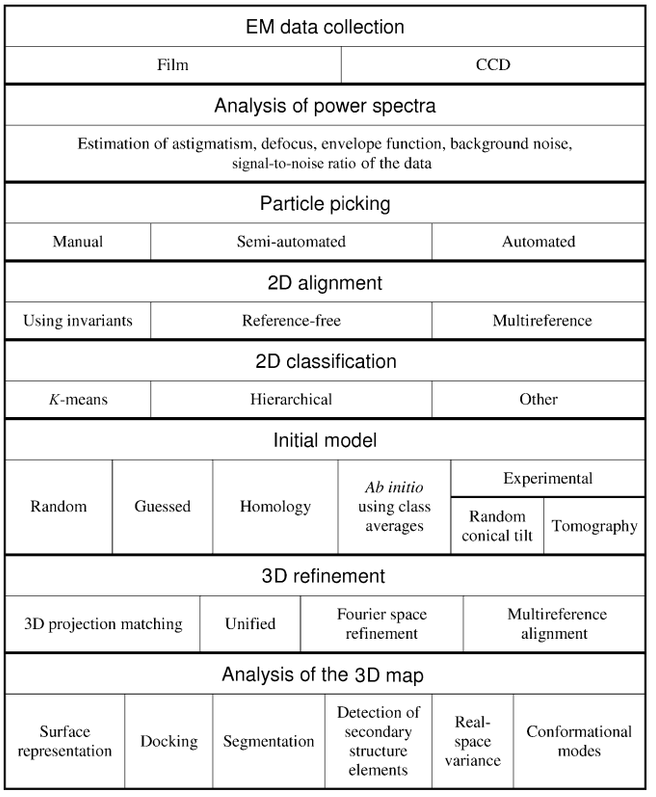


Fig. 2.5.7.1. Typical steps performed in a single-particle cryo-EM structure determination project.

between the orientation parameters $\psi_n, \theta_n, \varphi_n, t_{x_n}, t_{y_n}$ ($n$ = 1, 2, ..., $N$) and the 3D density $f$.

The parameters in (2.5.7.6) to be determined can be separated into two groups. (1) The orientation parameters $\psi_n, \theta_n, \varphi_n, t_{x_n}, t_{y_n}$ that have to be determined entirely by solving (2.5.7.6) and for which there are no initial guesses, and the structure $f$ itself, for which we may or may not have an initial guess. The number of parameters in this group is very large: $n^3 + 5m$. Note that in single-particle reconstruction, the number of projection data $m$ is far greater than the linear size of the data in pixels, *i.e.*, $m \gg n$. (2) Various parameters which we will broadly call the parameters of the image formation model (2.5.7.1)–(2.5.7.4): the defocus settings of the microscope $\Delta f_n$, the amplitude contrast ratio $q$ and, if analytical forms of the envelope function $E$, the power spectrum of the background noise $M$, or the structure $F$ are adopted, the parameters of these equations. Some of the parameters in the second group are usually known very accurately or can be estimated from micrograph data before one attempts to solve (2.5.7.6) (see Section 2.5.7.4), but they can also be refined during the structure determination process [for the method for correcting the defocus settings, see Mouche *et al.* (2001)].

Owing to the very large number of parameters in (2.5.7.6) and the nonlinearities present, one almost never attempts to solve the problem directly. Instead, structure determination using the single-particle technique involves several steps. (i) The macromolecular complex is prepared with a purity of at least 90%. (ii) The sample is flash-frozen in liquid ethane. Alternatively, cryo-negative stain techniques or traditional negative stain methods can be used. (iii) Pictures of the macromolecular complexes are taken. (iv) Exhaustive analysis of 2D particle images aimed at increasing the SNR of the data and evaluation of the homogeneity of the sample is performed. (v) An initial low-resolution model of the structure is established using either experimental techniques or computational methods. (vi) The initial structure is refined in order to increase the resolution using an enlarged data set. Only in this step does one attempt to minimize (2.5.7.6) more or less directly. (vii) Visualization and interpretation of the resulting 3D electron-density map is the last step; it often involves docking of X-ray structures of molecules into EM density maps in order to reveal the arrangement of known molecules within the EM envelope (Fig. 2.5.7.1). As within the weak-phase-object approximation of the image formation in EM the relation between densities in collected images and the 3D electron density of the imaged macromolecule is linear [(2.5.7.1)], all data-processing methods employed in the structure determination project should be linear, so the densities in the cryo-EM 3D model can be interpreted in terms of the electron density of the protein.

In the actual single-particle project not all the steps have to be executed in the order outlined above. The technique has proved to be particularly useful in studies of functional complexes of proteins whose base state is known to a certain resolution or even of functional complexes whose atomic (X-ray crystallographic) structure is known. In these cases, steps (iv) and (v) can be omitted and the structure of the functional complex (for examples with ligands bound to it) can be relatively easily determined using the native structure as a starting point for step (vi).

In addition to difficulties with obtaining good cryo-EM data, the technique is computationally intensive. The reason is that in order to obtain a sufficient SNR in the 3D structure, processing of hundreds of thousands of EM projection images of the molecule might be necessary. For each, five orientation parameters have to be determined, and this is in addition to determination of the image-formation parameters required for the optimization of correlation searches. In effect, it is not unusual for single-particle projects to consume weeks of the computer time of multiprocessing clusters. This also explains why the knowledge of the base structure simplifies the work to a large degree: when it is known, initial values of the orientation parameters can be easily established, reducing not only the computational time, but also possibilities of errors in the structure-determination process.

### *2.5.7.3. Electron microscopy and data digitization*

The electron microscope is a phase imaging system; *i.e.*, in order to create contrast in images, they have to be underfocused. Owing to the particular form of the CTF of the microscope [(2.5.7.4)], not only the amplitudes of the image in Fourier space are modified, but information in some ranges of spatial frequencies is set to zero and some phases have reversed sign. Therefore, in order to obtain possibly uniform coverage of Fourier space, the standard practice is to take pictures using different defocus settings and merge them computationally in order to fill gaps in Fourier space. The problem is compounded by the relation between underfocus and the envelope function of the microscope. Far-from-focus images have high contrast, but the envelope function has a relatively steep fall-off limiting the range of useful spatial frequencies. Conversely, close-to-focus images have little contrast, but the envelope function is decreasing, slowly extending useful information to high spatial frequencies. In effect, it is easier to process computationally far-from-focus data and to obtain accurate alignment of particles, but the results have severely limited resolution. Processing of close-to-focus data is challenging and results tend to be less accurate, but there is the potential to obtain high-resolution information.

The experimental techniques of initial structure determination (random conical tilt, tomography) require collection of tilt data. This is facilitated by dedicated microscope stages that can be rotated inside the microscope column yielding additional views of the same field. However, collection of high-quality tilt images is difficult. The quality of tilted images tends to be adversely affected by charging and drift effects. Moreover, as the stage is tilted the effective ice thickness increases (inversely proportionally to the cosine of the tilt angle, so at 60° the factor is two) and the contrast of the images decreases correspondingly. Finally,

the defocus in tilted micrographs varies depending on the position in the field, often forcing users to restrict the particle selection only to regions in the vicinity of the tilt axis. However, tilting establishes geometrical relations between different projections of the same particle, unambiguously allowing for robust determination of an initial 3D model and the handedness of the quaternary structure of the complex.

Electron microscope images can be either recorded on the film and subsequently converted to digital format, or they can be recorded using a charge-coupled device (CCD) camera in a digital format directly on a microscope. In either case, it is necessary to select the magnification of the microscope and the eventual pixel size of the digitized data before the data-collection session. High magnification can potentially yield high-resolution data, but at the same time it decreases the yield of particles. Lower magnification values can be used when images are recorded on film, which does not attenuate high spatial frequencies to the same extent as CCD cameras tend to do.

The pixel size has to be adjusted according to the expected resolution of the final structure. Although it is tempting to adopt a small pixel size (in the hope of achieving high resolution of the results), in most cases this is counterproductive, as it results in very large computer files that are difficult to handle and in excessively long data-processing times. Theoretically, the optimum pixel size is tied to the maximum frequency present in the data by Shannon's sampling theorem, which states that no information is lost if the signal is sampled at twice the maximum frequency present in the signal, and no additional information is gained by sampling using higher frequency. Thus, if the expected resolution is 12 Å, it should be sufficient to use a pixel size (on the specimen scale) of 6 Å. In practice, various image-processing operations performed during alignment of the data and 3D reconstruction of the complex significantly lower the range of useful frequencies. This is because in currently available single-particle reconstruction software packages rather unsophisticated interpolation schemes are employed, which were selected mainly for the speed of calculations. Therefore, it is advisable to oversample the data by a factor of 1.5 or even 3.0. For an expected resolution of 12 Å this corresponds to pixel sizes of 4 and 2 Å, respectively.

The windowed particles have to be normalized to adjust the image densities to a common framework of reference. The reason for this step is that microscopy conditions are never exactly the same and also within the same micrograph field the background densities can vary by a significant margin due to uneven ice thickness and other factors. A sensible approach to normalization is to assume that the statistical distribution of noise in areas surrounding particles should be the same (Boisset *et al.*, 1993). Hence a large portion of one of the micrographs from the processed set is selected and a reference histogram of its pixel values is generated. Next, assuming a linear transformation of pixel values, the two parameters of this transformation are found in such a way that the histogram of the transformed pixel values surrounding the particle optimally matches the reference histogram using $\chi^2$ statistics as a discrepancy measure.

### *2.5.7.4. Assessment of the data quality and estimation of the image formation parameters*

The initial assessment of the quality of the micrographs is usually performed during the data collection and in most cases before the micrographs are digitized. The micrographs are examined visually and those that have noticeable drift, astigmatisms, noticeable contamination or simply too low a number of particles to justify further analysis are simply discarded. After digitization of the accepted micrographs, the first step is estimation of the power spectrum, which will be examined for the presence of Thon rings (thus confirming that the micrograph is indeed usable) and astigmatism.

The method of averaged overlapping periodograms (Welch, 1967) is commonly used in EM to calculate the power spectrum. It is designed to improve the statistical properties of the estimate by taking advantage of the fact that when $K$ identically distributed independent measurements are averaged, the variance of the average is decreased with respect to the individual variance by the ratio $1/K$. Thus, instead of calculating a periodogram (squared moduli of the discrete Fourier transform) of the entire micrograph field, one subdivides it into much smaller windows, calculates their periodograms and averages them. Typically, one would chose a window size of $512 \times 512$ pixels and an overlap of 50%, which will result in the reduction of the variance of the estimate to few percent with respect to the variance of the periodogram of the entire field (Fernandez *et al.*, 1997; Zhu *et al.*, 1997). Further reduction of the variance is achieved by rotational averaging of the 2D power-spectrum estimate. The resulting one-dimensional (1D) profile is finally used in the third step of our procedure.

For a set of micrographs the power spectra can be evaluated either visually or computationally in an automated fashion. Of main concern are the presence of Thon rings, the astigmatism and the extent to which Thon rings can be detected. Although in principle astigmatic data could be used in subsequent analysis (in fact, astigmatism could be considered advantageous, as particles from the same micrograph would contain complementary information in Fourier space), in practice they are discarded as currently there is no software that can process astigmatic data efficiently. The extent of Thon rings indicates the 'resolution' of the data, *i.e.*, the maximum frequency to which information in the data can be present.

A number of well established programs can assist the user in the calculation of power spectra and automated estimation of defocus and astigmatism (Huang *et al.*, 2003; Mindell & Grigorieff, 2003; Sander *et al.*, 2003; Mallick *et al.*, 2005). Given the analytical form of the CTF [(2.5.7.4)], the problem is solved by a robust fitting of the CTF parameters such that the analytical form of the CTF matches the power spectrum of the micrograph. Usually, the steps employed are: (1) robust estimation of the power spectrum; (2) calculation of the rotational average of the power spectrum; (3) subtraction from this rotational average of the slowly decreasing background [roughly corresponding to $P_{\rm B}$ in (2.5.7.5)]; (4) fitting of the defocus value $\Delta f_n$ using known settings of the microscope (voltage, spherical aberration constant, ...) and usually assuming a constant and known value of the amplitude contrast ratio $q$ (for cryo-EM data, $q$ should be in the range 0.02–0.10); and (5) using the established defocus value $\Delta f_n$, analysis of the 2D power spectrum and fitting of the astigmatism amplitude and angle while refining the defocus. As long as the defocus value is not too small and there are at least two detectable zeros of the CTF, all available programs give very good and comparable results.

In some single-particle packages, the automated calculation of defocus is integrated with the estimation of additional characteristics of the image-formation parameters that are required for advanced application of a Wiener filter [(2.5.7.18)] (Saad *et al.*, 2001; Huang *et al.*, 2003), *i.e.*, the power spectra of two noise distributions $P_{\rm S}$ and $P_{\rm B}$ and the envelope function of the microscope $E_n$ for each micrograph. A possible approach is to select slowly varying functions and fit their parameters to match the estimates of $P_{\rm S}$, $P_{\rm B}$ and $P_d$ obtained from the data. Finally, it is necessary to have a description of the 1D rotationally averaged power spectrum of the complex $P_f$. One possibility is to carry out X-ray solution scattering experiments (Gabashvili *et al.*, 2000; Saad *et al.*, 2001) that yield a 1D power spectrum of the complex in solution. However, these experiments require large amounts of purified sample and the accuracy of the results in terms of the overall fall-off of the power spectrum can be disputed. For the purpose of cryo-EM, a simple approximation of the protein power spectrum by analytical functions is satisfactory.