International Tables for Crystallography (2010). Vol. B, Section 2.5.7.6, pp. 379–381.

### 2.5.7.5. 2D data analysis – particle picking

Depending on the properties of the imaged complex and the magnification used, a single micrograph can yield from a few to thousands of individual particle projections. The first step of the data processing is identification of particle projections in micrographs and their selection. The particles have to be windowed (boxed) using a window size exceeding the particle size by a 30–50% margin. Thus, for example, in order to determine the structure of a 550 kDa complex that has a diameter of ~120 Å to 12 Å resolution, it is appropriate to choose a pixel size of 3 Å and a window size of 60 pixels.

The selection of particles is a labour-intensive process; however, the quality of selected particle projections is a major factor in the subsequent steps of analysis and the inclusion of too many imperfect images may preclude successful determination of the 3D structure. There are three possible approaches: (1) manual selection; (2) semi-automated selection; (3) fully automated selection. In the early stages of analysis, particularly when little is known about the shape of the protein and the distribution of projection views, the manual approach is preferable. The researcher displays the micrograph on a computer screen (usually preprocessed by Fourier filtration and contrast-adjusted for better visibility of the protein) and interactively identifies locations of particle views. A trained and careful operator can yield much better results than automated approaches. The main risk is in inherent bias of a human operator – there is a tendency to focus on more familiar and more easily visible particle projections, omitting less frequently appearing orientations and in effect jeopardizing successful structure determination. In semi-automated approaches, an initial step in which putative particle projections in a micrograph are chosen is performed by a computer, all candidates are windowed and the user screens a gallery of possible particles instead of the full micrograph. Algorithms that perform the initial identification of particle views range from very simple (for example a band-pass filtration of a micrograph with subsequent selection of peaks that are no closer to each other than half of the expected particle size) to sophisticated nonlinear noise-suppression methods [for details on various algorithms see the Special Issue of the *Journal of Structural Biology* (Zhu *et al.*, 2004)]. Since the human operator will be responsible for the ultimate decision, preference is given to the faster method. In most cases, semi-automated methods are implemented within a framework of a user-friendly graphical user interface that can greatly facilitate the work. Fully automated methods are currently actively under development but, curiously, even for proteins whose high-resolution structure is known, the success rate cannot match that of a human operator (Zhu *et al.*, 2004).

The automated procedures can be divided into three groups: (1) those that rely on *ad hoc* steps of denoising and contrast enhancement followed by the search for regions of known size that emerge above the background level (Adiga *et al.*, 2004); (2) those that extract orientation-independent statistical features from regions of the micrograph that may contain particles and proceed with classification (Lata *et al.*, 1995; Hall & Patwardhan, 2004); and (3) those that employ templates, *i.e.*, either class averages of particles selected from micrographs or projections of a known 3D structure of the complex (Huang & Penczek, 2004; Sigworth, 2004).

The advantage of the first two approaches is that they do not require template images, *i.e.*, since they are based on a very broadly defined description of particles (general size, shape or abstract features derived from examples of typical particles), they are applicable in cases when no 3D structure of the complex is available. Methods from the second category usually require a training session for the algorithm to construct a set of weights for the predefined features. The methods that take advantage of the availability of templates vary greatly in complexity from straightforward cross-correlation with a generic shape (a Gaussian function, a low-passed circle) (Frank & Wagenknecht, 1984) to matched filters with large number of templates and parameters derived from the image-formation model of the micrograph (CTF and envelope functions) (Huang & Penczek, 2004; Sigworth, 2004). The motivation is clear: given an ideal object and image-formation parameters, it should be only a matter of sheer computer power and user's patience to have all particles matching the template selected. In practice, the problem is much more challenging and the success rate of template-based methods does not necessarily exceed the success rate of carefully tuned *ad hoc* methods.

One of the difficulties with the application of correlation techniques to the particle-picking problem is the unevenness of micrographs, which is caused by uneven illumination by the electron beam and, to a much larger degree, by the uneven thickness of the ice layer and, when used, the supporting carbon. A possible remedy is to calculate a 'locally normalized' cross-correlation function, in which the total variance of the micrograph is replaced by the local variance of the micrograph calculated within a window of $n$ pixels centred on the current location $l$. This method has a fast implementation in Fourier space (van Heel, 1982; Roseman, 2003). A faster method is to just apply a high-pass filtration of the micrograph using a high-pass Gaussian Fourier filter with a half-width $(1/np)$ Å$^{-1}$, where $p$ is the pixel size. This simple step will all but eliminate the unevenness of the micrograph background.

The main difficulty with the correlation technique is the computational complexity of the problem arising from the very large number of templates that have to be considered. The particles in the micrograph are projections of a 3D object with arbitrary in-plane rotations. In effect, to perform an exhaustive search, it is necessary to sample quasi-uniformly three Eulerian angles [equation (2.5.7.17) with $\Delta\psi = \delta\theta$]. For example, a very crude angular step of $\delta\theta = 10^\circ$ results in ~13 000 2D templates! A reduction in the number of templates can be achieved either using clustering techniques (Huang & Penczek, 2004; Wong *et al.*, 2004) or by exploring the eigenstructure of the whole set of templates (Sigworth, 2004).

### 2.5.7.6. 2D alignment of EM images

Alignment of pairs of 2D images is a fundamental step in single-particle reconstruction. It is aimed at bringing into register various particle projections by determining three orientation parameters (rotation angles and $x$ and $y$ translations) and is employed in 2D alignment of large sets of 2D noisy data and in 3D structure-refinement algorithms. The computational efficiency and numerical accuracy of this step are deciding factors in achieving high-quality structural results in an acceptable time.

All 2D alignment methods considered are aimed at finding transformation parameters such that the least-squares discrepancy between two images $f$ and $g$ is minimized,

$$\int \left| f(\mathbf{x}) - g(\mathbf{T}\mathbf{x}) \right|^2 \,\mathrm{d}\mathbf{x} \rightarrow \min, \qquad (2.5.7.7)$$

where $\mathbf{x} = [x \quad y \quad 1]^T$ is a vector containing the coordinates. $\mathbf{T}$ is the transformation matrix given by

$$\mathbf{T}(\alpha, x, y) = \begin{bmatrix} \cos\alpha & -\sin\alpha & t_x \\ \sin\alpha & \cos\alpha & t_y \\ 0 & 0 & 1 \end{bmatrix}, \qquad (2.5.7.8)$$

and is dependent on three transformation parameters: rotation angle $\alpha$ and two translations $t_x$ and $t_y$ . It has to be noted that a minimum of (2.5.7.7) can be found rapidly using the fast Fourier transform (FFT) algorithm if only the *xy* translation is sought (2D FFT), or if only the rotation angle is needed (1D FFT).

2D alignment methods can be divided into three classes: (1) those that employ exhaustive searches in order to find three orientation parameters; (2) those that perform exhaustive searches by using either simplifications (separate searches for translation and rotation parameters) (Penczek *et al.*, 1992) or by taking advantage of invariant image representations (Schatz & van Heel, 1990; Frank *et al.*, 1992 and the following discussion; Schatz & van Heel, 1992; Marabini & Carazo, 1996); or finally (3) those that are aimed at improvement of previously determined parameters and employ local searches.

In practice, as the windowed particles are approximately centred, the search for translation parameters can be restricted to relatively small values. A very efficient algorithm that takes advantage of the geometry is based on resampling to polar coordinates of the area of the image that roughly corresponds to the particle size. The resampling is done around centres placed on pixels located within a distance from the image centre that corresponds to a preset maximum translation (Joyeux & Penczek, 2002) (Fig. 2.5.7.2). For each translation, a 1D rotational cross-correlation function in polar coordinates is calculated. Overall, the alignment method based on resampling to polar coordinates comprises the following steps: (1) the image is resampled to polar coordinates; (2) 1D FFTs of various lengths are calculated, appropriately weighted and padded with zeros to equalize their lengths; (3) complex multiplications with 1D Fourier transforms of the similarly processed referenced image are calculated; (4) the inverse 1D FFT is calculated and the position of the maximum is found. The last step yields the rotation angle. Steps (1)–(4) are repeated with the image that is being aligned shifted to account for translations. In addition, the rotation angle for one of the images being mirrored is efficiently calculated in parallel with step (3) by repeating the multiplication with the 1D Fourier transforms of the reference image complex conjugated. This additional check is a necessity in the analysis of single-particle data sets, as usually one can expect on average half of the images to be mirrored versions of the other half in the data set. Overall, the method is very accurate, because only data under the circular mask enter the calculation.

For a set of $N$ images containing the same object in various orientations and corrupted by an additive noise, the problem of alignment would be relatively simple. For proteins that have strong preferred orientation and particularly when a staining technique is used for grid preparation, this is certainly the case. In the procedure called *reference-based alignment*, one of the images that appears ‘typical’ is selected and used as a reference to align the remaining images. After all available images are aligned their average is calculated and used as a reference in a repeated alignment of all images. The process is iterated until the orientations of the images stabilize (Frank *et al.*, 1982).

More formally, Frank *et al.* (1988) proposed the definition of a set of $N$ images $f_k$, $k = 1, \ldots, N$, aligned if a set of transformations $\mathbf{T}_k$, $k = 1, \ldots, N$, (rotation angles and translations) is found such that all pairs of images are mutually brought into register, so the expression

$$\begin{aligned}L_1(\{f\}, \{\mathbf{T}\}) &= \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \left\| f_k(\mathbf{T}_k\mathbf{x}) - f_l(\mathbf{T}_l\mathbf{x}) \right\|^2 \\ &= \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \left( \left\| f_k(\mathbf{T}_k\mathbf{x}) \right\|^2 + \left\| f_l(\mathbf{T}_l\mathbf{x}) \right\|^2 - 2 f_k(\mathbf{T}_k\mathbf{x}) f_l(\mathbf{T}_l\mathbf{x}) \right)\end{aligned} \tag{2.5.7.9}$$

is minimized. Although there is no simple way to minimize $L_1$, the interesting observation is that there is no requirement of the images to represent the same particle, not even a similar one. This leads to the conclusion that if the minimum of $L_1$ could be found, a set of diverse images could be aligned; moreover, upon alignment similar images would have similar orientation and subsequent classification of such an aligned data set would reveal subsets of similar images.

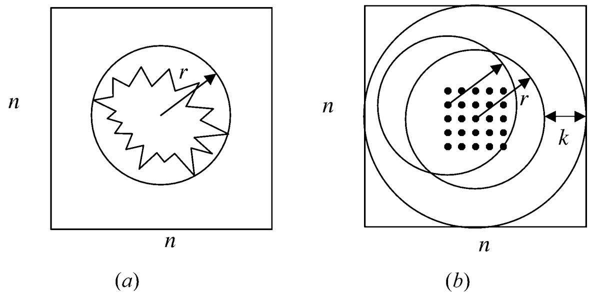

Fig. 2.5.7.2. The geometrical constraints of the 2D alignment problem. (*a*) The reference 2D particle is placed within a square image frame $n \times n$ pixels and its size is such that it can be bounded by a circle with a radius $r$ no larger than 0.9$n$. (*b*) The particle projection, the size of which is bounded by the same radius as the reference view, can be located within a circle centred on discrete locations within the image frame, such that the maximum translation is $k = (n/2) - r$. The number of possible translations is $(2k + 1)^2$. Reprinted from Joyeux & Penczek (2002) with permission from Elsevier.

A practical method of minimizing, called a *reference-free alignment*, was proposed by Penczek *et al.* (1992) by showing that minimization of $L_1$ is equivalent to maximization of

$$L_2(\{f\}, \{\mathbf{T}\}) = \sum_{k=1}^{N-1} \left\| f_k(\mathbf{T}_k\mathbf{x}) - \langle f \rangle_k \right\|^2, \tag{2.5.7.10}$$

where

$$\langle f \rangle_k = \frac{1}{N-1} \sum_{l=1, l\neq k}^{N} f_l(\mathbf{T}_l\mathbf{x}) \tag{2.5.7.11}$$

is the partial average of the set of images calculated with the exclusion of the $k$th image. The method is based on the observation that given a set of approximately aligned images, it should be possible to minimize $L_2$ by sequentially correcting alignments of individual images using the cross-correlation function between each image and the average of the remaining ones. On each step, depending whether the orientation of the image changes or not, (2.5.7.10) will decrease or remain constant.

The outcome of the reference-free alignment algorithm is an aligned set of $N$ images, so all particles that have similar shapes will have similar orientations. Thus, it is natural (and because of the alignment possible) to divide the data set into classes of images that have similar shapes and orientations, *i.e.*, to cluster them. A number of well known clustering algorithms have been adopted for EM applications (Frank, 1990). The general purpose of clustering is to organize objects (in the case of EM, images) into classes whose members are similar to each other, while dissimilar to objects from other classes.

Reference-free alignment with subsequent clustering works well as long as all particles share the same overall shape (*i.e.*, the very low frequency component), as is the case for ribosomes. However, some molecules yield projections that have quite different shapes, as for example is the case for barrel-like proteins GroEL (Roseman *et al.*, 1996) with rectangular views and circular end views or flat and rectangular hemocyanin (Boisset *et al.*, 1995). In this case, the reference-free alignment tends to be unstable, as (2.5.7.10) has multiple local minima, which in practice means that the global average of the whole data set can vary significantly depending on the initiation of the procedure. In general, reference-free alignment is an ‘alignment first, classification second' approach. It is possible to reverse this order by using invariants with the supporting rationale that once approximately homogeneous classes of images were found, it should be easy to align them subsequently as within each class they will share the same motif.

A practical approach to reference-free alignment known as *alignment by classification* (Dube *et al.*, 1993) is based on the observation that for a very large data set and centred particles one can expect that although the in-plane rotation is arbitrary, there is a high chance that at least some of the similar images will be in the same rotational orientation. Therefore, in this approach the images are first (approximately) centred, then subjected to classification, and subsequently aligned.

In its simplest form, the *multireference alignment* belongs to the class of *supervised classification* methods: given a set of templates (*i.e.*, reference images; these can be selected unprocessed particle projections, or class averages that resulted from preceding analysis, or projections of a previously determined EM structure, or projections of an X-ray crystallographic structure), each of the images from the available data sets is compared (using a selected discrepancy measure) with all templates and assigned to the class represented by the most similar one. Equally often multireference alignment is understood as a form of *unsupervised classification*, more precisely *K*-means classification, even if the description is not formalized in terms of the latter. Given a number of initial 2D templates, the images are compared with all templates and assigned to the most similar one. New templates are calculated by averaging images assigned to their predecessors and the whole procedure is repeated until a stable solution is reached.

### 2.5.7.7. *Initial determination of 3D structure using tilt experiments*

The 2D analysis of projection images provides insight into the behaviour of the protein on the grid in terms of the structural consistency and the number and shape of projection images. In order to obtain 3D information, it is necessary to find geometrical relations between different observed 2D images. The most robust and historically the earliest approach is based on tilt experiments. By tilting the stage in the microscope and acquiring additional pictures of the same area of the grid it is possible to collect projection images of the same molecule with some of the required Eulerian angles determined accurately by the setting of the goniometer of the microscope.

In random conical tilt (RCT) reconstruction (Radermacher *et al.*, 1987), two micrographs of the same specimen area are collected: the first one is recorded at a tilt angle of ~50° while the second one is recorded at 0° (Fig. 2.5.7.3). If particles have preferred orientation on the support carbon film (or within the amorphous ice layer, if no carbon support is used), the projections of particles in the tilted micrographs form a conical tilt series. Since in-plane rotations of particles are random, the azimuthal angles of the projections of tilted particles are also randomly distributed; hence the name of the method. The untilted image is required for two reasons: (i) the particle projections from the untilted image are classified, thus a subset corresponding to possibly identical images can be selected ensuring that the projections originated from similar and similarly oriented structures; and (ii) the in-plane rotation angle found during alignment corresponds to the azimuthal angles in three dimensions (one of the three Eulerian angles needed). The second Eulerian angle, the tilt, is either taken from the microscope setting of the goniometer or calculated based on geometrical relations between tilted and untilted micrographs. The third Eulerian angle corresponds to the angle of the tilt axis of the microscope stage and is also calculated using the geometrical relations between two micrographs. In addition, it is necessary to centre the particle projections selected from tilted micrographs; although various correlation-based schemes have been proposed, the problem is difficult as the tilt data tend to be very noisy and have very low contrast.

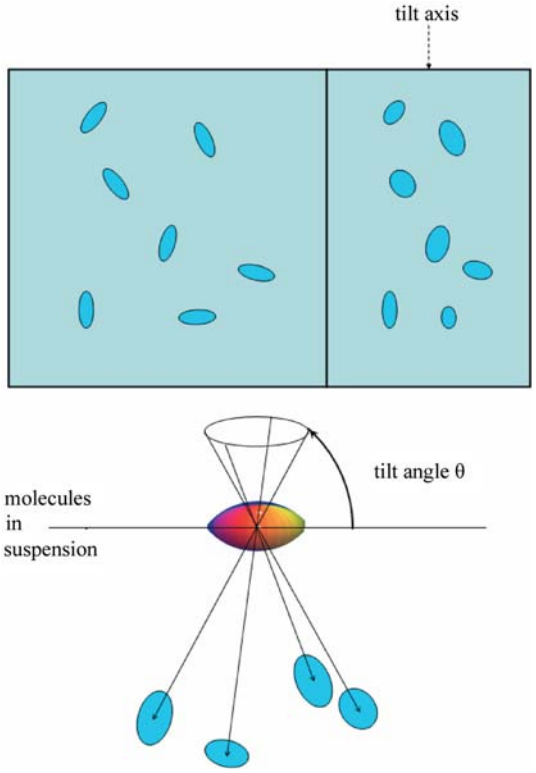

Fig. 2.5.7.3. Principle of random conical tilt reconstruction. A tilt pair of images of the same grid area is collected. By aligning the particle images in the untilted micrograph (left), the Eulerian angles of their counterparts in the tilted micrograph (right) are established. The particle images from the tilted micrograph are used for 3D reconstruction of the molecule (bottom). The set of projections form a cone in Fourier space; information within the cone remains undetermined.

Given three Eulerian angles and centred tilted projections, a 3D reconstruction is calculated. There are numerous advantages of the RCT method. (i) Assuming the sign of the tilt angle is read correctly (it can be confirmed by analysing the defocus gradient in the tilted micrographs), the method yields a correct hand of the structure. (ii) With the exception of the in-plane rotation of untilted projections, which can be found relatively easily using alignment procedures, the remaining parameters are determined by the experimental settings. Even if they are not extremely accurate, the possibility of a gross error is eliminated, which positively distinguishes the method from the *ab initio* computational approaches that use only untilted data. (iii) The computational analysis is entirely done using the untilted data, which have high contrast. (iv) The RCT method is often the only method of obtaining 3D information if the molecule has strongly preferential orientation and only one view is observed in untilted micrographs. The main disadvantage is that the conical projection series leaves a significant portion of the Fourier space undetermined. This follows from the central section theorem [equation (2.5.6.8) of Section 2.5.6]: as the tilt angle is less than 90°, the undetermined region can be thought to form a cone in three dimensions and is referred to as the missing cone. The problem can be overcome if the molecule has more than one preferred orientation. Subsets of particles that have similar untilted appearance (as determined by clustering) are processed independently and for each a separate 3D structure is calculated. If the preferred orientations are sufficiently different, *i.e.*, the orientations of the original particles in three dimensions are sufficiently different in terms of their angles with respect to the *z* axis, the 3D structures can be aligned and merged, all but eliminating the problem of the missing cone and yielding a robust, if resolution-limited, initial model of the molecule (Penczek *et al.*,

references