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 α 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