitudes, which already have two-fold symmetry as a consequence of the F.T.'s Friedel symmetry. Optical diffraction, which gives only amplitudes, is therefore useless for testing polarity. The phases, which are not necessarily two-fold symmetrical, provide a much better test; we obtain them from numerical F.T.s. It would seem possible to test

the particle's F.T. phases directly for the presence of two-fold axes. Unfortunately, however, it will often be impossible to decide whether the deviations from two-fold symmetry are caused by non-polarity or by noise. It is better to compare different particles. This was done by Amos and Klug (1975), whose approach is outlined in Section 7.4.5(c).

(j) *Deviations from strict helical symmetry*

The methods described above assume that the particle has strict helical symmetry. In practice, particles show various distortions, even before the artifacts of specimen preparation have occurred. Probably the most common is bending. This is often caused by particle adhesion and staining, but several helical structures have a built-in curvature. The classical case is the bacterial flagellum (for a recent study of which, see Trachtenberg and DeRosier, 1987).

However, even if the helix axis is straight, other distortions are possible. The twist angle is fixed by interaction of each subunit with other pairs of subunits. If these interactions are not sufficiently strong, the twist angle can drift, as has been found for actin filaments (Egelman *et al.*, 1983) and microtubules (Mandelkow and Mandelkow, 1985). In extreme cases, the helical parameters may change progressively from one end to the other (Moody, 1973). It cannot even be assumed that the helical lattice must be completely consistent, since a discontinuity or seam may exist (as in brain microtubules: Mandelkow and Mandelkow, 1985).

Most sensitive to distortion are the outer parts of helices. These may have been attached with a different symmetry, as in flagellar microtubules (see Warner, 1970). But even if the basic symmetry is that of the underlying helix, interactions with the surroundings may disturb it (e.g. Crowther *et al.*, 1985). Perturbations are possible even without such interactions. The classic case, the Dahlemense strain of tobacco mosaic virus, was found by X-ray diffraction (Caspar and Holmes, 1969). But similar perturbations in bacterial flagella are clearly visible with negative staining (Trachtenberg *et al.*, 1987).

The correction of helical distortions will be considered in Section 7.4.5(d).

7.4 IMAGE ANALYSIS BY NUMERICAL OPTIMIZATION

7.4.1 Introduction

(a) *Development of techniques for matching particle images*

As computers have improved, their applications to image analysis have extended beyond the simple calculation of functions (like the F.T.) that are analysed by the microscopist to deduce particle symmetries. This fails if the image cannot be made interpretable by periodic *translational* averaging (which is effectively what is done when the F.T. is calculated, and analysis focuses on reciprocal-lattice peaks or layer-lines). When this fails, we can pursue one of two general approaches. First, we can try to improve the signal-to-noise ratio by *rotational* averaging. However, particles with the wrong orientation, or several different rotation axes, can give rather confusing projections. In

7. Image Analysis of Electron Micrographs

that case we can adopt the second general approach: to invert our usual procedure. Instead of starting with the image, we can start with a possible symmetry of the model. We use it to predict significant features of the image; the symmetry that gives the best prediction is judged to be correct. The earliest applications of this approach followed after the quasi-equivalence theory of viral capsids (Caspar and Klug, 1962). To test that theory, it became necessary to predict the rather complex appearances of icosahedral viral capsids with large numbers of subunits. First shadowgraphs (Finch and Klug, 1965), and later computer graphics (Finch and Klug, 1967), were used to demonstrate a general agreement between the symmetry model and the images. But this model-building technique was limited by the subjective method used to judge the agreement. To make it objective, the predicted and experimental images should be compared by quantitative statistical tests.

Although that approach has not yet been fully implemented, recent progress has been made mostly in this general direction. Testing a symmetry prediction involves matching the predicted particle image (or some feature of this, or its F.T.) with actual micrographs. The matching is done numerically, either by maximizing a function which is big for good matches (e.g. the correlation function), or by minimizing a function which is big for bad matches (i.e. an "error-function"). Numerical optimization techniques are therefore used. These are also needed if the first approach (i.e. improving the signal-to-noise ratio by averaging images) is followed. For, to average the images, we must first determine their relative positions and orientations. These are found by matching the particle images or their F.T.s—again, using numerical optimization techniques.

(b) *Organization of this section*

All calculations require some essential numerical pre-processing, which is described in Section 7.4.2. Then attention turns (Section 7.4.3) to the question of comparing particle images, and considers general numerical methods for finding their relative positions and orientations. After this, we descend from the general to the particular. The simplest case (i.e. that requiring the least positional information) involves the averaging of a lattice which has been distorted. Conventional F.T. methods must be supplemented, or replaced, by the newer correlation methods (Section 7.4.4).

More positional information is needed to average images when the particles' positions are uncorrelated. Leaving aside the rather elementary problem of averaging different copies of the same two-dimensional lattice, the simplest of these cases involves the averaging of different images of helical particles (where there is one-dimensional translational symmetry). After considering these (Section 7.4.5), we look at the more difficult problem of comparing particles whose symmetry lacks any translation: they have only rotational symmetry. The simplest of these cases yield images with rotational symmetry. Here there are two problems: to find their centres, and to determine their rotational symmetries. These matters are considered in Section 7.4.6. A more difficult situation arises when a rotationally symmetric particle has several different rotation axes, of which an image can reveal no more than one. Then different projections can show different rotational symmetries, and we have the problem of combining these to give the particle's point-group (Section 7.4.7(a)). Even if the image lacks all rotational symmetry, it will preserve the imprint of the point-group. Specialized methods have been developed for revealing this (Sections 7.4.7(b) and (c)).

7.4.2 Numerical densitometry and processing

Densitometry is necessary for any numerical method of image analysis. Since it most commonly precedes one of the Fourier numerical methods (of image analysis, filtering or three-dimensional reconstruction), we shall describe densitometry from the Fourier viewpoint.

(a) Input

Densitometers. All densitometers scan light across a selected region, measure its intensity and convert it into a number, which is then sent to a computer (or direct to tape). Some densitometers accomplish the scanning movement with a rotating drum, around which the film must be bent. Others use a moveable flat stage, so that glass plates may be densitometered directly. If photographic copying of plates is to be avoided, another relevant factor is the size of the scanning raster (see below). Accuracy is important in measuring both the intensities and (even more) the positions. (Distortion of the scanning raster can cause complex artifacts in the Fourier transform, for example.)

Various satisfactory commercial microdensitometers exist; suitable ones are used by protein crystallographers, and crystallographic densitometry services would also be useful for electron microscopists. Microdensitometers with a particularly fine scanning raster are used by astronomers.

Densitometry. The raster size (which is partly fixed by the microdensitometer, but also partly adjustable) affects the sampling of the picture. Two factors set an upper limit to the permissible raster size. The first is the Wooster effect: the average transmittance of a large area cannot be converted into an average optical density.

The second factor is the need to preserve all the information in the image. The optical density array is like a half-tone picture, and too coarse a sampling array degrades its quality. This degradation can be expressed more quantitatively. The one-dimensional case is shown in Fig. 7.62. The sampled picture is equivalent to a continuous picture multiplied by a lattice of points (the raster). It follows that the sampled picture's transform equals the desired transform, convoluted with the transform of the raster lattice. This

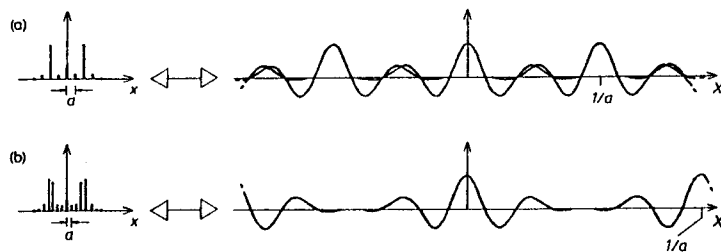


Fig. 7.62 (a) A one-dimensional density function (the same as in Fig. 7.26) is sampled coarsely. The F.T. (on the right) consists of many copies of the F.T. of the density function. These copies are placed so close together (on the reciprocal of the sampling lattice), that they overlap and interfere. (b) The density-function is sampled on a finer grid, so that the F.T. (right) consists of well-separated copies of the F.T. of the density-function itself.

convolution gives many copies of the continuous picture's transform, each copy centred on a point of the reciprocal lattice of the raster. If the picture's transform is too spread out, then overlap between the different picture transforms will introduce artifacts ("aliasing"). To avoid this overlap, the raster size should be smaller than the resolution of the picture. This resolution is most easily found from an optical diffraction pattern of the image, where it corresponds to the outermost diffraction spots (from a lattice) or layer-lines (from a helix). If the specimen lacks the translational order that would give such spots, then it is necessary to assume the best resolution found with similar specimens. In practice, a spacing corresponding to 6–8 Å is common for negatively-stained or frozen-hydrated specimens photographed at moderate electron doses. (A densitometer raster of 25 μm, and a micrograph magnification of 40 000, would be typical.)

Densitometry yields a computer file containing the array of optical densities measured at each pixel (element of picture area). The size of this file is of course the area scanned, divided by the area of the unit cell of the raster. Fourier transformation (see below) uses arrays in which the number of pixels in each dimension is composed only of very small factors. Usually it is a power of 2: 256 × 256 or 512 × 512 were typical in the early days of computer processing, though modern computers allow the unit cell length to be increased by an order of magnitude. Storage of such arrays is a less serious limitation than processing time; perhaps the complete displacement of analogue by numerical techniques must await the common use of highly-parallel computers (and densitometers). Meanwhile, the area to be densitometered should be chosen carefully, usually through the quality of its optical diffraction pattern.

(b) Preliminary calculations

Numerical pre-processing. The first stage of processing is to attempt to make each density number a linear function of the number of electrons reaching that pixel of the film. The optical density can be transformed by a suitable function, making use of the calibration of the electron microscope plate.

Next we shall need to examine the densitometered data and select the image of a particle for image processing. Failing a convenient electronic display, this has been done using line-printer output, with characters whose blackness (sometimes enhanced by overprinting) very roughly matches the optical density (MacLeod, 1970). Whatever display is used, the particle's boundaries must be determined and communicated to the program. A polygonal (usually four-sided) boundary (the "box") is drawn to contain the particle's segment of data. The program copies this, omitting (or setting to zero) all numbers that lie outside the box ("boxing": DeRosier and Moore, 1970).

"Boxing" can lead to artifacts when the Fourier transform is calculated. Viewed from a distance where the individual points are invisible, our boxed pixel array will resemble a parallelogram with sharp edges. The inner part of its transform will thus be dominated by the long "spikes" seen in the transform of a parallelogram (Fig. 7.23). These would be bad enough in themselves, but "aliasing" (see above) causes them to spread into neighbouring copies of the transform. They must be prevented by reducing the visibility of the edge of the pixel array ("floating"): from each of the optical densities within the box, we subtract the average optical density at its perimeter. This process is equivalent to surrounding the box with a uniform density (like uniform stain around the particle), and then subtracting that uniform density from both the particle and its surroundings.

So the particle, instead of having a weak positive density surrounded by more positive (denser) surroundings, now has negative density surrounded by virtually nothing.

The density array is now ready for Fourier or correlation analysis.

Fourier transformation. For many purposes, we need to calculate the pixel array's Fourier transform. This is an integral which must be computed as a sum, called the Discrete Fourier Transform (D.F.T.). Consider what determines the length of this calculation. To calculate any point in the transform, every pixel must be multiplied by both cosine and sine functions, specific for that pixel, and the terms for cosine and sine are summed separately to give the real and imaginary parts of the transform. (The real part is the horizontal component of the vector in Fig. 7.3, and the imaginary part is its vertical component.) The larger the picture, the more detailed is its transform; so the number of transform points we need (the output) should roughly equal the number of pixels in the picture (the input). Thus the total number of calculations is proportional to the square of the number of pixels. This means that a naïve method of calculation, sufficing for small arrays, can take prohibitively long with pictures that are not apparently very much larger. Thus, if a 64×64 pixel array took 1–2 min to process, a 512×512 pixel array (only 64 times bigger) would take days. Without a faster method of calculation, image processing could never have become widely used.

The faster method, the "Fast Fourier Transform" (F.F.T.), makes use of the representation of trigonometric functions as complex exponentials. Their simple properties allow the transform to be built up recursively from smaller transforms (see Brigham, 1974; or Elliott and Rao, 1982, for details). The method requires the number of pixels (N) to be composed of small factors; in the best case, it is simply a power of 2. In that case, the calculation time no longer increases as N^2 , but as $N \log_2(N)$, which increases only slightly more rapidly than N . To take the example in the previous paragraph, the 64×64 pixel array would (using the same computer) take less than a second with the F.F.T., and the 512×512 array would take less than a minute. Because of its importance and speed, the F.F.T. is generally considered the central workhorse of a numerical image-processing system. However, it imposes a restriction on the calculation. The number of input and output densities must usually be a power of 2. (The additional densities beyond the boxed array are set to zero: "zero-filling.") Being restricted to powers of 2, the (previously roughly equal) numbers of input and output points must be exactly equal. This implies that the reciprocal space coordinates represent a regular sampling of a single repeat of the computed transform.

The transform has Friedel symmetry, so only half of it need be calculated. Display of the transform (using the device that displays the picture) will usually show amplitudes, the phases being displayed only for selected spots or layer-lines. Interesting transform peaks will not usually fall exactly on points on the transform raster, so interpolation is needed to find the peak values.

The discrete transform (D.F.T.) we have calculated differs from the integral considered in Section 7.2.3(i). As pointed out earlier, it is periodic, so that it has "wrap-around" in both dimensions, like the screen of a computer game (i.e. it is mapped onto a torus). Aliasing artifacts occur if this "wraps" distant parts of the transform onto the central parts in which we are interested. Though very difficult to avoid this entirely, it should be minimized, and its effects not ignored. The D.F.T. has an exact inverse, which is also toroidal. It is this inverse, rather than the original picture, of which the D.F.T. gives us the transform.

7.4.3 Finding the orientation of particles: the general problem

(a) Comparing pairs of particles

The D.F.T. is usually the starting point for comparisons of the images of different particles, or of different parts of an image with itself. The details of calculating each comparison will depend on the particular symmetry being investigated (see Sections 7.4.4–7.4.7). First we examine the general question of how such comparisons can be used.

In general, images from several particles need to be combined together, and therefore compared together. However, it is simplest to begin with the comparison of pairs of particles, so that the essential geometrical problem can be viewed in isolation from the question of how to combine several images. In any case, that combination usually proceeds in practice through many pairwise combinations.

The transformation parameters. We consider first how many parameters need to be determined in order to specify the relative position and orientation of two particles. (For brevity, we shall refer to these as "transformation parameters".) Note that this is essentially the same as specifying the orientation and position of one particle, relative to an established coordinate system. For that system could be set up in particle 1, and we should then search for the transformation parameters needed to move this system into the corresponding parts of particle 2.

Frequently the particle has symmetry elements that define a standard position. For example, the axis of a helical particle would be made to coincide with the z -axis (and, if it is non-polar, a two-fold axis would coincide with the y -axis). Particles with point-group symmetry would have similar "standard" orientations to fit them to the symmetry of the (unsigned) coordinate axes. (For an irregular particle, however, the standard position is necessarily arbitrary.)

Any particle can be brought to the standard position by a combination of two kinds of movement. The first is a translation, specified by the three numbers (Δx , Δy , Δz), that brings the particle "centre" to the origin. The second kind of movement is a rotation that brings the particle to the standard orientation. Three angles are needed to specify this rotation. These could be successive rotations about different coordinate axes (Euler's angles; for which various conventions exist: Goldstein, 1980). Alternatively, we can use the fact that any rotation is equivalent to turning the particle, by an appropriate angle χ , about some axis. Specifying the orientation of that axis requires two more angles, making three angles altogether. Although the two angles specifying the rotation axis could be (θ, ϕ) , as in conventional polar coordinates, a more convenient set is shown in Fig. 7.63. They have the advantage that all three angles become zero when the particle is perfectly aligned.

These six transformation parameters specify how to bring an arbitrary particle into the standard position and orientation. The inverse transformation will move the "standard" particle so that it corresponds to some observed particle in the micrograph. The particle's position along the projection direction is probably irrelevant, so this parameter (say, Δy) can be omitted, leaving five transformation parameters (Δx , Δz , ψ , ω , χ) to be found. They specify the movement of the standard particle, so that its projection coincides with the observed image.

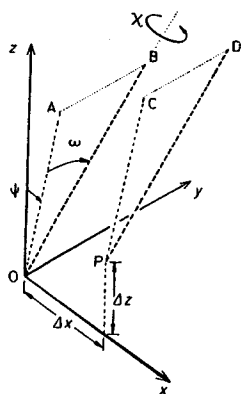


Fig. 7.63 Definition of the misorientation parameters Δx , Δz , ψ , ω and χ . The xz -plane is the plane of the grid, and PC is the projection (or image) of the particle axis PD, where P is the particle centre. After correcting for the displacements Δx , Δz of this centre from the origin, we have the translated particle axis OB and its projection OA. ψ defines the orientation of the particle axis relative to the grid plane, and χ is used to measure the relative orientations of different particles about their axes.

These five parameters can be further reduced if the particle has a preferential orientation. This usually originates in some internal symmetry. For example, the particle is likely to be very extended in the direction of any internal translation symmetry, and this direction is therefore likely to lie parallel to the surface of the grid. This is quite accurately true for two-dimensional sheets, and nearly true for helical particles. A rotation axis of fairly high order (perhaps 6 or more) can produce a ring-shaped structure with the same preferential attachment to the grid. Thus, for helices (Section 7.4.5), the particle axis (PD of Fig. 7.63) would lie (approximately) in the xz -plane, i.e. $\omega \doteq 0$. For sheets or ring-shaped structures, however, PD would lie approximately along y , i.e. $\omega \doteq 90^\circ$, so that ψ and χ measure virtually the same movement.

A different kind of reduction in the parameter space occurs when the particles form part of a lattice with disorder (Section 7.4.4). First, the disorder is usually confined to certain parameters. A distorted lattice commonly involves translational changes, but only very small rotations (and these may be exclusively about an axis perpendicular to the grid). Second, the disordering movements are often correlated, and therefore very similar for adjacent particles; this effectively reduces the information required to specify them.

Finding the transformation parameters: the general problem. The problem of finding the transformation parameters is conceptually simple when all the particles lie flat on the grid. Their projected images are identical, apart from orientation and position. One image can be chosen as the standard, and we need to find the transformation parameters (Δx , Δz , ψ) that will bring it into coincidence with any other image. To do this, a criterion of image matching is first set up in the form of an error-function. The data from the two particle images are processed in a way that depends on the assumed values of the transformation parameters, and the result (the error) measures the mismatch between

the transformed images. The value of this error will be a function of the three transformation parameters, which have to be adjusted until the error is minimized.

The problem is a little more complicated when the particles can assume any position on the grid. The standard particle image, defined by the orientation of the symmetry axes, must now be rotated in three dimensions to give the observed projection image. Again we need a criterion, in the form of an error-function, to discover how well the predicted and observed images match. This is often possible, even in the cases where our only information about the standard particle is its symmetry (Section 7.4.7(b)). In such a case, however, the error-function will depend on the assumed particle symmetry. Given a satisfactory error-function, we minimize it to find the transformation parameters. But now we can repeat the process, using a different assumed symmetry in the error-function. When the correct symmetry is used, the fit will be best. In this way, optimization of the particle match can provide an objective test of particle symmetry.

In either case, however, we have two different problems: defining (and calculating) the error-function; and adjusting the transformation parameters to minimize it. We now consider separately these two aspects of optimizing the particle position.

The error-function. All sensible error-functions reach a minimum when the transformation parameters are correctly chosen. However, that does not mean that there can be no significant differences between error-functions. They should satisfy two principal criteria. First, around the minimum, dependence on the transformation parameters should be as little as possible affected by image noise. Second, an error-function should be rapidly computed. Here the critical factor is not only the time taken for a given computation, but also how that time depends on the number of data values that enter the calculation. (For it is only too easy to devise algorithms that work well enough with a few values—image pixel densities—but become hopelessly slow when 10 or 100 times as many are used.)

Many different types of error-function have been used, but almost all can be classified into one of two groups: Fourier error-functions, and the correlation function (Equation (48)). (The latter must be maximized, so it behaves like the negative of an error-function.) Fourier methods use a variety of different error-functions, often called "residuals" (Amos, 1975). If the two transforms F_1 and F_2 are to be compared, one possibility is to use a difference function like $\Sigma|F_1 - F_2|$ or $\Sigma|F_1 - F_2|^2$. However, the correction for transformation parameters may mostly affect the phase, in which case a better error-function would be

$$\{\Sigma(\text{amplitude})|\text{phase difference}|\}/\{\Sigma(\text{amplitude})\}$$

called a "phase residual". Variants of this could be devised, e.g. where the phase difference is squared.

As indicated in Section 7.2.4(c), there are close connections between F.T.s and correlation functions. Whereas F.T.s can be calculated by the efficient method of the F.F.T., there is no rapid direct route for finding correlation functions. Consequently, many correlation function calculations proceed via a double F.T., especially when there are many data points.

Minimizing the error-function. When minimizing any error-function, it is a great advantage if that function is linear in the unknown parameters. (This is the case, for

example, with conventional least-squares optimization, and contributes much to its utility.) Such linearity ensures that there is only one minimum, and provides a direct means of finding it. Unfortunately, we are often forced to try to minimize error-functions that are not linear in the parameters. Such functions usually have several minima. Since the minima can fill the space of the non-linear parameters, their number can often increase exponentially with the number of these parameters. (When finding transformation parameters, this is still not prohibitive, though the computation time is not negligible.)

Surveying all the minima in parameter space is only the first part of a non-linear optimization. Having found the approximate position of the deepest minimum, we must next find its exact position. Indeed, in many cases this is the only real problem. For the rough position of the global minimum can often be estimated by matching the two pictures visually, or by matching by computer at progressively increasing resolutions. Fortunately, there exist several established techniques for finding "local" minima (Jacobs, 1977; Fletcher, 1980; Press *et al.*, 1987).

However, methods for speeding the global search are also useful. Unfortunately, it seems impossible to introduce any of the transformation parameters in a linear fashion (unless the particle is almost perfectly positioned), so we have up to five non-linear parameters to find. The problem of multiple minima would be simpler if, instead of having to search five-dimensional space, we could start by searching (say) a three-dimensional subspace in which the minimum fixes three of the parameters. Keeping these at their optimum values, we would then search the remaining two-dimensional space to optimize the remaining parameters. Such a decomposition is possible with rotational correlation functions (Section 7.4.6(f)).

This decomposition of the multidimensional search should be distinguished from the practice of finding the optimum for each parameter in turn. Here the adjustment of any parameter alters the optimum for all the others, so there is no reduction in the number of minima that must be checked to find the lowest. Indeed, this method is not even a good way to find the nearest ("local") minimum: the search moves through parameter space along a zig-zag path that can be very much longer than that necessary with better methods (Fletcher, 1980).

(b) *Comparing many particles*

Hitherto we have been concerned only with matching pairs of particles. However, we need an overall average from many particles. If there are s particles, $(s - 1)$ sets of transformation parameters are needed. So should we optimize these $(s - 1)$ sets using the micrograph data? Some overall error-function might be set up, to which each particle picture would contribute, with a weight dependent on its quality. The function would depend on all the parameters, and a search would be made through parameter space to find the global minimum.

Unfortunately, the very high dimensionality of that space would make this approach prohibitively slow. Instead, the problem is solved by the pairwise matching of particles. Each pairwise match will give a slightly different set of optimum transformation parameters, and the overall optimum could be calculated from all the matches. But even this is too large a task to perform thoroughly, since s particles have $s(s - 1)/2$ pairwise matches. So an even shorter cut is taken. The "best" two particles are matched, and their average is calculated. Then the "best" remaining particle is matched with this

average, giving a new average. The overall average "snowballs" in this way until it has incorporated the last particle.

This method is quick, and it seems to work quite well in practice. But, while it may be satisfactory when there is no doubt about the approximate match position of each particle, serious uncertainty on this point could render the method dangerously subjective. Consider the following case. From a random set of shapes, two "faces" could be chosen and averaged. Proceeding in this way, a meaningless "average face" could be produced, entirely by "objective" computer methods. This difficulty has been overcome, in the case of helix polarity tests, by using a scatter diagram (Fig. 7.65 and Section 7.4.5(c)). A similar statistical problem arises whenever the essential correctness of any match is open to doubt. Even when there is no such doubt, the "snowballing" method places undue reliance on the first particles, which can bias the final average. It might be instructive to repeat it, taking the particles in reverse order.

Clearly the accuracy of each pairwise match depends on the signal-to-noise ratio of the particles, which is increased if those particles are themselves symmetrical. Unfortunately, the molecular structure is preserved only at extremely low electron doses (Chapter 8), which greatly reduces the signal-to-noise ratio. Under these conditions, the particles or aggregates must be highly symmetrical to give sufficiently accurate data for pairwise matching. However, it is often difficult to persuade interesting cellular structures to form such aggregates. The question therefore arises as to whether, by some technique, the transformation parameters could be extracted from the extremely noisy low-dose images of particles with low (or no) symmetry. This is the subject of much current investigation (see especially the journal *Ultramicroscopy*), but it is not possible to pursue it further here.

7.4.4 Refining the parameters of distorted lattices

(a) *Classification of distortions*

The simplest type of distortion (*uniform distortion*) yields an exact lattice, but with changed lattice parameters. Usually this is detected by comparing the lattices of different samples of the same biological specimen. Sometimes, however, it reveals itself as departures from an underlying symmetry. For example, a pattern that apparently has the plane-group $p6$, and that gives diffraction spots with six-fold symmetry in their intensities, may nevertheless have a lattice that is not exactly hexagonal. This situation is often found when the lattice is curved to form part of a cylinder (e.g. bacteriophage "polyheads" or the cell walls of cylindrical bacilli).

The more common type of distortion is non-uniform, i.e. a spatial variation in the lattice parameters in a single sheet. This variation can be *random* (perhaps because the positions of the subunits are poorly fixed by the lattice forces), or it can vary smoothly. The latter variety, associated with elastic deformation, is particularly common, perhaps because of the strains experienced by specimens during preparation and irradiation. A smoothly varying distortion will cause points that were originally close together to remain so; we shall therefore refer to it as a *continuous distortion*. (Of course, this cannot apply exactly down to atomic dimensions, but it is often an adequate approximation at the resolutions achieved in practice.)

From the point of view of image analysis, all these distortions can be classified by the amount of information needed to specify them. This is least for uniform distortion (we need only the coordinates of two lattice vectors). It is also small for a continuous distortion, if that is negligible over the dimensions of a unit cell. For, though the lattice vectors are then functions of position, their slow variation would allow them to be approximated satisfactorily by a simple function with few arbitrary constants. Random fluctuations in the subunit positions, however, require the greatest detail to describe completely, and therefore make the greatest demands on the meagre information available from each unit cell. Continuous distortion is therefore easier to monitor (and correct) than random distortion. There is another reason for this. Continuous distortion can combine a substantial overall distortion (requiring correction) with a very small change in each unit cell. However, random distortions (if they are substantial enough to require correction) must produce a substantial movement in each unit cell, shifting the unit cell contents within it. This is hard to monitor unless we have independent evidence for the appearance of the contents, in isolation from the surrounding unit cells. Practical methods for monitoring distortion generally need to follow a combination of unit cells and their contents, and such methods are unlikely to prove satisfactory with random distortions.

(b) Methods for determining distortion

Uniform distortion, besides requiring the least information, has the further advantage of preserving exact translational symmetry. Therefore it can be satisfactorily analysed by Fourier methods (Aebi *et al.*, 1973). However, to obtain accurate values of the F.T. peaks from the F.F.T., it is necessary to re-sample (e.g. by bilinear interpolation) the optical density array at a regular sublattice of the crystal.

Non-uniform distortion cannot be handled satisfactorily within the framework of straightforward Fourier analysis. In the first attempt to meet this challenge (Crowther and Sleytr, 1977), the positions of the (distorted) lattice points were found by (computer) filtering with a mask that passed only the innermost diffraction peaks, so that the filtered image consisted of blurred peaks. To avoid averaging over many unit cells, however, the mask "holes" were made quite large. The centres of mass of the filtered image peaks were then calculated. Thus much of the analysis was concerned with the image, rather than with its F.T.

The logical conclusion of this approach was to apply correlation methods (developed in the early 1980s by the groups of Frank and Saxton). If a distorted lattice is to be scanned by a "reference patch", then that patch must be small to accommodate the distortions. However, it must be large enough to contain the information for specifying the unit cell. Its size (a compromise between these two factors) can vary from five unit cells (negatively-stained bacterial cell wall images of high contrast: Saxton and Baumeister, 1982) to 50 unit cells (unstained bacteriorhodopsin crystals of low contrast: Henderson *et al.*, 1986). To keep the "patch" as small as possible, it can first be averaged, over a fairly small area, by (computer) optical filtering.

To make the correlation method practical, the X.C.F. must be calculated rapidly. This is done using the F.F.T. (inversion of Equation (9)). For this purpose, the "reference patch" must be embedded in an area as large as the image. All its surrounding pixels are "floated" (Section 7.4.2(b)), i.e. set to the average of the "patch" boundary. Both this area, and the image, are Fourier-transformed, and one of the F.T.s is converted into

the complex conjugate (as required by Equation (9)), before they are multiplied and inverse-transformed. This yields the correlation coefficient as a function of displacements in two dimensions. Every broad correlation peak outlines a local minimum. From each of these, the exact coordinates of the minimum must be extracted, e.g. by averaging (Saxton and Baumeister, 1982) or by profile fitting (Henderson *et al.*, 1986). If the distortion is continuous, all the minima should lie approximately on a regular lattice (i.e. it should be clear which minimum derives from any given point in the undistorted lattice). For each minimum, therefore, we can calculate a distortion vector.

If we can assume the continuous nature of the distortion, then the distortion vectors can be expressed in terms of a relatively small number of parameters. This has been done (Henderson *et al.*, 1986) by fitting their coordinates to bicubic splines. (Obviously, some decision must be made as to the rapidity with which the distortion vectors can vary with position—i.e. as to the highest spatial frequency of the function they fit.) Given such a fit, the distortion vector can be calculated for any point, which facilitates resampling of the optical density array by bilinear interpolation.

7.4.5 Refining the orientation and symmetry of helical particles

(a) Positioning the helix axis

The position and orientation of the helix axis must be known accurately for determining helix polarity, for three-dimensional reconstruction (Section 7.6.4), or even for finding the parity of the layer-line order (Section 7.3.4(d)).

How many parameters shall we need to fix? A helical particle has a unique axis, which is coincident with the z-axis in the standard position. If the xz-plane is the plane of the micrograph, the xz-projection of the particle axis (PC in Fig. 7.63) is the "particle axis" seen in the micrograph. Unless the particle has obvious dihedral symmetry, there will be nothing to define Δz or χ for an isolated particle, though these will still be meaningful when comparing two different images of the same particle type.

We consider first the alignment of an isolated particle, for which the three transformation parameters that are needed are ψ , ω and Δx (Δz being put = 0).

The position and orientation of the projected particle axis is guessed when the image is densitometered; that guess is the z-axis. Even this will probably need to be refined by correcting ψ and Δx . In the F.T., the ψ -correction is the angle between the layer-lines and the X-axis. It is probably satisfactory to make these coincide by referring to a print-out of the F.T., and the only remaining problem is to interpolate the F.T. along these sloping layer-lines. This can be achieved by one of the interpolation techniques described in Section 7.6.4(d). After this, only two transformation parameters remain to be determined.

Phase asymmetry: Δx . The other two corrections do not affect the F.T. amplitudes; the layer-line phases must be used instead. These phases should either be equal at $F(X)$ and $F(-X)$, or differ by 180° .* If this situation does not hold, it must be brought about

* We are here assuming that the layer-line is not subject to interference from any other layer-line; if it is, we use the inner parts of the layer-line, where one order of Bessel function predominates; Crowther *et al.* (1985).

(as nearly as possible) by using the shifts Δx and ω . Δx causes the F.T. to be multiplied by a "complex wave" as shown in rule (d) of Fig. 7.12. Since the translation is along x , the lines of constant phase-shift lie perpendicular to X , i.e. along Z . So $F(X)$ and $F(-X)$ receive equal and opposite phase shifts.

Similar phase shifts are generated by ω on higher layer-lines, but not when $Z = 0$ (equator). Only Δx produces a (usually small) linear phase-shift there. It is therefore possible to estimate the translation Δx from a plot of the phase of the equatorial layer-line (Finch and Klug, 1971).

Phase asymmetry: ω . The phase correction resulting from particle tilt (ω) is slightly more complicated (DeRosier and Moore, 1970).* Tilting the particle gives its projection from a different direction. By the projection rule (Fig. 7.12(f)), this gives a different central section of its F.T. It is simplest to consider the particle F.T. as fixed in a (X, Y, Z) -coordinate frame, and to consider its values on the different (tilted) section-planes. Figure 7.64 shows three such section-planes (A, B, C) intersecting a layer-plane. Section-plane A is actually untilted, and corresponds to the XZ -plane. It intersects the principal maximum of the layer-plane at two points (arrowed at the ends of line A). As we have seen, the phases at these points will either be the same (n even) or will differ by 180° (n odd). The section-planes B and C are tilted with progressively larger values of ω . They intersect the principal maximum at different points (arrowed at the ends of lines B and C). Because these points are displaced from the diameter of the principal maximum, the simple phase relations no longer hold.

It is easy to estimate the phase differences, if there is no layer-line interference. The new phase relations can be seen in Fig. 7.36. The intersection of plane A is along the X -axis, and the phase difference appropriate to a layer-plane with $n = 3$ (as shown in Fig. 7.36) is 180° . Tilting the helix gives us F.T. sections on planes B and C, which intersect the layer-plane along the rows indicated. Here, the phase difference is rather smaller, since the vectors on the right have rotated clockwise, while those on the left have rotated anticlockwise. The change can be related quantitatively to the tilt angle ω (together with Z and n); see Equation (53) below.

The overall correction (DeRosier and Moore, 1970) will now be given. If the particle has been displaced by Δx and tilted by ω from its "perfect" position (Fig. 7.63), then, on a layer-line with the coordinate Z ,

$$\text{True phase } (F(X)) = \text{calculated phase } (F(X)) + n \arctan((Z \sin \omega)/X) \quad (53)$$

The application of this equation is connected with determining the helix hand, to which we now turn.

(b) Finding the helix hand, and confirming the helical lattice

Equation (53) can be used to correct the F.T. phases, provided we know Δx , ω and n . As explained, a very good estimate for Δx can be found from a plot of the phases on the equator. However, we do not know ω , and shall use the equation to find it. And, although we may know $|n|$ for each layer-line, from the (n, Z) plot, some ambiguity may remain. In any case, we shall not know its sign unless we have determined the hand

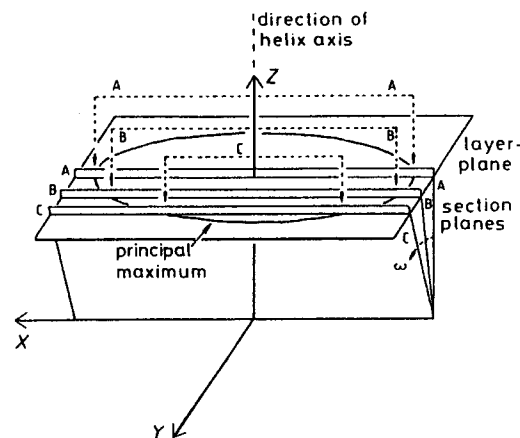


Fig. 7.64 Influence of particle tilt on layer-line phases in the helical F.T. Tilt by an angle ω (defined in Fig. 7.63) has the effect of tilting the section plane away from the Z -axis, causing rotations of the phases on both sides of the layer-line. (These are shown in Fig. 7.36.)

of the helix. These uncertainties can be removed with the help of Equation (53) (DeRosier and Moore, 1970; Finch, 1972a), in the following way.

First, we choose one of the possible helical lattices and chiralities. We calculate the appropriate (n, Z) plot, which fixes the values and signs of n on all layer-lines. For any given value of ω , we can now use Equation (53) to correct the phases at X and $-X$ on a layer-line. After correction, the phases at X and $-X$ should differ by $n(180^\circ)$. So we calculate the contribution of all pairs of points (X and $-X$) to an overall error-function of the phase-residual type (Section 7.4.3(a)). This function (E) will be determined by the helical lattice, of which the most important quantity is the twist angle Ω (to which we should perhaps add the order, N , of the rotation axis). It will also depend on the orientation parameters ω and Δx , which are used to find the corrected F.T. from the value given by the F.F.T. (The correction, using Equation (53), affects only its phase.) We now use E to refine the helical and orientation parameters.

Consider first the helical parameters Ω and N . Analysis of the helical diffraction pattern (Section 7.3.4) should give a sufficiently accurate estimate for $|\Omega|$, leaving only its sign to be determined. Thus refinement of the parameters Ω and N involves only a choice between alternatives. For each choice, E is a function of only ω and Δx ; and, since Δx can be found quite accurately by the method described in Section 7.4.5(a), ω is the only really uncertainty quantity. $|\omega|$ can be found by plotting E as a function of ω , for each symmetry choice. The minimum indicates the value of $|\omega| = |\omega_1|$. However, handedness cannot be determined from a single projection: the same fit will be given if the signs of ω and Ω are both reversed. To fix the hand, we need a second micrograph of the particle, obtained after slight tilting of the grid by a known angle $\Delta\omega$ in a known direction. We repeat the analysis on this second micrograph, and obtain a new pair of plots (for $+\Omega$ and $-\Omega$), with a new value $|\omega| = |\omega_2|$. Of the four possible quantities $(+\omega_2 + \omega_1)$, $(+\omega_2 - \omega_1)$, $(-\omega_2 + \omega_1)$, $(-\omega_2 - \omega_1)$, only one will equal $\Delta\omega$. This fixes the signs of ω_1 and ω_2 , and hence Ω .

* Note that these authors define ω as a tilt towards $-x$, rather than as shown in Fig. 7.63.

(c) Finding the polarity of the helix

As explained in Section 7.3.4(i), the noise level of individual helical F.T.s usually makes the detection of polarity dependent on a comparison of the F.T.s of different particles. It is first necessary to correct each F.T. for the particle's position and orientation, as described in the previous section. A further positional correction is needed, in the direction of each particle's axis. Moving any particle by Δz multiplies its F.T. by $\exp(2\pi i Z \Delta z)$; after this correction, its F.T. can be compared with that of the other particle. An error-function is used, similar to that employed in finding the particle orientation.

Before comparing the F.T.s, however, we need to decide whether the two particles are parallel or antiparallel (Amos and Klug, 1975). Lacking grounds for choice, it is best to try each orientation in turn. Thus, when optimizing the orientation, we obtain a minimum error-function ($E_m(\text{parallel})$) calculated for the parallel orientation, and another minimum ($E_m(\text{antiparallel})$) for the antiparallel orientation. The difference between them relates to the polarity of the particles. For, if they are non-polar, it cannot matter whether they are compared parallel or antiparallel, and any difference must be ascribed to noise. We therefore need to decide whether the difference between $E_m(\text{parallel})$ and $E_m(\text{antiparallel})$ is above this threshold value.

Our strategy will depend on whether the particles have any structures that distinguish their ends. Bacteriophage tails, for example, have heads at one end; and helices associated with cellular organelles may often be seen growing out from some root or origin structure. If the ends are distinguished in this way, then every comparison of two particles employs the same definition of "parallel". The data may be good enough to allow one particle to be taken as a comparison standard, as with the T4 extended sheath: Fig. 7.65a, b. Otherwise, it is necessary to get a better "standard" by averaging all F.T.s from particles in the best parallel orientation. This average F.T. can be compared with each individual F.T. to give $E_m(\text{parallel})$ and $E_m(\text{antiparallel})$. A convenient way to represent the comparison is to plot, for each particle, a point with coordinates [$E_m(\text{parallel})$, $E_m(\text{antiparallel})$]. For a non-polar particle, the points should cluster around the 45° line, as $E_m(\text{parallel})$ and $E_m(\text{antiparallel})$ should be equal. For a polar particle, the points should lie above it, as $E_m(\text{antiparallel}) > E_m(\text{parallel})$.

If, however, there is no visible distinction between the particle ends, then the polarity test is more difficult (as well as perhaps more relevant). Pairwise matching is done by the "snowballing" method described in Section 7.4.3(b). The apparent success of this procedure is not, of course, a guarantee of particle polarity; images might have been selected showing the same, random, deviations from two-fold symmetry, and their average will enhance these arbitrary deviations. So the final "averaged" F.T. must be compared with all particles in a plot like Fig. 7.65c. As before, polarity will be revealed by a clustering of points above the 45° line. This procedure, critically applied, can demonstrate that the particles are polar. Because of the limited resolution and high noise of electron micrographs, however, it cannot demonstrate non-polarity.

(d) Correcting distortions of the helix

Some of the possible distortions of helices were mentioned in Section 7.3.4(j). Distortions that affect only the surface structures of helices frequently lead to a superlattice, which can be determined as outlined in Section 7.3.4(g). Distortions of the helical geometry

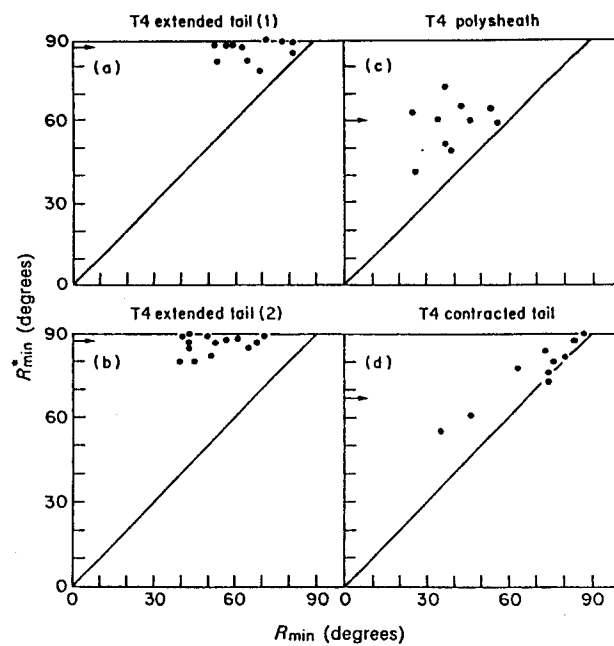


Fig. 7.65 Polarity plots (Amos and Klug, 1975); see text for explanation.

pose a more common problem. Usually they are simply avoided by finding conditions under which they are minimized (e.g. by selecting the few straight particles for analysis). But, as with distorted plane lattices (Section 7.4.4), methods are being developed to correct them.

Of all the distortions of helix geometry, the easiest to correct is helix bending (Fraser *et al.*, 1976; Egelman, 1986; Egelman and Stasiak, 1986). If the curvature is small, it causes few distortions in the local helical packing. The bend can be determined and used to define (orthogonal) curvilinear coordinates in which one of the axes follows the helix axis. A conformal mapping of the micrograph, from these coordinates to Cartesian coordinates, will straighten the helix.

It is rather more difficult to correct distortions of the twist angle. Bluemke *et al.* (1988) have developed a method in which the twist angles are adjusted by model-fitting. Models are constructed, in which the subunit structure derives from X-ray crystallography, the rise-distance is fixed, and the twist-angle of every subunit is adjustable. The model that correlates best with the micrograph will be accepted. To find it, a series of improved models is constructed by an iterative process. At each stage, the current estimates of the twist-angles are used to make a three-dimensional reconstruction of the micrograph. From this, new estimates are made of the subunit positions leading to new twist-angles and an improved model.

By allowing all the twist-angles to float independently, this method encounters a problem common to all attempts at correcting deviations from strict symmetry. Their

specification requires extra information, which must come from the micrograph, thereby reducing the amount of useful information it can yield. This problem was encountered when correcting distorted plane lattices (Section 7.4.4). We saw that, whereas the correction is difficult if the distortions are random, it becomes feasible if they are continuous; for, in the latter case, only a few parameters suffice to define the distorted coordinate system. It is because curvature of the helix axis is a continuous distortion that its correction is feasible. Similarly, the correct of twist-angle distortions will be most successful when they, too, are continuous. Indeed, when Bluemke *et al.* applied their method (Carragher *et al.*, 1988), they found that the twist-angle variations were continuous (see Fig. 4(a) of their paper) and could be interpolated from the values of a canonical set. These considerations suggest that, instead of devising *ad hoc* methods for correcting particular distortions, the entire helical lattice (including the rise-distance) should be allowed continuous distortions specified by a few parameters which could be determined by (non-linear) least-squares.

7.4.6 Rotational symmetry

(a) Rotational periodicities

As explained in the general introduction, all Fourier methods are essentially similar, in that they analyse a picture into a standard set of shapes or density-waves. For pictures showing rotational symmetry, we therefore need waves with rotational symmetry (as the density-waves considered in Section 7.2 have translational symmetry). But certain differences now arise. First, the centre of rotation must be specified. Second, there is the difference that translational waves are essentially similar in all directions, whereas there is a radical difference between rotational waves in the radial and tangential directions. Since we are interested only in the *rotational* symmetry of particles, there is no absolute need to use waves that vary in the radial direction at all. As a start, therefore, we divide up the micrograph into concentric narrow annuli (Fig. 7.66a), and analyse each annulus separately to determine its rotational symmetry.

As usual with the Fourier method, this is done by representing the density around each annulus as the sum of several different sinusoidal density-waves. That is, each annulus is the sum of annuli of the same radius (Fig. 7.66b, c, d) and having densities that vary sinusoidally with all necessary rotational periodicities. (An annulus with uniform density—zero-fold periodicity—is included to make the final density entirely positive.) As a result of this rotational Fourier analysis, one can find the contribution of each sinusoidal density-wave in terms of its amplitude (the “strength” of its contribution) and phase (the correct orientation of its positive peak). For an annulus of radius r , the contribution of the n -fold density-wave is denoted by $g_n(r)$ (see Section 7.2.3(k)). The total “power” of n -fold density-waves (Crowther and Amos, 1971) is found by squaring the amplitude of each $g_n(r)$, summing over all the annuli and (since the amount of information in an annulus is proportional to its radius) weighting each annulus according to its radius:

$$P_n = \int_0^{\infty} |g_n(r)|^2 2\pi r \, dr \quad (54)$$

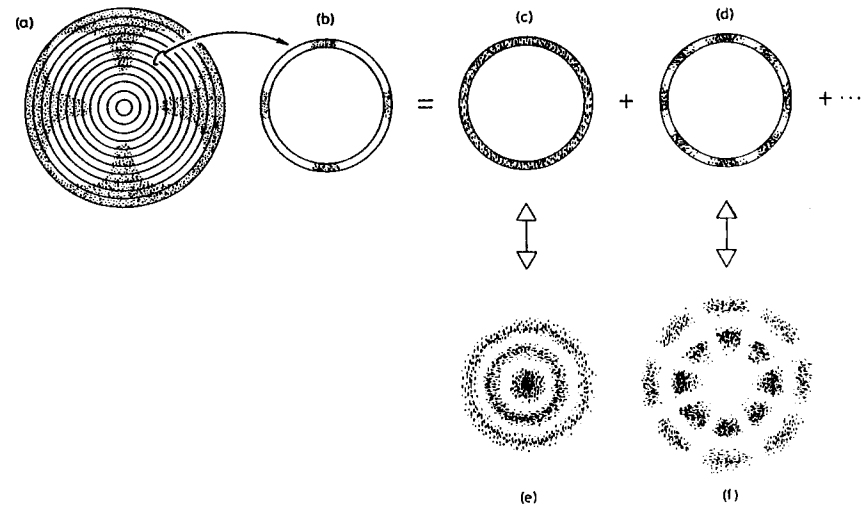


Fig. 7.66 The Fourier method for finding rotational symmetry. The image (a) is divided into annuli, of which one (b) is expressed as the sum of its rotational Fourier components ((c) and (d)). Their F.T.s are shown in (e) and (f).

(Note the similarity with Equation (43), Section 7.2.3(k). This power (denoted by P_n) is the rotational equivalent of the intensity of a spot in an optical diffraction pattern. The equivalent of the diffraction pattern itself is a plot of P_n against n (Fig. 7.67), and the peaks in this plot will indicate the principal rotational symmetries present in the micrograph, for the particular choice of rotation centre used in the calculation.

(b) Finding the rotation centre

To apply the method, we must be able to locate the correct rotation centre and then to calculate the power spectrum (i.e. the different P_n values). These tasks are interrelated. By the “correct” centre we mean the point about which the rotational symmetry of the particle is maximized. But the centre that maximizes a particle’s four-fold symmetry need not be the same as that which maximizes its five-fold symmetry. Consequently each prominent candidate for the “true” rotational symmetry has its own “correct” rotation centre and its own “correct” rotational power spectrum. We estimate which numbers are prominent candidates by looking at the micrograph or, if no symmetry is apparent, we calculate a trial rotational power spectrum using a centre located by visual inspection.

To decide between the candidates, we must locate the best particle centre. This is similar to locating the axis of a helical particle (Section 7.4.5(a)), except that we now have a section of the transform *perpendicular* to its rotation axis. We can therefore compare the transform phases at all points that should be equivalent through the assumed n -fold rotation axis. So we calculate an error-function, which measures the mean deviation of the transform at all these points. We calculate it for different positions of the rotation

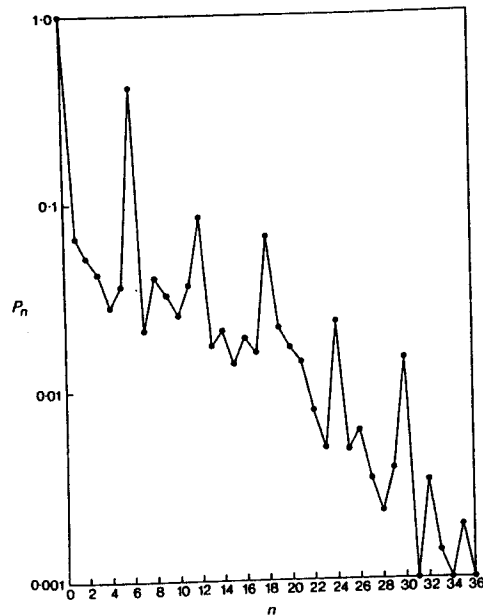


Fig. 7.67 Logarithmic plot of the rotational power spectrum of a T4 bacteriophage base-plate (Crowther and Amos, 1971).

centre. (The corresponding transforms required for this calculation can be obtained by application of the appropriate phase factor.) The correct position of the rotation centre (for an n -fold axis) minimizes this error-function. It could be found by some numerical method; but, since only two variables are involved, a two-dimensional plot suffices (Crowther and Amos, 1971).

(c) Finding the power spectrum

Different trial rotation axes (i.e. the trial values of n) are chosen in succession. For each, the particle centre is found by the method just described. Then the complete power spectrum $P_k(n)$ is calculated by one of the three methods outlined in Section 7.4.6(d). $P_k(n)$ is the power of the k -fold contribution to the particle structure, when the rotation centre appropriate to n -fold symmetry has been used. $P_k(n)$ is the rotational equivalent of an optical diffraction pattern. Just as, for a translational repeat a , the optical diffraction pattern has peaks at multiples of $1/a$, so the rotational power spectrum has peaks at multiples of n .

This procedure gives a different rotational power spectrum P_k for each candidate (n -fold) rotational symmetry. Naturally, the power spectrum appropriate to an n -fold axis will enhance the n -fold contribution as much as possible. But if, in each of these spectra, the dominant rotational contribution is always the same, then the particle has a unique rotation axis. Otherwise, its rotational symmetry (if any) is indeterminate.

(d) Details of calculating the power spectrum

For a given particle centre, there are three methods for calculating the power spectrum P_n . The first would divide the optical density of the micrograph into rings as in Fig. 7.66; the rotational Fourier component of each ring, $g_n(r)$, would be found by Equation (33), and from these quantities the power P_n would be found by Equation (54).

The other two methods would use the particle's F.T. (which was already calculated to locate the rotation centre). By either of these methods, we should calculate $G_n(R)$. One way to calculate it would be through an angular transform, as in Equation (34). However, this would require us to interpolate the F.T. at points required by the quadrature formula. So the third method would use Equation (39) (the first and last terms), that expresses the F.T. as a linear combination of $G_n(R)$ s, with known coefficients. These linear equations might be solved by the methods outlined in Section 7.6.4(d). By either of the two methods, we should have $G_n(R)$, from which P_n can be calculated by Equations (43) and (54).

(e) Effects of distortion and resolution

As with optical diffraction patterns, the information in a rotational power spectrum is restricted both by distortion and by limited resolution. *Distortion* broadens peaks so that they extend to other harmonics. Similar effects are seen in the optical diffraction patterns of distorted translational periodicities; but the effects of distortion would be expected to be particularly prominent in the case of rotational symmetry. There are two reasons for this. First, the effects of uniform stretching of the particle are far more serious than in the case of translational periodicities. Instead of merely producing a reciprocal stretching of the optical diffraction pattern, stretching presumably enhances even-fold rotational symmetries. The second reason derives from the particle shape. Whereas two-dimensional crystals are typically wide and thin, particles showing rotational symmetry are much more nearly isometric. In negatively stained preparations, such particles will be particularly liable to tilt or to the effects of anisotropic contraction (Section 7.3.4(h)). Tilting is also probably the most serious "distortion" artifact in sectioned or freeze-etched preparations.

The effect of *resolution* is essentially the same as in optical diffraction patterns, when due allowance is made for geometric factors. (Any given order of rotational harmonic requires less spatial resolution at large radii than at small ones.) In practice, the resolution found from rotational symmetry has been somewhat worse than that found from translational symmetry. Perhaps this is owing to astigmatism, or to the ease with which small isolated particles become distorted.

(f) Analysing rotational symmetry by correlation coefficients

The rotational symmetry of a particle's image can be found by comparing it with a similar image (or with itself), after rotation by different angles. If the images match, they should match again when the angle is increased by $360^\circ/N$ (N being the order of the rotation axis). As explained in Section 7.2.4, the best match can be taken to maximize their X.C.F., defined by Equation (48). The match will be a function of the two coordinates of the particle centre, as well as of the rotation angle ψ . However, when comparing two

images, the X.C.F. is a function of two additional transformation parameters ($\Delta x, \Delta z$). When comparing different particles, then, we have a five-parameter search. This can be facilitated (Frank *et al.*, 1978) by the subspace technique outlined in Section 7.4.3(a). First, the image's A.C.F. is calculated. This automatically has an exact two-fold axis, but every rotational symmetry of the image will also be present in the A.C.F., and with the same centre of rotation. Thus (with the exception of two-fold symmetry) the A.C.F. gives us an estimate of the rotation symmetry of the image. However, the A.C.F. has the additional useful property of possessing a unique centre. Consequently, the A.C.F.s of different particles can be compared (e.g. through a rotational X.C.F.) before the positions of the particles are known. This gives us the angular part (ψ) of the transformation relating two different images. Having corrected for ψ , the remaining transformation parameters ($\Delta x, \Delta z$) can be found by a two-dimensional translational X.C.F. The best rotation centre, for any rotation angle $360^\circ/n$, can be found from the X.C.F. of the rotated and unrotated images.

Both the rotational and translational X.C.F.s are more rapidly calculated by means of F.T.s (Frank *et al.*, 1988b), allowing use of the rapid F.F.T. algorithm (Section 7.4.2(b)). The rotational X.C.F. can be calculated from Equation (42) or, better, Equation (44); and the translational X.C.F. from the Fourier inversion of Equation (9).

7.4.7 Point-group symmetry

(a) Deductions from projection symmetries

If a particle has exact rotational symmetry, it must fit one of the point-groups mentioned in Section 7.2.1. The cyclic (C_N) and dihedral (D_N) groups constitute infinite series of similar point-groups, with $N = 1, 2, \dots$. Unless N is small, these groups generate ring-shaped structures. There are only three other point-groups, and these generate compact particles. Now it should be evident from micrographs whether a particle is compact or ring-shaped. So the determination of point-group symmetry is really only a choice between rather few (reasonable) alternatives. Biochemical evidence bears on this choice, since the number of subunits in the particle is the *order* of its point-group, if all the subunits are identical and equivalent. Knowing this number narrows the possible point-group symmetries of the particle; see Table 7.1.

Rotational analysis, by the methods discussed in the previous section, can reveal the rotational symmetry of a picture. But we are interested in this only as a means of finding the symmetry of the original particle. Since the picture probably represents a projection of the particle, what is the connection between the symmetries of a particle and of its projections?

By the projection rule (Section 7.2.2(d)), the symmetry of a projection must be the same as that of the corresponding central section of the F.T. Moreover (by the rotation rule of Section 7.2.2(d)), the point-group symmetry of the F.T. must be the same as that of the particle. So it might seem that the projection symmetry would be simply the same as the particle symmetry about an axis coinciding with the projection direction. However, there are two complications.

First, the shape and size of individual subunits may be such as to appear spherically symmetric at low resolution. Thus the subunits of oligomers may generate the appearance

7. Image Analysis of Electron Micrographs

Table 7.1 Point-groups, their orders and principal subgroups.

Group	Order	Axes	Principal subgroups
Cyclic (C_N)	N	N	C_M (M) [M divides N]
Dihedral (D_N)	$2N$	$N, 2$	C_N (N), C_2 (2)
Tetrahedral	12	2, 3	D_2 (4)
Octahedral	24	2, 3, 4	Tetrahedral (12), D_3 (6), D_4 (8)
Icosahedral	60	2, 3, 5	Tetrahedral (12), D_5 (10), D_3 (6)

Order equals the number of subunits in an aggregate with this point-group; *axes* are the rotational symmetry axes; and only the more interesting *subgroups* are listed. The numbers in brackets are the orders of the subgroups (which, by Lagrange's theorem, must be a factor of the order of the group), and subgroups of subgroups have been omitted.

of more symmetrical polyhedra. (For example, four subunits related by D_2 symmetry might assume the appearance of a regular tetrahedron.) In the most likely case, just one "spherical" subunit would generate, through the operations of the true point-group (G_1), a polyhedron with the pseudo-symmetry corresponding to a higher point-group G_2 (where G_1 is a subgroup of G_2). Each "spherical" subunit thereby becomes a vertex of the polyhedron, so the number of vertices is the number of subunits, i.e. the order of the true point-group G_1 . This gives us a criterion for pseudo-symmetry generation. The order of the subgroup (G_1) must equal the number of vertices in a polyhedron with the point-group symmetry G_2 . And this number equals G_2 divided by the order of one of its rotation axes (e.g. by 5, 3 or 2 if G_2 is the icosahedral point-group). Applying this criterion to the sub-groups of Table 7.1, we have only the cases list in Table 7.2.

Second, the projecting parts of the subunits (where the contrast is highest) may be clustered about symmetry axes. If the clusters are blurred into indistinguishable blobs, the appearance of a regular polyhedron can result. Such hypersymmetrical clusters are important only for the tetrahedral, octahedral and icosahedral point groups. They are listed in Table 7.3.

If an oligomer approximates one of the regular polyhedra in the third column of Table 7.2 or of Table 7.3, there will be some inevitable uncertainty about its true symmetry. (For example, an apparently octahedral aggregate could have D_3 symmetry, from Table 7.2, or else tetrahedral or octahedral symmetry, from Table 7.3.) This problem could exist even if the particle's three-dimensional structure had been determined by some reconstruction technique (because the resolution was inadequate); so there is even more uncertainty if one has to interpret only a few (or one) symmetrical

Table 7.2 Generation of polyhedra with pseudo-symmetries.

Pseudo-symmetry group	Axis	Polyhedron	Subgroup	Projection symmetry
Tetrahedral	2-fold	Octahedron	D_3	2, 4, 6
	3-fold	Tetrahedron	D_2	2, 3, 4
Octahedral	3-fold	Cube	D_4	2, 4, 6
	4-fold	Octahedron	D_3	2, 4, 6

The *polyhedron* has the symmetry of the *pseudo-symmetry* group, and is formed by joining points placed on the axis of given order. The *subgroup* is a point-group that can generate the *polyhedron* from a single point placed at some suitable position. (Only those cases are listed for which the subgroup differs from the pseudo-symmetry group.) An oligomer approximating one of the polyhedra in the third column could have, not the symmetry in the first column, but that in the last. The *projection symmetry* shows the possible symmetry of its micrographs.

Table 7.3 Regular polyhedra that can be generated from oligomers at low resolution.

Group	Cluster axis	Possible polyhedron	Projection symmetry
Tetrahedral	2-fold	Octahedron	2, 4, 6
Tetrahedral	3-fold	Tetrahedron	2, 3, 4
Octahedral	2-fold	Cuboctahedron	2, 4, 6
Octahedral	3-fold	Cube	2, 4, 6
Octahedral	4-fold	Octahedron	2, 4, 6
Icosahedral	2-fold	Icosidodecahedron	2, 6, 10
Icosahedral	3-fold	Dodecahedron	2, 6, 10
Icosahedral	5-fold	Icosahedron	2, 6, 10

The cluster axis is a rotation axis of the polyhedron about which the subunits are supposed to cluster. If the clusters are seen only as spheres, these spheres are arranged at the vertices of a polyhedron, the possible polyhedron in the third column. (Good descriptions and illustrations of the polyhedra are provided in Cundy and Rollett (1961).) An oligomer approximating one of the polyhedra in the third column could have the symmetry listed in the first column. The projection symmetry shows the possible symmetry of its micrographs.

projections. This is because the regular polyhedra of Table 7.3 show only a few types of rotational symmetry in projection (listed in the last columns of Tables 7.2 and 7.3). To illustrate these difficulties, suppose that a projection shows six-fold symmetry. It could derive from polyhedra which imply any of the following symmetry groups: D_3 , D_4 (from Table 7.2); or tetrahedral, octahedral or icosahedral (from Table 7.3); as well as its apparent symmetry, cyclic C_6 .

Nevertheless, if these problems of interpretation are borne in mind, the determination of point-group symmetry by electron microscopy can be very useful, especially if it is combined with other evidence (e.g. from molecular weights). The first stage must be the estimation of rotational symmetry from projections (Section 7.4.6). If the overall rotational symmetry cannot then be deduced, it might be found by a three-dimensional reconstruction, or by one of the methods used for icosahedral virus particles (next two sections).

(b) *Highly symmetrical particles: the "common lines" method*

Common lines. The "common lines" method (Crowther *et al.*, 1970a, b; Crowther, 1971) was the first, and has been the most widely used, Fourier method for finding the orientation of an isolated, highly symmetrical particle. In principle, the method could determine the particle's symmetry, but it was designed (and has been exclusively used) for finding the orientation of a particle of known symmetry (i.e. for finding the transformation parameters (ψ, ω, χ)).

We assume, as usual, that the image of a particle is approximately a two-dimensional projection of its structure. From this projection, we can calculate a section of its three-dimensional F.T.; the section plane (ABCD in Fig. 7.68) is oriented perpendicular to the direction of projection. Suppose that the particle has a three-fold rotation axis along the z-axis. The F.T. must possess the same rotational symmetry as the original particle, with the same orientation of the symmetry axes. The transform in Fig. 7.68 will consequently have a three-fold axis along its Z-axis (this is the direction along which the reader views Fig. 7.68). Since all points related by 120° rotation about this axis have identical values of the transform, the plane section A'B'C'D' (Fig. 7.68), derived from

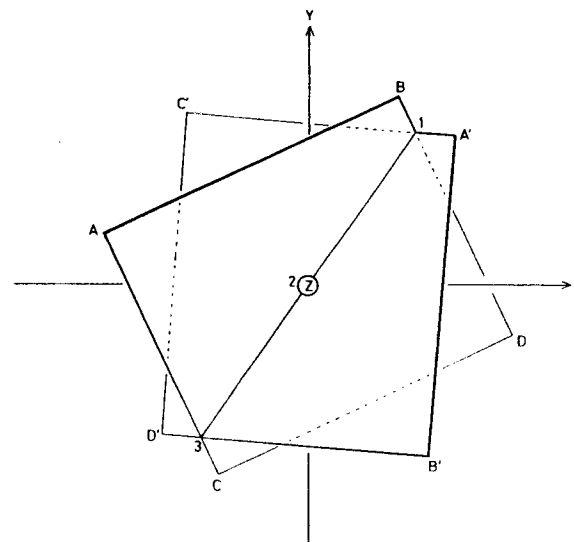


Fig. 7.68 The "common-lines" method for determining the orientation of a particle's rotation axes. ABCD is a section of the F.T., calculated from one image. The three-fold axis (along the Z-axis, pointing at the reader) generates from this two symmetry-related planes, of which only one (A'B'C'D') is shown. Both planes have common values along the line (1, 2, 3) of their intersection.

ABCD by 120° rotation about Z, must contain an identical copy of that part of the transform on ABCD. The planes ABCD and A'B'C'D' intersect at the line marked 123. Figure 7.69a shows the position of this line in the former plane, and Fig. 7.69b shows its position in the latter. Since A'B'C'D' is a copy of ABCD, each plane must contain the line 123 twice, i.e. once as in Fig. 7.69a and once as in Fig. 7.69b. Thus each plane contains two copies of the line (Fig. 7.69c). Along these two copies of the line, the transform has identical values on moving from points 1 to 2 to 3. These two copies of the line are called "common lines" in the original transform section.

Calculating the common lines' positions. The common lines must all pass through the centre of the transform, since all are generated by the intersection of planes that pass through that centre (Fig. 7.68). Moreover, it is clear from Figs 7.68 and 7.69 that, when there is a single rotation axis, the system of common lines on a transform plane is related by mirror symmetry about a line which is the projection of the rotation axis. When there are several rotation axes, the projection of each rotation axis on the transform plane will constitute a line of mirror symmetry for the system of common lines. Finally, the lines can be calculated conveniently in the following way. Represent the first transform plane by its normal \mathbf{n}_1 passing through the centre of the transform. From this normal, all the operations of the point-group generate new normals $\mathbf{n}_2, \dots, \mathbf{n}_N$, where N is the order of the point-group. Now we calculate all the vector (cross) products

$$\mathbf{p}_j = \mathbf{n}_1 \times \mathbf{n}_j \quad (j = 2, \dots, N) \quad (55)$$

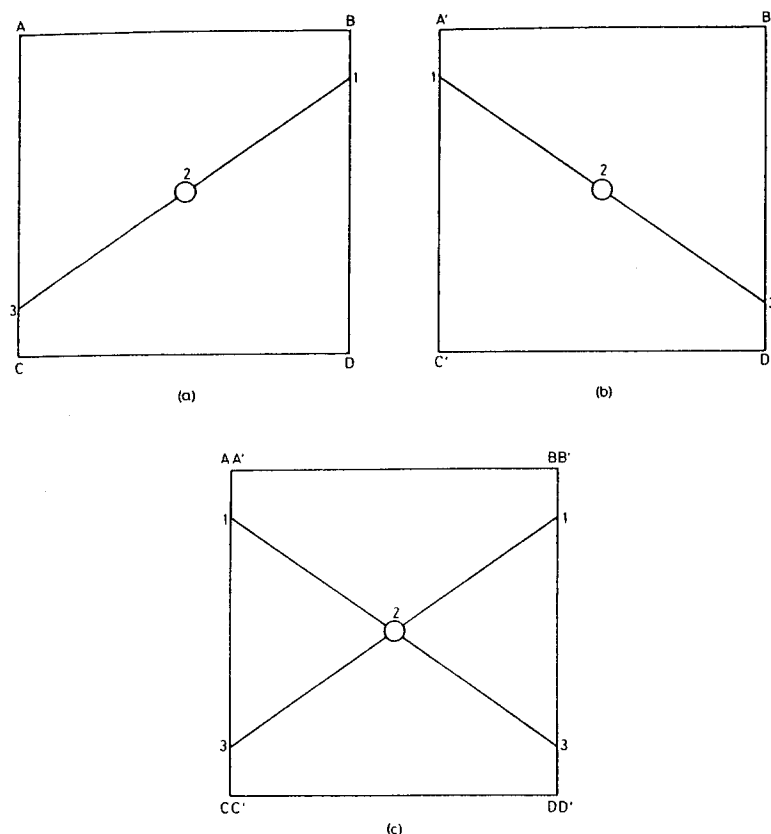


Fig. 7.69 The planes ABCD (a) and A'B'C'D' (b) of Fig. 7.67 laid flat. The common values lie along two lines, shown in (c).

Each p_j is perpendicular to n_1 , and therefore lies in the first transform plane; similarly, it also lies in the j th transform plane. The only line common to both planes is their intersection, which is a common line. The vectors p_j therefore define the directions of the common lines in the first transform plane.*

Using common lines to find the particle orientation. The orientation of the common lines depends on the mutual orientation of the intersecting planes ABCD and A'B'C'D', which depends on the orientation of the original plane ABCD with respect to the

* This method fails if the normal perpendicular or parallel to a rotation axis; the symmetry-related normals generated by this axis will then yield vector cross-products that are either coincident or null. The former situation corresponds to finding the rotation axis of a helix (Section 7.4.5(b)), and the latter to determining the rotational symmetry of an image (Section 7.4.6).

three-fold axis (Z-axis). Thus the orientation of the common lines depends on the direction of projection with respect to the three-fold axis. (For example, the closer that direction lies to the xy -plane, the more will the planes ABCD and A'B'C'D' be parallel to the Z-axis, and the closer together will be the points marked 1 in Fig. 7.69c.) Therefore, if the positions of the common lines were located in the transform section, one could deduce the orientation of the particle with respect to its three-fold axis.

This fact can be used to find the particle orientation (relative to the viewing direction) by the following search procedure. To begin with, we must choose (or guess) the particle's point-group symmetry. Then, for any trial viewing direction, the positions of the common lines can be calculated, and the values of the observed transform compared at corresponding points along them. From this comparison, we calculate an error-function that has a minimum when the trial viewing direction coincides with the one actually used. The error-function used is of the "phase residual" type (Section 7.4.3(a)).

Two angles (ω and χ of Fig. 7.63) completely determine the mutual orientations of the common lines. Having found them, a third angle (ψ) rotates the calculated particle transform with respect to the common lines. Although many different combinations of these three angles must be searched, the extent of the search can be reduced in two ways. First, the two angles (ω , χ) specifying the viewing direction need be searched only over an asymmetric unit of rotation space. Outside this region, the orientation of the intersections of the planes defining the common lines (i.e. the system of planes similar to those in Fig. 7.68) will be repeated again, and no new system of common lines will be derived. The symmetry of rotation space is thus the same as the symmetry of the particle, with the addition of a centre of symmetry (since the system of intersecting planes has a centre of symmetry). Second, the extent of the search can be reduced still further if the approximate orientation of the particle is clear from the micrograph. This situation usually applies in the case of negatively stained particles with icosahedral symmetry.

However, while determining the particle's orientation, it is also necessary to find its centre (i.e. Δx and Δz). (The situation is much the same as in the case of rotational symmetry (Section 7.4.6).) Two coordinates are needed to specify this centre, so the search must in principle be carried out over a total of five parameters. But the centre of a highly symmetrical particle can usually be chosen by eye with sufficient accuracy for preliminary analysis of the particle orientation. After finding this (from a search over the three angular parameters), the position of the centre can be refined by small adjustments of its two coordinates, using as a criterion the same error-function that determines the orientation. Thus, though the error-function is strictly a function of five parameters, the search can be started in a three-dimensional subspace (p. 230).

As we have seen, the error-function is calculated from transform values at corresponding points along the common lines. Before starting this calculation, it is useful to divide the transform into spherical shells. Then the error-function is found for each shell separately, and becomes a function of transform radius. This is useful for the following reason. If the resolution of the particle image is $d \text{ \AA}$, the particle transform beyond $1/d \text{ \AA}^{-1}$ is essentially noise, devoid of any precise symmetry. At such radii, the error-function must be relatively large. So a plot of the error-function against transform radius (i.e. particle resolution) shows the degree of particle preservation at various resolutions (Fig. 7.70). From this plot one can determine the resolution of the specimen. The best resolution so far obtained is 25 \AA , for bushy stunt virus (Crowther *et al.*, 1970a; Crowther, 1971)

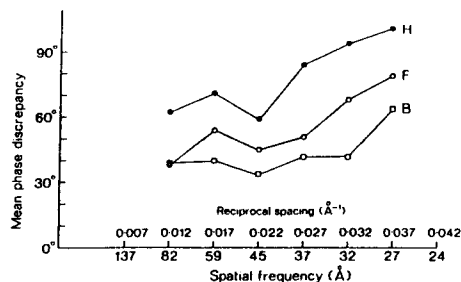


Fig. 7.70 The common-lines residual, for tomato bushy stunt virus, plotted as a function of resolution, to show how well the particle's symmetry is preserved (Crowther, 1971; reprinted with permission from the Royal Society). B, F and H refer to different particles; H, with a high residual (90° indicates no icosahedral correlation), is poorly preserved.

After the particle's orientation and position have been refined, we have the lowest value of the error-function. This value measures how well the particle fits the symmetry originally assigned to it. If the error-function were normalized to take account of the fact that the number of common lines depends on the order of the assumed symmetry group, then the error-function could in principle be used to determine the rotational symmetry of the particle.

(c) *Highly symmetrical particles: the method of functional expansions*

The common lines method uses the F.T. values only at corresponding points along the common lines. Determination of the particle orientation (or, perhaps, symmetry) thus ignores other parts of the F.T., which may also contain relevant information. So perhaps more reliable determinations may be possible by a method that makes use of the entire F.T.

Such a method has been developed (Provencher and Vogel, 1983, 1988). Since it was designed for the three-dimensional reconstruction of particles (not necessarily symmetrical) from projection data, more details are given under Section 7.6.5.

7.5 IMAGE ENHANCEMENT

7.5.1 General principles

Image enhancement techniques generate, from the micrograph, an improved two-dimensional picture. This can be of value in itself, but it is often just a stage in generating a three-dimensional picture of the specimen (Section 7.6). Image enhancement is needed in order to reverse, as far as possible, deficiencies arising from two causes: noise and imaging defects.

(a) *Noise*

The most serious problem derives from the very high noise level in electron micrographs. Before the fine structure of a particle can be revealed, we must first improve the

signal-to-noise ratio in its neighbourhood. Such an increase in information content can come only by taking information from other pictures of the same particle. Many different (noisy) copies of the image, usually in the same micrograph, are thus averaged to yield a single, less noisy, one.

First we must find the positions of the different copies; and methods have been described, in the previous two sections, for doing this. Then simple averaging can give us an unbiased estimate of the noise-free image. However, that might not be the most efficient procedure. Averaging corresponds to repeating an experiment many times (i.e. the many-image copies), and plotting all the results in a scatter diagram (i.e. the averaged image). If there are enough points, this diagram will show both the average trend, and also the average deviations from that trend. With fewer points, however, it is more efficient to fit theoretical models to the data (e.g. by least-squares). In the case of images, we have no "theory", so the "models" must be quite general. Nevertheless their number need not be very large. For it is limited by the resolution of the micrographs: we use only those density-waves (Sections 7.2.2 and 7.2.3) whose spatial frequencies are so low that they can be reliably determined. This approach leads to the important Fourier filtering techniques (optical or computer). An alternative approach, that is likely to find application, is based on the concept of maximum entropy (Section 7.5.3(d)).

(b) *Imaging defects*

In other fields of photography, where noise is a less serious problem, more complicated ways are used to rearrange the information in a picture, in order to correct for imaging errors. Blurring, for example, is equivalent to convolution with some simple function, such as a line of density. In theory, de-blurring can be achieved by deconvolution, a procedure (for which there are many algorithms) that processes the data from many adjacent pixels. In practice, the improvement is limited by the picture noise. For blurring redistributes the density information in ways that require very high accuracy to correct. Noise, which limits this accuracy, prevents blurring from being completely reversible. Thus blurring, like many physical processes, brings about some increase of entropy (equivalent to a loss of information).

Since specimen drift can usually be avoided,* "blurring" in the case of electron micrographs mostly takes the form of image aberrations, especially defocus. But projection along the direction of view also has some of the formal properties of "blurring". Neither defect can be corrected satisfactorily from the data in one micrograph. This irreversibility can be understood through the concept of "invisible functions". Suppose that we change the structure of the specimen by adding some density function to it. Usually the addition will also change the recorded image, but this will not always be the case. If, for example, the added function were such as to give a uniform projection in the direction of view, then (whatever its three-dimensional distribution might be) it would have no significant effect on the image. Let us call such a function an "invisible function".† After deducing the structure of the specimen, we could add to it any amount

* A possible exception is the situation with some low-temperature stages. Techniques for filtering micrographs with translational or rotational blurring have been developed by Carragher *et al.* (1986).

† See Rust and Bursus (1972). This concept is applicable to electron micrographs because the imaging process is nearly linear.

of any "invisible function", without changing the experimental data on which our deduction rests.

If we are ever to have an unambiguous specimen structure, we must somehow avoid these invisible functions. There are three ways to do this. Some invisible functions can be simply ignored. Thus functions with very high spatial frequencies are invisible because of the limited resolution of the imaging process, but they are ignored since (by a familiar convention) we represent the specimen as a smooth density distribution. Other invisible functions can be eliminated because of extraneous information (e.g. they may lack the assumed symmetry of the specimen, or they may extend outside its boundaries). Finally, any remaining invisible functions must be eliminated experimentally. So we need new micrographs, taken under new conditions for which the invisible functions are different. And we need sufficient micrographs, so that no significant invisible functions are common to them all. For example, to avoid the invisible functions that give uniform projections, we record images of tilted specimens (DeRosier and Klug, 1968). Another class of invisible functions is associated with defocus. At any given state of focus and astigmatism, there will be a class of functions yielding an image that—though not exactly zero—is weaker than the noise level of the micrograph. Such functions are, in practice, invisible. (Clearly, the higher the noise level, the larger is this class.) Again, we need additional micrographs, this time with different defocus (Schiske, 1968).

In general, therefore, "irreversible" blurrings of the image result from the existence of significant invisible functions; but they can be eliminated by additional micrographs with blurrings that are of a similar type, though different in detail. We discuss the example of defocus in Section 7.5.3(b), and that of projection in Section 7.6.

7.5.2 Optical filtering

(a) Principles

Fourier filtering has long been used in various ways to enhance ordinary photographs (see, for example, the review by Birch, 1972). Successful applications to electron microscopy (which are much more recent) have been restricted to structures with linear periodicity (one- or two-dimensional). They have the effect of fitting to the (approximately periodic) image a set of sine/cosine functions with the same periodicity. This fit is an approximation since, to make it reliable, functions with too high a spatial frequency (i.e. those beyond the "resolution" of the image) are omitted. This is accomplished by finding the image's F.T., and selecting only those parts that lie both on the reciprocal lattice of the image's periodic lattice, and within the circle defining the effective resolution of the image. Then the best estimates for the amplitudes and phases of the corresponding Fourier components are used to reconstruct a best-fit ("filtered") image. (It is important that the Fourier components should be chosen objectively from the F.T., and not varied arbitrarily until the best-looking filtered image is obtained.)

This approach is implemented by two general methods. Optical filtering accomplishes the Fourier transformation in an analogue fashion, with a lens. Only the reciprocal lattice points are allowed to pass through a selective mask, and the filtered image is obtained by a second Fourier transformation with another lens. However, it is not easy to modify the phase of a Fourier component, nor to extract its amplitude at the exact

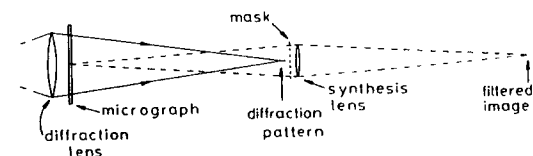


Fig. 7.71 The right half of an optical diffractometer (Fig. 7.49), adapted for optical filtering. Selected parts of the diffraction pattern are passed to the synthesis lens, which is positioned so that it could form a real image of the micrograph. At the position where that image would be formed, we have the filtered image.

position of the peak. Instead, portions of the transform around each reciprocal lattice point are passed to the second lens; so the filtered image is not exactly periodic.

The second method, numerical Fourier filtering, is described in the next section (7.5.3).

(b) Apparatus for optical analogue filtering

Early micrograph enhancement techniques, before computers became powerful and convenient, were almost exclusively analogue, and mostly based on the optical diffractometer (Section 7.3.2(b)), following Klug and DeRosier (1966). The optical diffractometer illustrated in Fig. 7.49 is easily adapted for filtering. The right half of the diffractometer, including the diffraction pattern, is shown in Fig. 7.71. We add a new lens (the "synthesis" lens) just after the diffraction pattern. The focal length of this lens is such that, if the micrograph on the left were illuminated with diffuse light, a real (inverted) image would be formed (using rays such as those indicated by the broken lines) on the extreme right of Fig. 7.71. This "filtered" image is made exclusively from light that had previously formed the optical diffraction pattern. It is the F.T. of the image's F.T. If the optical diffraction pattern is modified or trimmed, the filtered image will register the changes.

The arrangement shown in Fig. 7.71 is the simplest possible one (using just one lens to make the diffraction pattern, and another to re-form the image). More complicated arrangements are possible, and they can be "folded" with mirrors (though it should be noted that each reflection reverses the hand of the image).

The practice of optical filtering consists in constructing the correct mask, positioning it accurately at the diffraction pattern, and recording the filtered image. If the mask is transparent, the phases of the pattern will usually be changed. This is difficult to accomplish in a reliable way, and yet any change in these phases profoundly affects the filtered image. So it is easier to avoid phase artefacts by using an opaque mask containing holes which pass selected parts of the optical diffraction pattern. Such a mask is an amplitude filter in which the amount of light transmitted is not continuously variable, but must be either 0% or 100%.* (Since there are only two possible values for the transmittance, it is known as a binary filter.) Despite their obvious limitations, binary amplitude filters have proved very useful for processing electron micrographs. Technical details are described in reviews (Markham, 1968; Johansen, 1975; Erickson *et al.*, 1978)

* Variable transmittance filters, suitable for broader diffraction spots, are fine gauzes (e.g. electron microscope grids); Klug and DeRosier (1966).

(c) *Enhancing periodic features in the micrograph*

General. The binary amplitude filter is useful for clarifying electron microscope images if the structure of interest gives rise, in the optical diffraction pattern, to sharp spots or lines. A binary filter that transmits only those spots or lines will virtually eliminate, in the filtered image, irrelevant structures from the original micrograph. The efficiency of this filtering process will be greatest when as much as possible of the diffraction pattern is removed, without touching the part that conveys the picture information. (In that case the averaging will cover the largest number of unit cells.) Now all the picture information is in the diffraction spots or lines, so the filtering efficiency is greatest when these spots or lines have the smallest possible total area. They should therefore be as small and as few as possible. They will be small if the lattice is large and free of distortion; they will be few if the unit cell is small. Thus efficient filtering is promoted by a large, undistorted lattice with small unit cells. This implies that the number of unit cells in the lattice is a maximum (which also gives the maximum redundancy). Obviously this number is likely to be much greater with a two-dimensional repeat (as in a cell wall), than with a one-dimensional one (as in a helix).

Superposed periodic structures. Image imperfections are usually noise, but images of superposed structures can also be removed by filtering. The superposed structures can even be periodic. However, filtering is interfered with if some of the diffraction spots of the two periodic structures coincide. Such pairs of "common spots" cannot be separated by optical filtering. They can be omitted only by restricting the resolution of the filtered image. (Transmitting them with diminished strength is usually not satisfactory as the phases will probably be wrong.)

Diffraction spots are unlikely to coincide exactly unless the superposed sheets have the same structure. This occurs when a flat periodic sheet (e.g. a cell wall) is folded over (Section 7.3.3(b)), and the diffraction pattern is approximately the superposition of two mirror-related copies of the diffraction pattern of one sheet (Fig. 7.51). If any diffraction spots from the two copies coincide, the "upper" and "lower" diffraction spots can be regarded as belonging to a single, new lattice. This (fine) lattice is the reciprocal lattice of a large real lattice, the moiré pattern formed by the two sheets. If each unit cell of the moiré lattice contains u unit cells of the sheet, then there are only u different superpositions of unit cells with each other. When u is small, it is clear that the superposition pattern does not contain enough data to calculate the correct image of a unit cell, either by optical filtering or by any other technique. The coincidence of spots in the diffraction pattern is simply the way in which this limitation reveals itself in the Fourier transform.

This problem arises particularly when the two sheets form the two surfaces of a flattened cylinder, in which the subunits were originally arranged with helical symmetry. That symmetry determines the mutual orientation of the lattice directions in the two surfaces. In all flattened helices there must be a repeat vector along the equator (the "circumferential vector"), otherwise the helix will have a seam. Remembering that the reciprocal lattice has the shape of the real lattice rotated through a right-angle (Fig. 7.21), this means that there will be two superposed spots on the meridian (Z -axis) of the diffraction pattern (see, for example, Figs 7.53b, c). But optical filtering is nevertheless possible. The two diffraction spots on the Z -axis are really the same spot, deriving from

different sheets. (The Z -axis is the mirror-line relating the "upper" and "lower" diffraction patterns.) In this case, therefore, it is satisfactory to halve the amplitude of the superposed spots. Furthermore, the subunits sometimes aggregate to form different helices (as in the case of polyheads of phage T4: Yanagida *et al.*, 1970), which provide further data for disentangling the contributions of the two surfaces.

A more serious problem arises if the flattened helix has an exact repeat. This causes coincidence of lattice points in the direction of the helix axis (just as the circumferential vector causes their superposition in the perpendicular direction). These two independent superpositions define the moiré lattice of the helix, and hence the resolution of the filtered image. This restriction again expresses the fact that there is only a limited number of different superpositions of the subunits of the two sheets. The larger the repeat distance of the helix, the more different superposition patterns are available, and the higher the resolution to which the image can be filtered. (We shall encounter a very similar situation when we consider the problem of reconstructing the three-dimensional density distribution of a helical particle; Section 7.6.3(c).)

A further problem in filtering a flattened cylindrical (helical) particle arises from its restricted width. This causes the spots to broaden in the equatorial direction, so that they cannot always be separated cleanly from each other by optical filtering. Even when they can, the filtering is still less efficient than in the case of large flat sheets, since it is necessary to use larger mask apertures (slits instead of holes).

Despite the development of computer methods, analogue optical filtering can still be a most useful technique. However, the lattice should not be too distorted, so that the variations in different parts of the filtered image do not obscure the average picture. (Otherwise, numerical correlation methods will be needed to give a reliable average; see Section 7.4.4 and p. 254.)

7.5.3 Numerical filtering

The numerical F.T. gives us, at each reciprocal lattice point, not only the amplitude but also the phase. So we now have the possibility of modifying the phase. This permits new kinds of filtering, e.g. to correct for imaging defects (Section 7.5.3(b)).

After correction, we have an amplitude and phase at every observed reciprocal lattice point. These data are coefficients in a Fourier series, from which a periodic filtered image is calculated. By sampling the F.T. at precise reciprocal lattice points, this "crystallographic" method causes the filtered image to be the convolution of the original picture with its periodic lattice. That convolution superposes infinitely many different copies of the original picture, each copy being positioned at one of the points of the picture's lattice. Such a superposition can be approximated by multiple photographic copying (Markham *et al.*, 1964; Horne and Markham, 1972). Analogue optical filtering, in which the mask apertures are of finite size, is equivalent to multiple photographic copying with variable exposure times (Smith and Aebi, 1973).

(a) *Two-dimensional lattices*

Averaging: perfect lattices. The basic principles of Fourier filtering are much the same, whether it is done by numerical or by optical techniques. There are, however, certain

differences, conditioned partly by the needs, and partly by the opportunities, of numerical image processing.

We start by considering those differences imposed by its needs. The microdensitometer (Section 7.4.2(a)) measures the transmittance (transmitted intensity/incident intensity) at sampling points that are arranged on a regular lattice (the sampling lattice). This causes two differences with optical filtering (Aebi *et al.*, 1973). First, computer filtering can pre-process the transmittances to obtain numbers that are proportional to the intensity of electrons at that point of the image. This is not possible with optical filtering, in which non-linear effects can give rise to extra diffraction spots (which fit the reciprocal moiré lattice, if two sheets are superposed).

The second difference is much more important. The F.F.T. is necessarily calculated on a lattice that fits the reciprocal of the densitometer's sampling lattice. But that lattice will generally bear no simple relationship to the lattice of the periodic structure in the picture. Yet the F.T. values are required at the reciprocal lattice of the periodic structure. Because these reciprocal lattices are usually incommensurate, we shall not have exactly the values required for filtering. There are several ways to solve this problem. The filtering mask can, as with analogue (optical) filtering, be large enough to enclose all the relevant F.T. points. Alternatively, the original density data can be interpolated onto a new sampling lattice that is commensurate with the lattice of the periodic structure. (Interpolation methods are discussed in Section 7.6.4.) However, before any of these methods are applied, the lattice parameters should be refined by a least-squares method.

For further technical details associated with calculating the F.T.s of "perfect" lattices, see the detailed review by Amos *et al.* (1982).

Even though filtering is often merely an intermediate stage in three-dimensional reconstruction, it can be useful by itself, provided there is significant information in a two-dimensional image of the specimen. This holds, for example, for extremely thin sections of complex lattice structures such as muscle (Reedy and Reedy, 1985). It also holds for the details of subunit clustering in flattened helices, e.g. of "giant" bacteriophage heads (Yanagida, 1977). But it even holds for some undistorted helices that might seem to need three-dimensional reconstruction. In the tubular variants of papilloma-polyoma viruses, optical filtering revealed that the unit is a pentamer, both in the more obvious "pentamer" tubes (Kiselev and Klug, 1969) and (more surprisingly) in the "hexamer" tubes (Baker *et al.*, 1983). Filtering has also proved particularly useful for identifying different structures through their capacity to bind specific macromolecules such as *Fab* fragments. Since identification depends on the difference between two images, it is best accomplished by numerical methods. Moreover, it is necessary to use statistical techniques (Section 7.5.3(d)) to check the significance of the results.

Averaging: distorted lattices. Straightforward numerical Fourier filtering works satisfactorily only if the periodic lattice is relatively undistorted. Otherwise, a correlation method (Section 7.4.4) is more appropriate. If the distortion is continuous, the distortion vector can be determined for each point. Next, using bilinear interpolation, the density can be calculated at each point on an undistorted lattice. We then have a choice of filtering method. We can simply average the densities of the different (corrected) unit cells. Alternatively, we can use Fourier filtering, in which case we should calculate the F.T. of the corrected (interpolated) density distribution, and then proceed as for perfect lattices.

In some cases, however, the lattice distortions may be accompanied by systematic changes in the structure (i.e. in the unit cell contents). Suppose, for example, that the distortion is actually three-dimensional. The two components in the micrograph plane give the two-dimensional distortion discussed in Section 7.4.4; but the third component tilts the molecules and alters its projection. If such a micrograph is corrected for two-dimensional distortion and then averaged, the average will have degraded resolution through the incorporation of many slightly different projections. We need a way to distinguish the different projections, so that only those of the same type are averaged together.

A suitable technique has been developed by Frank *et al.* (1988a). The lattice is first corrected for distortion by the method of Henderson *et al.* (1986) (see Section 7.4.4(b)). Then small patches (20 × 20 unit cells) are averaged, and the resulting "patch averages" are subjected to correspondence analysis (Section 7.6.6(a)). This reveals the characteristic image types present, indicates whether they are interconvertible by continuous movements, and gives their locations in the corrected lattice. Such an approach is quite general but (as we have seen in other cases) generality must be paid for by a reduced signal-to-noise ratio. Perhaps, if the changes in appearance are caused by distortion-induced tilts, it might be more efficient to express the two tilt angles as a simple function of the lattice coordinates, and to determine that function by some minimization procedure.

(b) Correcting for defocus and astigmatism

The aberration phase shift. Numerical filtering makes it possible to correct for the state of focus and astigmatism in the image (Chapter 4). If the specimen is of low contrast, correction for imaging defects is simple enough to be performed by filtering (Erickson and Klug, 1971; Unwin and Henderson, 1975).

Let us suppose that, in a perfectly focused and corrected electron microscope, the electron wave at the image plane is (at a point specified by the vector \mathbf{x})

$$A(\mathbf{x}) = A_0 + \Delta A(\mathbf{x})$$

Here A_0 is a strong, constant wave and $\Delta A(\mathbf{x})$ is a small, varying wave that carries all the picture information. Simple Fourier imaging theory (Chapter 4; Goodman, 1968) shows that the electron wave at the back focal plane of the objective lens* is

$$\text{F.T.}[A(\mathbf{x})] = \text{F.T.}[A_0] + \text{F.T.}[\Delta A(\mathbf{x})]$$

Since A_0 is a constant, its F.T. is a peak at the origin, i.e. on the optical axis of the microscope. Around that axis, at the back focal plane, is $\text{F.T.}[\Delta A(\mathbf{x})]$. Thus the image, and the amplitude distribution around the back focal plane, are (when suitably scaled) simply F.T.s of each other.

Now suppose that the image has been changed by imaging defects, e.g. defocus or astigmatism. These can be supposed to affect the imaging process by changing the phases at the back focal plane. Consider a point at an angle α at the back focal plane (where α is a vector, since the angle is measured in a particular direction). Then the phase change is called the *aberration phase shift* $\chi(\alpha)$. Express α in terms of polar coordinates:

* Strictly, this is at the *reference sphere* (Goodman, 1968; Born and Wolf, 1980).

$|\alpha|$ and an azimuthal angle. $\chi(\alpha)$ can be developed as a power series in $|\alpha|$, which is useful since $|\alpha|$ is very small in an electron microscope. Then the different coefficients (which will be functions of the azimuthal angle) have different meanings. The coefficient of $|\alpha|$ refers to a simple translation of the image, and can be set at zero. The coefficient of $|\alpha|^2$ determines the state of focus and astigmatism. The coefficient of $|\alpha|^3$ determines the aberration coma, caused by residual misalignment of the microscope, and now realized to be significant at high resolution (Henderson *et al.*, 1986). The coefficient of $|\alpha|^4$ concerns the spherical aberration of (mostly) the objective lens.

Correcting imaging defects: the general problem. Correcting image defects is simple in principle. First we need to find F.T. $[\Delta A(x)]$ and the $\chi(\alpha)$ that applied when the micrograph was taken. Then we subtract $\chi(\alpha)$ from the phase of F.T. $[\Delta A(x)]$, to obtain the value of the F.T. under ideal imaging conditions. Finally, we inverse Fourier-transform it to obtain the "perfect" image. The problem is to find F.T. $[\Delta A(x)]$ and $\chi(\alpha)$ from the micrograph.

Start by considering F.T. $[\Delta A(x)]$. The micrograph records the probability distribution of electrons, i.e. the intensity of the electron wave function. This is

$$\begin{aligned} |A(x)|^2 &= [A_0 + \Delta A(x)][A_0^* + \Delta A^*(x)] \\ &= |A_0|^2 + A_0^* \Delta A(x) + A_0 \Delta A^*(x) + |\Delta A(x)|^2 \end{aligned}$$

Since we can choose the real axis to coincide with A_0 , this is

$$A_0^2 + A_0[\Delta A(x) + \Delta A^*(x)] + |\Delta A(x)|^2$$

If the specimen has very low contrast, we can neglect the last term, obtaining

$$|A(x)|^2 \doteq A_0^2 + 2A_0 \operatorname{Re}[\Delta A(x)]$$

We therefore face the problem that the micrograph gives us, not the required $\Delta A(x)$, but only its real part (symbolized by "Re").

Correcting imaging defects: pure phase objects. The solution is relatively simple if we can make the further assumption that the specimen is a pure phase object. (Actually, this is about 90% true of a specimen composed entirely of light atoms, and about 60% true of one composed entirely of heavy atoms.) Making this assumption, the (h, k) diffracted beam in the back focal plane will have the amplitude $i|F(h, k)| \exp[i\beta(h, k)]$, while the $(-h, -k)$ diffracted beam will have the amplitude $i|F(h, k)| \exp[-i\beta(h, k)]$. (Here $\beta(h, k)$ is the phase of the (h, k) reflection in the F.T. of the electron wave transmitted by the specimen, and the amplitude is a pure imaginary number since the specimen is supposed to be a weak pure phase object.) At the back focal plane, each beam will receive an aberration phase shift $\chi(h, k)$. For the most important aberrations (defocus, astigmatism and spherical aberration), $\chi(h, k)$ depends on an even power of $|\alpha|$, i.e. of $(h^2 + k^2)^{\frac{1}{2}}$. For these aberrations, $\chi(h, k) = \chi(-h, -k)$. Making this third assumption, the diffracted beams are

$$i \exp[i\chi(h, k)] |F(h, k)| \exp[i\beta(h, k)] \quad \text{and} \quad i \exp[i\chi(h, k)] |F(h, k)| \exp[-i\beta(h, k)]$$

Together, they generate at the image a wave of amplitude

$$\begin{aligned} & i \exp[i\chi(h, k)] |F(h, k)| \{ \exp[i\beta(h, k)] + \exp[-i\beta(h, k)] \} \\ &= 2i \exp[i\chi(h, k)] |F(h, k)| \cos[\beta(h, k)] \end{aligned}$$

As we have seen, however, the micrograph preserves only the real part of this wave, i.e.

$$-2 \sin[\chi(h, k)] |F(h, k)| \cos[\beta(h, k)]$$

$-2 \sin[\chi(h, k)]$ is called the "phase contrast transfer function".

To correct for imaging defects in this simple case, we have only to divide the calculated diffraction amplitudes by the phase contrast transfer function. This function can be estimated from the Thon rings in the diffraction pattern (Thon, 1966; Chapter 4). Its general shape, shown in Fig. 7.72, is (of course) mostly determined by the state of focus. When the defocus is big enough to affect the image seriously, it changes the sign of the outer diffraction spots. Division by the phase contrast transfer function restores the correct sign—the most important correction. However, this division is difficult to accomplish where the phase contrast transfer function is very small. Then we may need to combine data from several different micrographs with different focal states (Section 7.5.1(b)).

When our simplifying assumptions do not apply, the imaging correction is more complex; see Erickson (1973).

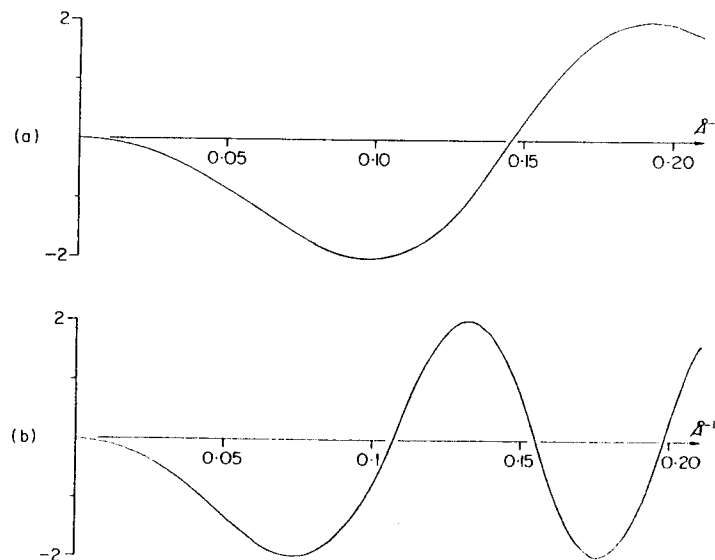


Fig. 7.72 The phase contrast transfer function appropriate to a Philips EM300 operated at 100 kV. It is plotted for two states of underfocus: (a) 1500 Å; (b) 2500 Å. (The first plot is adapted, and reprinted with permission, from *Progress in Biophysics and Molecular Biology*, Volume 39, L.A. Amos, R. Henderson and P. N. T. Unwin, "Three dimensional structure determination by electron microscopy". Copyright (1982), Pergamon Press PLC.)

(c) Structures with rotational symmetry

Fourier filtering applies, in principle, to rotational as well as to translational symmetry. Since optical diffraction works only for translational symmetry, however, rotational filtering must be performed by numerical techniques.

As usual, the first step is to find the symmetry; for rotational symmetry, this means the order (N) of the rotation axis. That can be determined by the methods outlined in Section 7.4.6. Those methods also give us the rotational Fourier components in either real or reciprocal space (i.e. either $g_n(r)$ or $G_n(R)$). Of these two quantities, the $G_n(R)$ would seem to be the more useful for rotational filtering. This is because they are easily terminated at a value of R corresponding to the resolution of the particle; we can then obtain a best-fit to density-waves in R at the same time as the rotational best-fit filtering. That filtering is performed in essentially the same way as translational optical filtering. Analogously to passing just the Fourier components corresponding to reciprocal lattice points, we now accept only the $G_n(R)$ for which n is a multiple of N . And whereas translational filtering eliminates all Fourier components whose spatial frequency exceed the picture's resolution, we now reject all multiples of n exceeding some upper limit m . m is found by examining the rotational power spectrum (Fig. 7.67); it corresponds to the highest order of any peak above the noise level. Having chosen the range of R and the required orders (n) of the rotational Fourier components $G_n(R)$, we substitute them into Equation (40) to calculate the filtered image.

Rotational filtering has not proved quite as useful as conventional translational filtering. Aggregates of subunits seldom have rotation axes of very high order; the largest appears to be 17-fold, for the disc aggregate of tobacco mosaic virus protein (rotational filtering: Crowther and Amos, 1971). This is far fewer than the number of unit cells in a typical translational filtering, so the improvement in signal-to-noise ratio is more modest. Nevertheless, even when the rotation axis is only six-fold (as with the T4 bacteriophage base-plate) filtering can yield sufficient improvement to reveal the location of protein components removed by mutations (Crowther *et al.*, 1977).

(d) Statistical aspects of filtering

Improvement and reliability of filtered images. As we have seen, filtering is roughly equivalent to averaging. Obviously the improvement resulting from averaging is the greater, the more images are averaged together. But exactly how big is the improvement in the signal-to-noise ratio?

The improvement factor is easily derived if the micrograph noise is "additive noise" (Section 7.3.3(a)), such as that caused by irregularities in the supporting film, or by "fog". In this case, the micrograph can be considered to be a perfect image to which noise has been added. Then the "signal" component of each micrograph is the same, so averaging n micrographs increases it by n . However, the "noise" components are uncorrelated, so averaging n micrographs increases the noise by only \sqrt{n} . Consequently, the signal-to-noise ratio is increased by $n/(\sqrt{n}) = \sqrt{n}$. Alternatively, the micrograph noise may be "multiplicative noise", such as that caused by using a very low electron dose, i.e. "shot noise"; then the electrons will have a Poisson distribution, which has a standard deviation of $\sqrt{(\text{number of electrons})}$. (The noise from irregularities in the

negative stain may fit this model approximately.) Averaging n micrographs multiplies the number of electrons by n , and therefore multiplies the standard deviation by \sqrt{n} . As before, the signal is multiplied by n , so the signal-to-noise ratio is increased by $n/(\sqrt{n}) = \sqrt{n}$. We thus find the same \sqrt{n} improvement in the signal-to-noise ratio, for both major sources of noise.

The improvement in the filtered image depends on removing light that corresponds to unwanted (e.g. noisy) features in the micrograph. Consequently, the improvement can be assessed (Aebi *et al.*, 1973; Smith *et al.*, 1976) by determining the fraction of the input light intensity that was lost by filtration. (When numerical filtration is used, one calculates the "power", a quantity proportional to the light intensity in an optical diffraction pattern.) The improvement factor is the fraction of the input intensity or power lost. This factor is used to assess the relative contributions of different kinds of averaging, e.g. of translational versus rotational averaging for a two-dimensional crystal. However, although a large power loss means that the micrograph has undergone a substantial amount of filtering (and, hopefully, improvement), it also implies that the micrograph was very noisy to begin with. Therefore power losses are most significant when used to compare the filterings of micrographs of similar initial quality.

Before features in a filtered image can be interpreted, it is necessary to determine their reliability. The variance $\sigma(\mathbf{x})^2$ of the averaged density $\langle \rho(\mathbf{x}) \rangle$ (at the point \mathbf{x} in the image) can be calculated from

$$\sigma(\mathbf{x})^2 = \langle \rho(\mathbf{x})^2 \rangle - \langle \rho(\mathbf{x}) \rangle^2 \quad (56)$$

If the averaging involves many different micrographs, the mean square density $\langle \rho(\mathbf{x})^2 \rangle$ could be calculated by Fourier-transforming the variance-covariance matrix* of the diffraction amplitudes:

$$\langle \rho(\mathbf{x})^2 \rangle = \sum_{\mathbf{h}} \sum_{\mathbf{h}'} \langle F(\mathbf{h}) \rangle \langle F(\mathbf{h}') \rangle \exp \{ -2\pi i \mathbf{x} \cdot (\mathbf{h} + \mathbf{h}') \} \quad (57)$$

Having calculated the variance, the reliability can be found from Student's t -test (Milligan and Flicker, 1987). $t(\mathbf{x})$ is calculated from

$$t(\mathbf{x}) = |\langle \rho(\mathbf{x}) \rangle| \sqrt{n} / \sigma(\mathbf{x}) \quad (58)$$

where n is the number of images averaged. $t(\mathbf{x})$ can be converted into probabilities using tables or various approximations (see, for example, Chapter 26 of Abramowitz and Stegun, 1964, *et seq.*). Contours of this probability can then be plotted (Trachtenberg and DeRosier, 1987). Unfortunately, there is some arbitrariness deriving from which optical density level is shown to be zero. (This might be the average density of the particle's perimeter, as in "floating"—see Section 7.4.2(b).) This arbitrariness does not apply, however, when difference images are calculated, since the "zero" level is the same for both images. Difference images are particularly susceptible to noise (since, although the densities are subtracted, their variances are added), so statistical tests are necessary in order to decide whether features in the difference map are significant.

* The errors in different Fourier coefficients may be largely uncorrelated, in which case this matrix would be nearly diagonal, so that Equation (57) could be approximately replaced by a single summation.

Maximum entropy. The maximum entropy method has not yet been used for processing electron micrographs; but such applications are to be expected soon, because of its value in de-blurring photographs, generating tomographic images from medical scanners, etc. (see Gull and Skilling, 1984, for a brief review). Moreover, although some aspects of the method remain unacceptable to many statisticians, the success of its practical applications suggests that, at least, it contains an important core of truth.

Maximum entropy (Jaynes, 1979; Skilling, 1984) is a development of probability theory originally applied to statistical mechanics. Probability, though often defined in terms of the frequency of an event in an infinite number of trials, is more usefully taken to represent our current state of knowledge. In estimating the probability that a given horse will win a race, for example, infinite series are irrelevant. If no information is available about the horses, then it is arbitrary not to give them equal odds. When some information is acquired, the odds should still remain as unbiased as is consistent with this information. According to the maximum entropy principle, this "most unbiased" distribution of probabilities is the one that can be realized in the greatest number of ways, consistent with the available information. That is, it is the distribution with the maximum entropy.

What is the relevance of this to de-blurring an image? Suppose, for instance, we know nothing about a micrograph except the total number (n) of electrons that were used to form it. Lacking any other knowledge, we have no restrictions on how these n electrons are to be distributed among the pixels. However, they would maximize their entropy by distributing themselves uniformly. In general, we need to assign a probability p_i that an electron would have reached the i th pixel. The maximum entropy method finds all the p_i by maximizing the picture's entropy $S = -\sum p_i \log(p_i)$, subject to constraints imposed by the blurred image data. Thus maximum entropy is another optimization technique, and makes correspondingly heavy demands on computer time.

7.6 THREE-DIMENSIONAL RECONSTRUCTION FROM IMAGES

The ultimate purpose of microscopy is to find the three-dimensional structure of the specimen. Starting from electron micrographs, that goal can be approached in a relatively straightforward manner, because of two factors. First, the severe aberrations of all electron lenses make it necessary to restrict their apertures to an angle very much smaller than that used in light microscopy. The depth of field is therefore considerable at high magnifications. All levels of an ordinary, thin, untilted biological specimen are simultaneously in focus, at least to within their effective resolution (usually no better than 15 Å). Second, many specimens for electron microscopy have very low contrast. This is most obviously true for unstained specimens, but it also applies to those areas of stained specimens that give high-resolution images. For the relevant changes in specimen contrast (either amplitude or phase) are then so small that they are effectively a linear function of the specimen mass thickness.

Because of these two factors, the image approximates to a projection of the density of scattering matter in the object. But there are substantial problems in interpreting these projections, even when we are presented with a range of views that is sufficient to lead to a unique solution. Everyday life presents us with few situations where we see the projection of an object, so that our visual system has had little need to evolve or acquire much facility in interpreting such data. Consequently, the intuitive methods that

were first tried with high-resolution micrographs encountered difficulties. It is instructive to look briefly at these, before discussing their present solutions.

7.6.1 Limitations of intuitive methods

(a) Stereo viewing

The most natural method uses our visual system, which is designed to perform a kind of three-dimensional reconstruction through stereoscopic vision. This works quite well in interpreting the stereo-micrographs of metal-shadowed surfaces, and other clear structures, at relatively low magnification (King, 1981). However, high-magnification stereo images of structures, showing detail near the microscope's resolution limit, produce an unsatisfactory and confusing effect. Partly this is because the finest detail, consisting mostly of imaging artifacts, is often uncorrelated in the two images. But it is easily shown that stereo-vision is certain to fail for many objects. For, if the depth coordinates of any three-dimensional structure could be found from one pair of stereo-photographs alone, then two projections of the structure would suffice to give the coordinates of all its points. But this is not possible if two or more isolated points lie on any plane perpendicular to the tilt axis. This is illustrated in Fig. 7.73: the different arrangements of points in (a) and (b) give the same pair of projections.

Though theoretically impossible, stereoscopic vision nevertheless works well in ordinary circumstances because it is used for interpreting a very restricted class of objects. These are groups of surfaces, like stage scenery, each possessing both continuity and texture. Because of the continuity, the depth coordinate changes only slowly and continuously on any given surface. Because each surface also has a texture, the two images of that surface can be brought into register, by a process akin to finding the X.C.F., in order to determine their depth coordinates. The shadowed surfaces given by

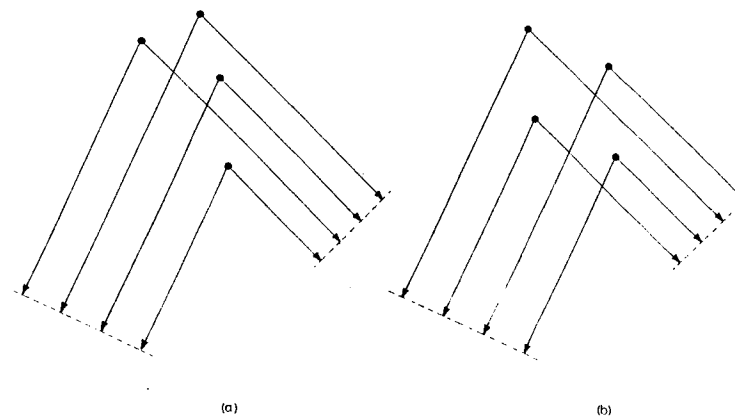


Fig. 7.73 The limitations of stereo-viewing as a technique for finding the depth coordinates of isolated points. The broken lines show the projections seen by the two eyes. (a) and (b) represent alternative interpretations of these projections, both satisfying them.

freeze-etch microscopy approximate fairly well to this type of structure, so they are satisfactory objects for stereo-viewing. (The vertical coordinates of freeze-etch surfaces could presumably be determined by the standard techniques of photogrammetry.) But high-resolution electron micrographs do not belong to this restricted class, and therefore require quite different techniques for their interpretation.

(b) *Model building*

The model-building approach is the next most natural one. Micrographs showing many different views of an identical novel particle present us with a confusing variety of projections. To find a single three-dimensional structure, that explains all of them, presents us with the challenge of a puzzle.

When we have accumulated galleries of characteristic views, each of known orientation, this puzzle must be faced. As with all puzzles, we try first to think of even a plausible solution. What type of structure might a macromolecular particle possess? Such particles are usually composed of subunits that are frequently arranged in a symmetrical way (Section 7.2.1). Symmetry profoundly affects the three-dimensional shape, and in a predictable way, so it should be determined first. (The procedure for doing so was outlined in Section 7.4.7.) Model-building has been particularly successful where it has had symmetry to help it, e.g. with icosahedral viruses (Finch and Klug, 1965, 1967), or clathrin coats (Crowther *et al.*, 1976).

If there is point-group symmetry of a high order, its deduction will probably account for most features of the characteristic projections, for the following reason. The volume of the repeating (or asymmetric) unit, equal to the particle volume divided by the point-group order, should be compared with the volume of a "resolution cell". This defines that volume of the specimen which is smeared, by defocussing and other imaging artifacts, into a uniform "blob". The number of resolution cells in the asymmetric unit measures the residual information in the image. If this is small, it may be feasible to calculate, display and compare all possible appearances of the particle in some orientation, and thereby to determine its structure without resorting to three-dimensional reconstruction (which will be described in Sections 7.6.2 and 7.6.3).

The asymmetric unit may be large, however; indeed, the particle may lack all symmetry, as in the case of the ribosome. Then an attempt to guess the particle structure faces the problem that, in theory, an astronomical number of structures must be tested against the characteristic views. It is far quicker to calculate the structure directly by three-dimensional reconstruction, if those views are undistorted and their orientations are accurately known. If they are not, however, there may be no alternative to the older method of guessing possible structures and assessing them by comparing predicted with observed views. Then it is necessary to reduce to a manageable size the number of possible structures.

Two methods have been used, though usually without being made explicit. The first is to reduce the resolution of the model by considering only the broad features of the particle structure. This is quite valid in principle, but difficult to apply rigorously in practice. A low-resolution structure is a blurred object that does not lend itself to representation in a model. So we adopt the second method: we impose on the solution the property of forming a model; and we then discover that we have thereby reduced the number of possible structures. For acceptable models must consist of a region of

uniform density bounded by a smooth surface. Now we have only to determine the shape of that surface, which needs all the fewer parameters to specify, since the model's resolution has been lowered.

By these two drastic simplifications of the model, the problem is no longer one of determining a density distribution from projections (a problem to be considered in more detail in Sections 7.6.2 and 7.6.3). Instead, it is one of determining a surface from its shadows. Whether or not this can be done depends on the clarity of the micrographs, which will obviously be poorer in the case of smaller particles. If the "shadows" are clear, is the problem soluble? It would seem so, provided that the surface contains no hollows (which can never affect the shadows, and are therefore "invisible functions", as defined in Section 7.5.1(a)). Granted this condition, the convex surface could be found as follows. Back-projection (see Section 7.6.3 below) from each shadow generates a prism, and the superposed prisms of all the back-projected shadows contain the required surface within their common volume. We can estimate this surface from the smoothest approximation to that volume. (However, published models do sometimes contain concave areas, e.g. where three or more bulges join; but these seem to be required by the low resolution of the model, which imposes smoothness constraints on the surface.)

Through techniques such as these, the model-building technique has been able to obtain useful and reliable structural details of the main features of the ribosome (Lake, 1976).

7.6.2 Reconstruction from projections: the back-projection method

An electron micrograph is approximately a projection of the specimen in the direction of the electron beam. So the problem of finding the specimen structure from micrographs is (nearly) equivalent to the mathematical problem of reconstructing a density distribution from its projections. Model-building guesses a density distribution and tests it by calculating the projections. Projections are easily calculated from a density distribution: but how should we set about calculating a density distribution directly from its projections?

Since the whole problem stems from the microscope's large depth of field, the most direct solution would calculate the image to be expected from a microscope with a smaller depth of field. The depth of field is determined by the lens aperture. For consider the image formed by a lens of large aperture. Its depth of field is small; but, if the lens is covered by a card with a pinhole, the depth of field becomes very large: an object at any distance gives an equally sharp (pinhole) image. This is true irrespective of where the pinhole is placed, but different positions of the pinhole give different pinhole images (Fig. 7.74). When the entire lens aperture is used, the image is the sum of all the different pinhole images (ignoring diffraction effects). Figure 7.74 shows how, with three simultaneous pinholes, the three different pinhole images from any object combine exactly at only one distance. Objects (A, B or C) at different distances from the lens give pinhole images that superpose exactly at different field depths (A', B', C').

The narrow depth of field of a wide-aperture lens is thus a consequence of combining many different pinhole images. Each shows a projection of the scene from the viewpoint of the corresponding pinhole. So the wide-aperture lens is an analogue device for recombining many different projections to yield a three-dimensional image. It should

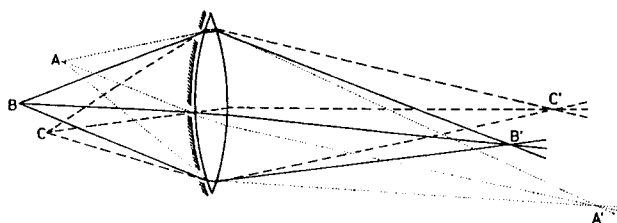


Fig. 7.74 A wide aperture lens has a narrow depth of field for each point of the object, and hence preserves depth information about the object's points (A, B, C) in the image (A', B', C'). The lens achieves this by superposing a series of different pinhole images; three such images are isolated by the mask covering the left of the lens.

be quite feasible to simulate this process using different electron micrograph projections. But one feature of each pinhole image is unnecessary: the divergence of the rays as they leave the pinhole. This divergence could be neglected if the image being reconstructed were very small. So the reconstruction could just as well be accomplished with parallel rays. That would give us the "back-projection" method.

Since this method is a calculation rather than an instrument, it can avoid the limited range of projections admitted by ordinary lenses (even by wide-aperture light-microscope objectives). So this calculation can produce a three-dimensional image that is superior to that of any light microscope. Nevertheless it is imperfect. This can be appreciated by considering the image it would give of a point object, were this reconstructed from projections over 180° (2π steradians). The incoming rays converging onto the image of the point would arrive from the surface of a hemisphere, and the outgoing rays would diverge to fill the other hemisphere. Thus the point-image would be surrounded by a multitude of radiating lines, like the lines of force around a point charge. The average density of these lines would vary inversely as the distance from the point-image ($1/r$ dependence). Every point of an object would be imaged in the same way, so the reconstructed image would (apart from magnification) be the object convoluted with a spherically symmetric $1/r$ density distribution (the "point spread function"). The relation of the reconstruction to the object would be that of an electrical or gravitational potential distribution to the assemblage of point charges or masses.

Are these imperfections serious? The point-spread function will not obscure the image of a point very much. Furthermore, a line of density will, after convolution, remain a line (though with a more diffuse cross-section). The images of distant points will be so blurred that they cause only slight interference. This explains the success of the light microscope in surveying the three-dimensional structure of thick Golgi sections containing a few, dark, widely separated cell processes. But more diffuse objects (such as most specimens for high-resolution electron microscopy) are poorly reconstructed.

Thus we see that even an optically perfect electron microscope would not give the best possible images, and some form of image processing is inescapable for determining the three-dimensional structure of the object. But it is quite easy to correct for the imperfections of back-projection. In the case of a tilt-series, the sections in Fourier space fan out, so that their separation increases proportionately to the radius R from their common intersection. Thus, for a back-projected image, the density in Fourier space drops off as $1/R$; so multiplying by R can restore the correct density. This "R-weighted

back-projection" method (Gilbert, 1972) is the most common method in computer-assisted tomography (Herman, 1980; Natterer, 1986). It is also the method of choice when reconstructing the three-dimensional structure of particles from tilt-series. It has been used, for example, with negatively-stained clathrin cages (Vigers *et al.*, 1986), and with sectioned ribo-nucleoprotein particles (Skoglund and Daneholt, 1986; Skoglund *et al.*, 1986).

There are many other ways of solving the projection equations, e.g. by iteration, by matrix inversion or by eigenfunction expansions. (See the books on computer-assisted tomography listed above.) However, in electron microscopy, the main alternative to R-weighted back-projection, as might be expected, also depends on Fourier transforms.

7.6.3 Reconstruction from projections: the Fourier method

(a) General principles

Fourier transform theory (Section 7.2.2) provides a simple approach to finding the three-dimensional structure of an object from its projections. We can calculate the required three-dimensional structure if we know its three-dimensional transform (Section 7.2.2(c)). But we can obtain this three-dimensional transform by using the projection rule (Section 7.2.2(d)). This connects the transform of the three-dimensional object with the transforms of its projections. These two-dimensional transforms are plane sections through the required three-dimensional transform. Each plane section passes through the centre of that transform, and is oriented parallel to the plane of the projection.

Thus the projection rule changes the problem of reconstructing the three-dimensional object from its *projections* into the intuitively much simpler problem of reconstructing a different three-dimensional structure (the transform) from its *sections*. This simpler problem is not however completely free of difficulties. The two-dimensional sections are infinitely thin, so it might appear impossible to construct a three-dimensional transform from any finite number of them. In fact it is possible, but only because the transform we are seeking has a texture composed of smoothly changing "regions", all of very roughly the same size and shape. This texture allows the transform to be interpolated from its values sampled at the points of some (appropriately fine) lattice. All the necessary sampled values of the three-dimensional transform can therefore be determined from two-dimensional sections, provided these are nowhere too far apart. Exactly how far apart they can be depends on the size and shape of the "regions" of the transform. And these, in turn, depend on the size and shape of the object whose three-dimensional transform we are attempting to reconstruct. This question needs more detailed consideration.

(b) How many projections are needed to reach a desired resolution?

If there is too big a gap between the two-dimensional sections, it must be narrowed by including additional sections, i.e. by obtaining additional projections. The effect of this additional information can be seen by considering a simple case. Suppose that all the projections are obtained by rotating the object, through equal angular increments, about a single axis (a "tilt series"). A section through the transform, cut perpendicular to this

axis, will look something like Fig. 7.75. This section contains numbered lines showing the positions of the transform sections obtained from projections. The separation between the lines naturally increases with the distance from the transform centre. So long as this separation is smaller than the size of the transform "regions", we have sufficient data to interpolate the transform. Beyond a certain radius, however, the separation of the lines becomes too great, and the transform can no longer be determined. Thus we can always determine the transform out to some maximum radius (determined by the size of the "regions" and the number of sections). Our calculation of the object is thereby restricted to using only that part of the transform within this radius. The size of this radius limits the resolution of our reconstruction, in which the finest spacing is roughly the reciprocal of this radius.

This approach allows us to obtain a useful rule of thumb relating the greatest dimension of the object (D), and the number of projections (n), to the resolution of the reconstruction (d) (Klug, 1971). For the maximum radius to which the transform can be determined is $1/d$ (Fig. 7.75). At this radius, the separation of the section planes (corresponding to projections) is approximately the radius ($= 1/d$) times the angle between the planes ($= \pi/n$). That separation must equal the size of a transform "region", which is the reciprocal of the object size (D). (This follows from the sampling theorem; see Section 7.2.2(h).) So we have $\pi/(nd) = 1/D$, showing that the resolution of the

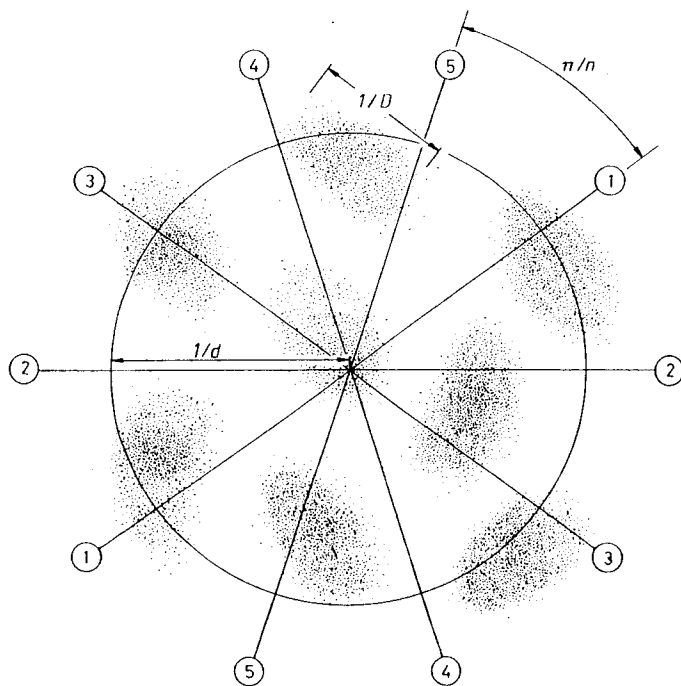


Fig. 7.75 Diagram for deducing the rule connecting the particle diameter (D) and the number of equally-spaced tilt images (n), with the resolution of the three-dimensional reconstruction (d). (See text.)

reconstruction $d = \pi D/n$. Alternatively, the number (n) of projections needed to reach some desired resolution (d) in the reconstruction is $n = \pi D/d$.

This formula applies only if the tilt-series are at equal angles. In practice, this can be difficult to achieve. When a grid is tilted by more than about 60° , various factors (e.g. a strong defocus gradient) start to limit the value of the images. A tilt-series is therefore likely to have a wedge-shaped gap, though it might be possible to fill this if the particle has rotational symmetry (with the rotation axis along the plane of the grid). The missing region's shape depends on the tilting scheme used. For conical tilt-series (Section 7.6.6(b)), the missing region fills a cone; indeed, the problem of the missing region is usually referred to as the "missing cone". Various schemes have been proposed for restoring the missing data, but there is really no alternative to measuring them, through particle symmetry, or through making specimens where the particles have different orientations. However, if the missing region is really a cone of rather small semi-angle, its volume can be quite a small fraction of the volume of the sphere (of radius equal to the reciprocal of the resolution). Consequently, the imperfection resulting from omission of the cone data may be tolerable.

(c) Two ways to obtain the projections

We still have the problem of obtaining these projections. There are two general ways of doing this. The direct way is to use a goniometer stage to obtain a "tilt-series" of micrographs. There are, however, severe imaging problems at high angles of tilt, so that part of the tilt-series will be missing. Moreover, radiation damage in the specimen accumulates with each successive micrograph, and soon reaches unacceptable levels. New specimens will be needed to complete the tilt-series. Despite these problems, tilt-series have usually given the best three-dimensional reconstructions. That is because tilting is the only method that will obtain the necessary projections from two-dimensional crystals. These contain more unit cells than any other type of electron microscope specimen, and consequently yield the highest signal-to-noise ratio. The techniques for analysing tilt-series have been extensively reviewed (Fuller, 1981; Amos *et al.*, 1982), and need not be described here.

The second general method avoids (at a price) the problems of tilt-series. A particle's rotational symmetry can give, in one picture, many different views of the repeating unit. This means that a single projection immediately yields the appearance of a number of symmetry-related projections, and several projections are obtained for the price of one. What sort of symmetry is best for this? Rotations (rather than translations) cause the object's projection to be repeated from different directions, so the symmetry group must contain either rotations or screw axes (which combine a rotation with a translation). In looking for symmetries that contain these, the plane- and space-groups are poor choices, since they restrict the highest permissible order of a rotation or screw axis to 6. We are left with point-groups or helical line-groups (Section 7.2.1). The largest number of different rotations in any useful point-group is present in the icosahedral group (order = 60; this group is of considerable practical importance, since it applies to most small viruses). But there are even more different rotations in those cases of helical symmetry where the repeat distance is long. (All the asymmetric units within a helix repeat must have different orientations, since any two units with identical orientations can differ only by a translation, i.e. the helix repeat.)

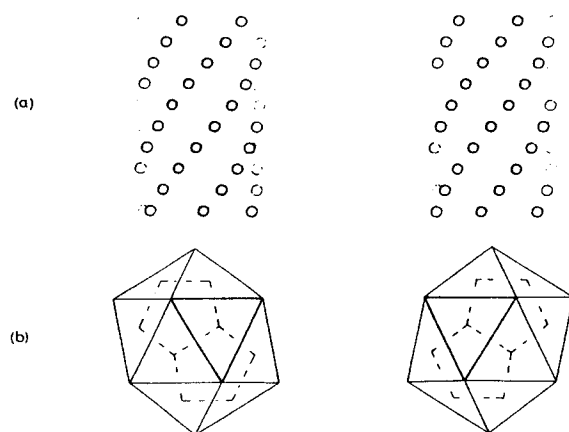


Fig. 7.76 Two stereo pairs illustrating how, for objects with helical or icosahedral symmetry, one projection is sufficient to convey depth information. (View the left image with the left eye, etc.) (a) From one view of a helical particle (top part of Fig. 7.28), a group of 10 annuli has been isolated, but the groups on the left and right differ by a displacement of one annulus. This generates stereo images because of the rotation by the twist angle Ω (about 17°) relating the displaced groups. (b) One view of an icosahedron (down a two-fold axis), and containing part of an inscribed dodecahedron, is repeated with a mirror plane. This generates a stereo image because the two copies represent views with a rotation of 36° . (Try fusing the two dark triangles.)

Thus the most favourable particle symmetries for three-dimensional image reconstruction are either icosahedral symmetry, or helical symmetry with a long repeat and (ideally) a parallel rotation axis as well. This can be illustrated as follows. Since one view of the particle has the same appearance as does a view taken from a nearby direction, one picture of a structure with either symmetry can be repeated exactly in such a way that the two (identical) images form a stereo pair (Fig. 7.76).

How is this particle symmetry utilized by the Fourier method? The high rotational symmetry of the particle is present also in the transform (rotation rule, Section 7.2.2(d)). One view gives us, not merely one section of the transform, but all the symmetry-related sections. In favourable cases, those sections go far towards filling up the transform.

d) Using the transform sections

Given the data, in the form of a tilt-series of the image of a symmetrical particle, there remain several tasks.

First, the transform must be calculated and corrected for defocus and distortions. Correcting distortions has been considered in Sections 7.4.4 (lattices) and 7.4.5(d) (helices). Correcting for defocus, etc. (i.e. for the phase contrast transfer function appropriate to the micrograph) was outlined in Section 7.5.3(b). This is a particularly important correction if high resolution (10 \AA or better) is to be attained. Extra information, from electron or X-ray diffraction, concerning the F.T. amplitudes helps to refine this correction.

By such means, a resolution of around $6\text{--}7 \text{ \AA}$ is obtainable from (sufficiently ordered) plane crystals, and better than 10 \AA has recently been obtained from helical tobacco mosaic virus particles (Jeng *et al.*, 1989). Though far short of atomic resolution, this can reveal the orientation of α -helices and other broad features of the domain structure. To approach atomic resolution, many new problems must be solved. Among these is the need to correct the phase contrast transfer function for coma (Henderson *et al.*, 1986).

The second task is to find the exact orientation of the projections (i.e. of the transform sections). For a tilt-series, the angles are given by the goniometer stage (or they can be deduced from the Thon rings given by the regions at the four corners of the micrograph), but particles used in different tilt-series must have their mutual orientations determined. For a symmetrical particle, this means determining the orientation of its symmetry axes (see Section 7.4(b),(c)). Each different particle orientation gives us a different section of the particle's transform.

The third task is to estimate this transform. We need its values at the points of a regular lattice, so that we can calculate an undistorted inverse; but very few of the desired lattice points will lie exactly on one of the observed transform sections. This problem receives detailed discussion in the next section (7.6.4).

Finally, the particle density must be calculated, and displayed; a few comments about display are added in Section 7.6.8(b).

7.6.4 Finding the three-dimensional F.T. from sections

(a) Reverse interpolation

To calculate the particle's density, we need its F.T. We shall invert this by an F.F.T. routine, obtaining (from this Fourier series) a periodic density distribution—copies of the particle arranged on a three-dimensional lattice. Neighbouring copies of the particle should not overlap; this implies that the F.T. must be sampled at a fine three-dimensional reciprocal lattice, whose size is prescribed by a three-dimensional version of the sampling theorem (Section 7.2.2(h)). Although we need the F.T. at these lattice points, we have experimental values for it only on certain central sections. It would seem, therefore, that we must calculate the values at the lattice points by some kind of interpolation. In effect, the sections present us with tabulated values of the F.T., at unequal intervals (and without differences). Although this is not a favourable situation for interpolation, there exist applicable methods (such as Aitken's method: Acton, 1970).

The difficulty with this approach to find the F.T. is that we cannot easily tell how accurate the results are. Obviously the accuracy will depend on the positions of our transform sections, but exactly how do we estimate it? How are we to be sure that we have enough section data to get a reliable particle density? It turns out (Crowther *et al.*, 1970b) that such questions can be answered more satisfactorily if we consider the *reverse* of our present problem. If we had found what we are now looking for (i.e. the values of the F.T. sampled at the lattice), then we could calculate our experimental data (i.e. F.T. values on the sections). For the lattice-sampled F.T. values suffice, by the sampling theorem, to calculate the F.T. anywhere (including at our experimental points on the sections). This calculation is an interpolation from data at equal intervals, and

it is quite reliable. Since, therefore, the reverse problem is soluble through interpolation, the original problem itself should be soluble by the reverse operation, i.e. by "reverse interpolation".

(b) *A simple example*

Exactly how this works is best appreciated from an example. Suppose we are trying to determine a one-dimensional F.T.; we need a set of equally spaced sampled values that we shall refer to as "determining values". These should be distinguished from the experimental F.T. values, from which we are trying to deduce them. To simplify the example, we suppose that the F.T. is symmetric about the origin (because the corresponding picture has a two-fold axis), and that the resolution of the picture is so low that only two "defining values" are required to reconstruct its F.T. If we knew these values, the reconstruction could be achieved as in Fig. 7.77a. The unit of X is the reciprocal of the picture's width, and the defining values (shown as vertical arrows) lie at exact multiples of these units. Each arrow-tip defines a sinc-function (Section 7.2.2(e)) with nodes where the other arrows must be positioned. To reconstruct the F.T., we simply add together all these sinc-functions, obtaining the curve in Fig. 7.77b. However, we do not yet know the "defining values", but only the experimental ones.

How much information does a knowledge of just *one* experimental value give us? The answer is shown in Fig. 7.77b and c. The heights of the required defining values are $F(0)$ and $F(1)$ [$= F(-1)$], so these are the axes of Fig. 7.77c. Knowledge of any experimental value allows us to plot a straight line on this diagram. If the experimental value were at the origin (i.e. point A in Fig. 7.77b), we should know $F(0)$ exactly, but have no information about $F(1)$; so the corresponding line in Fig. 7.77c is vertical. Similarly, if the experimental value were at $X = 1.0$ (point D in Fig. 7.77b), we should know $F(1)$, so line D in Fig. 7.77c is horizontal. Other points (B, C, E) in Fig. 7.77b give straight lines with appropriate gradients in Fig. 7.77c. The position of each point determines the gradient of the corresponding line. However, the line is not fixed completely until one of its intercepts is also known; this is found from the experimental value of the F.T. All the lines of Fig. 7.77c, if their experimental values are correctly measured, must intersect at one point, whose $F(0)$ and $F(1)$ coordinates give us the "defining values" from which the F.T. can be reconstructed.

In practice, of course, the experimental values will contain errors, which will affect their magnitudes rather than their positions. Consequently, although we shall still know the gradients of the lines accurately (from the positions of the experimental values), there will be errors in their intercepts (from the magnitudes). Instead of a single line, each experimental value now gives us a pair of parallel lines separated by the expected error. If we try to calculate $F(0)$ and $F(1)$ from two such lines, we shall encounter the situation shown in Fig. 7.77d. The intersection of the two sets of parallel lines gives us an error ellipse. (All points on this ellipse have the same expected error.) We see that the error is much greater in one direction than in another. These directions are simply the directions of the ellipse's axes. Thus the information contained in the error ellipse can be more briefly summarized by a pair of perpendicular error bars, tilted to lie parallel to the ellipse's axes (Fig. 7.77e).

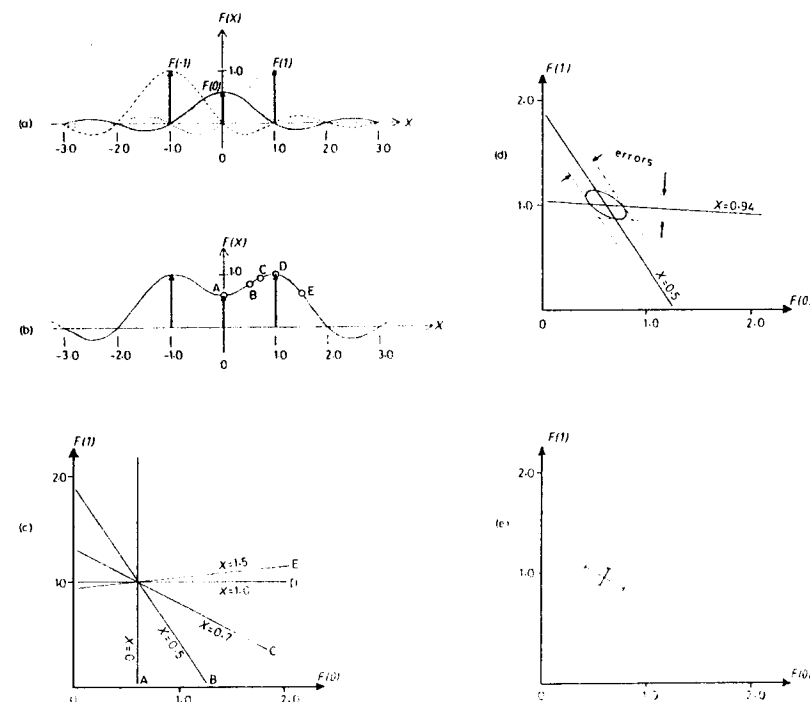


Fig. 7.77 A series of diagrams to explain how an extremely simple F.T. can be deduced from two experimental sampled values, and how the accuracy depends on their positions. (a) The F.T. can be reconstructed from three defining values, $F(-1)$, $F(0)$ and $F(1)$; but only two values are needed here, as $F(-1) = F(1)$. (b) The F.T. is sampled at different points (A, B, C, D, E). (c) The information concerning $F(0)$ and $F(1)$ which is derived from each of these experimental values. Lines A, ... refer to the corresponding points in (b). The correct values of $F(0)$ and $F(1)$ are obtained at the intersection of all the lines. (d) Trying to reconstruct the F.T. from two experimental values, $X = 0.5$ (point B in (a)), and $X = 0.94$ (a point between C and D in (a)). Errors in measuring the F.T. at these points lead to an ellipse of uncertainty in determining $F(0)$ and $F(1)$. (e) The ellipse of uncertainty can be replaced by its axes, which are error bars.

(c) *Mathematics of reverse interpolation*

How would the calculations of the previous section be programmed for a computer? In this section, we continue to suppose that the F.T. is one-dimensional. (Three-dimensional F.T.s pose the additional problem of sampling schemes, discussed in the next section.) However, the one-dimensional case is not only a simplification. It is of practical importance in three-dimensional reconstruction from two-dimensional lattices, where the F.T. consists of lines of continuous amplitude positioned at the reciprocal lattice points; reconstructing the F.T. along such lines involves one-dimensional reverse interpolation: Amos *et al.*, 1982; Henderson *et al.*, 1986.)

We start with the formulation of Section 7.6.4(a), that each experimental value could be calculated from the (as yet undetermined) defining values, by using an interpolation formula. This formula may be based on the sampling theorem (Section 7.2.2(h)) and use sinc-functions, or it may use other interpolation methods and functions. In almost all cases it will express the interpolation as a convolution. Let the defining values be F_j at $X = j\Delta X$ ($j = 0, +1, +2, \dots, +J$). Let the experimental values be $G(X_k)$, where $k = 1, \dots, K$; and let the interpolation function be $q(X)$, which will depend on the interpolation interval ΔX . Then we have, by the convolution Equation (6),

$$G_k \equiv G(X_k) = \int_{-\infty}^{\infty} \left\{ \sum_j F_j \delta(u - j\Delta X) \right\} q(X_k - u) du \\ = \sum_j q(X_k - j\Delta X) F_j = \sum_j q_{kj} F_j \quad (59)$$

Thus, if all the experimental values were viewed as a vector G , we could calculate them from the matrix equation

$$G = qF \quad (60)$$

Our problem is to reverse this, i.e. to calculate the defining values vector from the matrix q and the vector G . We cannot simply use the reciprocal matrix, as q is not square; it has K rows and J columns. Moreover, straightforward matrix inversion would tell us nothing about the reliability of the vector F . If the matrix q were square, however, the reliability of its inverse could be found from the eigenvalue spectrum of the inverse matrix; so we would diagonalize q . This is excluded since q is not square, but there is a decomposition of a rectangular matrix that corresponds to diagonalization. This is called a "singular value decomposition" (usually given the abbreviation SVD: Lawson and Hanson, 1974, and Chapter 2 of Press *et al.*, 1987, give explanations and programs for this generally useful technique). As part of the decomposition, SVD gives the eigenvalues of q . Those eigenvalues below a critical value are set to zero, which avoids the large errors associated with inverting a near-singular matrix. (The use of SVD for exactly this sort of least-squares fit is described in Chapter 14 of Press *et al.*, 1987.) In the case of our reverse interpolation problem, additional transform sections should then be added, until the smallest eigenvalues become large enough. However, matters are a little more complicated since the matrix q is, in general, complex. Nevertheless, it is surprising that SVD has not been applied to our problem. Instead, a closely related eigenvalue method is used (Crowther *et al.*, 1970a,b). The matrix $q^H q$ is formed. It is square, symmetric and real. Its eigenvalues are the squares of the corresponding eigenvalues of q (as is easily seen by multiplying the SVD expressions for q^H and q). However, these can still tell us when we have sufficient transform section data to make the solution reliable. Moreover we can calculate, from the eigenvalue spectrum, estimates for the errors in the "determining values" of the F.T., and hence in the calculated particle density. (See Crowther *et al.*, 1970b; or, for the errors in SVD, Chapter 14 of Press *et al.*, 1987.)

(d) *Sampling schemes*

Hitherto we have supposed that we have only a one-dimensional F.T. to reconstruct. However, a three-dimensional density reconstruction obviously needs a three-dimensional

F.T. In principle, nothing changes with the extra dimensions; the sampling theorem still applies (using three-dimensional sinc-functions, or—if a non-rectangular lattice is used—the F.T.s of Voronoi polyhedra, obtainable by Laue transformation: Hosemann and Bagchi, 1962). In practice, however, the number of "determining values" is very greatly increased, so that the eigenvalue analysis of the matrix q becomes quite lengthy. Because of the relative slowness of computers during the years when these methods were developed, special F.T. sampling techniques were invented to shorten the calculations; and these will continue to be useful, until array processors and super-computers become easily available.

The problem hardly exists for helical particles, where the F.T. consists of discrete thin layer-planes (Section 7.2.3(b)). Unless some of the layer-planes interfere, the F.T. on each plane is determined by its value along any line that is in the layer-plane, and also intersects the Z -axis. (See Section 7.2.3(b) and Equation (19).) In this case, there is only one Bessel function, so the interpolation problem actually disappears; the calculated F.T. on each layer-line can be taken to define $G_n(R, Z)$ (Equation (20)). If, however, some layer-planes should interfere, then there will be several different Bessel functions. So we shall have to find several different $G_n(R, Z)$ s, each with a different value of n . (The values of n will be available through the (n, Z) plot; Sections 7.2.3(c) and 7.3.4.) To find these $G_n(R, Z)$ s, the layer-plane is divided into concentric annuli, centred on the Z -axis and of regularly increasing radii. Along each annulus, we now have the sort of one-dimensional reverse interpolation problem discussed in the previous section; the only difference lies in the use of $G_n(R, Z)$ as interpolation functions.

Cylindrical polar coordinates have also been used for reconstructing the F.T.s of particles with point-group symmetry (Crowther *et al.*, 1970b), in order to simplify the computations. However, there are now no discrete layer-planes; Z is a continuous variable which must be discretized for the purpose of evaluating the Fourier integral. Z is usually sampled at equal increments which are spaced somewhat more finely than the reciprocal of the particle diameter. On these closely-spaced Z -planes, the F.T. is further sampled on annuli of the sort used for helical F.T.s. If the point-group has an N -fold axis, this is positioned along the Z -axis and each $G_n(R, Z)$ has its n equal to a multiple of N . This procedure has also been used for higher point-groups, especially the icosahedral group, although that group cannot be expressed directly in the least-squares equations. In this case, it would appear to be more satisfactory to use the method of functional expansion (next section).

7.6.5 Reconstruction from projections: functional expansions

(a) *Functions appropriate to point-group symmetry*

Most of the techniques surveyed in this chapter are based on a representation of particle structure as the sum of a (relatively small) number of density-functions. The simplest of these are the familiar trigonometric functions (sine and cosine). By using them, we have the Fourier transform (Section 7.2.2), and the Fourier projection-section reconstruction method (Section 7.6.3). Though quite general, it is most appropriate when the structure is crystalline, as the translational repeats of a crystal match those of the

trigonometrical functions. If the particle has any form of rotational symmetry, then other functions are more convenient. ("Convenience" means needing fewer functions to approximate, to a given accuracy, any structure with this symmetry.) The same functions can also be used (like the trigonometric functions) for 3-dimensional reconstruction (Crowther *et al.*, 1972).

A structure with one axis of rotational (or screw) symmetry is best represented with cylindrical polar coordinates (Fig. 7.48). As the structure is periodic in ϕ , it is appropriate to use a Fourier series in this coordinate. (With helical structures, the Fourier series couples ϕ and z ; see Section 7.2.3(j).) The particle cannot be periodic in the radial coordinate r , so a quite different set of functions must be used here. Extensive application to helical structures has made the Bessel functions (Section 7.2.3(j)) familiar. However, other (and perhaps superior) functions could be used (Zeitler, 1974; Smith and Aebi, 1974).

When particles have point-group symmetry with more than one rotation axis, the spherical polar coordinate system (Fig. 7.78) is more convenient. All the rotational periodicity is associated with ϕ and θ . Every particle is periodic in ϕ , so that a Fourier series is again the most appropriate in this coordinate. But the Fourier series is now coupled with functions that depend on θ , giving the well-known "spherical harmonics" (described in almost any textbook on mathematical physics or quantum mechanics). Convenient formulae exist for computing them, and also for transforming them when the spherical coordinate system is rotated.

The spherical harmonics describe the angular variation of the wave functions of the hydrogen atom. Through this, their characteristic symmetries have become familiar: spherical (s), cubic (p), etc. Combinations of them can generate other symmetries; this too is familiar from the "hybridization" of s- and p-orbitals to give tetrahedral sp^3 orbitals. Only certain combinations will yield any particular symmetry. Thus, when using these harmonics to represent the angular features of a particle's structure, the symmetry shows up through restrictions on the functions used. (This minimizes the number of functions needed, adding to the convenience of the method.) The restrictions, as in the case of helical symmetry, take the form of "selection rules" (see Finch and Holmes's (1967) review for the case of icosahedral symmetry).

As with helices, we need special non-periodic density-functions to represent the density in the radial direction. Various choices are possible. For example, X-ray scattering theory uses the "spherical Bessel" functions (Abramowitz and Stegun, 1964, *et seq.*, provide tables and formulae for these and other relevant functions). These have also been used

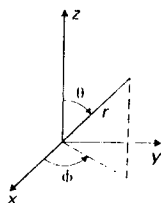


Fig. 7.78 Spherical coordinate system, defining the variables θ , ϕ and r .

in an ambitious effort to reconstruct particle structures from extremely noisy micrographs of disoriented particles (Kam, 1980; Kam and Gafni, 1985).

(b) Reconstruction by series expansion

This general approach has been implemented by Provencher and Vogel (1983, 1988) as an alternative to the common-lines technique. They use spherical harmonics for the angular components of the particle density, but choose Hermite functions for the radial component. (Hermite functions have the advantage of being almost unchanged by Fourier transformation; the Gaussian distribution—see Fig. 7.18—is the simplest of them.)

Since the particle is three-dimensional, the coefficients of the expansion must depend on three (integer) variables. These coefficients, which define the particle's structure, correspond to its F.T.; our task is to use projection data to determine them. If we knew them, then the projected particle density could be calculated from a series, which would also involve the appropriate density-functions and the rotations needed to define the particle's orientation. Such an equation, which essentially allows the projection to be calculated from the particle density-distribution and orientation, needs to be inverted. As it stands, this is impossible, since a three-dimensional distribution cannot generally be found from one projection. With sufficient projections, however, the inversion of the whole set of equations becomes possible in principle.

Nevertheless there are technical problems. The inversion involves a multidimensional integral over the density-distributions of each projection. To reduce the computation to a manageable size, these density-distributions are first "compressed" by Fourier transformation, etc. Of course, the inversion cannot be effected without knowing all the particle orientations, and we have yet to find them. If we knew them, we could calculate the particle's density-distribution and, from it, a theoretical estimate for each experimental projection. The agreement between theory and experiment can be measured by a variance, which will reach a minimum value when the correct particle orientations are chosen. Consequently, we can find these orientations by minimizing the variance.

Besides these techniques, and the usual procedures for finding the position/orientation of each particle's image in the plane of the micrograph (Section 7.4.3) and for correcting for the state of focus (Section 7.5.3(b)), the method employs statistical tests both for the resolution limit and for the distribution of errors in the reconstruction. It has been applied to 50S ribosomal subunits (unsymmetrical particles) and to the icosahedral virus capsids of tomato bushy stunt (TBSV) and Semliki forest (SFV) viruses (Vogel *et al.*, 1986; Vogel and Provencher, 1988). The best test is provided by TBSV, which has been previously reconstructed by the common-lines technique, and whose structure has been determined by X-ray crystallography. Both comparisons yield good agreement: the latter confirms the reliability of the method; the former indicates the similarity of the results obtained by the two reconstruction methods (and both methods also find the same resolution, 25 Å, for TBSV micrographs). But the conclusions concerning SFV were disputed by Fuller (1987), who reconstructed the closely related Sindbis virus using a modification of the common-lines method (Section 7.4.7(b)). (The modifications correct for lack of independence in the common-lines data and improve convergence when

determining the orientation.) It is to be hoped that the source of this discrepancy will be found, so that both techniques can be used with confidence.

7.6.6 Reconstruction from projections of isolated unsymmetrical particles

Isolated unsymmetrical particles pose a special problem, since there is no symmetry to help in finding the position and orientation of the particle (Section 7.4). Unless the specimen permits all the exposures necessary when collecting tilt-series data, we may be forced to interpret single micrograph images. An unsymmetrical particle gives projections with many different appearances, which are present as a jumble in its micrographs. The first step towards finding its structure is to sort this jumble into a gallery of characteristic views.

(a) Characteristic views

Visual selection. Any classification into characteristic views must start by selecting the clearest of the images, and making some preliminary allowance for artifact in them. Established distortion phenomena may well be present: the shrinkage of negative stain can flatten and anisotropically contract small particles embedded in pools of stain, stretch particles in thin films of stain spanning holes in the supporting grid, and flatten particles between a carbon film sandwich. Clues to the particle shape can emerge from the general spatial relations between the objects in a micrograph—which particles are superposed on, or leaning against, other particles, and which are probably lying flat. When all such factors have been taken into account, it may be possible to classify the images into several characteristic projections (accumulating a “gallery” of micrographs for each), and perhaps to form some initial views about their relative orientations.

Correspondence analysis. However, classifying particle images is a subjective process. Objective, statistical, methods are preferable. General classification methods have been developed by statisticians under the names “multivariate statistical analysis” or “classification” (see, e.g., Gordon, 1981). From these, one technique has been extensively applied to the classification on unsymmetrical particle images (van Heel and Frank, 1981; Frank and van Heel, 1982). This technique, correspondence analysis (Greenacre, 1984), is unfortunately both mathematical and complicated. For details of its applications, readers are referred to the review by Frank *et al.* (1988b); or (for a well-illustrated introductory review) to Bretauière and Frank (1986). Here it is only possible to give a brief account of how the technique tackles the classification problem.

To begin with, the particle images must be clear enough, not only for recognition, but also for orientation and matching. Then, by the methods discussed in Section 7.4, we find, for each image, the transformation parameters that will bring it to a standard position. When in that standard position, every image is scanned in exactly the same way. Thus each image gives a string of pixels, such that the corresponding pixels of different strings refer to corresponding parts of the particle images. All our data can now be collected into a rectangular array of pixel densities. Each row is the scanned image of a different particle; each column is the set of corresponding pixels from the same part of each particle. This rectangular array, after scaling by division by the total

pixel density, is subjected to an eigenvalue analysis (presumably along very similar lines to SVD; see Section 7.6.4(c)). For each eigenvalue, a corresponding “eigen-image” can be calculated. The size of the eigenvalue gives the relative contribution of its eigen-image to the total set of particle images.

The situation is roughly analogous to Fourier analysis of a noisy lattice (Section 7.2.2). There, the size of the Fourier component (corresponding to eigenvalue) indicates the relative contribution of the appropriate density-wave (corresponding to eigen-image) in the original data. In Fourier analysis, when the “size” of the Fourier component (i.e. the amplitude at the corresponding reciprocal lattice point) falls below the noise level, nothing is gained by including the corresponding density-wave in the filtered image; and we therefore exclude all Fourier components beyond a certain spatial frequency (i.e. beyond a certain resolution). Similarly, in correspondence analysis, eigenvalues below a certain cut-off value are ignored. Only the largest eigenvalues (in practice, about the first eight) are considered statistically significant; these contain the lowest-resolution information.

The (~ eight) acceptable eigenvalues, and their corresponding eigen-images, define significant “factors” in the particle images. Each particle image can be represented as a linear combination of the eigen-images, and the coefficients (which can be calculated) show the contribution of that “factor” to the particular particle image. For each particle, the set of eight coefficients constitute the coordinates of a point in eight-dimensional space (“factor space”). After each particle has had its linear combination coefficients calculated, there will be a point in factor space for that particle. The total set of points constitutes a “cloud” whose distribution conveys information about the types of particle image present in the micrographs. To assess this “cloud”, two-dimensional projections of the eight-dimensional space are plotted. Such a projection shows the distribution of all particle images with respect to the two chosen “factors”. If the “cloud” shows two separate concentrations in the projection, this indicates that there are two distinct classes of particle. The particle image typical of each cluster is calculated and examined. Two such images might, for example, be nearly mirror-images; this would suggest that the particles are attaching by opposite surfaces. On the other hand, the “cloud” may form a continuous band without any local concentrations; this would indicate that particle images can vary continuously with respect to these factors, perhaps as a consequence of different orientations about some preferred axis. In some cases this can be confirmed by tilting experiments (e.g. the “O” and “R” images of negatively-stained ribosomes; Verschoor *et al.*, 1986). The local concentrations of points—“clusters”—may be connected with each other in ways that relate the (random) particle images to the three-dimensional structure of the particle. Attempts are therefore being made, by various statistical techniques, to classify these clusters into “super-clusters”, etc., so as to construct a tree-like hierarchy.

Before correspondence analysis can be performed, the particle images must be aligned accurately so that, after they have been scanned, the 1st, 2nd, etc. pixels of different images all correspond as nearly as possible. This requirement restricts the technique to images that are very similar—perhaps identical. After analysis, a single cluster of points in factor-space would confirm that they are indeed identical, and they could then be averaged with confidence. Two distinct clusters, on the other hand, would imply two classes of particle image, and averaging could then proceed with each class separately; the class members would be identified by the positions of the corresponding points in

factor-space. Such applications make correspondence analysis a technique of image enhancement.

However, it is discussed here, as a technique of three-dimensional reconstruction, because much of the variation in particle image is the consequence of particle orientation. Correspondence analysis can provide some objective information about the types of orientation present. For example, two clusters of points might indicate two attachment surfaces. On the other hand, elongated particles might always lie with the long axis in the plane of the grid, but with all possible rotations about that axis; then the points in factor space would (if the images permitted correspondence analysis) all cluster around a single closed curve. However, as a technique contributing to three-dimensional reconstruction, correspondence analysis seems to suffer under two disadvantages. First, different particle orientations must often yield images that are so different that the image alignment step can easily fail. In such a case, the images subjected to analysis may need previous visual sorting. Second, correspondence analysis is a general technique that makes no special assumptions about the images to be classified. This is a strength, in that no *a priori* assumptions are built into the analysis. But it can also be a weakness when analysing particle orientations, since it omits the common features present in images that are all projections of the same solid. We turn now to the general problem of reconstructing the three-dimensional structure from images of isolated, unsymmetrical particles.

(b) *Determining the particle orientation*

Experimental limitations. If many different projections are obtained, along known directions, from the same solid structure, then it is straightforward to reconstruct that structure by (for example) the back-projection method (Section 7.6.2) or the Fourier method (Section 7.6.3). Unfortunately, the necessary experimental data are difficult or impossible to obtain, for two reasons. The first is that the preparation procedure distorts the particles. The distortion is usually correlated with the position of the supporting film, so that particles with different orientations have their structures distorted in different directions. The second reason is that radiation damage causes significant image degradation in even one exposure. Multiple exposures, necessary to obtain different views of the same particle, impose correspondingly greater image degradation. Radiation damage can be minimized in two ways, but each carries a penalty. Negatively stained preparations are somewhat more resistant than ice-embedded preparations, but cause more particle distortion. The other way to reduce radiation damage is to reduce the exposure; but the reduction necessary to eliminate damage leaves the "image" with so few electrons that the particle orientation is virtually impossible to determine. The effect of these limitations is to reduce the resolution of the reconstructions, and also to leave some uncertainty about possible distortions in them. The limitations also lead to a choice of strategy for collecting the images (i.e. the projection data); each strategy poses different problems in determining the orientation of the projection direction.

Obtaining and recombining projection data. The different strategies for collecting image data can be classified according to the number of exposures exacted from each particle. Easiest to analyse are the images of a single particle tilted through angles that are known

(e.g. from a goniometer stage), so that we know the orientations of all the F.T. sections. The only remaining problem is to find the position of the particle's centre in each image, so that all F.T. sections will have their phases calculated relative to a common origin. This translational alignment can be performed by finding the X.C.F. of images with similar orientations (Hoppe and Tietz, 1986). (This approach has been used to find the structure of a 50S ribosomal subunit; Oettl *et al.*, 1983; Hoppe *et al.*, 1986.) But it is probably more accurate to use colloidal gold particles in the micrograph as reference points (Skoglund *et al.*, 1986).

At the other extreme, radiation damage may need to be reduced to the minimum, so that only one exposure is possible. The nature of the projection data obtainable from this exposure will depend on the particle shape. Suppose, for example, that the particle is elongated and lies with its long axis in the plane of the grid, but has no other preferred orientation. The different particles in a micrograph will then provide a random tilt-series, with the projection direction rotating about the long axis. (Let us refer to this situation as a "random one-axis tilt-series".) However, many types of particle (like most objects placed on a table) come to rest in one of a small number of orientations. In some cases, there are (as with a coin) only two stable orientations, giving essentially the same projection if the grid is untilted. If, however, the grid is tilted through (say) 50° , a wide range of projections is obtainable ("random conical-tilt-series"). For each image then gives a section of the F.T. on a plane tilted 50° relative to the grid-plane normal. And, since all the F.T. section-planes are (like the particles) rotated randomly about the grid-plane normal, we can sample all of the F.T., except for the portion within a cone of semi-angle 40° ($= 90^\circ - 50^\circ$).

How shall we perform the three-dimensional reconstruction, i.e. how shall we orient the random F.T. sections? For each image, we shall have to find the correct translation (as with the conventional tilt-series). But we shall also need an orientation angle: with the random one-axis tilt-series, that angle is the rotation about the unique axis; with the random conical tilt-series, it is the rotation about the grid normal. If all the projections were very clear and undistorted, the whole problem would seem to be soluble in principle. For the F.T. sections could be rotated and their amplitudes compared pairwise until a unique angular sequence was obtained. Converting the angular sequence into numerical rotations would be more difficult, but might be approximately soluble by using the criterion that the "chunks" of F.T. amplitude must have roughly similar sizes and shapes (because of the sampling theorem). Finally, the translations could be estimated from the centres of mass, and refined by bringing the phases of adjacent F.T. sections into harmony.

Presumably, however, such a scheme is chimerical when the images are noisy and distorted projections. Then more data are needed to render the problem soluble. A method has been developed and applied to the three-dimensional reconstruction of ribosomal subunits (Carazo *et al.*, 1988; Frank *et al.*, 1988b, c; Verschoor *et al.*, 1989). Two exposures are obtained from a grid; the first, when the grid is tilted (giving a random conical-tilt-series); the second, when it is in the untilted position. The first exposure, which is less degraded by radiation damage, yields the projection data actually used in the reconstruction. The second exposure helps to establish the orientation angle for each particle (and hence of the corresponding F.T. section).

The orientation of particles in random tilt-series could be determined either by orienting the section within the F.T., or by correlating the particle images directly. The

latter procedure was developed by Guckenberger (1982), and has been adapted for the random conical-tilt-series by Carazo and Frank (1988).

7.6.7 Reconstruction of sectioned symmetric structures

(a) Deficiencies of images of sections

Finding structures from projections is a relatively new development in electron microscopy. The traditional approach is to cut thin sections of complicated structures. However, this technique does not always solve the problem. A section of finite thickness degrades the resolution (to something approaching the section thickness), because the entire contents of the section are projected onto the image. When the structure of interest is much smaller than the section thickness, additional data are necessary to reconstruct it from this projection.

Obviously, sections can be tilted, and the micrographs taken at different tilts can be used for three-dimensional reconstruction by one of the methods described in Sections 7.6.2–7.6.6. For example, sections of isolated particles have been reconstructed (Skoglund and Daneholt, 1986; Skoglund *et al.*, 1986) using the *R*-weighted back-projection method (Section 7.6.2). On the other hand, sections showing a two-dimensional lattice would be better reconstructed by the Fourier method (Section 7.6.3–7.6.4; see the review by Amos *et al.*, 1982). This approach has been used to reconstruct the structure of the M-band of fish muscle (Luther and Crowther, 1984).

However, reconstructions from sectioned material face the problem that the plastic embedding material is much more radiation-sensitive than negative stain. The exposure necessary to take one micrograph can thin the section to 50–80% of its original thickness, depending on the embedding plastic (Bennett, 1974). (Along with the thinning, there is also some shrinkage in the plane of the grid, but this is relatively small, and its effects can be corrected.) The magnitude of this thinning has stimulated the development of methods for finding the three-dimensional structure from just one untilted micrograph.

The problem is that, with only one untilted micrograph, there are many “invisible functions” (Section 7.5.1(b)): any density function whose projection is uniform may be added to the derived structure, without changing the micrograph. Invisible functions can be restricted only by having additional data in the same micrograph: the specimen must be symmetrical. Since the highest symmetry order involves translations, we should expect most success in two cases: one-dimensional repeating structures (which, in general, are helices), and two-dimensional lattices. Sectioned helices were reconstructed by Lake and Slayter (1972), and sectioned two-dimensional lattices by Crowther and Luther (1984).

In each case, the images consist of the original structure modified by two consecutive operations. First, we have only a section of it; that is, its density distribution has been multiplied by a “section-function” that is unity within the section, and zero outside. Second, the resulting density distribution has been projected perpendicular to the section plane. Both operations must be reversed before the original density distribution can be obtained; but these reversals are not of equal difficulty. The more difficult reversal is

the first operation in the case of helices, and the second in the case of two-dimensional lattices.

(b) Helices

The reconstruction of a helical structure from one projection is fairly straightforward (Sections 7.6.3(c), 7.6.4(d)), provided the helix repeat is long. For, in the case, the F.T. is confined to layer-planes on which its amplitude has circular symmetry and its phase rotates by some number of complete revolutions, that number being the order of the layer-plane (Section 7.2.3(b)). After sectioning, the structure has been multiplied by a “section-function” (unity within the section, zero outside); so the helical F.T. has been convoluted with the section-function’s F.T. Consider the case where the section-plane is parallel to the helix axis. Then the section-function’s F.T. is a “spike” perpendicular to the helix axis. Convolution with such a “spike” keeps all the layer-planes intact, and merely redistributes the density within them. This spoils the F.T.’s circular symmetry; but the very simplicity of that symmetry, and the unaltered layer-plane arrangement, make deconvolution a practical operation. The deconvolution integral was converted into a sum, i.e. a matrix equation, which was solved by methods very similar to those discussed in Section 7.6.4(c). However, there is a further complication. Although the form of the section-function’s F.T. is known (the “spike” is a sinc-function), the details depend on the exact positions of the section surfaces, which have yet to be determined. Therefore trial section-functions are tested and judged by the quality of the deconvoluted helical F.T. (Lake, 1972).

(c) Two-dimensional lattices

The problems were different with the two-dimensional lattice structure. Such a structure, lacking any rotational symmetry axis (except perhaps a two-fold axis) within the lattice plane, is impossible to reconstruct from only one projection. In this case, therefore, the section-function is not a hindrance, but the factor that makes reconstruction possible. Instead, it is the projection along the section thickness that needs to be reversed. This projection can be viewed as a convolution (with a projection-function, a line perpendicular to the section plane), so reversal again involves deconvolution. Unlike the situation with helices, however, the deconvolution is now in real space. Two deconvolution procedures were tried (Crowther, 1984). The first used F.T.s to convert the convolution into multiplication, so that deconvolution involves division by the F.T. of the projection-function. That F.T. is a sinc-function, whose zeroes frustrate the attempted division and limit the extent of the deconvolution. The second deconvolution procedure (like Lake’s) converted the integral into a sum, rather as in the “reverse interpolation” of Section 7.6.4. Thereafter, the solution followed closely that of Section 7.6.4(c). A matrix equation was obtained, in which the matrix is ill-conditioned (this is the way in which the SVD equation also limits the extent of the deconvolution). However, instead of using SVD, Crowther calculated the square, symmetric matrix product. Its eigenvalue spectrum indicated the exact limits of the deconvolution, and the corresponding eigenvalues showed how the reliability of features varied inversely with their detail. Finally, as in the case of sectioned helices, the section function had to be refined by means of the criterion that it should give the best deconvolution.

7.6.8 Reliability and display of structures

(a) Reliability

Quantitative measures of reliability are most important when interpreting the structures produced by three-dimensional reconstruction. This is because the structures have no in-built test of their reliability, such as the quality of an electron-density map in protein crystallography. On the other hand, three-dimensional structures are meaningful in relation to biochemical data, so there is a real danger of misinterpretation if they are unreliable.

All the considerations affecting the reliability of filtered images (Section 7.5.3(d)) apply also in the case of three-dimensional constructs. Thus averaging increases the signal-to-noise ratio, and is consequently a most important factor in three-dimensional reconstruction. This is why most of the successful work has employed specimens with lattice or helical symmetry, or else has used particles whose images were so clear (e.g. through intrinsic symmetry) that they could be aligned accurately for averaging. The most reliable way to estimate the errors is to repeat the reconstruction several times, using different data, and to calculate from the variability of the results some statistical measure of the error. Thus the original data set is partitioned into a reasonable number of subsets, and the calculated densities used to find the standard deviation and (via Student's *t*) the probability of that density, as explained in Section 7.5.3(d). It would be very instructive (though perhaps depressing) if two structures were displayed, one at a contour level of 1 standard deviation above, and the other below, the mean density.

However, it is useful if errors can also be assessed during the course of the calculation. Presumably this can be done through a comparison of the processed data subsets at each stage. However, this is not always necessary. It can be predicted, in advance of the calculation, how the errors in the data will be increased or reduced during the calculation, as a consequence of the positions of the sampling sections of the F.T. This can be done roughly by rules of thumb (for an equi-angular tilt-series, see Section 7.6.3(b)). It can be done more precisely by examining the eigenvalue spectrum obtained during the least-squares reverse interpolation (Sections 7.6.4(b), (c)).

One of the most insidious potential sources of error is the programming mistake ("bug"). Three-dimensional reconstruction programs are quite complex, with several different steps, so it is not easy to avoid errors. Although the obvious ones demand correction, the programmer may not notice more subtle errors, e.g. in weighting factors or in the handedness of the calculated structure. The best way to test them would be to project the calculated structure, giving pseudo-images, and then to see if the program reconstructs them correctly. (Of course, the correctness of the projection program would also need testing!) Instead of this, the projections calculated from the final structure are sometimes compared with those obtained experimentally. But this is less satisfactory, since it is difficult to make satisfactory comparisons between such images by eye.

(b) Display of three-dimensional structures

After calculating a three-dimensional structure, we shall wish to represent it for publication or further analysis. Now the three-dimensional structure is a continuous function of spatial position. It is sometimes published as a series of contour maps; but,

though these can be useful to clarify specific points, they give no clear impression of the overall shape. For the human visual system cannot perceive such a complicated function as a three-dimensional object. Instead, it perceives such objects by their reflecting surfaces. We saw earlier (Section 7.6.1(b)) that this limitation causes model-builders to make their structures akin to sculptures. Thus, although the computational limitations of old-fashioned model-building can be circumvented, its display limitations are immovable. Consequently we must convert our continuous three-dimensional density distribution into a sculptural model. The common procedure is to plot sections at short intervals of the *z*-coordinate, cut out sheets along some chosen contour level on each section, and stack the sheets to form a model. Alternatively, there are computer programs that calculate the surface at some chosen contour level and display it on a graphics terminal. Some of these use "wire netting" ("chicken-wire") representations of the surface, but more recent programs suitable for raster-graphics systems give fairly realistic images of the contour surface.

ACKNOWLEDGEMENTS

The preparation of this chapter was assisted by access to various unpublished notes written for EMBO courses by Drs U. Aebi, L. A. Amos, R. A. Crowther, J. T. Finch, R. Henderson, A. Klug, W. O. Saxton and P. R. Smith. I also wish to thank Drs J. Frank and M. Stewart for preprints and reprints, Drs L. A. Amos, R. A. Crowther and J. Lepault for original figures, and Drs R. A. Crowther and D. J. DeRosier for comments on this chapter.

REFERENCES

- Abramowitz, M. and Stegun, I. A. (1964. *et seq.*) *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. National Bureau of Standards (Applied Mathematics Series, Vol. 55). U.S. Government Printing Office, Washington D.C.
- Acton, F. S. (1970). *Numerical Methods That Work*. Harper and Row, New York.
- Aebi, U., Smith, P. R., Dubochet, J., Henry, C. and Kellenberger, E. (1973). *J. Supramol. Struct.* **1**, 98-522.
- Aebi, U., Fowler, W. E. and Smith, P. R. (1982). *Ultramicroscopy* **8**, 191-206.
- Aebi, U., Fowler, W. E., Buhle, E. L. Jr and Smith, P. R. (1984). *J. Ultrastruct. Res.* **88**, 143-176.
- Amos, L. A. (1975). *J. Mol. Biol.* **99**, 65-73.
- Amos, L. A. (1979). In *Microtubules* (K. Roberts and J. S. Hyams, eds), pp. 1-64. Academic Press, London.
- Amos, L. A. (1982). In *Electron Microscopy of Proteins* (J. R. Harris, ed.), Vol. 3, pp. 207-250. Academic Press, London.
- Amos, L. A. and Klug, A. (1974). *J. Cell Sci.* **14**, 523-549.
- Amos, L. A. and Klug, A. (1975). *J. Mol. Biol.* **99**, 51-64.
- Amos, L. A., Henderson, R. and Unwin, P. N. T. (1982). *Progr. Biophys. Mol. Biol.* **39**, 183-231.
- Baker, T. S. (1981). In *Electron Microscopy in Biology* (J. D. Griffith, ed.), Vol. 1, Ch. 6, pp. 189-290. Wiley-Interscience, New York. (Complete up to Dec. 1979.)
- Baker, T. S., Caspar, D. L. D. and Murakami, W. (1983). *Nature (Lond.)* **303**, 446-448.
- Bennett, P. (1974). *J. Cell Sci.* **14**, 693-701.
- Berger, J. E. and Harker, D. (1967). *Rev. Sci. Instrum.* **38**, 292-293.
- Birch, K. G. (1972). *Rep. Progr. Phys.* **34**, 1265-1314.
- Blumke, D. A., Carragher, B. and Josephs, R. (1988). *Ultramicroscopy* **26**, 255-270.

- Born, M. and Wolf, E. (1980). *Principles of Optics*. Pergamon Press, Oxford.
- Bracewell, R. (1986). *The Fourier Transform and its Applications*. McGraw-Hill, New York.
- Bretaudière, J.-P. and Frank, J. (1986). *J. Microscopy* **144**, 1-14.
- Brigham, E. O. (1974). *The Fast Fourier Transform*. Prentice-Hall, Englewood Cliffs, N. J.
- Buerger, M. J. (1959). *Vector Space*. J. Wiley, New York.
- Carazo, J. M. and Frank, J. (1988). *Ultramicroscopy* **24**, 13-22.
- Carazo, J. M., Wagenknecht, T., Rademacher, M., Mandiyan, V., Boublik, M. and Frank, J. (1988). *J. Mol. Biol.* **201**, 393-404.
- Carragher, B., Bluemke, D. A., Potel, M. J. and Josephs, R. (1986). *Proteins* **1**, 176-187.
- Carragher, B., Bluemke, D. A., Gabriel, B., Potel, M. J. and Josephs, R. (1988). *J. Mol. Biol.* **199**, 315-331.
- Caspar, D. L. D. and Holmes, K. C. (1969). *J. Mol. Biol.* **46**, 99-133.
- Caspar, D. L. D. and Klug, A. (1962). *Cold Spring Harbor Symp. Quant. Biol.* **27**, 1-24.
- Cochran, W., Crick, F. H. C. and Vand, V. (1952). *Acta Crystallogr.* **4**, 581-586.
- Crane, H. R. (1950). *Sci. Monthly* **70**, 376-389.
- Crowther, R. A. (1971). *Phil. Trans. Roy. Soc. Lond. B* **261**, 221-230.
- Crowther, R. A. (1972). In *The Molecular Replacement Method* (M. G. Rossmann, ed.), pp. 173-178. Gordon & Breach, New York.
- Crowther, R. A. (1982). In *Structural Molecular Biology* (D. B. Davies, W. Saenger and S. S. Danyluk, eds), Plenum Press, New York.
- Crowther, R. A. (1984). *Ultramicroscopy* **13**, 295-304.
- Crowther, R. A. and Amos, L. A. (1971). *J. Mol. Biol.* **60**, 123-130.
- Crowther, R. A. and Klug, A. (1975). *Ann. Rev. Biochem.* **44**, 161-182.
- Crowther, R. A. and Luther, P. K. (1984). *Nature (Lond.)* **307**, 569-570.
- Crowther, R. A. and Sleytr, U. B. (1977). *J. Ultrastruct. Res.* **58**, 41-49.
- Crowther, R. A., Amos, L. A., Finch, J. T. and Klug, A. (1970a). *Nature (Lond.)* **226**, 421-425.
- Crowther, R. A., DeRosier, D. J. and Klug, A. (1970b). *Proc. Roy. Soc. Lond. A* **317**, 319-340.
- Crowther, R. A., Amos, L. A. and Klug, A. (1972). *Proceedings of the 5th European Congress on Electron Microscopy*, pp. 593-597. Institute of Physics, London.
- Crowther, R. A., Finch, J. T. and Pearse, B. M. F. (1976). *J. Mol. Biol.* **103**, 785-788.
- Crowther, R. A., Lenk, E. V., Kikuchi, Y. and King, J. (1977). *J. Mol. Biol.* **116**, 489-523.
- Crowther, R. A., Padron, R. and Craig, R. (1985). *J. Mol. Biol.* **184**, 429-439.
- Cundy, H. M. and Rollett, A. P. (1961). *Mathematical Models*, 2nd edn. Clarendon Press, Oxford.
- DeRosier, D. J. and Klug, A. (1968). *Nature (Lond.)* **217**, 130-134.
- DeRosier, D. J. and Moore, P. B. (1970). *J. Mol. Biol.* **52**, 355-369.
- Dodson, E. J. (1985). In *Molecular Replacement* (P. A. Machin, ed.), pp. 33-45. S.E.R.C., Daresbury Laboratory, Daresbury, Warrington WA4 4AD.
- Egelman, E. H. (1986). *Ultramicroscopy* **19**, 367-374.
- Egelman, E. H. and Stasiak, A. (1986). *J. Mol. Biol.* **191**, 677-698.
- Egelman, E. H., Francis, N. and DeRosier, D. J. (1983). *J. Mol. Biol.* **166**, 605-623.
- Elliott, D. F. and Rao, K. R. (1982). *Fast Transforms: Algorithms, Analyses, Applications*. Academic Press, New York.
- Erickson, H. P. (1973). *Adv. Opt. Electron Microsc.* **4**, 163-199.
- Erickson, H. P. and Klug, A. (1971). *Phil. Trans. Roy. Soc. Lond. B* **261**, 105-118.
- Erickson, H. P., Voter, W. A. and Leonard, K. (1978). *Methods Enzymol.* **49**, 39-63.
- Finch, J. T. (1972a). *J. Mol. Biol.* **66**, 291-294.
- Finch, J. T. (1972b). *Proceedings of the 5th European Congress on Electron Microscopy*, pp. 578-579. Institute of Physics, London.
- Finch, J. T. and Holmes, K. C. (1967). In *Methods in Virology* (K. Maramorosch and H. Kaprowski, eds), Vol. 3, pp. 351-474. Academic Press, New York and London.
- Finch, J. T. and Klug, A. (1965). *J. Mol. Biol.* **11**, 403-423.
- Finch, J. T. and Klug, A. (1967). *J. Mol. Biol.* **24**, 289-302.
- Finch, J. T. and Klug, A. (1971). *Phil. Trans. Roy. Soc. Lond. B* **261**, 211-219.
- Finch, J. T. and Klug, A. (1972). In *The Generation of Subcellular Structures*. First John Innes Symposium, Norwich (R. Markham and J. B. Bancroft, eds), pp. 167-177. North Holland Publishing Co., Amsterdam.

- Finch, J. T., Klug, A. and Nermut, M. V. (1967). *J. Cell. Sci.* **2**, 587-590.
- Fiskin, A. M. and Beer, M. (1968). *Science* **159**, 1111-1113.
- Fletcher, R. (1980). *Practical Methods of Optimization*, Vol. 1: Unconstrained Optimization. John Wiley, New York.
- Frank, J. (1981). In *Methods in Cell Biology* (J. N. Turner, ed.), Vol. 22, Ch. 12, pp. 199-213. Academic Press, New York.
- Frank, J. and van Heel, M. (1982). *J. Mol. Biol.* **161**, 134-137.
- Frank, J., Goldfarb, W., Eisenberg, D. and Baker, T. S. (1978). *Ultramicroscopy* **3**, 283-290.
- Frank, J., Shimkin, B. and Dowse, H. (1981). *Ultramicroscopy* **6**, 343-358.
- Frank, J., Chiu, W. and Degn, L. (1988a). *Ultramicroscopy* **26**, 345-360.
- Frank, J., Rademacher, M., Wagenknecht, T. and Verschoor, A. (1988b). *Methods Enzymol.* **164**, 3-35.
- Frank, J., Verschoor, A., Wagenknecht, T., Rademacher, M. and Carazo, J. M. (1988c). *Trends Biochem. Sci.* **13**, 123-127.
- Fraser, R. D. B., MacRae, T. P., Suzuki, E. and Davey, C. L. (1976). *J. Microscopy* **108**, 343-348.
- Fuller, S. D. (1981). In *Methods in Cell Biology* (J. N. Turner, ed.), Vol. 22, Ch. 14, pp. 251-295. Academic Press, New York.
- Fuller, S. D. (1987). *Cell* **48**, 923-934.
- Gibbs, A. J. and Rowe, A. J. (1977). In *Principles and Techniques of Electron Microscopy* (A. J. Hayat, ed.), Vol. 7, Ch. 6, pp. 202-230. Van Nostrand Reinhold Company, New York.
- Gilbert, P. F. C. (1972). *Proc. Roy. Soc. B* **182**, 89-102.
- Goldstein, H. (1980). *Classical Mechanics*, 2nd edn. Addison-Wesley, Reading, MA.
- Goodman, J. W. (1968). *Introduction to Fourier Optics*. McGraw-Hill, San Francisco, CA.
- Gordon, A. D. (1981). *Classification*. Chapman & Hall, London.
- Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.
- Grimstone, A. V. and Klug, A. (1966). *J. Cell Sci.* **1**, 351-362.
- Guckenberger, R. (1982). *Ultramicroscopy* **9**, 167-174.
- Gull, S. F. and Skilling, J. (1984). *I.E.E. Proc.* **131** (Part F), 646-659.
- Hegerl, R. and Altbauer, A. (1982). *Ultramicroscopy* **9**, 109-116.
- Henderson, R., Baldwin, J. M., Dowling, K. H., Lepault, J. and Zemlin, F. (1986). *Ultramicroscopy* **19**, 147-178.
- Herman, G. T. (1980). *Image Reconstruction from Projections*. Academic Press, New York.
- Holser, W. T. (1958). *Z. Kristallogr.* **110**, 266-281.
- Hoppe, W. and Tietz, H. R. (1986). *Ultramicroscopy* **20**, 217-237.
- Hoppe, W., Oetli, H. and Tietz, H. R. (1986). *J. Mol. Biol.* **192**, 291-322.
- Horne, R. W. and Markham, R. (1972). In *Practical Methods in Electron Microscopy* (A. M. Glauret, ed.), Vol. 1, pp. 361-379. North Holland, Amsterdam.
- Hosemann, R. and Bagchi, S. N. (1962). *Direct Analysis of Diffraction by Matter*. North Holland, Amsterdam.
- Huxley, H. E. (1963). *J. Mol. Biol.* **7**, 281-308.
- International Tables for Crystallography (1983). Vol. A. Reidel Publishing Company, Dordrecht, Holland.
- Jacobs, D. A. H. (1977). *The State of the Art in Numerical Analysis*. Academic Press, London.
- Jaynes, E. T. (1979). In *The Maximum Entropy Formalism* (R. D. Levine and M. Tribus, eds), pp. 15-118. MIT Press, Cambridge, MA.
- Jeng, T.-W., Crowther, R. A., Stubbs, G. and Chiu, W. (1989). *J. Mol. Biol.* **204**, 251-257.
- Johansen, B. V. (1975). In *Principles and Techniques of Electron Microscopy* (A. J. Hayat, ed.), Vol. 5, Ch. 4, pp. 114-173. Van Nostrand Reinhold, New York.
- Kam, Z. (1980). *J. Theoret. Biol.* **82**, 15-39.
- Kam, Z. and Gafni, I. (1985). *Ultramicroscopy* **17**, 251-262.
- King, M. V. (1981). In *Methods in Cell Biology* (J. N. Turner, ed.), Vol. 22, Ch. 2, pp. 13-22. Academic Press, New York.
- Kisilev, N. A. and Klug, A. (1969). *J. Mol. Biol.* **40**, 155-171.
- Klug, A. (1971). *Phil. Trans. Roy. Soc. Lond. B* **261**, 173-180.
- Klug, A. and Berger, J. E. (1964). *J. Mol. Biol.* **10**, 565-569.

- Klug, A. and DeRosier, D. J. (1966). *Nature (Lond.)* **212**, 29–32.
- Klug, A., Crick, F. H. C. and Wyckoff, H. W. (1958). *Acta Crystallogr.* **11**, 199–213.
- Krimm, S. and Anderson, T. F. (1967). *J. Mol. Biol.* **27**, 197–202.
- Lake, J. A. (1972). *J. Mol. Biol.* **66**, 255–269.
- Lake, J. A. (1976). *J. Mol. Biol.* **104**, 131–159.
- Lake, J. A. and Slayter, H. (1972). *J. Mol. Biol.* **66**, 271–282.
- Lawson, C. L. and Hanson, R. J. (1974). *Solving Least-Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ.
- Lepault, J. and Leonard, K. (1985). *J. Mol. Biol.* **182**, 431–442.
- Lipson, H. and Taylor, C. A. (1958). *Fourier Transforms and X-ray Diffraction*. G. Bell & Sons, London.
- Luther, P. K. and Crowther, R. A. (1984). *Nature (Lond.)* **307**, 566–568.
- MacLeod, I. D. (1970). *IEEE Trans: Au-19*, 160–162.
- Mandelkow, E.-M. and Mandelkow, E. (1985). *J. Mol. Biol.* **181**, 123–136.
- Markham, R. (1968). In *Methods in Virology* (K. Maramorosch and H. Kaprowski, eds), Vol. 4, pp. 503–529. Academic Press, New York and London.
- Markham, R., Frey, S. and Hills, G. J. (1963). *Virology* **20**, 88–102.
- Markham, R., Hitchborn, J. H., Hills, G. J. and Frey, S. (1964). *Virology* **22**, 342–359.
- McLachlan, D. Jr. (1958). *Proc. Natl. Acad. Sci. Wash.* **44**, 948–956.
- Milligan, R. and Flicker (1987). *J. Cell Biol.* **104**, 29–39.
- Misell, D. L. (1978). *Image Analysis, Enhancement and Interpretation*, Vol. 7 of Practical Methods in Electron Microscopy (A. M. Glauert, ed.), pp. 1–305. North Holland, Amsterdam.
- Moody, M. F. (1967a). *J. Mol. Biol.* **24**, 167–200.
- Moody, M. F. (1967b). *J. Mol. Biol.* **24**, 201–208.
- Moody, M. F. (1971). *Phil. Trans. Roy. Soc. Lond. B* **261**, 181–195.
- Moody, M. F. (1973). *J. Mol. Biol.* **80**, 613–635.
- Natterer, F. (1986). *The Mathematics of Computerized Tomography*. John Wiley & Sons, Ltd., Chichester.
- Oettl, H., Hegerl, R. and Hoppe, W. (1983). *J. Mol. Biol.* **163**, 431–450.
- Pauling, L. and Corey, R. B. (1951). *Proc. Natl. Acad. Sci. Wash.* **37**, 205–211.
- Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1987). *Numerical Recipes. The Art of Scientific Computing*. Cambridge University Press, Cambridge.
- Provencher, S. W. and Vogel, R. W. (1983). *Progress in Scientific Computing* (S. Abarbanel, R. Glowinski, G. Golub and H.-O. Weiss, eds), Vol. 2, pp. 310–319. Birkhäuser, Stuttgart.
- Provencher, S. W. and Vogel, R. W. (1988). *Ultramicroscopy* **24**, 209–222.
- Ramachandran, G. N. (1960). *Proc. Indian Acad. Sci.* **52**, 240–254.
- Reedy, M. K. and Reedy, M. C. (1985). *J. Mol. Biol.* **184**, 145–176.
- Rust, B. W. and Burrus, W. R. (1972). *Mathematical Programming and the Numerical Solution of Linear Equations*. Elsevier, Amsterdam.
- Saxton, W. O. (1978). *Computer Techniques for Image Processing in Electron Microscopy*. Academic Press, New York.
- Saxton, W. O. and Baumeister, W. (1982). *J. Microscopy* **127**, 127–138.
- Saxton, W. O., Pitt, T. J. and Horner, M. (1979). *Ultramicroscopy* **4**, 343–354.
- Schiske, P. (1968). In *Electron Microscopy 1968* (D. S. Bocciarelli, ed.), Vol. 2. Tipografia Poliglotta Vaticana, Rome.
- Seymour, J. and DeRosier, D. J. (1987). *J. Microscopy* **148**, 195–210.
- Skilling, J. (1984). *Nature (Lond.)* **309**, 748–749.
- Skoglund, U. and Daneholt, B. (1986). *Trends Biochem. Sci.* **11**, 499–503.
- Skoglund, U., Anderson, K., Strandberg, B. and Daneholt, B. (1986). *Nature (Lond.)* **319**, 560–564.
- Smith, P. R. (1978). *Ultramicroscopy* **3**, 153–160.
- Smith, P. R. and Aebi, U. (1973). *J. Supramol. Struct.* **1**, 516–521.
- Smith, P. R. and Aebi, U. (1974). *J. Phys. A: Math., Nucl. Gen.* **7**, 1627–1633.
- Smith, P. R. and Aebi, U. (1976). *J. Mol. Biol.* **106**, 271–276.
- Smith, P. R., Aebi, U., Josephs, R. and Kessel, M. (1976). *J. Mol. Biol.* **106**, 243–270.
- Stewart, M. (1988). *J. Electron Microsc. Tech.* **9**, 325–358.
- Taylor, C. A. and Lipson, H. (1964). *Optical Transforms*. G. Bell & Sons, London.

- Thon, F. (1966). *Z. Naturforschung* **21a**, 476–478.
- Trachtenberg, S. and DeRosier, D. J. (1987). *J. Mol. Biol.* **194**, 581–602.
- Trachtenberg, S., DeRosier, D. J. and Macnab, R. M. (1987). *J. Mol. Biol.* **194**, 603–620.
- Trus, B. L. and Steven, A. C. (1981). *Ultramicroscopy* **6**, 383–386.
- Unwin, P. N. T. and Henderson, R. (1975). *J. Mol. Biol.* **94**, 425–440.
- van Heel, M. and Frank, J. (1981). *Ultramicroscopy* **6**, 187–194.
- van Heel, M. and Keegstra, W. (1981). *Ultramicroscopy* **7**, 113–130.
- Verschoor, A., Frank, J., Wagenknecht, T. and Boublik, M. (1986). *J. Mol. Biol.* **187**, 581–590.
- Verschoor, A., Zhang, N.-Y., Wagenknecht, T., Obrig, T., Radermacher, M. and Frank, J. (1989). *J. Mol. Biol.* **209**, 115–126.
- Vigers, G. P. A., Crowther, R. A. and Pearse, B. M. F. (1986). *EMBO J.* **4**, 529–534.
- Vogel, R. W. and Provencher, S. W. (1988). *Ultramicroscopy* **24**, 223–240.
- Vogel, R. W., Provencher, S. W., von Bonsdorff, C. H., Adrian, M. and Dubochet, J. (1986). *Nature (Lond.)* **320**, 533–535.
- Warner, F. D. (1970). *J. Cell Biol.* **47**, 159–182.
- Watson, G. N. (1958). *A Treatise on the Theory of Bessel Functions*, 2nd edn. Cambridge University Press, Cambridge.
- Yanagida, M. (1977). *J. Mol. Biol.* **109**, 515–537.
- Yanagida, M., Boy de la Tour, E., Alif-Steinberger, C. and Kellenberger, E. (1970). *J. Mol. Biol.* **50**, 35–58.
- Zeitler, E. (1974). *Optik* **39**, 396–415.

NOTE ADDED IN PROOF

Image analysis of electron micrographs of bacteriorhodopsin sheets (Section 7.4.4(b)) has recently yielded an interpretable atomic model (Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E. and Downing, K. H. (1990), *J. Mol. Biol.* **213**, 899–929). Under sufficiently favourable conditions, therefore, the methods reviewed in this chapter can provide an alternative to protein crystallography and multi-dimensional nuclear magnetic resonance.