3. Alignment Methods

3.1. Quantitative Definitions of Alignment

Alignment is initially understood as an operation that is performed on two or more images with the aim of bringing a common motif contained in those images into register. Implicit in the term "common motif" is the concept of homogeneity of the image set: the images are deemed "essentially" the same; they differ only in the noise component, and perhaps the presence or absence of a relatively small ligand, a small change in conformation, or orientation. That the difference be small is an important stipulation; it ensures, in all alignment methods making use of the cross-correlation function (see section 3.3), that the contribution from the correlation of the main component with itself (the "autocorrelation term") is very large compared to the contribution stemming from its correlation with the difference structure. Alignment so understood is directly related to our visual concept of likeness; realizing it in the computer amounts to the challenge of making the computer perform as well as a 3-year-old child in arranging building blocks of identical shapes into a common orientation.

The introduction of dissimilar images, occurring in a heterogeneous image set, forces us to generalize the term alignment: according to this more expanded meaning, even dissimilar motifs occurring in those images are considered "aligned" when they are positioned such as to minimize a given functional. The generalized Euclidean distance (i.e., the variance of the difference image) is often

used as the functional. With such a quantitative definition, the precise relative position (relating to both shift and orientation) between the motifs after digital "alignment" may or may not agree with our visual assessment, which relies on the perception of edges and marks in both images rather than a digital pixel-by-pixel comparison.

The concept of homogeneous versus heterogeneous image sets is fundamental in understanding averaging methods, their limitations and the ways these limitations can be overcome. These topics will form the bulk of the remainder of this chapter and chapter 4.

3.2. Homogeneous Versus Heterogeneous Image Sets

3.2.1. Alignment of a Homogeneous Image Set

Assume that we have a micrograph that shows N "copies" of a molecule in the same view. With the use of an interactive selection program, these molecule images are separately extracted, normally within a square "window," and stored in a stack of arrays:

$$\{p_{ij},\ i = 1 \ldots, N;\ j = 1 \ldots, J\} \tag{3.9}$$

Within the selection window, the molecule is roughly centered, and it has normally random azimuthal "in-plane" orientations. We then seek coordinate transformations $\mathbf{T}_i$ such that their application to p_{ij} results in the precise superimposition of all realizations of the molecule view. In such a superimposition, any pixel indexed j in the transformed arrays $p'_{ij} = \mathbf{T}_i p_{ij}$ refers to the same point in the molecule projection's coordinate system. When this goal is achieved, it is possible to form a meaningful average[4] (figures 3.8 and 3.3a,c).

$$\bar{p}_j = \frac{1}{N}\sum_{i=1}^{N} p'_{ij} \tag{3.10}$$

In contrast, the average would be meaningless if the different pixels with the same index j originated from different points of the coordinate system of the molecule projection, resulting from a failure of alignment, or from the molecules having different conformation (as the proverbial "apples and oranges"), resulting from structural or orientational heterogeneity.

If the deviations are small, however, then the resulting average will at least be similar to the ideal average that would be obtained without alignment error: the former will be a slightly blurred version of the latter. For random translational

[4]Throughout this book, ensemble averages are denoted by a bar over the symbol that denotes the observed quantity; for example, $\bar{p}_j$ denotes the average over multiple measurements of the pixel j. In contrast, averages of a function over its argument range will be denoted by angle brackets, as in the following example: $\langle p \rangle = \frac{1}{J}\sum_{j=1}^{J} p_j$.

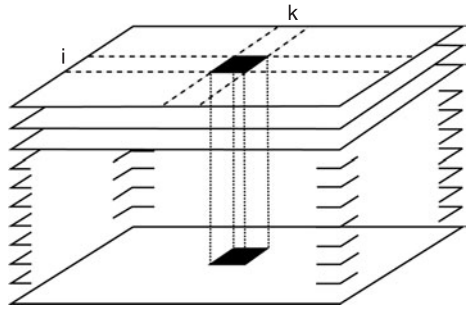

Figure 3.8 Definition of the average image. We imagine the images to be stacked up. For each pixel (indexed i, k), the column average is computed and stored in the element (i, k) of the resulting average image. At the same time, the variance of the pixel is computed and stored in the element (i, k) of the variance map. From Frank (1984b), reproduced with permission of the Electron Microscopy Foundation.

deviations, the blurring can be described in Fourier space by a Gaussian function, analog to the temperature factor of X-ray crystallography:

$$\mathfrak{F}\{\hat{p}\} = \mathfrak{F}\{\bar{p}\} \exp[-k^2/k_0^2] \tag{3.11}$$

where $\mathfrak{F}\{\cdot\}$ stands for the Fourier transform of the term within the bracket, k is the spatial frequency, and $k_0^2 = 1/r_0^2$ is a "temperature" parameter (the Debye–Waller factor of X-ray crystallography) due to random translations characterized by the size of the r.m.s. (root mean square) deviation r_0 (see also chapter 5, section 9.3, where this subject is discussed in greater detail).

3.2.2 Alignment of a Heterogeneous Image Set

In the case of a heterogeneous image set, such as a set comprising molecules presenting different views, alignment does not have the clear and unambiguous meaning as before. Rather, it must be defined in an operational way: as a procedure that establishes a defined geometrical relationship among a set of images by minimizing a certain functional. A well-behaved alignment algorithm will have the effect that particles within a homogeneous subset are "aligned" in the same sense as defined above for homogeneous sets, while particles belonging to different subsets are brought into geometrical relationships that are consistent with one another. To put it more concretely, in the alignment of 50S ribosomal subunits falling into two views, the crown view and the kidney view, all particles in the crown view orientation will be oriented consistently, and the same will be true for all particles in the kidney view orientation. At the same time, the orientation between any of the crown-view and any of the kidney-view particles will be fixed, but the exact size of the cross-group angle will depend on the choice of the alignment algorithm. Although the size of this relative, fixed angle is irrelevant, a fixed spatial relationship is required for an objective, reproducible, and meaningful characterization of the image set by multivariate data analysis

and classification. Exceptions are those methods that produce an alignment implicitly ("alignment through classification" of Dube et al., 1993; Marabini and Carazo, 1994a) or make use of invariants that lend themselves to classification (Schatz and van Heel, 1990, 1992; Schatz, 1992).

Heterogeneity may exist because molecules principally identical in structure may still be different in shape because of the different extents to which a structural component is flexed. Particles or fibers that are thin and extended may bend or deform without changing their local structure, often the true object of the study. From the point of view of studying the high-resolution structure, the diversity of overall shape may be seen as a mere obstacle and not in itself worthy of attention. Bacterial flagella present a good example of this situation. In those cases, a different approach to alignment and averaging may be possible, one in which the idealized overall particle shape is first restored by computational means. Structural homogeneity can thus be restored. The general approach to this problem is the use of curvilinear coordinate transformations. Such "unbending" methods have been introduced to straighten fibers and other linear structures in preparation for processing methods, thus far covered here, that assume rigid body behavior in all rotations and translations. The group of Alasdair Steven at the National Institutes of Health has used these methods extensively in their studies of fibrinous structures (Steven et al., 1986, 1988, 1991; Fraser et al., 1990). Geometrical unbending enables the use of helical reconstruction methods on structures whose shapes do not conform to the architecture of the ideal helix (Egelman, 1986; Steven et al., 1986; Hutchinson et al., 1990).

Yet a different concept of alignment comes to bear when an attempt is made to orient different projections with respect to one another and to a common 3D frame of reference; see section 3 in chapter 5. We will refer to this problem as the *problem of 3D projection alignment*. It is equivalent to the search for the common phase origin in the 3D reconstruction of 2D crystal sheets (Amos et al., 1982; Stewart, 1988a). An even closer analogy can be found in the common-lines methods used in the processing of images of spherical viruses (Crowther et al., 1970; Cheng et al., 1994).

3.3. Translational and Rotational Cross-Correlation

3.3.1. The Cross-Correlation Function Based on the Generalized Euclidean Distance

The cross-correlation function is the most important tool for alignment of two images. It can be derived in the following way: we seek, among all relative positions of the images (produced by rotating and translating one image with respect to the other), the one that maximizes a measure of similarity. The two images, represented by J discrete measurements on a regular grid, $\{f_1(\mathbf{r}_j), j=1\ldots,J\}$, $\{f_2(\mathbf{r}_j), j=1\ldots,J\}$, may be interpreted as vectors in a J-dimensional Cartesian coordinate system (see also chapter 4, where extensive use will be made of this concept). The length of the difference vector, or the generalized Euclidean distance between the vector end points, is often used as a measure of their dissimilarity or as an inverse measure of their similarity. By introducing search

parameters for the rotation and translation, $(\mathbf{R}_\alpha, \mathbf{r}')$, we obtain the expression to be minimized:

$$E_{12}^2(\mathbf{R}_\alpha, \mathbf{r}') = \sum_{j=1}^{J} [f_1(\mathbf{r}_j) - f_2(\mathbf{R}_\alpha \mathbf{r}_j + \mathbf{r}')]^2 \tag{3.12}$$

The rotation matrix $\mathbf{R}_\alpha$ denotes a rotation of the function $f_2(\mathbf{r})$ by the angle α, while the vector $\mathbf{r}'$ denotes a shift of the rotated function. In comparing two images represented by the functions $f_1(\mathbf{r})$ and $f_2(\mathbf{r})$, we are interested in finding out whether similarity exists for *any* combination of the search parameters. This kind of comparison is similar to the comparison our eyes perform—almost instantaneously—when judging whether or not shapes presented in arbitrary orientations are identical. By writing out the expression (3.12) explicitly, we obtain

$$E_{12}^2(\mathbf{R}_\alpha, \mathbf{r}') = \sum_{j=1}^{J} [f_1(\mathbf{r}_j)]^2 + \sum_{j=1}^{J} [f_2(\mathbf{R}_\alpha \mathbf{r}_j + \mathbf{r}')]^2 - 2 \sum_{j=1}^{J} f_1(\mathbf{r}_j) f_2(\mathbf{R}_\alpha \mathbf{r}_j + \mathbf{r}') \tag{3.13}$$

The first two terms are invariant under the coordinate transformation:

$$\mathbf{r}_j \rightarrow \mathbf{R}_\alpha \mathbf{r}_j + \mathbf{r}_k \tag{3.14}$$

The third term is maximized, as a function of the search parameters $(\mathbf{R}_\alpha, \mathbf{r}_k)$, when the generalized Euclidean distance E_{12} assumes its minimum. This third term is called the *cross-correlation function*:

$$\Phi_{12}(\mathbf{R}_\alpha, \mathbf{r}_k) = \sum_{j=1}^{J} f_1(\mathbf{r}_j) f_2(\mathbf{R}_\alpha \mathbf{r}_j + \mathbf{r}_k) \tag{3.15}$$

In practical application, the use of equation (3.15) is somewhat clumsy because determination of its maximum requires a 3D search (i.e., over the ranges of one rotational and two translational parameters). Functions that separately explore angular and translational space have become more important. Two additional impractical features of equation (3.15) are that (i) $\Phi_{12}(\mathbf{R}_\alpha, \mathbf{r}_k)$ is not normalized, so that it is not suitable for comparing images that originate from different experiments, and (ii) it is dependent on the size of the "bias" terms $\langle f_1 \rangle$ and $\langle f_2 \rangle$, that is, the averaged pixel values, as well as the size of the variances, which should all be irrelevant in a meaningful measure of similarity. These shortcomings are overcome by the introduction of the translational cross-correlation function and a suitable normalization, detailed in the following.

For later reference, we also introduce the *autocorrelation function* (ACF), which is simply the cross-correlation function of a function $f_1(\mathbf{r}_j)$ with itself (see appendix 1).

3.3.2. Cross-Correlation Coefficient and Translational Cross-Correlation Function

3.3.2.1. Definition of the cross-correlation coefficient The cross-correlation coefficient is a well-known measure of similarity and statistical interdependence. For two functions represented by discrete samples, $\{f_1(\mathbf{r}_j); j = 1 \ldots, J\}$ and

$\{f_2(\mathbf{r}_j); j=1 \ldots, J\}$, the cross-correlation coefficient is defined as

$$\rho_{12} = \frac{\sum_{j=1}^{J} [f_1(\mathbf{r}_j) - \langle f_1 \rangle][f_2(\mathbf{r}_j) - \langle f_2 \rangle]}{\left\{ \sum_{j=1}^{J} [f_1(\mathbf{r}_j) - \langle f_1 \rangle]^2 \sum_{j=1}^{J} [f_2(\mathbf{r}_j) - \langle f_2 \rangle]^2 \right\}^{1/2}} \tag{3.16}$$

where

$$\langle f_i \rangle = 1/J \sum_{j=1}^{J} f_i(\mathbf{r}_j); \; i = 1, 2 \tag{3.17}$$

Note that $-1 \leq \rho_{12} \leq 1$. A high value of ρ_{12} means that the two functions are very similar for a particular choice of relative shift and orientation. However, in comparing images, it is more relevant to ask whether ρ_{12} is maximized for *any* "rigid body" coordinate transformation applied to one of the images. In the defining equation (3.16), only the numerator is sensitive to the quality of the alignment between the two functions, while the denominator contains only the variances, which are invariant under shift or rotation of the image. When a variable coordinate transformation $\mathbf{r}_j \rightarrow \mathbf{R}_\alpha \mathbf{r}_j + \mathbf{r}_k$ is applied to one of the functions, the numerator becomes identical (except for the subtraction of the averages $\langle f_1 \rangle$ and $\langle f_2 \rangle$, which only changes the constant "bias" terms) with the cross-correlation function introduced in the previous section.

3.3.2.2. Translational cross-correlation function Specifically, the translational CCF is obtained by restricting the coordinate transformation to a 2D translation $\mathbf{r}_j \rightarrow \mathbf{r}_j + \mathbf{r}_k$. However, for any $\mathbf{r}_k \neq 0$, the images no longer overlap completely, so that the summation can now be carried out only over a subset J' of the J pixels. This is indicated by the stipulation $\mathbf{r}_j + \mathbf{r}_k \in \mathbf{A}$ under the summation symbol, meaning that only those indices j should be used, for which the sum vector $\mathbf{r}_j + \mathbf{r}_k$ is still within the boundary of the image:

$$\Phi_{12}(\mathbf{r}_k) = \frac{1}{J'} \sum_{j=1[r_j + r_k \in A]}^{J} [f_1(\mathbf{r}_j + \mathbf{r}_k) - \langle f_1 \rangle][f_1(\mathbf{r}_j + \mathbf{r}_k) - \langle f_1 \rangle] \tag{3.18}$$

By letting the "probing vector" $\mathbf{r}_k$ assume all possible positions on the grid, all relative displacements of the two functions are explored, allowing the position of best match to be found: for that particular $\mathbf{r}_k$, the CCF assumes its highest value (figure 3.9).

The formulation of the overlap contingency is somewhat awkward and indirect in the lexicographic notation, as used in equation (3.18). However, it is easy to write down a version of this formula that uses separate indexing for the x- and y-coordinates. In such a revised formula, the index boundaries become explicit. Because the practical computation of the CCF is normally done by fast

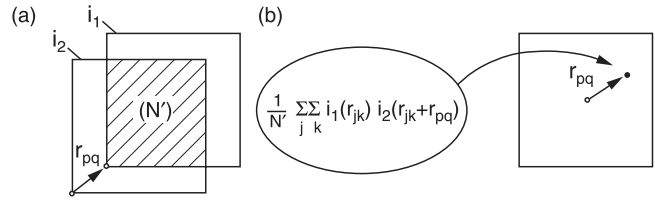

Figure 3.9 Definition of the cross-correlation function (CCF). For the computation of the CCF for a single argument $\mathbf{r}_{pq}$, (a) is shifted with respect to (b) by vector $\mathbf{r}_{pq}$. In this shifted position, the scalar product of the two images arrays is formed and put into the CCF matrix at position (p, q). For the computation of the entire function, the vector $\mathbf{r}_{pq}$ is allowed to assume all possible positions on the sampling grid. In the end, the CCF matrix has an entry in each position. From Frank (1980), reproduced with permission of Springer Verlag.

Fourier transformations (see below), we will refrain from restating the formula here. However, an issue related to the overlap will resurface later on in the context of fitting and docking in three dimensions by "globally" versus "locally" normalized cross-correlation (see chapter 6, section 6.2.2).

3.3.2.3. Computation using the fast Fourier transform In practical application, the computation of $\Phi_{12}(\mathbf{r}_k)$ is speeded up through the use of a version of the convolution theorem (appendix 1). For notational convenience, this is restated here in the continuous form:

The convolution product of two functions,

$$C_{12}(\mathbf{r}') = \iint f_1(\mathbf{r}) f_2(\mathbf{r} - \mathbf{r}')\, \mathrm{d}\mathbf{r} \tag{3.19}$$

is equal to the inverse Fourier transform of the product $\mathfrak{F}\{f_1\}\mathfrak{F}\{f_2\}$, *where* $\mathfrak{F}\{\cdot\}$ *denotes the Fourier transform of the function in the bracket*. The definition of the CCF differs from that of the convolution equation (3.19) only in the substitution of a plus sign for the minus in the vector argument, and this is reflected in complex conjugation when forming the product in Fourier space. The resulting recipe for fast computation of the CCF is

$$\Phi_{12}(\mathbf{r}') = \mathfrak{F}^{-1}\{\mathfrak{F}\{f_1\}\mathfrak{F}^*\{f_2\}\} \tag{3.20}$$

where $\mathfrak{F}^{-1}$ denotes the inverse Fourier transform, and * denotes complex conjugation. Thus, the computation involves three discrete fast Fourier transformations and one scalar matrix multiplication.

The usual choice of origin in the digital Fourier transform (at element [1,1] of the array) requires an adjustment: for the origin to be in a more convenient position, namely in the center of $\Phi_{12}(\mathbf{r}')$, a shift by half the image dimensions in the x- and y-directions is required. For even array dimensions (and thus for any power-of-two based algorithms) this is easily accomplished by multiplication of the Fourier product in equation (3.20) with the factor $(-1)^{(k_x+k_y)}$, where k_x, k_y are the integer grid coordinates in Fourier space.

For convenience, we will sometimes make use of the notation

$$\Phi_{12}(\mathbf{r}') = f_1 \otimes f_2 \tag{3.21}$$

in analogy to the use of

$$C_{12}(\mathbf{r}') = f_1(\mathbf{r}) \circ f_2(\mathbf{r}) \tag{3.22}$$

for the convolution product.

The computation via the Fourier transformation route implies that the 2D image array $p(l, m)$ (switching back for the moment to a 2D argument formulation for clarity) is replaced by a circulant array, with the periodic properties

$$p(l + L, m + M) = p(l, m) \tag{3.23}$$

This has the consequence that the translational CCF computed in this way, $\Phi_{12}(l', m')$ in discrete notation, contains contributions from terms $p_1(l + l' - L, m + m' - M)\ p_2(l, m)$ when $l + l' > L$ and $m + m' > M$. We speak of a "wrap-around effect": instead of the intended overlap of image areas implied in the definition of the CCF [equation (3.18) and figure 3.9], the bottom of the first image now overlaps the top of the second, etc.

To deal with this problem, it is common practice to extend the images two-fold in both directions, by "padding" them with their respective averages (i.e., filling in the average value into the array elements unoccupied by the image):

$$\langle p_1 \rangle = \frac{1}{J} \sum_{j=1}^{J} p_1(\mathbf{r}_j);\ \langle p_2 \rangle = \frac{1}{J} \sum_{j=1}^{J} p_2(\mathbf{r}_j) \tag{3.24}$$

Alternatively, the images may be "floated," by subtraction of $\langle p_1 \rangle$ and $\langle p_2 \rangle$, respectively, then "padded" with zeros prior to the fast Fourier transform calculation, as described by DeRosier and Moore (1970). The result is the same as with the above recipe, because any additive terms have no influence on the CCF, as defined in equation (3.18). In fact, programs calculating the CCF frequently eliminate the Fourier term F_{00} at the origin in the course of the computation, rendering the outcome insensitive to the choice of padding method.

Another difference between the CCF computed by the Fourier method and the CCF computed following its real-space definition, equation 3.18, is that the normalization by $1/J'$ (J' being the varying number of terms contributing to the sum) is now replaced by $1/J$ [or, respectively, by $1/(4J)$ if twofold extension in both dimensions by padding is used].

3.3.3. Rotational Cross-Correlation Function

The rotational CCF (analogous to the *rotation function* in X-ray crystallography) is defined in a similar way as the translational CCF, but this time with a rotation as probing coordinate transformation. Here, each function is represented by

samples on a polar coordinate grid defined by Δr, the radial increment, and $\Delta\phi$, the azimuthal increment:

$$\{f_i(l\Delta r, m\Delta\phi);\ l = 1\ldots, L;\ m = 1\ldots, M\};\quad i = 1, 2 \tag{3.25}$$

We define the discrete, weighted, rotational CCF in the following way:

$$\begin{aligned} C(k) &= \sum_{l=l_1}^{l_2} w(l) \sum_{m=0}^{M} f_1(l\Delta r,\ \mathrm{mod}\,[m+k, M]\Delta\phi) f_2(l\Delta r, m\Delta\phi)\Delta\phi l\Delta r \\ &= \sum_{l=l_1}^{l_2} w(l) c(l,k) l\Delta r \end{aligned} \tag{3.26}$$

If weights $w(l) \equiv 1$ are used, the standard definition of the rotational CCF is obtained. The choice of nonuniform weights, along with the choice of a range of radii $\{l_1 \ldots l_2\}$, is sometimes useful as a means to place particular emphasis on the contribution of certain features within that range. For example, if the molecule has a peripheral protuberance that is flexible, we might want to prevent it from contributing to the rotational correlation signal.

The computation of the 1D inner sums $c(l, k)$ along rings normally takes advantage (Saxton, 1978; Frank et al., 1978a, 1986) of the *Fourier convolution theorem*, here in discrete form:

$$c(l,k) = \sum_{m'=0}^{M-1} F_1(l\Delta r, m'\Delta\phi) F_2(l\Delta r, m'\Delta\phi)\Delta\phi' \exp[2\pi i m'(\Delta\phi - \Delta\phi')] \tag{3.27}$$

where $F_i(l\Delta r, m'\Delta\phi)$, $i = 1, 2$ are the discrete Fourier transforms of the lth ring of the functions f_i. A further gain in speed is achieved by reversing the order of the summations over rings and over Fourier terms in equations (3.26) and (3.27), as this reduces the number of inverse 1D Fourier transformations to one (Penczek et al., 1992).

3.3.4. Peak Search

The search for the precise position of a peak is a common feature of all correlation-based alignment techniques. The following rule of thumb has been found effective in safeguarding against detection of spurious peaks in aligning two particle images: not just the highest, but at least the three highest-ranking peaks are determined. Let us assume their respective values are p_1, p_2, and p_3. For a significant peak, one would intuitively expect that the ratio p_1/p_2 is well above p_2/p_3, on the assumption that the subsidiary peaks p_2 and p_3 are due to noise.

Computer programs designed for the purpose of searching peaks are straightforward: the array is scanned for the occurrence of relative peaks, that is, elements that stand out from their immediate neighbors. In a 1D search

(e.g., of a rotational cross-correlation function), each element of the array is compared with its two immediate neighbors. In a 2D search (typically of a 2D translational CCF), each element is compared to its eight immediate neighbors. (In a 3D search, to be mentioned for completeness, the number of next neighbors is $8+9+9=26$ as in a 3D tic-tac-toe). Those elements that fulfill this criterion are put on a stack in ranking order. At the end of the scan, the stack will contain the desired list of highest peaks.

The peak position so found is given only as an integer multiple of the original sampling distance. However, the fact that the peak has finite width and originates mathematically from many independent contributions all coherently "focused" on the same spot means that the position can be found with higher accuracy by the use of some type of fitting. First, the putative peak region is defined as a normally circular region around the element with highest value found in the peak search. Elements within that region can now be used to determine an effective peak position with noninteger coordinates. Methods widely used are parabolic fitting and the computation of the center of gravity.

The accuracy in determining the peak position is determined by the SNR (see section 4.3), not by the resolution reflected in the peak width. As the linear transfer theory shows [equation (3.33) in section 3.4.1], the peak shape is determined by the (centrosymmetric) *autocorrelation* function (ACF) of the point spread function. Its precise center can be found by a fitting procedure, provided that the SNR is sufficiently large.

3.4. Reference-Based Alignment Techniques

3.4.1. Principle of Self-Detection

Reference-based alignment techniques were developed primarily for the case of homogeneous image sets, that is, images originating from particles containing identical structures and presenting the same view. In that case, all images of the set $\{p_n(\mathbf{r});\ n=1\ldots,N\}$ have a "signal component" in common—the projection $p(\mathbf{r})$ of the structure as imaged by the instrument—while differing in the noise component $n_i(\mathbf{r})$. Any of the images can then act as reference for the rest of the image set (principle of *self-detection*; see Frank, 1975a). In formal terms:

$$p_1(\mathbf{r}) = p(\mathbf{r}) + n_1(\mathbf{r}) \tag{3.28}$$

$$p_2(\mathbf{r}) = p(\mathbf{r}) + n_2(\mathbf{r}) \tag{3.29}$$

so that, using the notation for correlation introduced in section 3.3.2,

$$\Phi_{12}(\mathbf{r}') = p_1(\mathbf{r}) \otimes p_2(\mathbf{r}) = p(\mathbf{r}) \otimes p(\mathbf{r}) + n_1(\mathbf{r}) \otimes p(\mathbf{r}) + p(\mathbf{r}) \otimes n_2(\mathbf{r}) + n_1(\mathbf{r}) \otimes n_2(\mathbf{r}) \tag{3.30}$$

The first term is the ACF of the structure common to both images, which has a sharp peak at the origin, while each of the other three terms is a CCF of two statistically unrelated functions. The shape of the peak at the center is determined solely by the ACF of the point-spread function associated with the contrast transfer function: according to equation (2.11) with $p(\mathbf{r}) = I(\mathbf{r})$ and $p_0(\mathbf{r}) = C_p(\mathbf{r})$ and, by virtue of the convolution theorem, one obtains

$$p(\mathbf{r}) \otimes p(\mathbf{r}) = [h(\mathbf{r}) \circ p_0(\mathbf{r})] \otimes [h(\mathbf{r}) \circ p_0(\mathbf{r})] = [h(\mathbf{r}) \otimes h(\mathbf{r})] \circ [p_0(\mathbf{r}) \otimes p_0(r)] \quad (3.31)$$

The term $[p_0(\mathbf{r}) \otimes p_0(\mathbf{r})]$ is, in the terminology of X-ray crystallography, the *Patterson function* of the projection of the original structure. Its most important feature in equation (3.31) is that it acts, for all practical purposes, as a delta-function, due to the appearance of the sharp peak stemming from exact overlap of the atomic object with itself for zero displacement vector; that is, in the center of the function. Since the convolution of a function with the delta-function simply acts to reproduce the function (see appendix 1), we obtain the result that equation (3.31) is essentially the same as $h(\mathbf{r}) \otimes h(\mathbf{r})$; that is, the ACF of the instrument's point spread function.[5] Its value at the origin, important for the ability to detect the peak at low SNR, is determined by the size of the "energy" integral:

$$E_B = \int_B ||H(\mathbf{k})||^2 d\mathbf{k} \quad (3.32)$$

where B is the resolution domain and $H(\mathbf{k})$ is the contrast transfer function. This follows from Parseval's theorem, which is simply a statement of the invariance of the norm of a function on transforming this function into Fourier space (see section 4.3).

In this context, it should be noted that the alignment of images of the same object taken at different defocus settings leads to a CCF peak whose shape is determined by

$$\Phi_{h_1 h_2} = h_1(\mathbf{r}) \otimes h_2(\mathbf{r}) \quad (3.33)$$

with $h_1(\mathbf{r})$ and $h_2(\mathbf{r})$ being the point-spread functions representing the action of the instrument for the two defocus settings. In this situation, the peak height is determined by the size of the "cross-energy" integral (i.e., its discrete-valued equivalent):

$$E_B = \int_B H_1(\mathbf{k}) H_2(\mathbf{k}) d\mathbf{k} \quad (3.34)$$

which is critically dependent on the relative positions of contrast transfer zones with different polarity (Frank, 1972b; Al-Ali, 1976; Al-Ali and Frank, 1980).

[5]This property was previously exploited by Al-Ali and Frank (1980), who proposed the use of the cross-correlation function of two micrographs of the same "stochastic" object to obtain a measure of resolution.

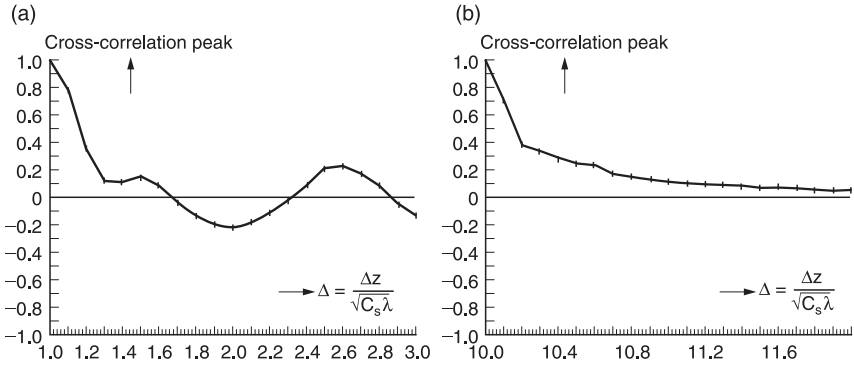

Figure 3.10 Peak value of the cross-correlation function $\Phi_{h_1h_2}$ between two point spread functions with different defocus, plotted as a function of their defocus difference. Theoretically, the cross-correlation function of two micrographs of the same structure is obtained as a convolution product between $\Phi_{h_1h_2}$ and the Patterson function of the structure, see equation (3.31). (a) $\Delta\hat{z} = 1$ (Scherzer focus) used as reference; (b) $\Delta\hat{z} = 10$ used a reference. It is seen that in the case of (a), as the defocus difference increases, the peak drops off to zero and changes sign twice. From Zemlin (1989b), reprinted with permission of John Wiley & Sons, Inc.

[As an aside, this integral is Linfoot's measure of *correlation quality* (Linfoot, 1964).] In fact, the value of the integral in equation (3.34) is no longer positive-definite, and unfortunate defocus combinations can result in a CCF with a peak that has inverted polarity or is so flat that it cannot be detected in the peak search as it is drowned by noise (Frank, 1980; Zemlin, 1989b; Saxton, 1994; see figure 3.10).

Saxton (1994) discussed remedies for this situation. One of them is the obvious "flipping" of transfer zones in the case where the transfer functions are known, with the aim of ensuring a positive-definite (or negative-definite, whatever the case may be) integrand (Typke et al., 1992). A more sophisticated procedure suggested by Saxton (1994) is to multiply the transform of the CCF with a factor that acts like a Wiener filter:

$$W(\mathbf{k}) = \frac{H_1(\mathbf{k})H_2(\mathbf{k})}{|H_1(\mathbf{k})|^2|H_2(\mathbf{k})|^2 + \varepsilon} \tag{3.35}$$

where H_1 and H_2 are the known CTFs of the two images and ε is a small quantity that ensures the boundedness of $W(\mathbf{k})$, as it keeps the noise amplification in any spectral domain within reasonable margins.

Apart from the degradation of the CCF peak due to the mismatch in CTF polarities, which can be fixed by "flipping" the polarity of zones in the Fourier domain, there are other effects that diminish the size of the peak, and thus may lead to difficulties in the use of the CCF in alignment. Typke et al. (1992) identified magnification changes and local distortions of the specimen as sources

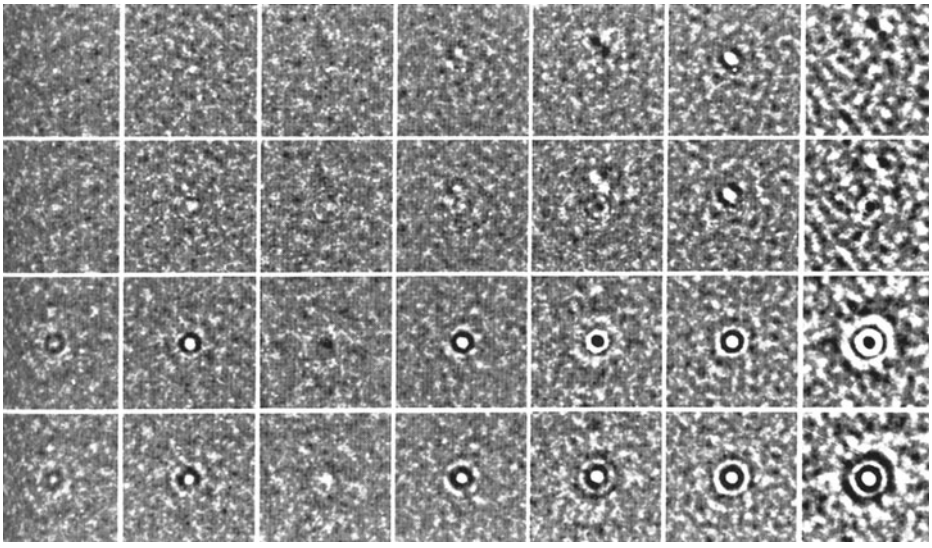

Figure 3.11 The effect of image restoration and magnification correction on the cross-correlation function (CCF) of two micrographs of a defocus series. One of eight images is cross-correlated against the remaining seven images. Only the central 128×128 portion of the CCFs is shown. Top row: without correction. Second row: sign reversal applied to CTFs to make the CCF Fourier integral in equation (3.34) positive-definite. Third row: micrographs corrected only for magnification changes and displacements. The CCF peak comes up strongly now, but in three cases with reversed sign. Bottom row: micrographs additionally corrected for sign reversal. All CCF peaks are now positive. From Typke et al. (1992), reproduced with permission of Elsevier.

of those problems and demonstrated that, by applying an appropriate compensation, a strong CCF peak can be restored (see figure 3.11). However, these effects come into play mainly in applications where large specimen fields must be related to one another, while they are negligible in the alignment of small (typically in the range of 64×64 to 150×150) single-particle arrays. The reason for the insensitivity in the latter case is that all local translational components are automatically accounted for by an extra shift, while the small local rotational components of the distortions (in the range of maximally 1°) affect the CCF Fourier integral underlying the CCF [equation (3.30)] only marginally in the interesting resolution range (see appendix in Frank and Wagenknecht, 1984).

3.4.2. ACF/CCF-Based Search Strategy

3.4.2.1. Rationale In order to avoid a time-consuming 3D search of the parameter space spanned by $\alpha, \Delta x, \Delta y$, translation and rotation searches are normally performed separately. One search strategy that accomplishes this separation, introduced by Langer et al. (1970), makes use of the translation invariance of the ACF (figures 3.12a–d). In fact, this method goes back to search techniques in X-ray crystallography involving the Patterson function. According

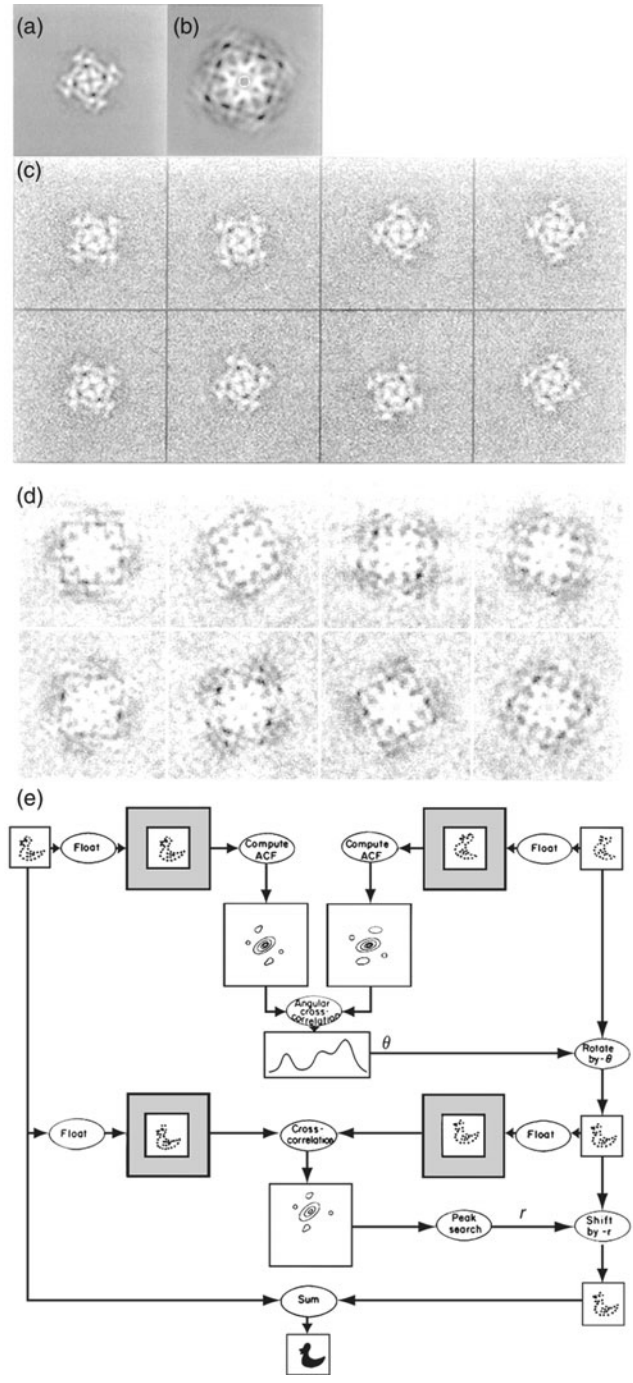

Figure 3.12 Alignment using the autocorrelation function (ACF). (a) A projection of the calcium release channel, padded into a 128×128 field. (b) ACF of (a). (c) Eight noisy realizations of (a), shifted and rotated randomly. (d) ACFs of the images in (c). Two

to this scheme (figure 3.12e), the orientation between the images is first found by determining the orientation between their ACFs, and subsequently the shift between the correctly oriented molecules is found by translational cross-correlation.

The ACF of an image $\{p(\mathbf{r}_j); j=1\ldots, J\}$, represented by discrete samples on a regular grid, $\mathbf{r}_j$, is obtained by simply letting

$$f_1(\mathbf{r}_j) \equiv f_2(\mathbf{r}_j) \equiv p(\mathbf{r}_j)$$

in the formula for the translational CCF [equation (3.18)]:

$$\Phi(\mathbf{r}_k) = \frac{1}{J'} \sum_{j=1(\mathbf{r}_j+\mathbf{r}_k \in A)}^{J} p(\mathbf{r}_j)p(\mathbf{r}_j + \mathbf{r}_k) \tag{3.36}$$

The property of shift invariance is immediately clear from the defining formula, since the addition of an arbitrary vector to $\mathbf{r}_j$ will not affect the outcome (except possibly for changes due to the boundary terms). Another property (not as desirable as the shift invariance; see below) is that the two functions $p(\mathbf{r}_j)$ and $p'(\mathbf{r}_j) = p(-\mathbf{r}_j)$ have the same ACF. As a result, the ACF is always centrosymmetric.

For fast computation of the ACF, the *convolution theorem* is again invoked as in the case of the CCF (section 3.3.2): the ACF is obtained by inverse Fourier transformation of the squared Fourier modulus, $|F(\mathbf{k})|^2$, where $F(\mathbf{k}) = \mathfrak{F}^{-1}\{p(r)\}$. Here, the elimination of "wrap-around" artifacts, by twofold extension and padding of the array prior to the computation of the Fourier transform, is

---

important properties of the ACF can readily be recognized: it is always centered at the origin, irrespective of the translational position of the molecule, and it has a characteristic pattern that reflects the (in-plane) orientation of the molecule in (c) from which the ACF was derived. Thus, it is seen that the ACF can be used to determine the rotation of unaligned, uncentered molecules. (e) Scheme of two-step alignment utilizing the translation invariance of the ACF. Both the reference and the image to be aligned (images on the top left and top right, respectively) are "floated," or padded into a larger field having the same average density, to avoid wrap-around artifacts. Next, the ACFs are computed and rotationally cross-correlated. The location of the peak in the rotational CCF establishes the angle θ between the ACFs. From this location it is inferred that the image on the top right has to be rotated by $-\theta$ to bring it into the same orientation as the reference. (Note, however, that this conclusion is correct only if the motif contained in the images to be aligned is centrosymmetric. Otherwise, the angle θ between the ACFs is compatible with the angle being either θ or $\theta + 180°$, which means both positions have to be tried in the following cross-correlation; see text.) Next, the correctly rotated, padded image is translationally cross-correlated with the reference, yielding the CCF. The position $\mathbf{r}$ of the peak in the CCF is found by a 2D peak search. Finally, the rotated version of the original image is shifted by $-\mathbf{r}$ to achieve complete-alignment. From Frank and Goldfarb (1980), reproduced with permission of Springer-Verlag.

particularly important, since these artifacts are peripherally located and may lead to incorrect angles in the rotation search.

We consider an idealized situation, namely two images of the same motif, which differ by a shift, a rotation α, and the addition of noise. The ACFs of these images are identical in appearance, except that they are rotated by α relative to each other. As in section 3.3.2, we now represent the ACFs in a polar coordinate system and determine the relative angle by computing their rotational CCFs:

$$C(k) = \sum_{l=l_1}^{l_2} w(l)\Phi(l\Delta r, m\Delta\phi + \alpha)\Phi(l\Delta r, \Delta\phi(m+k))\Delta\phi l\Delta r \qquad (3.37)$$

The rotational CCF will have a peak centered at the discrete position $k_{\max} = \text{int}(\alpha/\Delta\phi)$ [where int(·) denotes the integer closest to the argument], which can be found by a maximum search over the entire range of the function.

The position may be found more accurately by using a parabolic fit or center-of-gravity determination in the vicinity of $k_{\max}$ (see section 3.3.4). The weights $w(l)$, as well as the choice of minimum and maximum radii l_1, l_2, are used to "focus" the comparison on certain ring zones of the autocorrelation function. For instance, if weights are chosen such that a radius $r = r_w$ is favored, then this has the effect that features in the molecule separated by the distance $|\mathbf{r}_1 - \mathbf{r}_2| = r_w$ will provide the strongest contributions in the computation of the rotational CCF. The weights can thus be used, for instance, to select the most stable, reliable (i.e., reproducible) distances occurring in the particle.

3.4.2.2. Ambiguity One problem of the ACF-based orientation search is the above-mentioned symmetry, which makes the ACFs of the two functions $p_1(\mathbf{r})$ and $p_2(\mathbf{r}) = p_1(-\mathbf{r})$ indistinguishable. Consequently, the fact that the ACFs of two images optimally match for $k\,\Delta\Phi = \alpha$ may indicate that the images match with the relative orientations (i) α, (ii) $\alpha + 180°$, or (iii) both. (The last case would apply if the images themselves are centrosymmetric.)

Zingsheim et al. (1980) solved this problem by an "up/down cross-correlation test," in the following way (figure 3.12):

(i) Rotate image #2 by α
(ii) Compute the CCF between #1 and #2, which results in a peak maximum ρ_α at the location $\{\Delta x_\alpha, \Delta y_\alpha\}$
(iii) Rotate image #2 by $\alpha + 180°$
(iv) Compute the CCF between #1 and #2, which results in a peak maximum $\rho_{\alpha+180°}$ at $\{\Delta x_{\alpha+180°}, \Delta y_{\alpha+180°}\}$
(v) If $\rho_\alpha > \rho_{\alpha+180°}$ then shift #2 by $\{\Delta x_\alpha, \Delta y_\alpha\}$, else by $\{\Delta x_{\alpha+180°}, \Delta y_{\alpha+180°}\}$; i.e., both ways are tried, and the orientation that gives maximum cross-correlation is used

Later the introduction of multivariate data analysis and classification (see chapter 4) provided another way of distinguishing between particles lying in 180°-rotation related positions. In the multivariate data analysis of such

a mixed data set, it is clear that the factor accounting for the "up" versus "down" mass redistribution will predominate, unless the particle view is close to centrosymmetric.

3.4.3. Refinement and Vectorial Addition of Alignment Parameters

Inevitably, the selection of an image for reference produces a bias in the alignment. As a remedy, an iterative refinement scheme is applied (figure 3.13), whereby the average resulting from the first alignment is used as reference in the next alignment pass, etc. (Frank et al., 1981a). The resolution of averages improves in each pass, up to a point of saturation.

However, when multiple steps of rotations and shifts are successively applied to an image, the discrete representation of images poses practical difficulties. As the image is subjected to several steps of refinement, the necessary interpolations successively degrade the resolution. As a solution to this problem, the final image is directly computed from the original image, using a rotation and shift combination that results from vectorial additions of rotations and shifts obtained in each step.

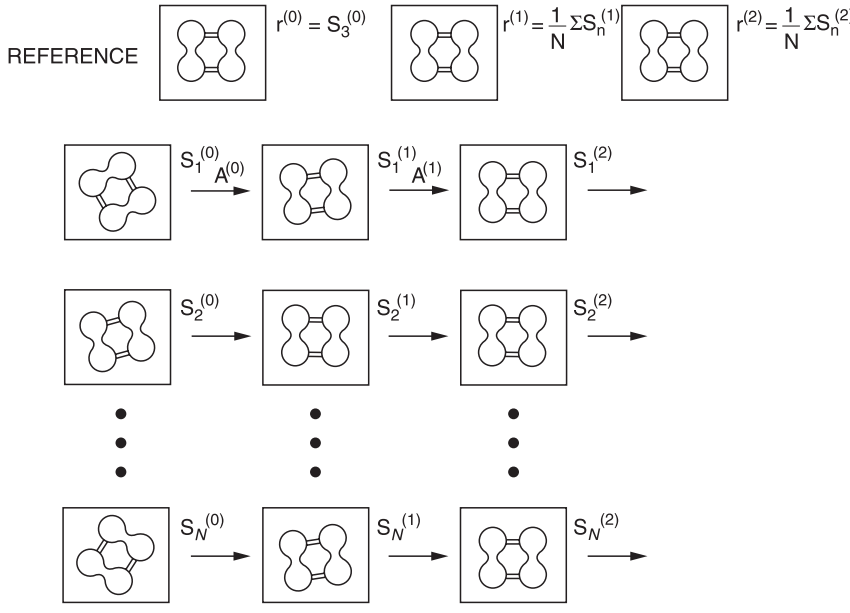

Figure 3.13 Reference-based alignment with iterative refinement. Starting with an arbitrary reference particle (picked for its "typical" appearance), the entire set of particles is aligned in the first pass. The aligned images are averaged, and the resulting average image is used as reference in the second pass of alignment. This procedure is repeated several times, until the shifts and rotation angles of the entire image set stabilize, that is, remain virtually unchanged. Adapted from Frank (1982).

For any pixel with coordinates $\mathbf{r}$ in the original image, we obtain new coordinates in the first alignment step as

$$\mathbf{r}' = \boldsymbol{\alpha}_1 \mathbf{r} + \mathbf{d}_1 \tag{3.38}$$

where $\boldsymbol{\alpha}_1$ and $\mathbf{d}_1$ are, respectively, the 2×2 rotation matrix and translation vector found in the alignment. In the next step (i.e., the first step of refinement), we have similarly

$$\mathbf{r}'' = \boldsymbol{\alpha}_2 \mathbf{r} + \mathbf{d}_2 \tag{3.39}$$

where $\boldsymbol{\alpha}_1$ and $\mathbf{d}_2$ are, respectively, the rotation matrix and translation vector expressing the adjustments. Instead of using the image obtained by applying the second transformation, one must apply a consolidated coordinate transformation to the original image. This consolidated transformation is

$$\mathbf{r}'' = \boldsymbol{\alpha}_2(\boldsymbol{\alpha}_1 \mathbf{r} + \mathbf{d}_1) + \mathbf{d}_2 = \boldsymbol{\alpha}_2 \boldsymbol{\alpha}_1 \mathbf{r} + \boldsymbol{\alpha}_2 \mathbf{d}_1 + \mathbf{d}_2 = \boldsymbol{\alpha}_{\text{res}} \mathbf{r} + \mathbf{d}_{\text{res}} \tag{3.40}$$

with the resulting single-step rotation $\boldsymbol{\alpha}_{\text{res}} = \boldsymbol{\alpha}_2 \boldsymbol{\alpha}_1$ and single-step translation $\mathbf{d}_{\text{res}} = \mathbf{d}_1 + \mathbf{d}_2$.

3.4.4. Multireference Methods

Multireference methods have been developed for situations in which two or more different motifs (i.e., different views, particle types, conformational states, etc.) are present in the data set (van Heel and Stöffler-Meilicke, 1985). Alignment of a set of N images with L references (i.e., prototypical images showing the different motifs) leads to an array of $L \times N$ correlation coefficients, and each image is put into one of L bins, depending on which of its L correlation coefficients is maximum. In a second round, averages are formed over the subsets so obtained, which are lower-noise realizations of the motifs and can take the place of the initial references. This refinement step produces improved values of the rotational and translational parameters, but also a migration of particles from one bin to another, on account of the now-improved discriminating power of the cross-correlation in "close-call" cases. This procedure may be repeated several times until convergence is achieved, as indicated by the fact that the parameter values stabilize and the particle migration stops. Multireference alignment, by its nature, is intertwined with classification, since the initial choice of references is already based on a presumed inventory of existing views, and the binning of particles according to highest CCF achieved in the alignment is formally a case of supervised classification (chapter 4, section 4.10).

Elaborate versions of this scheme (e.g., Harauz et al., 1988) have incorporated multivariate data analysis and classification in each step. Earlier experience of van Heel and coworkers (Boekema et al., 1986; Boekema and Böttcher, 1992; Dube et al., 1993) indicated that the multireference procedure is not necessarily stable; initial preferences may be amplified, leading to biased results, especially for small particles. The failure of this procedure is a consequence of an intrinsic problem

of the reference-based alignment approach. In fact, it can be shown that the multireference alignment algorithm is closely related to the *K*-means clustering technique, sharing all of its drawbacks (P. Penczek, personal communication, 1995).

3.5. Reference-Free Alignment Techniques

3.5.1. Introduction

As we have seen, reference-based alignment methods fail to work for heterogeneous data sets when the SNR (section 4.3) drops below a certain value. The choice of the initial reference can be shown to bias the outcome (see also Penczek et al., 1992). For instance, van Heel et al. (1992a,b) and Grigorieff (2000) demonstrated that, in extreme cases, a replica of the reference emerges when the correlation averaging technique is applied to a set of images containing pure random noise. A similar observation was earlier reported by Radermacher et al. (1986b) for correlation averaging of crystals: when the SNR is reduced to 0 (i.e., only a noise component remains in the image), an average resembling the reference motif is still formed. This phenomenon is easily understood: maximum correlation between a motif and an image field occurs, as a function of translation/rotation, when the image has maximal similarity. If (as often is the case) the correlation averaging procedure employs no correlation threshold, such areas are always found in a noise field. By adding up those areas of maximum similarity, one selectively reinforces all those noise components that optimally replicate the reference pattern. This becomes quite clear from an interpretation of images as points in a multidimensional space (see chapter 4, section 1.4.2).

The emergence of the signal out of "thin air," an inverse Cheshire cat phenomenon, is in some way analogous to the outcome of an experiment in optical filtering: a mask that passes reflections on the reciprocal grid is normally used to selectively enhance Fourier components that build up the structure repeating on the crystal lattice. If one uses such a mask to filter an image that consists of pure noise, the lattice and some of the characteristics of the structure are still generated, albeit with spatially varying distortions because the assignment of phases is left to chance. Application of rigorous statistical tests would of course dismiss the spurious correlation peaks as insignificant. However, the problem with such an approach is that it requires a detailed statistical model (which varies with many factors such as object type, preparation technique, imaging conditions, and electron exposure). Such a model is usually not available.

We will see later on (chapter 5, section 8.3) that cross-validation methods can be used to safeguard against noise building up to form an average (or a 3D reconstruction) that simply reproduces the reference. These methods can also be used to detect model bias, but they do not lend themselves readily to a recipe for how to avoid it.

Attempts to overcome these problems have led to the development of reference-free methods of alignment. Three principal directions have been

taken; one, proposed by Schatz and van Heel (1990, 1992; see also Schatz et al., 1990; Schatz, 1992) eliminates the need for alignment among different classes of images by the use of invariants; the second, proposed by Penzcek et al. (1992), solves the alignment problem by an iterative method; and the third, developed for larger data sets (Dube et al., 1993; Marabini and Carazo, 1994a), is based on the fact that among a set of molecules sufficiently large and presenting the same view, subsets with similar "in-plane" orientation can always be found.

In explaining these approaches, we are occasionally forced to make advance reference to the subject of multivariate data analysis and classification, which will be described in some detail in chapter 4. For a conceptual understanding of the present issues, it may be sufficient to think of classification as a "black box" procedure that is able to sort images (or any patterns derived from them) into groups or classes according to their similarities.

3.5.2. Use of Invariants: The Double-Antocorrelation Function

Certain functions that are invariant both under rotation and translation of an image, but still retain distinct structural properties, can be used to classify a heterogeneous image set. The ACF of an image is invariant under a translation of a motif contained in it. This property has been used in the reference-based alignment method originally proposed by Langer et al. (1970; see also Frank, 1975a; Frank et al., 1978a; and section 3.4.2). The *double-autocorrelation function* (DACF) has the additional property that it is invariant under rotation, as well; it is derived from the normal translational ACF by subjecting the latter to an operation of rotational autocorrelation (Schatz and van Heel, 1990).

Following Schatz and van Heel's procedure (figure 3.14), the ACF is first resampled on a polar grid to give $\hat{\Phi}(r_n, \phi_1)$. From this expression, the rotational ACF is obtained by computing 1D fast Fourier transforms (FFTs) and conjugate scalar Fourier products along rings, using the same philosophy as in the computation of the rotational CCF; see equations (3.26) or (3.37). The resulting function has the remarkable property of being invariant under *both* translation and rotation, because the addition of an arbitrary angle to ϕ_1, in the above expression $\hat{\Phi}(r_n, \phi_1)$, again does not affect the outcome.

Later, Schatz and van Heel (1992; see also van Heel et al., 1992b) introduced another function, which they termed the *double self-correlation function* (DSCF), that avoids the "dynamic" problem caused by the twofold squaring of intensities in the computation of the DACF. The modification consists of a strong suppression of the low spatial frequencies by application of a band-pass filter. Through the use of either DACF or DSCF, classification of the aligned, original images can be effectively replaced by a classification of their invariants, which are formed from the images without prior alignment. In these schemes, the problem of alignment of a heterogeneous image set is thus entirely avoided; alignment needs only to be applied separately, after classification of the invariants, to images within each homogeneous subset.

This approach of invariant classification poses a principal problem, however, which is obvious from its very definition (Schatz and van Heel, 1990; Frank et al.,

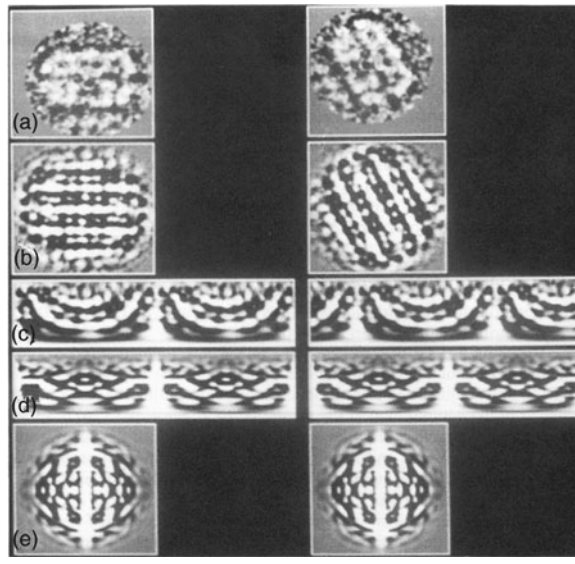

Figure 3.14 Demonstration of the invariance of the double autocorrelation function (DACF) under rotation and shift of an image. (a) Image of worm hemoglobin in side view and the same image rotated and shifted. (b) Autocorrelation functions of the images in (a). The ACF of the image on the left is seen to be identical to that on the right, but rotated by the same angle as the image. (c) The ACFs of (b) in polar coordinates (horizontal axis: azimuthal angle, from 0° to 360°; vertical axis: radius). The rotation between the ACFs is reflected by a horizontal shift of the polar coordinate representation. (d) A second, 1D autocorrelation in the horizontal direction produces identical patterns for both images: the DACF in a polar coordinate representation. (e) The DACF mapped into the Cartesian coordinate system. From Schatz (1992), reproduced with permission of Verlag Hänsel-Hohenhausen.

1992; Penczek et al., 1992): the DACFs and DSCFs are twofold degenerate representations of the initial images. Forming the ACF (or its "self-correlation" equivalent) of an image corresponds to eliminating the phase information in its Fourier transform. In the second step, in forming the DACF or DSCF from the ACF, the phases describing the rotational features in the 1D Fourier transforms along rings are also lost. Therefore, a classification of the invariants does not necessarily map into a correct classification of the images themselves. Theoretically, a very large number of patterns have the same DACF; the saving grace is that they are unlikely to be realized in the same experiment. Another problem is the reduction in SNR due to the loss of phases. Experience will tell whether the problem of ambiguity is a practical or merely an academic problem.

For completeness, another alignment approach making use of invariants, which has not gained practical importance because of inherent problems, should be mentioned. This approach is based on the use of moments (Goncharov et al., 1987; Salzman, 1990). Moments require a choice of an integration boundary and are extremely noise sensitive. They are, therefore, not well suited for alignment of raw data with very low SNRs.

3.5.3. Exhaustive Sampling of Parameter Space*

These are approaches based on the premise that groups of particles with closely related rotations can be found by applying multivariate data analysis (chapter 4) to the particle set that has been merely *translationally* aligned. If a data set is sufficiently large, then the number of particles presenting a similar appearance and falling within an angular range $\Delta\phi$ is on the average:

$$n_v = (\Delta\phi/(2\pi) \times n_{\text{tot}} \times p_v \tag{3.41}$$

where n_{tot} is the total number of particles, and p_v is the probability of encountering a particular view. For example, if the particle presents five views with equal probability ($p_v = 0.2$) and $n_{\text{tot}} = 10,000$, then $n_v = 60$ for $\Delta\phi = 10^\circ$.

Any particles that have the same structure and occur in similar azimuths will then fall within the same region of factor space. In principle, corresponding averages could be derived by the following "brute force" method: the most significant subspace of factor space (which might be just 2D) is divided according to a coarse grid, and images are averaged separately according to the grid square into which they fall. The resulting statistically well-defined averages could then be related to one another by rotational correlation, yielding the azimuthal angles. The within-group alignment of particles can be obviously refined so that eventually an angle can be assigned to each particle.

The method of "alignment through classification" by Dube et al. (1993) proceeds somewhat differently from the general idea sketched out above, taking advantage of the particle's symmetry, and in fact establishing the symmetry of the molecule (in this case, the head-to-tail connector protein, or "portal protein," of bacteriophage $\phi 29$) as seen in the electron micrograph. The molecule is rosette-shaped, with a symmetry that has been given variously as 12- or 13-fold by different authors. In the first step, molecules presenting the circular top view were isolated by a multireference approach. For unaligned molecules appearing in the top view with N-fold symmetry, a given pixel at the periphery of the molecule is randomly realized with high or low density across the entire population. If we proceed along a circle, to a pixel that is $360^\circ/(2N)$ away from the one considered, the pattern of variation across the population has consistently shifted by 180°: if the pixel in a given image was dark, it is now bright, and vice versa. Dube et al. (1993) showed how these symmetry-related patterns of variation are reflected in the eigenimages produced by multivariate data analysis (chapter 4).

3.5.4. Iterative Alignment Method*

Penczek et al. (1992) introduced a method of reference-free alignment based on an iterative algorithm that also avoids singling out an image for "reference." Its rationale and implementation are described in the following.

We first go back to an earlier definition of what constitutes the alignment of an entire image set. By generalizing the alignment between two images to a set of N images, Frank et al. (1986) proposed the following definition: *a set of N images is*

aligned if all images are pairwise aligned. In the notation by Penczek et al. (1992), alignment of such a set $\mathbf{P} = \{p_i; i = 1, \ldots, N)$ is achieved if the functional

$$L(\mathbf{P}, \mathbf{S}) = \int \sum_{i=1}^{N-1} \sum_{k=i+1}^{N} [p_i(\mathbf{r}; s_\alpha^i, s_x^i, s_y^i) - p_k(\mathbf{r}; s_\alpha^i, s_x^i, s_y^i)]\mathrm{d}\mathbf{r} \tag{3.42}$$

is minimized by appropriate choice of the set of the $3N$ parameters:

$$\mathbf{S} = \{s_\alpha^i, s_x^i, s_y^i; i = 1, \ldots, N\} \tag{3.43}$$

The actual computation of all pair-wise cross-correlations, as would be required by the minimization of equation (3.42), would be quite impractical; it would also have the disadvantage that each term is derived from two raw images having very low SNR. Penczek and coworkers show, however, that minimization of equation (3.42) is equivalent to minimization of another functional:

$$\bar{L}(\mathbf{P}, \mathbf{S}) = \int \sum_{i=1}^{N-1} [p_i(\mathbf{r}; s_\alpha^i, s_x^i, s_y^i) - \bar{p}_i(\mathbf{r})]^2 \mathrm{d}\mathbf{r} \tag{3.44}$$

where

$$\bar{p}_i(\mathbf{r}) = \frac{1}{N-1} \sum_{k=1; k\neq i}^{N} p_k(\mathbf{r}; s_\alpha^i, s_x^i, s_y^i) \tag{3.45}$$

In this reformulated expression, each image numbered i is aligned to a partial average of all images, created from the total average by subtracting the current image numbered i. This formula lends itself to the construction of an iterative algorithm involving partial sums, which have the advantage of possessing a strongly enhanced SNR when compared to the raw images.

To go into greater detail, the algorithm consists of two steps (figure 3.15). The first step, which is the random approximation of the global average, proceeds as follows (for simplicity, the argument vector $\mathbf{r}$ is dropped):

(i) Pick two images p_i and p_k at random from the set $\mathbf{P}$ of images
(ii) Align p_i and p_k (using any algorithm that minimizes) $||p_i - p_k||$)
(iii) Initialize the global average (general term to be denoted by a–m) by setting $a_2 = p_i \oplus p_k$, where the symbol $\oplus$ is used in this and the following description to denote the algebraic sum of the two images *after application of the rotation and shifts found*
(iv) Set a counter m to 3
(v) Pick the next image p_i from the set of $N - m + 1$ remaining images
(vi) Align p_i and a_{m-1}
(vii) Update the global average to give $a_m = [p_i \oplus (m-1)a_{m-1}]/k$
(viii) Increase counter m by 1. If $m = N$ then stop, else go to step (v)

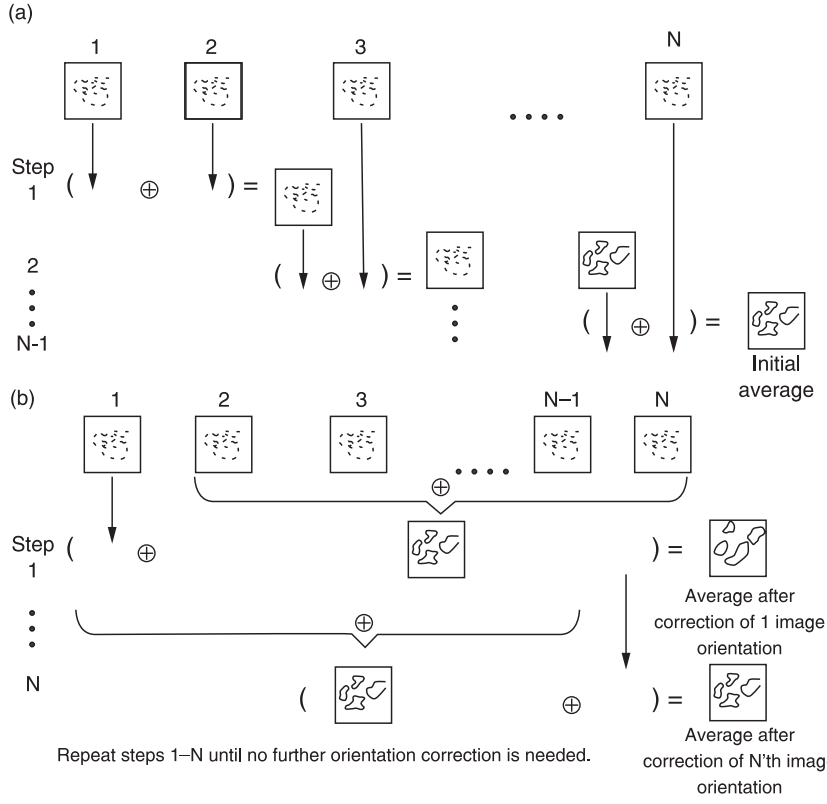

Figure 3.15 Scheme for reference-free alignment by Penczek et al. (1992). The algorithm consists of two parts (a and b); $\oplus$, symbol to indicate formation of the sum after the best orientation is found. For details, see text. From Penczek et al. (1992), reproduced with permission of Elsevier.

The second part of the algorithm performs an iterative refinement of the average $A = a_N$:

(i) Set counter $m = 1$
(ii) Create the modified average A' by subtracting the current image p_k in its current position:

$$A' = (NA - p_k)/(N - 1);$$

(iii) Align p_k with A'
(iv) Update the global average A as follows:

$$A = [p_k \oplus (N - 1)A']/N;$$

(v) Increase m by 1. If $m < N$ then go to (2)
(vi) If in step (3) any of the images changed its position significantly, then go back to step (1), else stop

The algorithm outlined has the following properties:

(i) No reference image is used, and so the result of the alignment does not depend on the choice of a reference, although there is some degree of dependency on the sequence of images being picked.
(ii) The algorithm is necessarily suboptimal since it falls short (by a wide margin) of exploring the entire parameter space.
(iii) Experience has shown that, when a heterogeneous image set is used, comprising dissimiliar subsets (e.g., relating to different particle orientations), then the images within each subset are aligned to one another on completion of the computation.

3.6. Alignment Using the Radon Transform

The 2D Radon transform (also known under the name *sinogram* [e.g., van Heel, 1987)] in its continuous form is defined for a 2D function $g(\mathbf{r})$ as

$$\hat{g}(p,\xi) = \int g(\mathbf{r})\delta(p - \xi^T\mathbf{r})\mathrm{d}\mathbf{r} \tag{3.45a}$$

where $\mathrm{r} = (x.y)^T$ and $\delta(p - \xi^T\mathbf{r})$ represents a line defined by the direction of the (normal) unit vector ξ. We can think of the 2D Radon transform as a systematic inventory of 1D projections of the image as a function of angle, with the angle being defined by the direction of the vector ξ. This can be seen in the relationship between the image in figure 3.16a and its Radon transform (figure 3.16c).

In this space, the extensive theory of Radon transforms and their applications in 2D and 3D image processing cannot be covered. It should merely be noted that the Radon transform yields an effective method for simultaneous translational and rotational 2D alignment of images (Lanzavecchia et al., 1996). Another brief section (chapter 5, section 7.4) is devoted to the use of the Radon transform in 3D projection matching.

4. Averaging and Global Variance Analysis*

4.1. The Statistics of Averaging

After successful alignment of a set of images representing a homogeneous population of molecules, all presenting the same view, each image element j in the coordinate system of the molecule is represented by a set of measurements:

$$\{p_i(\mathbf{r}_j);\, i = 1, \ldots, N\} \tag{3.46}$$

which can be characterized by an average:

$$\bar{p}_{(N)}(\mathbf{r}_j) = \frac{1}{N}\sum_{i=1}^{N} p_i(\mathbf{r}_j) \tag{3.47}$$

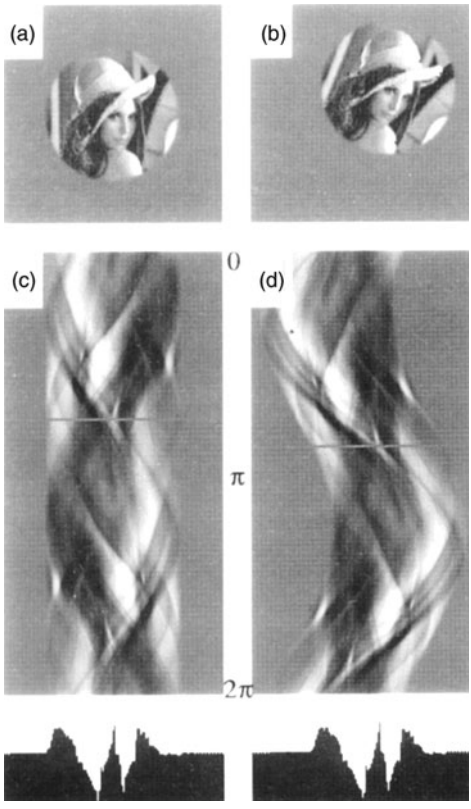

Figure 3.16 Alignment of two identical images using the Radon transform. (a) Image in reference position; (b) the same image, rotated by $\pi/8$ and shifted diagonally. (c, d) Radon transforms (sinograms) of (a, b), computed in the full interval $\{0, 2\pi\}$. Cross-correlation of the sinograms, line by line, reveals that the line marked is identical in the two Radon transforms, revealing both relative angle and shift between the two images. From Lanzavecchia et al. (1996), reproduced with permission of Oxford University Press.

and a variance:

$$v_{(N)}(\mathbf{r}_j) = s_{(N)}(\mathbf{r}_j) = \frac{1}{N-1}\sum_{i=1}^{N}[p_i(\mathbf{r}_j) - \bar{p}_{(N)}(\mathbf{r}_j)]^2 \tag{3.48}$$

Both $\{\bar{p}_{(N)}(\mathbf{r}_j); j = 1, \ldots, J\}$ and $\{V_{(N)}(\mathbf{r}_j); j = 1, \ldots, J\}$ can be represented as images, or maps, which are simply referred to as "average map" and "variance map."

The meaning of these maps depends on the statistical distribution of the pixel measurements. If we use the simple assumption of additive noise with zero-mean Gaussian statistics,

$$p_i(\mathbf{r}_j) = p(\mathbf{r}_j) + n_i(\mathbf{r}_j) \tag{3.49}$$

then $\bar{p}_{(N)}(\mathbf{r}_j)$ as defined above is an unbiased estimate of the mean and thus represents the structural motif $p(\mathbf{r})$ more and more faithfully as N is being increased. The quantity $V_{(N)}(\mathbf{r}_j) = s^2_{(N)}(\mathbf{r}_j)$ is an estimate of $\sigma^2(\mathbf{r}_j)$, the squared standard deviation of the noise.

A display of the variance map was first used in the context of image averaging in EM by Carrascosa and Steven (1979) and in single-particle averaging by Frank

et al. (1981a). The variance is a function that is spatially varying, for two reasons: (i) the noise statistics varies with the electron dose, which in turn is proportional to the local image intensity, and (ii) part of the observed variations are "signal associated"; that is, they arise from components of the structure that differ from one particle to the other in density or precise location. Usually, no statistical models exist for these variations.

The variance map is particularly informative with regard to the signal-associated components as it allows regions of local inconsistency to be spotted. For instance, if half of a set of molecules were carrying a ligand, and the other half not, then a region of strong inconsistency would show up in the variance map precisely at the place of the ligand.

4.2. The Variance Map and the Analysis of Statistical Significance

We have seen in the foregoing that one of the uses of the variance map is to pinpoint image regions where the images in the set vary strongly. The possible sources of interimage variability are numerous:

(i) Presence versus absence of a molecule component, for example, in partial depletion experiments (Carazo et al., 1988)
(ii) Presence versus absence of a ligand, for example, in immunoelectron microscopy (Gogol et al., 1990; Boisset et al., 1993b), ribosome-factor binding complex (Agrawal et al., 1998; Valle et al., 2002), calcium release channel bound with a ligand (Wagenknecht et al., 1997; Sharma et al., 1998; Samsó et al., 1999, 2000; Liu et al., 2002)
(iii) Conformational change, that is, movement of a mass (Carazo et al., 1988, 1989) or of many thin flexible masses (Wagenknecht et al., 1992)
(iv) (Compositional heterogeneity
(v) Variation in orientation, for example, rocking or flip/flop variation (van Heel and Frank, 1981; Bijlholt et al., 1982)
(vi) Variation in stain depth (Frank et al., 1981a, 1982; Boisset et al., 1990a)
(vii) Variation in magnification (Bijlholt et al., 1982)

Striking examples are found in many studies of negatively stained molecules, where the variance map often reveals that the stain depth at the boundary of the molecule is the strongest varying feature. The 40S ribosomal subunit of eukaryotes shows this behavior quite clearly (Frank et al., 1981a); see figure 3.17.

An example of a study in which structural information is gleaned from the variance map is found in the paper by Wagenknecht et al. (1992) (figure 3.18). Here, a core structure (E2 cores of pyruvate dehydrogenase) is surrounded by lipoyl domains which do not show up in the single particle average because they do not appear to assume fixed positions. Their presence at the periphery of the E2 domain is nevertheless reflected in the variance map by the appearance of a strong white halo of high variance. These early findings are particularly interesting in the light of the high-resolution structure emerging now by a combination of results from cryo-EM and X-ray crystallography, and indications of the dynamical behavior of this molecule (Zhou et al., 2001c).

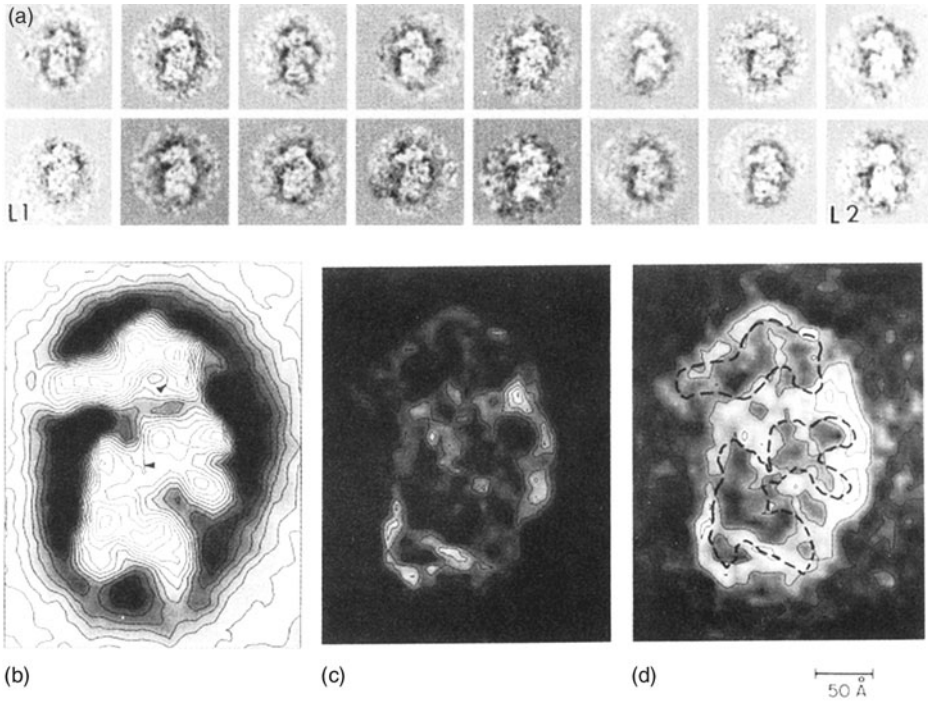

Figure 3.17 Average and variance map obtained from an aligned set of macromolecules. (a) Sixteen of a total set of 77 images showing the 40S ribosomal subunit of HeLa in L-view orientation; (b) average image; (c) variance map; and (d) standard deviation map, showing prominent variations mainly at the particle border where the amount of stain fluctuates strongly (white areas indicate high variance). From Frank et al. (1981), with permission of the American Association for the Advancement of Science.

However, this "global" variance analysis made possible by the variance map has some obvious shortcomings. While it alerts us to the presence of variations and inconsistencies among the images of a data set, and gives their location in the image field, it fails to characterize the different types of variation and to flag those images that have an outlier role. For a more specific analysis, the tools of multivariate data analysis and classification must be employed (see chapter 4).

Another important use of the variance map is the assessment of significance of local features in the average image (Frank et al., 1986), using standard methods of statistical inference (e.g., Cruickshank, 1959; Sachs, 1984): each pixel value in that image, regarded as an estimate of the mean, is accompanied by a confidence interval within which the true value of the mean is located with a given probability. In order to construct the confidence interval, we consider the random variable:

$$t = \frac{\bar{p}(\mathbf{r}_j) - p(\mathbf{r}_j)}{s(\mathbf{r}_j)} \tag{3.50}$$

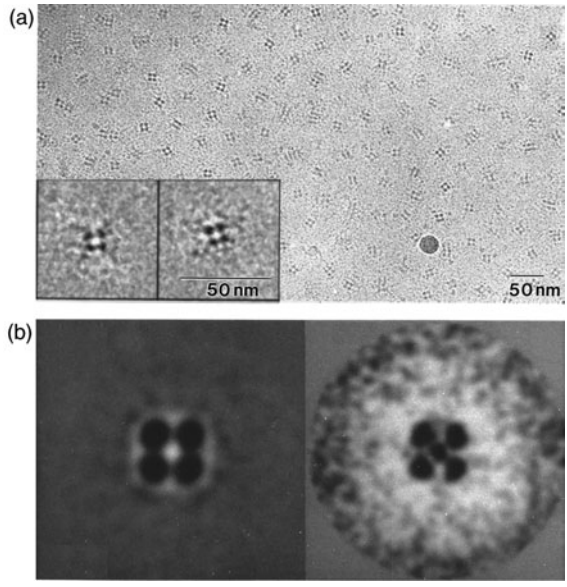

Figure 3.18 Example of the type of information contained in a variance map: visualization of a "corona" of highly flexible lipoyl domains surrounding the E2 core of puryvate dehydrogenase complex of *Escherichia coli*. (a) Electron micrograph of frozen-hydrated E2 cores presenting fourfold symmetric views. Each is surrounded by fuzzy structures believed to be lipoyl domains bound to the molecule. Due to their changing positions, the average map (b, left) depicts only the E2 core, but the variance map (b, right) shows a ring-shaped region of high variance. From Wagenknecht et al. (1992), reproduced with permission of Elsevier.

where $s(\mathbf{r}_j) = [V(\mathbf{r}_j)/N]^{1/2}$ is the standard error of the mean, which can be computed from the measured variance map expressed by equation (3.48) (see figure 3.17). The true mean lies in the interval:

$$\bar{p}_{\text{true}}(\mathbf{r}_j) = \bar{p}(\mathbf{r}_j) \pm ts(\mathbf{r}_j) \tag{3.50a}$$

the *confidence interval*, with probability P, if t satisfies

$$P = \int_{-1}^{+1} S_{N-1}(\tau)\mathrm{d}\tau \tag{3.51}$$

where $S_{N-1}(\tau)$ is the *Student distribution* with $N-1$ degrees of freedom. Beyond a number of images $N=60$ or so, the distribution changes very little, and the confidence intervals for the most frequently used probabilities become $t=1.96$ ($P=95\%$), $t=2.58$ ($P=99\%$), and $t=3.29$ ($P=99.9\%$).

It must be noted, however, that the use of the variance map to assess the significance of local features in a difference map should be reserved for regions

where no signal-related inconsistencies exist. In fact, the statistical analysis outlined here is strictly meaningful only for a homogeneous image set, in the sense discussed in section 3.2. Another caveat stems from the fact that the test described is valid under the assumption that the pixels are independent. This is not the case in a strict sense, because the images are normally aligned to a common reference.

Let us now discuss an application of statistical hypothesis testing to the comparison between two averaged pixels, $\overline{p}_1(\mathbf{r}_j)$ and $\overline{p}_2(\mathbf{r}_j)$, which are obtained by averaging over N_1 and N_2 realizations, respectively (Frank et al., 1988a). This comparison covers two different situations: in one, two pixels $j \neq k$ from the *same* image are being compared, and the question to be resolved is if the difference between the values of these two pixels, separated by the distance $|\mathbf{r}_j - \mathbf{r}_k|$, is significant. In the other situation, the values of the same pixel $j = k$ are compared as realized in two averages resulting from different experiments. A typical example might be the detection of extra mass at the binding site of the antibody in an immunolabeled molecule when compared with a control. Here, the question posed is whether the density difference detected at the putative binding site is statistically significant.

In both these situations, the standard error of the difference between the two averaged pixels is given by

$$s_d[p_1(\mathbf{r}_j), p_2(\mathbf{r}_j)] = [V_1(\mathbf{r}_j)/N_1 + V_2(\mathbf{r}_j)/N_2]^{1/2} \tag{3.52}$$

Differences between two averaged image elements are deemed significant if they exceed the standard error by at least a factor of three. This choice of factor corresponds to a significance level of 0.2% [i.e., $P = 98\%$ in equation (3.52)]. In single-particle analysis, this type of significance analysis was first done by Zingsheim et al. (1982), who determined the binding site of bungarotoxin on the projection map of the nicotinic acetylcholine receptor molecule of *Torpedo marmorata*. Another example was the determination, by Wagenknecht et al. (1988a), of the anticodon binding site of P-site tRNA on the 30S ribosomal subunit as it appears in projection. An example of a *t*-map showing the sites where an undecagold cluster is localized in a 2D difference map is found in the work of Crum et al. (1994). Figure 3.19 shows the difference mapping of a GFP (green fluorescent protein) insert in the ryanodine receptor/calcium release channel, and a map of statistically significant regions, displayed at the 99.9% level of confidence.

The same kind of test is of course even more important in three dimensions, when reconstructions are compared that show a molecule with and without a ligand, or in different conformational states. In single-particle reconstruction, variance analysis in three dimensions is handicapped by the fact that an ensemble of 3D reconstructions does not exist; hence, a 3D variance map cannot be computed in a straightforward way following expressions that would correspond to equations (3.47) and (3.48). However, other, approximate routes can be taken, as will be detailed in chapter 6, section 2.

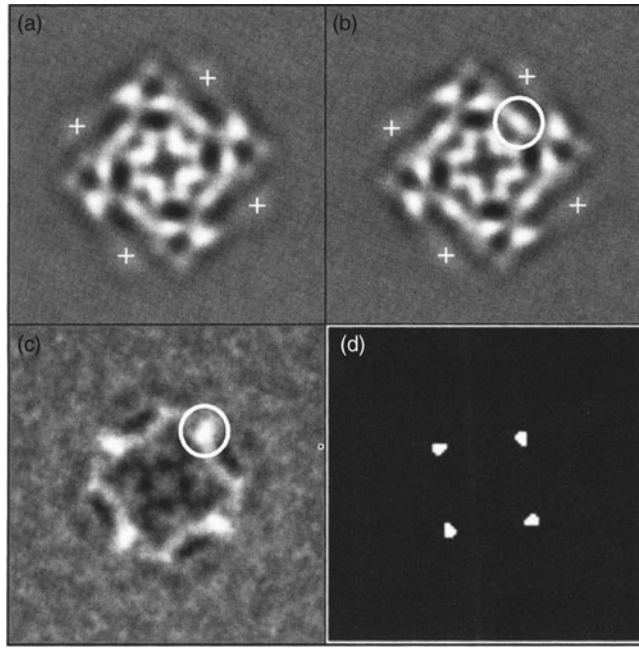

Figure 3.19 Localization of a GFP insert in the ryanodine receptor (RyR) by difference mapping and significance analysis. (a) Average of RyR wild-type ($N = 269$ particles); (b) average of RyR · GFP ($N = 250$ particles); (c) difference map b−a. The largest difference is located in the region encircled, and in the three regions related to it by fourfold symmetry. (d) Map of statistically significant regions, in the difference map, obtained by the *t*-test. The map is displayed at a $>99.9\%$ level of confidence. From Liu et al. (2002), reproduced with permission of The American Society for Biochemistry and Molecular Biology, Inc.

4.3. Signal-to-Noise Ratio

4.3.1. Concept and Definition

The signal-to-noise ratio (SNR) is the ratio between the variance of the signal and the variance of the noise in a given image.[6] This measure is extremely useful in assessing the quality of experimental data, as well as the power of 2D and 3D averaging methods. Unfortunately, the definition varies among different fields, with the SNR often being the square root of the above definition, or with the numerator being measured peak to peak. In the following we will use the ratio of variances, which is the most widely used definition in digital signal processing.

[6]Note that in the field of electrical engineering, the signal-to-noise ratio is defined as the ratio of the signal power to the noise power. This ratio is identical with the variance ratio only if the means of both signal and noise are zero.

The sample variance of an image indexed i, $\{p_{ij}; j = 1, \ldots, J\}$ is defined as

$$\mathrm{var}(p_i) = \frac{1}{J-1}\sum_{j=1}^{J}[p_{ij} - \langle p_i\rangle]^2 \tag{3.53}$$

with the sample mean

$$\langle p\rangle = \frac{1}{J}\sum_{j=1}^{J} p_{ij} \tag{3.54}$$

According to Parseval's theorem, the variance of a band-limited function can be expressed as an integral (or its discrete equivalent) over its squared Fourier transform:

$$\mathrm{var}(p_i) = \int_{\mathbf{B}_\square} |P_i(\mathbf{k})|\mathrm{d}\mathbf{k} \tag{3.55}$$

where $\mathbf{B}_\square$ denotes a modified version of the *resolution domain* $\mathbf{B}$, that is, the bounded domain in Fourier space representing the signal information. The modification symbolized by the $\square$ subscript symbol is that the integration exempts the term $|P(\mathbf{0})|^2$ at the origin. (Parseval's theorem expresses the fact that the norm in Hilbert space is conserved on switching from one set to another set of orthonormalized basis functions.)

When we apply this theorem to both signal and additive noise portions of the image, $p(\mathbf{r}) = o(\mathbf{r}) + n(\mathbf{r})$, we obtain

$$\mathrm{SNR} = \frac{\mathrm{var}(o)}{\mathrm{var}(n)} = \frac{\int_{\mathrm{B}_\square} |O(\mathbf{k})|^2|H(\mathbf{k})|^2\mathrm{d}\mathbf{k}}{\int_{\mathrm{B}_\square} |N(\mathbf{k})|^2\mathrm{d}\mathbf{k}} \tag{3.56}$$

Often the domain, $\mathbf{B}'$, within which the signal portion of the image possesses appreciable values, is considerably smaller than $\mathbf{B}$. In that case, it is obvious that low-pass filtration of the image to band limit $\mathbf{B}'$ leads to an increased SNR without signal being sacrificed. For uniform spectral density of the noise power up to the boundary of $\mathbf{B}$, the gain in SNR on low-pass filtration to the true band limit $\mathbf{B}'$ is, according to Parseval's theorem, equation (3.55), equal to the ratio area$\{\mathbf{B}\}$/area$\{\mathbf{B}'\}$. Similarly, it often happens that the noise power spectrum $|N(\mathbf{k})|^2$ is uniform, whereas the transform of the signal transferred by the instrument, $|O(\mathbf{k})|^2|H(\mathbf{k})|^2$, falls off radially. In that case, the elimination, through low-pass filtration, of a high-frequency band may boost the SNR considerably without affecting the interpretable resolution.

4.3.2. Measurement of the Signal-to-Noise Ratio

Generally, the unknown signal is mixed with noise, so the measurement of the SNR of an experimental image is not straightforward. Two ways of measuring the SNR of "raw" image data have been put forth: one is based on the dependence of the sample variance [equation (3.53)] of the average image on N (= the number of

images averaged) and the other on the cross-correlation of two realizations of the image.

4.3.2.1. N-*dependence of sample variance* We assume that the noise is additive, uncorrelated, stationary (i.e., possessing shift-independent statistics), and Gaussian; and further, that it is uncorrelated with the signal (denoted by p). In that case, the variance of a "raw" image p_i is, independently of i,

$$\mathrm{var}(p_i) = \mathrm{var}(p) + \mathrm{var}(n) \tag{3.57}$$

The variance of the average $\bar{p}_{[N]}$ of N images is

$$\mathrm{var}(\bar{p}_{[N]}) = \mathrm{var}(p) + \frac{1}{N}\mathrm{var}(n) \tag{3.58}$$

that is, with increasing N, the proportion of the noise variance in the variance of the average is reduced. This formula suggests the use of a plot of $\mathrm{var}(\bar{p}_{[N]})$ versus $1/N$ as a means to obtain the unknown quantities $\mathrm{var}(p)$ and $\mathrm{var}(n)$ (Hänicke, 1981; Frank et al., 1981a; Hänicke et al., 1984). If the assumptions made at the beginning are correct, the measured values of $\mathrm{var}(\bar{p}_{[N]})$ should lie on a straight line whose slope is the desired quantity $\mathrm{var}(n)$ and whose intersection with the $\mathrm{var}(p)$ axis (obtained by extrapolating it to $1/N = 0$) gives the desired quantity $\mathrm{var}(p)$. Figure 3.20 shows such a variance plot obtained for a set of 81 images of the negatively stained 40S ribosomal subunit of HeLa cells (Frank et al., 1981a). It is seen that the linear dependency predicted by equation (3.58) is indeed a good approximation for this type of data, especially for large N.

4.3.2.2. Measurement by cross-correlation Another approach to the measurement of the SNR makes use of the definition of the cross-correlation coefficient (CCC). The CCC of two realizations of a noisy image, p_{ij} and p_{kj}, is defined as [see equation (3.16)]

$$\rho_{12} = \frac{\sum_{j=1}^{J} [p_{ij} - \langle p_i \rangle][p_{kj} - \langle p_k \rangle]}{\left\{ \sum_{j=1}^{J} [p_{ij} - \langle p_i \rangle]^2 \sum_{j=1}^{J} [p_{kj} - \langle p_k \rangle]^2 \right\}^{1/2}} \tag{3.59}$$

where $\langle p_i \rangle$ and $\langle p_k \rangle$ again are the sample means defined in the previous section.

When we substitute $p_{ij} = p_j + n_{ij}, p_{kj} = p_k + n_{kj}$, and observe that according to the assumptions both noise functions have the same variance $\mathrm{var}(n)$, we obtain the simple result (Frank and Al-Ali, 1975):

$$\rho_{12} = \frac{\alpha}{1 + \alpha} \tag{3.60}$$

from which the SNR is obtained as

$$\alpha = \frac{\rho_{12}}{1 - \rho_{12}} \tag{3.61}$$

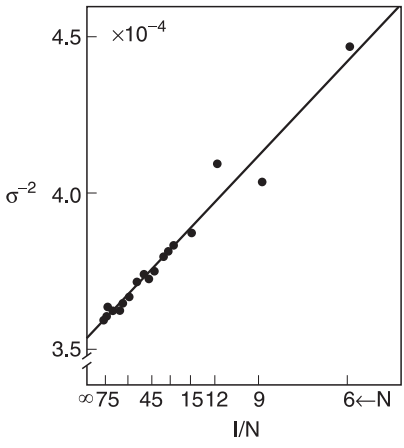

Figure 3.20 Decrease in the variance of the average image as a function of the number of images averaged, N, or (linear) increase in the variance with $1/N$. Extrapolation to $1/N = 0$ allows the variance of the signal to be measured. Averaged were $N = 81$ images of the 40S ribosomal subunit from HeLa cells negatively stained and showing the so-called L-view. From Frank et al. (1981a), reproduced with permission of The American Association for the Advancement of Science.

Thus, the recipe for estimating the SNR is quite simple: choose two images from the experimental image set, align them and filter the images to the resolution deemed relevant for the SNR measurement. Then compute the CCC and use equation (3.61) to obtain the SNR estimate. Because of the inevitable fluctuations of the results, it is advisable to repeat this procedure with several randomly picked image pairs and use the average of the different SNR measurements as a more reliable figure.

In practice, the measurement of the SNR in the image (as opposed to the SNR of the *digitized* image) is somewhat more complicated since the microdensitometer measurement will introduce another noise process which, unchecked, would lead to overly pessimistic SNR estimates. To account for this additional contribution, Frank and Al-Ali (1975) used a control experiment, in which two scans of the same micrograph were evaluated with the same method, and a corresponding SNR for the microdensitometer α_d was found. The true SNR is then obtained as

$$\alpha_{\text{true}} = \frac{1}{(1 + (1/\alpha))/(1 + (1/\alpha_d)) - 1} \tag{3.62}$$

where α is the SNR deduced from the uncorrected experiment, following equation (3.61).

5. Resolution

5.1. The Concept of Resolution

Optical definitions of resolution are based on the ability of an instrument to resolve two points separated by a given distance, d. When the two points get closer and closer, their initially separate images merge into a single one, first enclosing a "valley" between them, which gradually fills, until it entirely disappears. The definition usually stipulates that the ratio between valley and peak should be of a certain minimum size for the points to be resolved.

Two criticisms can be raised against this definition; one is quite general, the other relates to its application in EM. The first criticism (Di Francia, 1955) concerns an information-theoretical aspect of the experiment: if it is known *a priori* that the object consists of two points, then measurement of their mutual distance in the image is essentially a pattern-recognition problem, which is limited by noise, not by the size of d. (For instance, if the image of a single point, i.e., the point spread function, is a circularly symmetric function, the image of two points with nonzero distance will no longer be rotationally symmetric, and the exact distance could be simply obtained by fitting, even when the distance is much less than the diameter of the point-spread function.) The other criticism is that the resolution criterion formulated above does not lend itself to a suitable experiment when applied to high-resolution EM. On the nanometer scale, any test object, as well as the support it must be placed on, reveals its atomic makeup. Both considerations suggest that the definition and measurement of resolution require a fundamentally different approach. In fact, the most useful definitions and criteria all refer to Fourier representations of images.

In both crystallography and statistical optics, it is common to define resolution by the orders of (object-related) Fourier components available for the Fourier synthesis of the image. This so-called crystallographic resolution R_c and Raleigh's point-to-point resolution distance d for an instrument, which is diffraction limited to R_c, are related by

$$d = 0.61/R_c \tag{3.63}$$

In electron crystallography, the signal-related Fourier components of the image are distinguished from noise-related components by the fact that the former are concentrated on the points of a regular lattice, the *reciprocal lattice*, while the latter form a continuous background. Thus, resolution can be specified by the radius of the highest diffraction orders that stand out from the background. What "stand out" means can be quantified by relating the amplitude of the peak to the mean of the background surrounding it (see below).

In single-particle averaging, on the other hand, there is no distinction in the Fourier transform between the appearance of signal and noise, and resolution estimation must take a different route. There are two categories of resolution tests; one is based on the comparison of two independent averages in the Fourier domain (*cross-resolution*), while the other is based on the multiple comparison of the Fourier transforms of all images participating in the average. The differential phase residual (Frank et al., 1981a), Fourier ring correlation (Saxton and Baumeister, 1982; van Heel et al., 1982c), and the method of Young's fringes (Frank et al., 1970; Frank, 1972a) fall into the first category, while the spectral SNR (Unser et al., 1987, 1989) and the Q-factor (van Heel and Hollenberg, 1980; Kessel et al., 1985) fall into the second.

Averaging makes it possible to recover a signal that is present in the image in very small amounts. Its distribution in Fourier space, as shown by the signal power spectrum, usually shows a steep falloff. For negatively stained specimens, this falloff is due to the imperfectness of the staining: on the molecular scale, the stain salt forms crystals and defines the boundary of the molecule only within a

margin of error. In addition, the process of air-drying causes the molecule to change shape, and one must assume that this shape change is variable as it depends on stain depth, orientation, and properties of the supporting carbon film. For specimens embedded in ice, gross shape changes of the specimen are avoided, but residual variability (including genuine conformational variability) as well as instabilities of recording (drift, charging, etc.) are responsible for the decline of the power spectrum. Some of these resolution-limiting factors will be discussed below (section 5.4); for the moment, it is important to realize that there is a practical limit beyond which the signal power makes only marginal contributions to the image.

However, all resolution criteria listed below in this section have in common that they ignore the falloff of the signal in Fourier space. Thus, a resolution of $1/20\,\text{Å}^{-1}$ might be found by a consistency test, even when the signal power is very low beyond $1/30\,\text{Å}^{-1}$. An example of this kind of discrepancy was given by van Heel and Stöffler-Meilicke (1985), who studied the 30S ribosomal subunits from two eubacterial species by 2D averaging: they found a resolution of $1/17\,\text{Å}^{-1}$ by the Fourier ring correlation method, even though the power spectrum indicated the presence of minimal signal contributions beyond $1/20\,\text{Å}^{-1}$, or even beyond $1/25\,\text{Å}^{-1}$, when a more conservative assessment is used. The lesson to be learned from these observations is that, for a meaningful statement about the information actually present in the recovered molecule projection, a resolution assessment ideally should be accompanied by an assessment of the range and falloff of the power spectrum.

The same concern about the meaning of "resolution" in a situation of diminishing diffraction power has arisen in electron crystallography. Henderson et al. (1986) invented a quality factor (IQ, for *image quality*) that expresses the SNR of each crystal diffraction spot and applied a rigorous cutoff in the Fourier synthesis depending on the averaged size of this factor, with the averaging being carried out over rings in Fourier space. Glaeser and Downing (1992) have demonstrated the effect of including higher diffraction orders in the synthesis of a projection image of bacteriorhodopsin. It is seen that, with increasing spatial frequency, as the fraction of the total diffraction power drops to 15%, the actual improvement in the definition of the image (e.g., in the sharpness of peaks representing projected alpha-helices) becomes marginal.

5.2. Resolution Criteria

It should be noted that the treatment in this section applies to both 2D and 3D data: the formation of 2D averages, as well as 3D reconstruction. To avoid duplication, we will talk about both types of data below, and refer back to this section when the subject of resolution comes up again in chapter 5.

5.2.1. Definition of Region to be Tested

The resolution criteria to be detailed below all make use of the discrete Fourier transform of the images to be analyzed. It is of crucial importance for a meaningful application of these measures that no correlations are unwittingly

introduced when the images are prepared for the resolution test. It is tempting to use a mask to narrowly define the region in the image where the signal—the averaged molecule image—resides, as the inclusion of surrounding material with larger inconsistency might lead to overly pessimistic results. However, imposition of a binary mask, applied to both images that are being compared, would produce correlation extending to the highest resolution. This is on account of the sharp boundary of a binary mask, with a 1-pixel falloff, which requires Fourier terms out to the Nyquist limit to be utilized in the representation. Hence, the resolution found in any of the tests described below would be falsely reported as the highest possible—corresponding to the Nyquist spatial frequency. To avoid this effect, one has to use a "soft" mask whose falloff at the edges is so slow that it introduces correlations at low spatial frequencies only. Gaussian-shaped masks are optimal for this purpose since all derivatives of a Gaussian are again Gaussian, and thus continuous.

Fourier-based resolution criteria are, therefore, governed by a type of uncertainty relationship: precise localization of features for which resolution is determined makes the resolution indeterminate; and, on the other hand, precise measurement of resolution is possible only when the notion of localizing the features is entirely abandoned.

5.2.2. Comparison of Two Subsets Versus Analysis of the Whole Data Set

5.2.2.1. Criteria based on two equally large subsets Many criteria introduced in the following test the reproducibility of a map obtained by averaging (or, as we will later see, by 3D reconstruction) when based on two randomly drawn subsets of equal size. For example, the use of even- and odd-numbered images of the image set normally avoids any systematic trends such as related to the origin in different micrographs or different areas of the specimen. Each subset is averaged, leading to the average images $\bar{p}_1(\mathbf{r}), \bar{p}_2(\mathbf{r})$ ("subset averages").

Let $F_1(\mathbf{k})$ and $F_2(\mathbf{k})$ be the discrete Fourier transforms of the two subset averages, with the spatial frequency $\mathbf{k}$ assuming all values on the regular Fourier grid (k_x, k_y) within the Nyquist range. The Fourier transforms are now compared, and a measure of discrepancy is computed, which is averaged over rings of width Δk and radius $k = |\mathbf{k}| = [k_x^2 + k_y^2]^{1/2}$. The result is then plotted as a function of the ring radius. This curve characterizes the discrepancy between the subset averages over the entire spatial frequency range.

In principle, a normalized version of the generalized Euclidean distance, $|F_1(\mathbf{k}) - F_2(\mathbf{k})|$ could be used to serve as a measure of discrepancy, but two other measures, the *differential phase residual* and the *Fourier ring correlation*, have gained practical importance. These will be introduced in sections 5.2.3 and 5.2.4. Closely related is the criterion based on Young's fringes (section 5.2.5), which will be covered mainly because of its historical role and the insights it provides.

Even though it is clear that the information given by a curve cannot be condensed into a single number, it is nevertheless common practice to derive a single "resolution figure" for expediency.

5.2.2.2. Truly independent versus partially dependent subsets. The iterative refinement procedures that are part of the reference-based alignment (section 3.4) as well as the reference-free alignment (section 3.5) complicate the resolution estimation by halfset criteria since they inevitably introduce a statistical interdependency between the two halfset averages being compared. This problem, which is closely related to the problem of model bias, will resurface in the treatment of angular refinement of reconstructions by 3D projection matching (chapter 5, section 7.2). In the 3D case, the problem has been extensively discussed (Grigorieff, 2000; Penczek, 2002a; Yang et al., 2003). Grigorieff's (2000) suggestion, if applied to the case of 2D averaging, would call for the use of two markedly different 2D references at the outset, and for a total separation of the two randomly selected image subsets throughout the averaging procedure. The idea is if the data are solid and self-consistent, then they can be expected to converge to an average that is closely reproducible. If, on the other hand, the noise is so large that the reference leaves a strong imprint on the average, the resulting deviation between the subset averages will be reported as a lack of reproducible resolution. We come back to the idea of using independent subset averages in the discussion of the spectral signal-to-noise ratio (SSNR), since it leads to the formulation of a straightforward relationship between the SSNR and the Fourier ring correlation.

5.2.2.3. Criteria based on an evaluation of the whole data set The criteria of the first kind have the disadvantage of large statistical uncertainty, and this is why criteria based on a statistical evaluation of the total set are principally superior, although the advantage diminishes as the numbers of particles increases, which is now often in the tens of thousands. The Q-factor and the SSNR are treated in sections 5.2.7 and 5.2.8, respectively.

5.2.3. Differential Phase Residual

The differential phase residual measures the root mean square (r.m.s.) deviation of the phase difference betweeen the two Fourier transforms, weighted by the average Fourier amplitude. If $\Delta\phi(\mathbf{k})$ is the phase difference between the two Fourier transforms for each discrete spatial frequency $\mathbf{k}$, then the differential phase residual (DPR) is defined as follows:

$$\Delta\bar{\phi}(k, \Delta k) = \left| \frac{\sum\limits_{[k, \Delta k]} [\Delta\phi(\mathbf{k})]^2 [|F_1(\mathbf{k})| + |F_2(\mathbf{k})|]}{\sum\limits_{[k, \Delta k]} [|F_1(\mathbf{k})| + |F_2(\mathbf{k})|]} \right|^{1/2} \tag{3.64}$$

The sums are computed over Fourier components falling within rings defined by spatial frequency radii $k \pm \Delta k; k = |\mathbf{k}|$ and plotted as a function of k (figure 3.21). In principle, as in the case of the Fourier ring correlation to be introduced below, the entire curve is needed to characterize the degree of consistency between the two averages. However, it is convenient to use a single

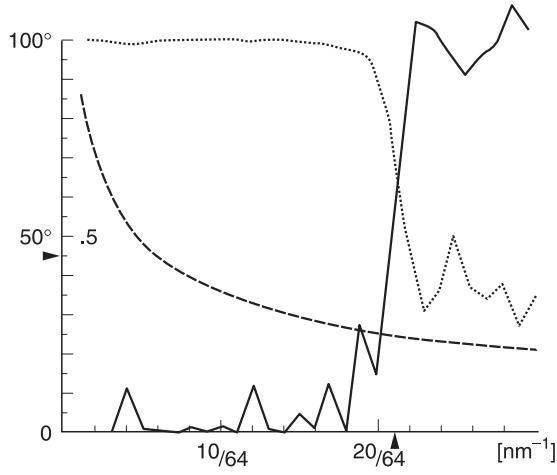

Figure 3.21 Resolution assessment by comparison of two subaverages (calcium release channel of skeletal fast twitch muscle) in Fourier space using two different criteria: differential phase residual (DPR, solid line, angular scale $0^\circ \ldots 100^\circ$) and Fourier ring correlation (FRC, dashed line, scale $0 \ldots 1$). The scale on the x-axis is in Fourier units, denoting the radius of the rings over which the expressions for DPR or FRC were evaluated. The DPR resolution limit ($\Delta\bar{\Phi} = 45^\circ$; see arrowhead on y-axis) is $1/30\,\text{Å}^{-1}$ (arrowhead on x-axis). For FRC resolution analysis, the FRC curve was compared with twice the FRC for pure noise (dotted curve). In the current example, the two curves do not intersect within the Fourier band sampled, indicating an FRC resolution of better than $1/20\,\text{Å}^{-1}$. From Radermacher et al. (1992a), reproduced with permission of the Biophysical Society.

figure, k_{45}, the spatial frequency for which $\Delta\bar{\phi}(k, \Delta k) = 45^\circ$. As a conceptual justification for the choice of this value, one can consider the effect of superimposing two sine waves differing by $\Delta\bar{\phi}$. If $\Delta\bar{\phi}$ is less than 45°, the waves tend to enforce each other, whereas for any $\Delta\bar{\phi} > 45^\circ$, the maximum of one wave already tends to fall in the vicinity of the zero of the other, and destructive interference starts to occur.

It is of crucial importance in the application to EM that the phase residual (and any other Fourier-based measures of consistency, to be described in the following) be computed "differentially," over successive rings or shells, rather than globally, over the entire Fourier domain with a circle of radius k. Such global computation is often used, for instance, to align particles with helical symmetry (Unwin and Klug, 1974). Since $|F(\mathbf{k})|$ falls off rapidly as the phase difference $\Delta\phi$ increases, the figure k_{45} obtained with the global measure would not be very meaningful in our application; for instance, excellent agreement in the lower spatial frequency range can make up for poor agreement in the higher range and thus produce an overoptimistic value for k_{45}. The differential form of the phase residual, equation (3.64), was first used by Crowther (1971) to assess the preservation of icosahedral symmetry as a function of spatial frequency. It was first used in the context of single-particle averaging by Frank et al. (1981a).

One can easily verify from equation (3.64) that the DPR is sensitive to changes in scaling between the two Fourier transforms. In computational implementations, equation (3.64) is therefore replaced by an expression in which $|F_2(\mathbf{k})|$ is dynamically scaled, that is, replaced by $s|F_2(\mathbf{k})|$, where the scale factor s is allowed to run through a range from a value below 1 to a value above 1, through a range large enough to include the minimum of the function. The desired DPR then is the minimum of the curve formed by the computed residuals. One of the advantages of the DPR is that it relates to the measure frequently used in electron and X-ray crystallography to assess reproducibility and the preservation of symmetry.

5.2.4. Fourier Ring Correlation

The Fourier ring correlation (FRC) (Saxton and Baumeister, 1982; van Heel et al., 1982) is similar in concept to the DPR as it is based on a comparison of the two Fourier transforms over rings:

$$\mathrm{FRC}(k, \Delta k) = \frac{\mathrm{Re}\{\sum\limits_{[k,\,\Delta k]} F_1(\mathbf{k})F_2^*(\mathbf{k})\}}{\{\sum\limits_{[k,\,\Delta k]} |F_1(\mathbf{k})|^2 \sum\limits_{[k,\,\Delta k]} |F_2(\mathbf{k})|^2\}^{1/2}} \tag{3.65}$$

Again, as in the definition of the DPR in the previous section, the notation under the sum refers to the terms that fall into a ring of certain radius and width. The FRC curve (see figure 3.21) starts with a value of 1 at low spatial frequencies, indicating perfect correlation, then falls off more or less gradually, toward a fluctuating flat region of the curve with values that originate from chance correlation.

Here, the resolution criterion is derived in two different ways: (i) by comparison of the FRC measured with the FRC expected for pure noise, $\mathrm{FRC}_{\mathrm{noise}} = 1/(N_{[k,\,\Delta k]})^{1/2}$, where $N_{[k,\,\Delta k]}$ denotes the number of samples in the Fourier ring zone with radius k and width Δk, or (ii) by comparison of the FRC measured with an empirical threshold value. The term $\mathrm{FRC}_{\mathrm{noise}}$ is often denoted as "σ," even though it does not have the meaning of a standard deviation. Using this notation, the criteria based on the noise comparison that have been variably used are 2σ (van Heel and Stöffler-Meilicke, 1985), 3σ (e.g., Orlova et al., 1997), or 5σ (Radermacher et al., 1988, 2001). As empirical threshold, in the second group of FRC-based criteria, the value 0.5 (Böttcher et al., 1997) is most frequently used.

The $\mathrm{FRC} = 3\sigma$ criterion invariably gives much higher numerical resolution values than $\mathrm{FRC} = 0.5$. A critical comparison of the 0.5 and the 3σ criteria, and a refutation of some of the comments in Orlova et al. (1997) are found in Penczek's appendix to Malhotra et al. (1998). A data set is considered that is split into halves, and the FRC is calculated by comparing the halves. It is shown here that $\mathrm{FRC} = 0.5$ corresponds to $\mathrm{SNR} = 1$. In other words, in the corresponding shell, the noise is already as strong as the signal. Therefore, the inclusion of data beyond this point appears risky. More about the application of resolution criteria to 3D

reconstructions, as opposed to 2D averages considered here, will be found in chapter 5.

DPR Versus FRC Regarding the relationship between FRC and DPR, experience has generally shown that the FRC $= 3\sigma$ gives consistently a more optimistic answer than DPR $= 45°$. In order to avoid confusion in comparisons of resolution figures, some authors have used both measures in their publications. Some light on the relationship between DPR and FRC has been shed by Unser et al. (1987), who introduced another resolution measure, the SSNR (see section 5.2.8). Their theoretical analysis confirmed the observation that always $k_{\mathrm{FRC}[3\sigma]} > k_{\mathrm{DPR}[45°]}$ for the same model data set. Further illumination of the relative sensitivity of these two measures was provided by Radermacher (1988) who showed, by means of a numerical test, that an FRC $= 2 \times 1/N^{1/2}$ cutoff (the criterion initially proposed, before a factor of 3 was adopted) (Orlova et al., 1997) is equivalent to a SNR of 0.2, whereas the DPR 45° cutoff is equivalent to SNR $= 1$. Thus, the FRC cutoff, and even the DPR cutoff with its fivefold increased SNR seem quite optimistic; on the other hand, for well-behaved data the DPR curve is normally quite steep, so that even a small increase in the FRC cutoff will often lead to a rapid increase in SNR.

Another observation by Radermacher (1988), later confirmed by de la Fraga et al. (1995), was that the FRC cutoff of FRC $= 2\sigma$ corresponds to a DPR cutoff of 85°.

5.2.5. Young's Fringes

Because of its close relationship with the DPR and the FRC, the method of Young's fringes (Frank et al., 1970; Frank, 1972a, 1976) should be mentioned here, even though this method is nowadays rarely used to measure the resolution of computed image averages. However, Young's fringes shed an interesting light on the relationship between common information and correlation. The method is based on the result of an optical diffraction experiment: two micrographs of the same specimen (e.g., a thin carbon film) are first brought into precise superimposition, and then a small relative translation Δx is applied. The diffraction pattern (figure 3.22) shows the following intensity distribution:

$$\begin{aligned} I(\mathbf{k}) &= |F_1(\mathbf{k}) + F_2(\mathbf{k}) \exp(2\pi i k_x \Delta x)|^2 \\ &= |F_1(\mathbf{k})|^2 + |F_2(\mathbf{k})|^2 \\ &\quad + 2|F_1(\mathbf{k})||F_2(\mathbf{k})| \cos[2\pi i k_x \Delta x + \phi_1(\mathbf{k}) - \phi_2(\mathbf{k})] \end{aligned} \tag{3.66}$$

The third term, the Young's fringes term proper, is modulated by a cosine pattern whose wavelength is inversely proportional to the size of the image shift, and whose direction is in the direction of the shift. Since a fixed phase relationship holds only within the domain where the Fourier transform is dominated by the signal common to both superimposed images, while the relationship is random outside of that domain, the cosine fringe pattern induced by the shift can be used to visualize the extent of the resolution domain.

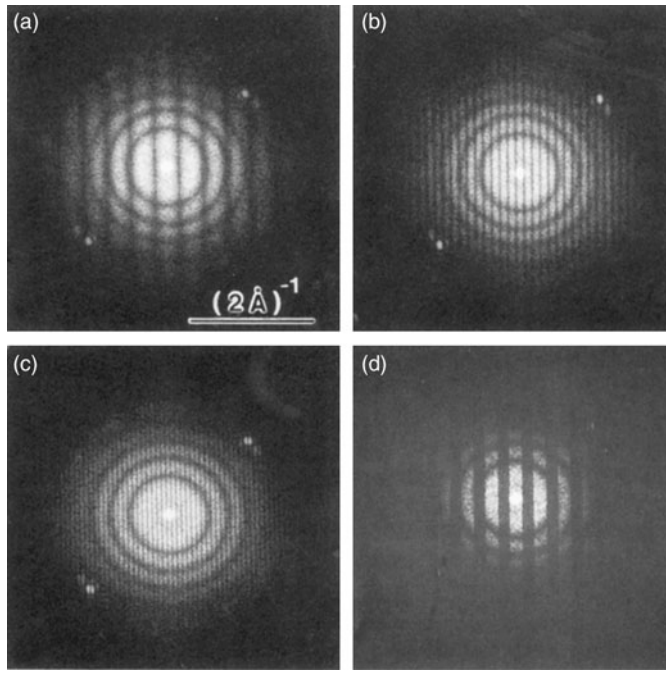

Figure 3.22 Computed diffraction patterns showing Young's fringes. The patterns are obtained by adding two different electron micrographs of the same specimen area and computing the power spectrum of the resulting image. (a–c) The micrographs are added with different horizontal displacements (corresponding to 10, 30, and 50 Å). The intensity of the resulting pattern follows a cosine function. (d) A pattern with approximately rectangular profile is obtained by linear superposition of the patterns (a–c) using appropriate weights. From Zemlin and Weiss (1993), reproduced with permission of Elsevier.

Moreover, the position of the Young's fringes is sensitive to the phase difference between the two transforms. When two images of an object are obtained with the same defocus setting, the phases are the same. Consistent shifts affecting an entire region of Fourier space show up as shifts of the fringe system. When two images of an object are obtained with different defocus settings, then the fringe system shifts by 180° wherever the contrast transfer functions differ in polarity (Frank, 1972a).

Using digital Fourier processing, the waveform of the fringes can be freely designed, by superposing normal cosine-modulated patterns with different frequencies (Zemlin and Weiss, 1993). The most sensitive detection of the band limit is achieved when the waveform is rectangular. Zemlin and Weiss (1993) obtained such a pattern of modulation experimentally (figure 3.22d).

5.2.6. Statistical Limitations of Halfset Criteria

Criteria based on splitting the data set in half in a single partition (as opposed to doing it repeatedly for a large number of permutations) are inferior because of large statistical fluctuations. The results of an evaluation by de la Fraga et al.

(1995) are interesting in this context. By numerical trial computations, these authors established confidence limits for DPR and FRC resolution tests applied to a data set of 300 experimental images of DnaB helicase. The data set was divided into halfsets using many permutations, and corresponding averages were formed in each case. The resulting DPR and FRC curves were statistically evaluated. The confidence limits for both the DPR determination of 10 Fourier units and the FRC determination of 13 units were found to be ± 1 unit. This means that even with 300 images, the resolution estimates obtained by DPR_{45} or $FRC_{0.5}$ may be as much as 10% off their asymptotic value. Fortunately, though, the number of particles in typical projects nowadays go into the thousands and tens of thousands, so that the error will be much smaller.

Another obvious flaw has not been mentioned yet: the splitting of the data set in half worsens the statistics, and will inevitably underreport resolution. Since the dependence of the measured resolution on the number of particles N entering the reconstruction is unknown, there is no easy way to correct the reported figure. Morgan et al. (2000) chose to determine this relationship empirically, by extrapolating from the trend of a curve of $FSC_{0.5}$ as a function of $\log(N)$, which the authors obtained by making multiple reconstructions with increasing N.

5.2.7. Q-Factor

This and the following section deal with criteria based on an evaluation of the whole data set. The Q-factor (van Heel and Hollenberg, 1980; Kessel et al., 1985) is easily explained by reference to a vector diagram (figure 3.23b) depicting the summation of equally-indexed (i.e., relating to the same discrete spatial frequency $\mathbf{k}$) Fourier components $P_i(\mathbf{k})$ in the complex plane, which takes place when an image set is being averaged. Because of the presence of noise (figure 3.23a), the vectors associated with the individual images zigzag in the approximate direction of the common signal. The Q-factor is simply the ratio between the length of the sum vector and the length of the total pathway of the vectors contributing to it:

$$Q(\mathbf{k}) = \frac{|\sum_{i=1}^{N} F_i(\mathbf{k})|}{\sum_{i=1}^{N} |F_i(\mathbf{k})|} \tag{3.67}$$

Obviously, from its definition, $0 \leq Q \leq 1$. For pure noise, $Q(\mathbf{k}) = 1/\sqrt{N}$, since this situation is equivalent to the random wandering of a particle in a plane under Brownian motion (Einstein equation).

The Q-factor is quite sensitive as an indicator for the presence of a signal component, because the normalization is specific for each Fourier coefficient. A map of $Q(\mathbf{k})$ (first used by Kessel et al., 1985; see figure 3.23c) readily shows weak signal components at high spatial frequencies standing out from the background and thus enables the ultimate limit of resolution recoverable (*potential resolution*) to be established. Again, a quantitative statement can be obtained by averaging

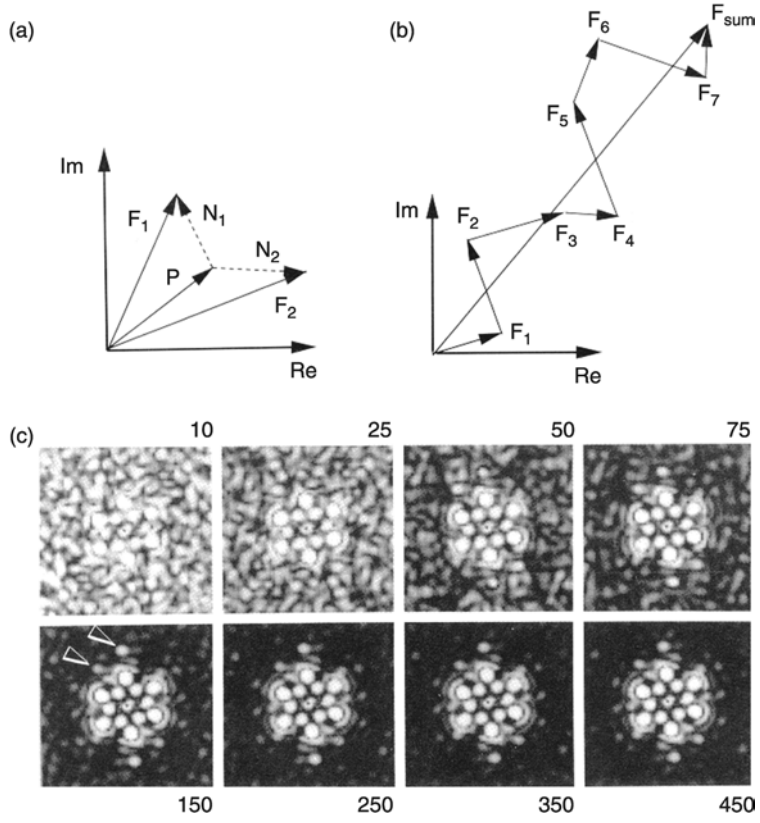

Figure 3.23 Averaging of signals in the complex plane. (a) Corresponding Fourier components (i.e., same-**k**) F_1, F_2 of two aligned images that represent the same signal (P) but differ by the additive noise components N_1 and N_2. (b) Definition of the Q-factor: it relates to the addition of vectors (in this case $N=7$) representing same-**k** Fourier components of seven images in the complex plane. F_{sum}/N is the Fourier component of the average image. The Q-factor is now defined as the ratio between the length of F_{sum} and the sum of the lengths of the contributing vectors F_i. Only in the absence of noise, can the maximum $Q=1$ be reached. (c) Q-factor obtained in the course of averaging over an increasing number of repeats of a bacterial cell wall. As N increases, the Fourier components belonging to the noise background perform a random walk, while the signal-containing Fourier components all add up in the same direction, as illustrated in (b). As a result, the signal-related Fourier components stand out in the Q-factor map when N is sufficiently high. From Kessel et al. (1985), reproduced with permission of Blackwell Science Ltd.

this measure over rings in the spatial frequency domain, and plotting the result, $Q'(k)$, as a function of the ring radius $k=|\mathbf{k}|$. The stipulation that $Q'(k)$ should be equal to or larger than $3/\sqrt{N_{[k,\,\Delta k]}}$ can be used as a resolution criterion. For some additional considerations regarding the statistics of the Q-factor, see Grigorieff (1998).

Sass et al. (1989) introduced a variant of the Q-factor, which they called the S-factor:

$$S(\mathbf{k}) = \frac{|\frac{1}{N}\sum_{i=1}^{N} F_i(\mathbf{k})|^2}{\frac{1}{N}\sum_{i=1}^{N} |F_i(\mathbf{k})|^2} \tag{3.68}$$

This factor relates to the structural content (or, in the parlance of signal processing, the *energy*) of the images being averaged. The expectation value of the S-factor for Fourier coefficients of pure noise is $1/N_{[k,\Delta k]}$. Thus, a resolution criterion can be formulated by stipulating that the ring zone-averaged value of $S(\mathbf{k}) \geq 3/N_{[k,\Delta k]}$.

5.2.8. Spectral Signal-to-Noise Ratio

The spectral signal-to-noise ratio (SSNR) was introduced by Unser et al. (1987; see also Unser et al., 1989, and corresponding 3D forms introduced by Grigorieff, 2000, and Penczek, 2002a) as an alternative measure of resolution. It is based on a measurement of the SNR as a function of spatial frequency and has the advantage of having better statistical performance than DPR and FRC. Unser et al. (1987) also pointed out that the SSNR relates directly to the Fourier-based resolution criteria commonly used in crystallography. An additional advantage is that it allows the *improvement* in resolution to be assessed that can be expected when the data set is expanded, provided that it follows the same statistics as the initial set. This prediction can be extended to the asymptotic resolution reached in the case of an infinitely large data set. Finally, as we will show below, the SSNR provides a way to relate the DPR and the FRC to each other in a meaningful way.

Following Unser's treatment, the individual images q_{i}, which represent single molecules, are modeled assuming a common signal component $[p(\mathbf{r}_{\mathrm{j}}); j = 1 \ldots, J]$ and zero-mean, additive noise:

$$q_i(\mathbf{r}_j) = p(\mathbf{r}_j) + n_i(\mathbf{r}_j) \qquad (i = 1 \ldots, N) \tag{3.69}$$

An equivalent relationship holds in Fourier space:

$$Q_i(\mathbf{k}_l) = P(\mathbf{k}_l) + N_i(\mathbf{k}_l) \qquad (i = 1..., N) \tag{3.70}$$

where $\mathbf{k}_l$ are the spatial frequencies on a discrete 2D grid indexed with l. Both in real and Fourier space, the signal can be estimated by averaging:

$$\bar{p}(\mathbf{r}_j) = \sum_{i=1}^{N} q_i(\mathbf{r}_j); \;\; \bar{P}(\mathbf{k}_l) = \sum_{i=1}^{N} Q_i(\mathbf{k}_l) \tag{3.71}$$

The definition of the SSNR, $\alpha_{\mathbf{B}}$, is based on an estimate of the SNR in a local region **B** of Fourier space. It is given by

$$\alpha_{\mathbf{B}} = \frac{\sigma_{s\mathrm{B}}^2}{\sigma_{n\mathrm{B}}^2/N} - 1 \tag{3.72}$$

The numerator, $\sigma_{s\mathrm{B}}^2$, is the local signal variance, which can be estimated as

$$\sigma_{s\mathrm{B}}^2 = \frac{1}{n_{\mathrm{B}}} \sum_{l \in \mathrm{B}} |\bar{P}(\mathbf{k}_l)|^2 \tag{3.73}$$

where $n_{\mathbf{B}}$ is the number of Fourier components in the region **B**. The denominator in equation (3.72), $\sigma_{n\mathrm{B}}^2$, is the noise variance, which can be estimated as

$$\sigma_{s\mathrm{B}}^2 = \frac{\sum_{l \in \mathrm{B}} \sum_{i=1}^{N} |P_i(\mathbf{k}_l) - \bar{P}(\mathbf{k}_l)|^2}{(N-1)n_{\mathrm{B}}} \tag{3.74}$$

By taking the regions **B** in successive computations of $\alpha_{\mathbf{B}}$ to be concentric rings of equal width in Fourier space, the spatial frequency dependence of $\alpha_{\mathbf{B}}$ can be found, and we obtain the curve $\alpha(\mathbf{k})$. Generally, the SSNR decreases with increasing spatial frequency. The resolution limit is taken to be the point where $\alpha(\mathbf{k})$ falls below the value $\alpha(\mathbf{k}) = 4$ (figure 3.24). Consideration of a numerical model has shown that this limit is roughly equivalent to $\mathrm{DPR} = 45°$. The statistical analysis given by Unser et al. (1987) also allows upper and lower confidence intervals for the $\alpha = 4$ resolution limit to be established. For $N = 30$ images of the herpes simplex virus particles used as a test data set, the resolution was estimated as $1/29\,\text{Å}^{-1}$ ($\alpha = 4$), but the confidence intervals ranged from 1/27 to $1/30\,\text{Å}^{-1}$. This indicates that for such small data sets, resolution estimates obtained with any of the measures discussed should be used with caution.

These authors also address the question of "How much is enough?," referring to the number of noisy realizations that must be averaged to obtain a satisfactory result. From the SSNR curve obtained for a given data set, the resolution improvement available by increasing N to N' can be estimated by shifting the threshold from $\alpha_N = 4$ to $\alpha_{N'} = 4N/N'$ (see figure 3.24).

The statistical properties of the SSNR are summarized in Penczek (2002a). For an FRC derived from independently processed halfsets (see section 5.2.2), the following simple relationships can be shown to hold:

$$FRC = \frac{SSNR}{SSNR + 1} \tag{3.74a}$$

and, conversely,

$$SSNR = \frac{FRC}{1 - FRC} \tag{3.74b}$$

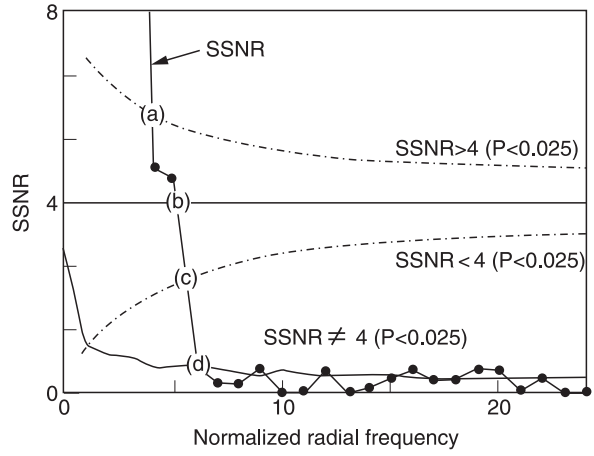

Figure 3.24 Experimental spectral signal-to-noise ratio (SSNR) curve obtained for a set of 30 images of the herpes virus Type B capsomer. With increasing spatial frequency, the curve drops rapidly from an initially high value (>8) to values close to 0. The SSNR resolution limit (here $1/29\,\text{Å}^{-1}$) is given by the spatial frequency where the experimental curve (b) intersects $\text{SSNR} = 4$. The dashed lines represent the (a) upper ($+2\sigma$) and (c) lower (-2σ) confidence limits for a measurement of $\text{SSNR} = 4$. The solid line at the bottom (d) represents the upper (2σ) confidence limit for a measurement of $\text{SSNR} = 0$. From Unser et al. (1987), reproduced with permission of Elsevier.

which needs to be modified into

$$SSNR = \frac{2FRC}{1 - FRC} \tag{3.74c}$$

taking into account that the FRC here reflects the statistics of the halfset. We see that the relationship between the two measures is the same as the one between the CCF and the SNR (section 4.3; Frank and Al-Ali, 1975).

5.3. Resolution and Cross-Resolution

At this point it will be useful to show how Fourier-based measures of resolution can be used to measure "cross-resolution." Although not rigorously defined, the concept originates with attempts to cross-validate results that come from different experiments or even different techniques of imaging. For instance, a structure might be solved both by X-ray crystallography and cryo-EM, and then the question can be asked to what extent, and to which resolution, the latter reproduces the former. It is immediately evident that this question can be answered by applying any of the *halfset criteria* (sections 5.2.3, 5.2.4, and 5.2.5) to the two different maps in question, since these will report the spatial frequency up to which structural information common to both experiments extends. An example for such an assessment was provided by Penczek et al. (1999), who compared density maps obtained from cryo-EM of ribosomes with those

obtained by X-ray crystallography (Ban et al., 1999) and by Roseman et al. (2001), who compared a cryo-EM map of GroEL with a density representation derived from the X-ray structure. We will return to this point in chapter 6, section 3.3.

5.4. Resolution-Limiting Factors

Although the different resolution criteria emphasize different aspects of the signal, it is possible to characterize the resolution limit qualitatively as the limit, in Fourier space, beyond which the signal becomes so small that it "drowns" in noise. Beyond a certain limit, no amount of averaging will retrieve a trace of the signal. This "practical resolution limit" is the result of many adversary effects that diminish the contrast at high spatial frequencies. The effects of various factors on the contrast of specimens embedded in ice have been discussed in detail (Henderson and Glaeser, 1985; Henderson, 1992). Useful in this context is the so-called relative *Wilson plot*, defined in regions where $\mathrm{CTF} \neq 0$ by

$$w(k) = \frac{\text{image Fourier amplitudes}}{\text{electron diffraction amplitudes} \times \mathrm{CTF}} \tag{3.75}$$

where both the expressions in numerator and denominator are derived by averaging over a ring in Fourier space with radius $k = |\mathbf{k}|$. This plot is useful because it allows those effects to be gauged that are *not* caused by the CTF but nevertheless affect the image only, not the diffraction pattern of the same specimen recorded by the same instrument. In the parlance of crystallography, these would be called *phase effects* since they come to bear only because diffracted rays are recombined in the image, so that their phase relationship is important. The most important effects in this category are specimen drift, stage instability, and charging. [Here, we consider effects expressed by classical envelope terms associated with finite illumination convergence and finite energy spread (see chapter 2, section 3.3.2) as part of the CTF.] For instance, drift will leave the electron diffraction pattern unchanged, while it leads to serious deterioration in the quality of the image since it causes an integration over images with different shifts during the exposure time.

Therefore, an ideal image would have a Wilson plot of $w(k) = \text{const}$. Instead, a plot of $w(k)$ for tobacco mosaic virus, as an example of a widely used biological test specimen, shows a falloff in amplitude by more than an order of magnitude as we go from $1/\infty$ to $1/10\,\text{Å}^{-1}$. This cannot be accounted for by illumination divergence (Frank, 1973a) or energy spread (Hanszen, 1971; Wade and Frank, 1977), but must be due to other factors. Henderson (1992) considered the relative importance of contributions from four physical effects: radiation damage, inelastic scattering, specimen movement, and charging. Of these, only the last two, which are difficult to control experimentally, were thought to be most important. Both effects are locally varying, and might cause some of the spatial variation in the power spectrum across the micrograph described by Gao et al. (2002) and Sanders et al. (2003a).

Concern about beam-induced specimen movement led to the development of spot scanning (Downing and Glaeser, 1986; Bullough and Henderson, 1987; see section 3.2 in chapter 2). Charging is more difficult to control, especially at temperatures close to liquid helium where carbon becomes an insulator. Following Miyazawa et al. (1999), objective apertures coated with gold are used as a remedy, since their high yield of back-scattered electrons neutralizes the predominantly positive charging of the specimen.

In addition to the factors discussed by Henderson (1992), we have to consider the effects of conformational variability that is intrinsically larger in single molecules than in those bound together in a crystal. Such variability reduces the resolution not only directly, by washing out variable features in a 2D average or 3D reconstruction, but it also decreases the accuracy of alignment since it leads to reduced similarity among same-view images, causing the correlation signal to drop (see Fourier computation of the CCF, section 3.3.2).

5.5. Statistical Requirements following the Physics of Scattering

So far, we have looked at signal and noise in the image of a molecule, and the resolution achievable without regard to the physics of image formation. If we had the luxury of being able to use arbitrarily high doses, as is the case for many materials science applications of EM, then we could adjust the SNR to the needs of the numerical evaluation. In contrast, biological specimens are extremely radiation sensitive. Resolution is limited both on the sides of high and low doses. For high doses, radiation damage produces a structural deterioration that is by its nature stochastic and irreproducible; hence, the average of many damaged molecules is not a high-resolution rendition of a damaged molecule, but rather a low-resolution, uninformative image that cannot be sharpened by "deblurring." For low doses, the resulting statistical fluctuations in the image limit the accuracy of alignment; hence, the average is a blurred version of the signal.

The size of the molecule is a factor critical for the ability to align noisy images of macromolecules (Saxton and Frank, 1977) and bring single-particle reconstruction to fruition. The reason for this is that for a given resolution, the size determines the amount of structural information that builds up in the correlation peak ("structural content" in terms of Linfoot's criteria; see section 3.3 in chapter 2). Simply put, a particle with 200 Å diameter imaged at 5 Å resolution covers $200^2/5^2 = 1600$ resolution elements, while a particle with 100 Å diameter imaged at the same resolution covers only 400.

By comparing the size of the expected CCF peak with the fluctuations in the background of the CCF for images with Poisson noise, Saxton and Frank (1977) arrived at a formula that links the particle diameter D to the contrast c, the critical dose p_{crit}, and the sampling distance d. Accordingly, the minimum particle diameter D_{min} allowing significant detection by cross-correlation is

$$D = \frac{3}{c^2 d p_{\mathrm{crit}}}$$

Henderson (1995) investigated the limitations of structure determination using single-particle methods in EM of unstained molecules, taking into account the yield ratio of inelastic to elastic scattering, which determines the damage sustained by the molecule, versus the amount of useful information per scattering event in bright-field microscopy. He asked the question "What is the smallest size of molecule for which it is possible to determine from images of unstained molecules the five parameters needed to define accurately its orientation (three parameters) and position (two parameters) so that averaging can be performed?" The answer he obtained was that it should be possible, in principle, to reconstruct unstained protein molecules with molecular mass in the region of 100 kDa to ~3 Å resolution.

However, by a long shot, this theoretical limit has not been reached in practice, for two main reasons: there is likely a portion of noise not covered in these calculations, and secondly, the signal amplitude in Fourier space falls off quite rapidly on account of experimental factors not considered in the contrast transfer theory, a discrepancy observed and documented earlier on (see Henderson, 1992). Experience has shown that proteins should be ~400 kDa or above in molecular mass for reconstructions at resolutions of 10 Å to be obtained. The spliceosomal U1 snRP (molecular mass of ~200 kD) reconstructed by Stark et al. (2001) is one of the smallest particles investigated with single-particle methods without the help of stain, but it does have increased contrast over protein on account of its RNA.

Another conclusion of the Henderson study that may be hard to reconcile with experiments is that the number of images required for a reconstruction at 3 Å resolution is ~19,000, independent of particle size. The reconstruction of the ribosome to 11.5 Å already required 73,000 images (Gabashvili et al., 2000), though later, due to improvements in EM imaging and image processing, the same resolution could be achieved with half that number (e.g., Valle et al., 2003a). Still, this means that 19,000 images is just half the number necessary to achieve a resolution that is lower by a factor of 3 than predicted. Only when the problems of specimen and stage instability and charging are under better control will the estimations by Henderson (1995) match with the reality of the experiment.

5.6. Noise Filtering

The reproducible resolution determined by using one of the methods described in section 3.2 of this chapter is a guide for the decision what part of the Fourier transform contains significant, signal-related information to be kept, and what part contains largely noise to be discarded. Since some signal contributions are found even beyond the nominal resolution boundary, low-pass filtration (or any nonlinear noise filtering techniques) should normally not be applied until the very end of a project, when a final 2D average or 3D reconstruction has been obtained. The reason is that low-pass filtration at any intermediate step would remove the opportunity to utilize the more spurious parts of the signal, and nonlinear filtering would prohibit the use of any linear Fourier-based processing afterwards.

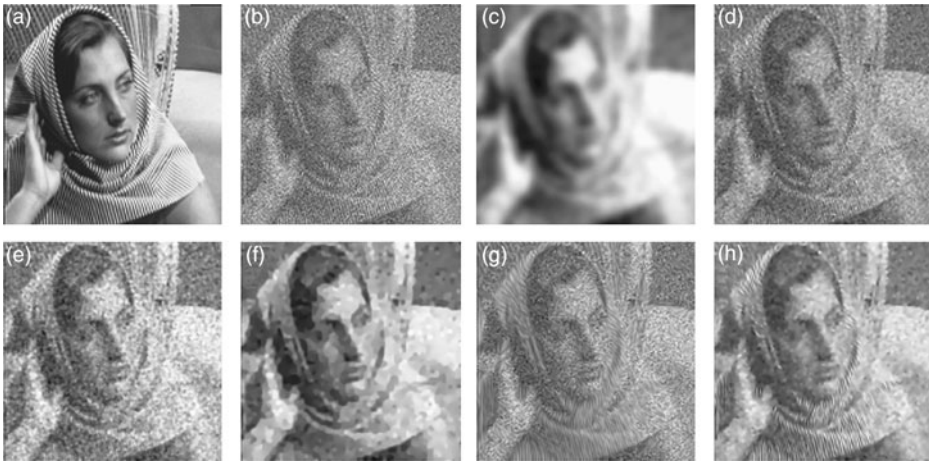

Figure 3.25 Demonstration of low-pass filtration and other means of noise reduction. (a) Test image (256×256); (b) test image with additive "white" noise (SNR = 0.8); (c) Gaussian filter; (d) median filter (5×5 stencil); (e) adaptive Wiener filter; (f) edge enhancing diffusion; (g) coherence enhancing diffusion; (h) hybrid diffusion. From Frangakis and Hegerl (2001), reproduced with permission of Elsevier.

In the following, the profiles of three linear filters for low-pass filtration are given. Other, nonlinear filters are discussed by Frangakis and Hegerl (2001). A demonstration of both types of filters, reproduced from that work, is given in figure 3.25.

Low-pass filters must be designed such that transitions in Fourier space are smooth, to avoid "ringing" or "overshoot" artifacts (also known as *Heaviside phenomenon*) that appear in real space along edges of objects. In those terms, a filter with a top-head profile would have the worst performance. A Gaussian filter, defined as

$$H(k) = \exp[-k^2/2k_0^2] \tag{3.75a}$$

is ideal in this regard, but it has the disadvantage of attenuating, as a function of spatial frequency, too soon while coming to a complete blocking too late relative to the desired spatial frequency radius.

Two other Fourier filters are more frequently used as they offer a choice of width for the transition zone: the Fermi filter (Frank et al., 1985) and the Butterworth filter (Gonzales and Woods, 1993).

The Fermi filter follows the distribution of particles following the Fermi statistics at a given temperature; hence, the width of the transition zone is specified by the "temperature" parameter T:

$$H(k) = \frac{1}{1 + \exp[(k - k_0)/T]} \tag{3.75b}$$

The Butterworth filter has the following profile:

$$H(k) = \frac{1}{1 + |(k/k_0)|^{2\kappa}} \tag{3.75c}$$

where κ is a parameter that determines the order of the filter.

Profiles of some of the filter functions, along with the associated point-spread functions, and examples for their application, are contained in appendix 2.

6. Validation of the Average Image

Questions of significance and reproducibility of features have been addressed above in this chapter (section 4.2). The problem with such treatments is that they must be based on assumptions regarding the statistical distribution of the noise. In the following we will discuss a technique that requires no such assumptions.

The question of significance of features in the average image can be addressed by a method of multiple comparison that makes no assumptions regarding the statistics: the rank sum method (Hänicke, 1981; Hänicke et al., 1984). Given a set of N images, each represented by J pixels:

$$p_{ij} = \{ p_i(\mathbf{r}_j);\ j = 1 \ldots, J;\ i = 1 \ldots, N \} \tag{3.76}$$

The nonparametric statistical test is designed as follows: each pixel p_{ij} of the ith image is ranked according to its value among the entire set of pixels $\{ p_{ij};\ j = 1 \ldots, J\}$: the pixel with the lowest value is assigned rank 1, the second lowest, rank 2, and so forth. Finally, the pixel with the largest value receives rank J. In the case of value ties, all pixels within the equal-value group are assigned an average rank. In the end, the image is represented by a set of rank samples $\{r_{i1}, r_{i2}, \ldots, r_{iJ}\}$, which is a permutation of the set of ordered integers $\{1, 2, \ldots, J\}$ (except for those rank samples that result from ties). After forming such rank representations for each image, we form a *rank sum* for each pixel:

$$R_j = \sum_{i=1}^{N} r_{ij}, \qquad 1 \le j \le J \tag{3.77}$$

In order to determine whether the difference between two pixels, indexed j and k, of the average image,

$$\bar{p}_j = 1/N \sum_{i=1}^{N} p_{ij}, \qquad \bar{p}_k = 1/N \sum_{i=1}^{N} p_{ik} \tag{3.78}$$

is statistically significant, on a specified significance level α, we can use the test statistic

$$c_{jk} = |R_j - R_k| \tag{3.79}$$

Now the following rule holds: *whenever c_{jk} is greater than a critical value $D(\alpha, J, N)$, the difference between the two pixels is deemed significant*. This critical value has been tabulated by Hollander and Wolfe (1973) and, for large N, by Hänicke (1981).

In addition to answering questions about the significance of density differences, the rank sum analysis also gives information about the spatial resolution. Obviously, the smallest distance (*critical distance*) between two pixels that fulfills

$$c_{jk} > D(\alpha, J, N) \tag{3.80}$$

is closely related to the local resolution in the region around those pixels. The nature of this relationship and a way to derive a global resolution measure from this are discussed by Hänicke et al. (1984).