Chapter 11

Semiautomatic, High-Throughput, High-Resolution Protocol for Three-Dimensional Reconstruction of Single Particles in Electron Microscopy

Carlos Oscar Sorzano, J.M. de la Rosa Trevín, J. Otón, J.J. Vega, J. Cuenca, A. Zaldívar-Peraza, J. Gómez-Blanco, J. Vargas, A. Quintana, Roberto Marabini, and José María Carazo

Abstract

In this chapter we describe the steps needed for reconstructing the three-dimensional structure of a macromolecular complex starting from its projections collected in electron micrographs. The concepts are shown through the use of Xmipp 3.0, a software suite specifically designed for the image processing of biological structures imaged with electron or X-ray microscopy. We illustrate the image processing workflow by applying it to the images of Bovine Papilloma virus published in Wolf et al. (Proc Natl Acad Sci USA 107:6298–6303, 2010). We show that in the case of high-quality, homogeneous datasets with a priori knowledge about the initial volume, we can have a high-resolution 3D reconstruction in less than 1 day using a computer cluster with only 32 processors.

Key words: Single particle analysis, Electron microscopy, Image processing, 3D reconstruction, Workflows

1. Introduction

The study of the structure of protein and protein complexes by Transmission Electron Microscopy (TEM) provides key insight into the way that these macromolecules perform their function in the cell (1–3). One of the techniques available to perform these studies is called Single Particle Analysis (SPA). In SPA, thousands of projections of different copies of the same molecule are computationally combined in a single volume. It is assumed that all particles being analyzed correspond to exactly the same conformation. If this is not the case, projection images are separated in different classes of homogeneous populations. The quality of the images

Alioscka A. Sousa and Michael J. Kruhlak (eds.), *Nanoimaging: Methods and Protocols*, Methods in Molecular Biology, vol. 950, DOI 10.1007/978-1-62703-137-0_11, © Springer Science+Business Media, LLC 2013

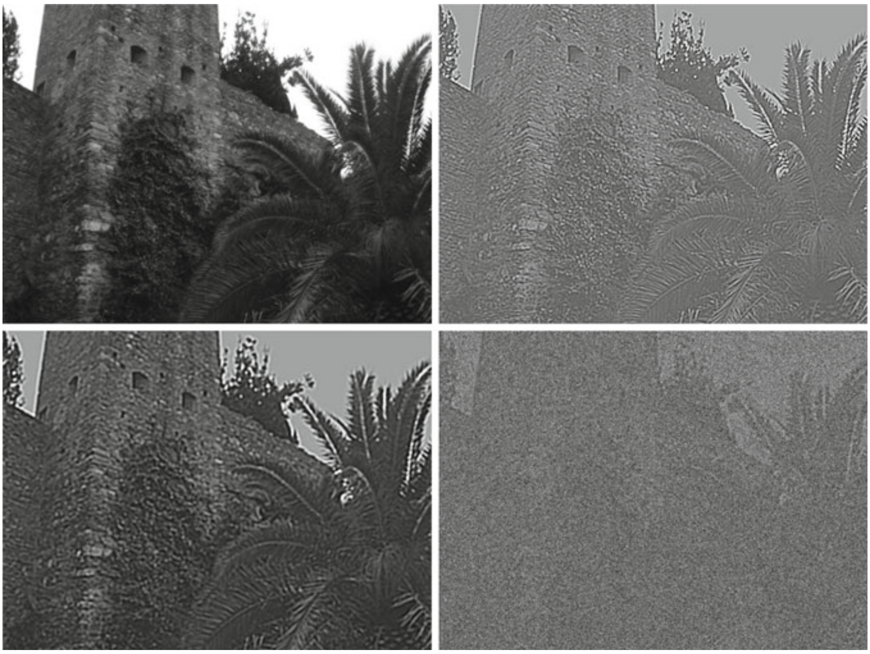


Fig. 1. *Top-left*: Original image. *Top-right*: Image affected by CTF with no envelope decay. *Bottom-left*: Image affected by CTF with envelope decay (the decay acts as a lowpass filter). *Bottom-right*: Image affected by CTF with envelope decay and noise.

returned by the microscope is rather low: first, the overall image contrast is lost by the TEM microscope (1) (see Fig. 1); second, the structural information is severely degraded at high frequencies by aberrations of the microscope; finally, the macromolecule is not isolated in the vacuum but depending on the sample preparation it is surrounded by a thin layer of amorphous ice (cryo-EM) or a thicker layer of carbon and a heavy metal salt (negative staining), which introduces an important background signal that interferes with the signal from the complex under study. Altogether, the Signal-to-Noise Ratio (SNR) is well below 1/10, i.e., there is 10 times more noise than signal.

In this chapter, we introduce the standard image processing workflow needed to produce a volume from a collection of micrographs. The workflow is presented with the use of Xmipp 3.0 software. Note that other software packages may define different workflows, but in any case, the final structure of a complex must be obtained through a user-defined sequence of the standard steps introduced in this chapter.

In short, the standard image processing workflow starts by screening the micrographs to check that they are not astigmatic or drifted and realize the maximum frequency available. Then particles are selected from the micrographs either manually or semiautomatically. Particles are extracted from the micrographs into a gallery and they are again screened to find possible wrongly picked particles. The set of selected particles is aligned and classified attempting to identify possible 2D inhomogeneities. Possible contaminants or alternative structures are removed from the dataset. Next, an initial model is constructed either from the particles themselves or from a priori knowledge about the particle being reconstructed. Finally, the initial model is further refined using the projection images and its resolution estimated. At this level of model refinement it is still possible to have a mixture of different structural populations, and there are methods to sort them into different homogeneous classes.

2. Materials

1. *Software*. SPA image processing is normally performed via software packages like Spider (4), Eman (5), Imagic (6), or Xmipp (7) among others. These software suites allow image processing starting from the raw micrographs and ending at the final reconstructed three-dimensional (3D) structure (see Note 1).
2. *Hardware*. The whole process is rather demanding of computational resources and it is normally performed in computer clusters or supercomputers (cloud computing is an obvious choice for the future but at present it is not in place). The operating system of this kind of computers is always Unix-like, and therefore, Unix is the natural environment for these software packages. An average configuration of Xmipp uses a cluster with 8–16 Gb of RAM memory per node, 8–32 cores per node, and several nodes. Most Xmipp programs scale well up to 128 processors. Beyond this point, inter-process communications and disk access may become a bottleneck, although the optimal performance is rather system dependent and has to be tested on each cluster configuration.

3. Methods

In the following, we describe the mainstream protocols needed to perform a 3D reconstruction starting from the electron micrographs.

3.1. Micrograph Screening

The first step is to check the quality of the collected micrographs. Only high-quality micrographs should progress to further analysis in a high-resolution analysis. For medium-low resolution analysis, one might include not so good micrographs depending on the

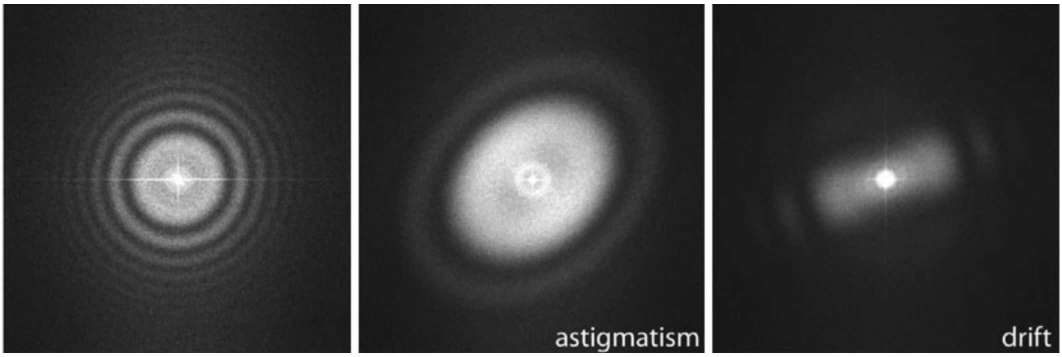


Fig. 2. *Left*: Example of a good micrograph, the PSD has circularly symmetric Thon rings. The presence of many Thon rings is normally associated to the preservation of structural information in high frequency. *Middle*: Example of astigmatic micrograph, Thon rings are elliptical. *Right*: Example of drifted micrograph, Thon rings appear incomplete.

resolution loss that one is willing to tolerate. The Xmipp protocols produce a number of criteria that may help to screen good from bad micrographs.

3.1.1. Identifying Astigmatic and Drifted Micrographs

Good micrographs have a homogeneous background level (without any smooth gradient along the micrograph), are not astigmatic, have no drift, and have high-resolution structural information (visible Thon rings in high frequencies) (see Fig. 2). Astigmatic micrographs could, in principle, be processed. However, in practice they are avoided since most programs cannot track correctly the astigmatism angle through all the sequence of iterative alignments.

Micrographs can be semi-automatically screened by estimating their Power Spectrum Density (PSD) and their Contrast Transfer Function (CTF) (8). The PSD is an estimate of the energy distribution of the micrograph over frequency. Astigmatic micrographs have elliptical rings, while non-astigmatic micrographs have circular rings. Drifted micrographs have masked Thon rings (they do not appear to be complete). The CTF is normally described by a number of parameters, among which the most important are the microscope operating voltage and the defocus (9). However, Xmipp provides a full 2D characterization of the CTF, as well as, the background noise (8), which also plays a role in the accurate determination of the defoci and cannot simply be "filtered out."

In order to correctly estimate the PSD it is important that the digital micrograph does not have empty borders or labels as in Fig. 3.

3.1.2. Determining the Maximum Resolution of the Micrographs

The information content of the micrograph is an important issue to consider. We need to estimate at which resolution the information content of the micrograph fades into the noise (let us call $f_{\max}$ to this frequency). This frequency is characterized by a strong

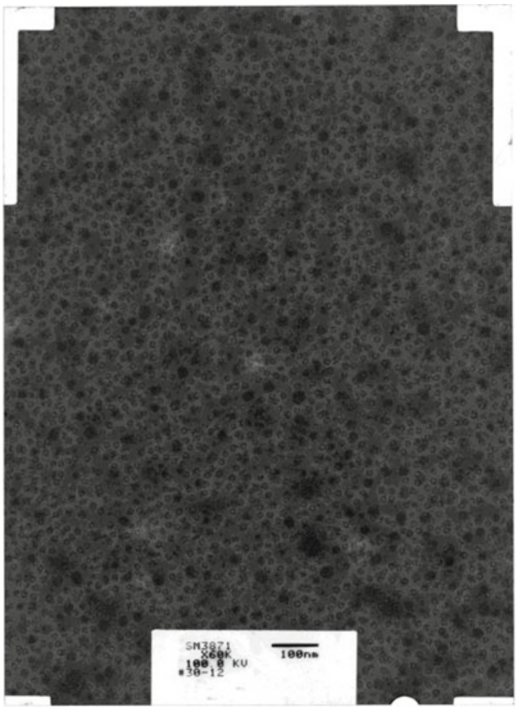


Fig. 3. Example of digitized micrograph with blank areas in the corners and a label for identification. The PSD may not be correctly estimated if these elements are present in the digital micrograph.

decay of the envelope of the CTF (Xmipp reports the frequency at which the CTF envelope drops below 1% of its maximum value). 3D Reconstruction algorithms may recover a few Angstroms more of resolution beyond this frequency, but not many, and the final volume resolution is strongly determined by the maximum resolution visible in the micrographs.

Related to the maximum attainable resolution are the sampling rate and the downsampling factor. The sampling rate relates spatial frequencies measured in 1/Angstroms with digital frequencies measured in cycles/sample (the maximum digital frequency is 0.5 cycles/sample (10)). The relationship is $\omega = fT$, where $\omega$is the digital frequency, *f*is the spatial frequency in 1/Angstroms, and *T*is the sampling rate in Angstroms/pixel. When we downsample the images, we increase the pixel size. Downsampling brings in two important benefits: first, images are obviously smaller with the subsequent gain in disk space and computing time; second, downsampling reduces the noise in the images by cutting out in Fourier space a region where no signal is present.

In principle, we can increase the pixel size by a factor *K*as long as $0.5 \geq f_{\max} KT$ (downsampling by a factor larger than this introduces an effect called aliasing in which high-frequency components may be strongly degraded). In practice, it is normally preferred to have a safety region such that $0.2 \leq f_{\max}KT \leq 0.3$ (this is so to avoid the aliasing of the signal with the noise). For instance, if $f_{\max} = \frac{1}{6}\mathring{A}^{-1}$, and $T = 1\mathring{A}/\text{pixel}$, then we can downsample by a factor $1.2 \leq K \leq 1.8$. Note that the downsampling factor may not be an integer number. In the Xmipp protocols, downsampling is performed in Fourier space (and that is why it can accept non-integer factors). Downsampling in Fourier space provides the best accuracy even for integer downsampling factors since the micrograph PSD is multiplied by a rectangular window (11). On the other side, downsampling by simply averaging neighboring pixels (an operation often called *binning*) introduces strong distortions in the high-frequency components (11).

Considering $\omega_{\max} = f_{\max}T$ (T being the effective sampling rate after downsampling), ideal micrographs have $0.2 \leq \omega_{\max} \leq 0.3$. Micrographs with $\omega_{\max} \leq 0.2$ (also called oversampled) can be safely downsampled since no special gain in resolution will be obtained by so finely sampled images and the noise present may hinder the quality of the final reconstruction. Micrographs with $\omega_{\max} \geq 0.3$ may suffer from aliasing resulting in reconstruction artifacts (they are said to be undersampled).

3.1.3. Criteria to Identify Bad Micrographs Based Only on the Micrograph PSD

1. *PSD correlation at 90°*. The PSD of non-astigmatic micrographs correlates well with itself after rotating the micrograph 90°. This is so because non-astigmatic PSDs are circularly symmetrical, while astigmatic micrographs are elliptically symmetrical. High correlation when rotating 90° is an indicator of non-astigmatism. In Xmipp, this criterion is computed on the enhanced PSD (12).
2. *PSD radial integral*. This criterion reports the integral of the radially symmetrized PSD. This criterion can highlight differences among the background noises of micrographs. This criterion is computed on the enhanced PSD.
3. *PSD variance*. The PSD is actually estimated by averaging different PSD local estimates in small regions of the micrograph. This criterion measures the variance of the different PSD local estimates. Untilted micrographs have equal defoci all over the micrograph, and therefore, the variance is due only to noise. However, tilted micrographs have an increased PSD variance since different regions of the micrograph have different defoci. Low variance of the PSD is indicative of non-tilted micrographs.
4. *PSD Principal Component Variance*. When considering the local PSDs previously defined as vectors in a multidimensional space, we can compute the variance of their projection onto the first principal component axis. Low variance of this projection is indicative of a uniformity of local PSDs, for example, this is another measure of the presence of tilt in the micrograph.

5. *PSD PCA Runs test*. When computing the projections onto the first principal component, as discussed in the previous criterion, one should expect that the sign of the projection is random for untilted micrographs. Micrographs with a marked nonrandom pattern of projections are indicative of tilted micrographs. The larger the value of this criterion, the less random the pattern is.

3.1.4. Criteria to Identify Incorrectly Fitted CTFs

1. *Fitting score*. The CTF is computed by fitting a theoretical model to the experimentally observed PSD. This criterion is the fitting score. Smaller scores correspond to better fits. A complete description of Xmipp fitting score is given at Sorzano, Jonic (8).
2. *Fitting correlation between the first and third zeroes*. The region between the first and third zeroes is particularly important since it is where the Thon rings are most visible. This criterion reports the correlation between the experimental and theoretical PSDs within this region. High correlations indicate good fits.
3. *First zero disagreement*. If the CTF has been estimated by two different methods (normally Xmipp and Ctffind (13)), then this criterion measures the average disagreement in Angstroms between the first zero in the two estimates. Low disagreements are indicative of a correct fit.

3.1.5. Criteria to Identify Bad Micrographs Based on the Fitted CTF

1. *Damping*. This is the envelope value at the border of the PSD. Micrographs with a high envelope value at border are either wrongly estimated or strongly undersampled.
2. *First zero average*. This is the average in Angstroms of the first zero over all possible directions. Normally, this value should be between 4 and 10 times the effective sampling rate in Angstroms/pixel.
3. *First zero ratio*. This measures the astigmatism of the CTF by computing the ratio between the largest and smallest axes of the first zero ellipse. Ratios close to 1 indicate no astigmatism.

3.2. Particle Picking

The next step is to pick particles from those micrographs passing the previous screening step. Particle picking can be performed manually or semiautomatically. Manual particle picking simply lets the user to choose those particles he is interested in from the electron micrographs. The ease of picking particles depends on the particle size, image contrast, contamination level, etc. The task can be facilitated by displaying the micrograph with a moderate zoom-out factor.

Image filters can also help to better visualize the particles of interest. In Xmipp we provide the following filters: color dithering, bandpass filter, anisotropic diffusion, mean shift, background subtraction, and contrast/brightness enhancement. Particularly useful are color dithering and bandpass filtering. Color dithering (14) is an algorithm that reduces differences between adjacent gray values. Bandpass filter (15) filters out gray-scale variations that are either too small or too large with respect to the size of the particle we are interested in.

3.2.1. Manual Particle Picking

Manual picking is sometimes criticized for biasing the dataset towards those shapes that the user better recognize or has in mind as the possible projections of her structure. In order to avoid this bias, automatic or semiautomatic particle picking algorithms are sometimes preferred as an objective way of choosing particles. At the same time, the particle picking process, which in general is a tedious and time-consuming task, is computationally assisted and accelerated. Algorithmically chosen datasets contain a nonnegligible amount of wrongly picked particles (contaminants, true particles on carbon, particle conglomerates, etc.). These datasets have to be carefully scrutinized to remove wrongly picked particles from the onset. This can be done by a manual revision of the automatic picking results, and/or by classifying in 2D the selected particles and eliminating those classes corresponding to wrongly picked particles.

3.2.2. Semiautomatic Particle Picking

Xmipp allows semiautomatic particle picking (16). The algorithm has been designed to keep a low false-positive rate (FPR), i.e., picking as few wrongly picked particles as possible. In order to keep the FPR as low as possible, the algorithm must be trained by the user in the kind of particles he is interested in. The first step of the training requires the user to manually pick about 100–200 particles. Then, the algorithm learns a set of features describing the particles being picked. In the next step, the algorithm tries to automatically select particles from a micrograph that has not been manually processed. The first attempt will pick a number of true particles, but many other true particles may be left. The user is required to pick those "unseen" particles. He is also required to correct the algorithm by removing those particles that have been wrongly picked automatically. This correction information helps the algorithm to distinguish true particles from other objects looking alike. This process of the algorithm trying to correctly pick particles and the user correcting for errors is repeated on more micrographs until the user is satisfied with the algorithm results. At each step, the algorithm learns from its errors and next time it will try to better distinguish between particles and nonparticles. No automatic particle picking algorithm is absolutely infallible. As a rule of thumb, they can be applied when the imaging conditions are not specially challenging.

3.3. Particle Extraction, Screening, and Preprocessing

The next step is to extract the particles from the micrographs and form a stack of projections of our particle. This stack of projections is further analyzed in 2D or 3D in subsequent steps. Projection screening helps to identify projections that are not typical (in the statistical sense). Nontypical particles may correspond to underrepresented projection directions or states of our particle, but also to wrongly picked particles and outliers. If an automatic particle picking algorithm has been used, nontypical particles normally correspond to wrongly picked particles, such as particles on edges. Finally, particle preprocessing helps to highlight or concentrate specific features of our dataset.

3.3.1. Particle Extraction

When extracting the particle projections, there are a number of actions we can take:

1. *Phase flipping*. Use the information from the estimated CTF to compensate for phase reversals introduced by the microscope at different frequencies. This phase correction is better performed on the micrographs (not the projections), since we have enough information to perform a deconvolution.
2. *Taking logarithm*. Depending on your acquisition system you may have to take the logarithm of the data in order to have a linear relationship between the gray values in the image and those in the volume.
3. *Contrast inversion*. Most image processing algorithms expect to see the particle as a white object over a dark background. However, some imaging conditions produce just the opposite. At this moment, you may invert the contrast if your particles are black over a white background.
4. *Normalization*. The same projection in different micrographs may have different gray values. Even within the same micrograph there might be a light gradient causing the gray values to be different. In order to eliminate a local gradient, a ramp in the gray values is fitted for each projection image and then subtracted from the image. Then, the image values are linearly transformed so that in the background there is zero mean and standard deviation equal to one. Noise statistics should be similar in all projections after this normalization step.
5. *Dust removal*. Sometimes dust, hot or cold spots can be seen in a projection. These pixels are identified by noting that their gray values are normally far from the mean of the rest of the image. You should choose to fill these pixels with a random value from a Gaussian with zero-mean and unity-standard deviation.

3.3.2. Particle Screening

Automatic picking algorithms have a nonnegligible FPR (i.e., they pick locations in the micrograph that do not actually correspond to true particles). Particle screening is a successful way of identifying them.

For each image, we calculate the gray values histogram, square the gray values and compute the radial average of these squared values. The gray histogram and the radial average are stacked into a multivariate vector associated to each projection. Now, we perform a PCA analysis of the whole set of projections and project the multivariate vector onto the PCA space spanned by the first two eigenvectors. Then, we analyze the multivariate normality of these projections. The normality is measured as the Mahalanobis distance of the PCA projection to the space origin. Typical projections have a small distance, while nontypical projections have a larger distance. The dataset is sorted according to their distance (called *z*-score). Normally, wrongly picked particles, particles on edges, contaminants, etc., have a large distance and are sorted to the end of the list. At this point, the user may discard the particles he considers that do not correspond to the structure under study, or that for some reason do not follow the general trend.

3.3.3. Particle Preprocessing

Sometimes, we may want to apply an image processing filter to our images in order to reduce noise, highlight certain features, mask out the background, etc. In Xmipp protocols the following choices are available.

1. *Scaling*. Change the size of the projection images. Normally a size reduction is performed so that image processing is faster. The amount of noise is also reduced.
2. *Fourier lowpass, highpass, and bandpass filtering*. Low frequencies correspond to slow variations in the image; too high frequencies normally come from noise or very fine details. You may use these filters to remove any frequency band that you are not interested in. Cutoff frequencies are normalized to 0.5 (they are called digital frequencies). The digital frequency of an object whose diameter is D pixels is $\omega = \frac{1}{D}$. The Xmipp protocols offer a wizard that allows interactive selection of the cutoff frequencies so that we may preview the effect of the filter in our images.
3. *Fourier Gaussian filtering*. We may implement multiresolution approaches by a series of Gaussian filters in Fourier space. Coarse representations are obtained by multiplying in Fourier space by a Gaussian of small bandwidth. Finer representations are obtained by enlarging the bandwidth of the Gaussian filter.
4. *Dust removal*. Similar to the dust removal option in the previous section.
5. *Normalization*. Besides the noise normalization with background subtraction described in the previous section (the recommended option), Xmipp also offers to normalize the images to have zero mean and unitary standard deviation, or to normalizing without background subtraction.
6. *Masking*. Apply a mask to the images to concentrate the analysis on a specific region. A wizard helps to choose among a wide variety of masks.

3.4. 2D Analysis

Two-dimensional analysis is the most common way of getting acquainted with the structural information contained in the projection images. In a way, it is an Exploratory Data Analysis (EDA) that is very much used in other data analysis contexts. The whole set of images is classified into different groups as homogenous as possible. Inside each group, projections are aligned and (normally) averaged producing a class representative. The class representative has much less noise (thanks to the averaging or equivalent operation) than the raw projections and allows a better visual identification of the structural features. The price for this improved visualization is the possible blurring introduced by averaging nonidentical projection images. If the group is small enough (between 50 and 150 projections), then this blurring effect is minimized.

2D analysis allows identifying wrongly picked particles, contaminants if they are sufficiently dissimilar from the structure under study, particle aggregations, damaged particles, etc. We should select among the classes all those that are likely to come from the structure we are studying in order to choose a projection population as homogeneous as possible in terms of biochemical species and conformational state. The rest of the classes may be interesting by themselves and worthy of further analysis, although they should not be mixed with the "good" classes in order to avoid contamination in the 3D reconstruction process. This class selection process is tricky in the sense that the user might bias the final 3D reconstruction by removing projections that a priori do not fit with his preconceived idea about the structure being reconstructed. The boundary between removal of contaminants, damaged particles, and wrongly picked particles and the removal of valid projection images is a fuzzy, fine line difficult to characterize.

Class representatives can also be used as input images to a common-lines algorithm (17, 18). This kind of algorithms addresses the initial reference problem. Most 3D reconstruction algorithms need an initial volume as starting point. The initial volume must share some general features with the structure under study. When such a starting volume is not available, common-lines algorithms are capable of producing one by identifying common lines in Fourier space. This identification is rather error prone and sensitive to noise. That is why class representatives, with less noise, are particularly well suited as input for these algorithms.

Xmipp offers several possibilities to perform this 2D EDA, which are briefly described below.

3.4.1. CL2D (Clustering in 2D) (19)

The algorithm performs alignment and classification of the input images. It allows the identification of a large number of classes (the larger the number of classes, the fewer particles will take part in each class; there must be a balance between avoiding blurring by averaging dissimilar projections and having sufficient images in each class to significantly reduce the noise and allowing better visualization). First, the algorithm split the data in a small number of classes that are subsequently subdivided into more and more classes until the desired number of classes is reached. Images are allowed to change class at any moment. This divisive structure has been proved to yield more robust results than directly starting with a large number of classes. The algorithm requires relatively few parameters (the most important is the final number of classes). We can also control how to measure the distance between a raw projection and the class representative (correlation or correntropy (20)) and how to decide to which class a raw projection belongs (classical multireference assignment or robust assignment (19)). Two setups are normally used: (1) correlation and classical multireference assignment, which normally works well with good quality images and it is standard in other software packages; (2) correntropy and robust assignment, which works well with good and bad quality images and it is particular to Xmipp. The output classes are sorted so that one image is similar to the next. Additionally, thanks to the hierarchical nature of the algorithm, we can compute the core of any class (we define the core of the class as the subset of images that were always together in the classification hierarchy). We further refine the core by eliminating outliers in the PCA space spanned by the first two eigenvectors (in the same way as in the particle screening). We refer to these refined cores as stable cores.

3.4.2. ML2D (Maximum Likelihood in 2D) (21)

Maximum Likelihood is a different classification framework. Instead of assigning each projection image to a single class, it is assigned to all classes with different probabilities. Classes are then updated with the weighted average of all particles (weighted by probability). The algorithm solves at the same time the alignment and classification problems and it is well suited to work with a relatively small number of classes (less than 20–30). The algorithm can be applied in Fourier space instead of real-space. Fourier space has the advantage of letting noise be correlated, which is actually the case in Electron Microscopy due to the Contrast Transfer Function (CTF). Another advantage of Maximum Likelihood in Fourier space is that the CTF needs not be estimated.

3.4.3. KerDenSOM (Kernel Density Estimation Self-Organizing Map) (22)

This algorithm projects the input images onto an output space of class representatives. Nearby classes are similar to each other (as in the output of CL2D but in a 2D topology). The algorithm does not align images to classes, so that they have to be previously aligned. The algorithm is useful to perform a careful analysis of a given region in a set of images that may be the output of a previous classification (CL2D or ML2D) or alignment.

3.4.4. Rotational Spectra (23)

The rotational spectrum of an image measures its rotational symmetry. A twofold symmetric image can be rotated 180° (=360/2) and we obtain the same image. A threefold symmetric image can be rotated 120° (=360/3) and 240° (=2 × 360/3) and we obtain again the same image. The rotational spectrum measures the strength of each symmetry in a particular image. The rotational spectrum is rotationally invariant (i.e., if we rotate the image by any angle, the rotational spectrum remains the same), but it is not translationally invariant (i.e., if we shift the image, the rotational spectrum changes). For this reason, it is important to have all the images translationally aligned, and compute the rotational spectrum with rotations around the image center. The analysis of the rotational spectrum may reveal heterogeneities in particles whose projections are relatively symmetric (24) in a particular projection direction. The most important parameters of this step are the two radii between which the symmetry is analyzed (for instance, a ring-shaped projection can have symmetry only between the two radii where the ring is inscribed).

3.5. 3D Analysis

The two main steps in 3D are the analysis of homogeneous populations (all projections belong to a single structure) and heterogeneous populations (in the dataset there are projections coming from different structures).

The homogenous population analysis is normally referred to as 3D model refinement. Starting from an initial guess of the structure being reconstructed, a workflow of image processing steps is taken to refine this initial guess. The initial guess may be obtained by the same protein in a slightly different conformation, from a similar protein, from an atomic model, from random conical tilt, from a common-lines algorithm, etc. The initial reference chosen may bias the final result, and it is customary to strongly filter the initial volume so that all details are removed, and only the general shape at very low resolution remains.

The heterogeneous population analysis can be seen as the simultaneous refinement of several volumes with homogeneous population. The model refinement problem is coupled to the problem of classifying the input raw projections into different classes (the homogeneous populations).

3.5.1. 3D Model Refinement

The most basic input of a model refinement algorithm is a starting volume (normally at low or very low resolution) and a set of projections supposedly from a single 3D structure and ideally covering all possible projection directions. The starting volume is projected in all possible directions (reference projections) and the experimental images are compared (after alignment) with all of them. The projection direction of the best matching reference projection is assigned to the corresponding experimental image and the in-plane alignment parameters (in-plane rotation and shift) are annotated. These alignment parameters are then used by a 3D reconstruction algorithm to produce a better estimate of the 3D structure of the macromolecular complex. This process is iterated till convergence or a fixed number of iterations have been performed.

The most important parameter of this algorithm is the plan for 3D alignment search. To align each experimental projection with respect to the current estimate of the volume being reconstructed, we need 5 parameters (3 Euler angles defining the projection direction and the in-plane rotation, and 2 shifts defining an in-plane displacement). We can perform a truly 5D search for these parameters (trying all possible combinations of Euler angles and shifts within a bounded range) or perform a 3D + 2D parameter search (looking for the best projection direction keeping the shift fixed, and then looking for the best shift using the projection direction just found). A full 5D search provides more accurate results at the cost of a significant increase of computation time. We can use both strategies during the process: it is custom to start with a 5D search for the first few (23, 25) iterations and then switch to a 3D + 2D strategy. The bounded search regions for each parameter (angles and shifts) are also normally diminished in size so that the parameter search is performed in a narrower area (this helps the process to converge while saving time by not looking for the parameters in areas that are rather unlikely to produce a good match). An exhaustive search is performed within the bounded region using a sampling step. As the region becomes narrower, the sampling step is also diminished so that the alignment is more accurate. Typical search ranges for the first iterations allow any angle and shift in the first 4–5 iterations with an angular sampling of 10° (the shift sampling is not needed since the shift is searched in Fourier space by making use of its correlation property, which allows for a continuous shift search). From the fourth to fifth iteration, the angular range is reduced to about twice the angular sampling, and this is progressively diminished towards a sensible value (the smallest identifiable angle is given by $\arctan\frac{1}{R}$, where R is the radius of the object being reconstructed in pixels).

During the alignment and reconstruction process we may also correct for the amplitude effects of the CTF. Xmipp does so by grouping the input projections according to their defocus values. Those images belonging to the same defocus group are reconstructed together, and then the different volumes are merged in 3D by using a Wiener filter (26). Alternatively, a B-factor correction can be employed (25).

Another important issue is how to generate the reference volume for the next iteration from the current reconstruction estimate. Using the raw estimate may produce overfitting since the very high frequencies of the volume may be fitting mostly noise components; this is especially true during the first iterations since the angular sampling is still large and the reconstruction estimate is rather rough. The raw estimate can be masked in real space (using a fixed radius or a user-defined mask) and Fourier space (by lowpass filtering the volume to a certain frequency; normally this cutoff frequency is gradually increased so that the alignment uses a progressively improved reference volume).

3.5.2. 3D Heterogeneity Analysis

In case we suspect that our population of projections may come from different 3D structures (different conformations, different oligomeric states, different binding states, etc.) we can simultaneously solve the 3D reconstruction and classification problem. The idea is to sort the input projections into different classes so that each class is homogenous, and then a 3D reconstruction is performed within each class. This is the standard idea behind multireference 3D reconstruction. In Xmipp we have extended the Maximum Likelihood framework to 3D (each experimental projection belongs to all classes and all projection directions with different probabilities; then, the volumes representing each class are updated considering the relative weights of each projection) (27). The algorithm is rather time consuming and, because of this, its applicability is limited to relatively large angular samplings (in the order 5–10° depending on the image size and number of images). At the end of the algorithm, we may assign each experimental image to the class with maximum likelihood. Then, within each class we can run a model refinement (see previous Section) in order to improve the resolution of the class model.

3.6. Example 3D Reconstruction

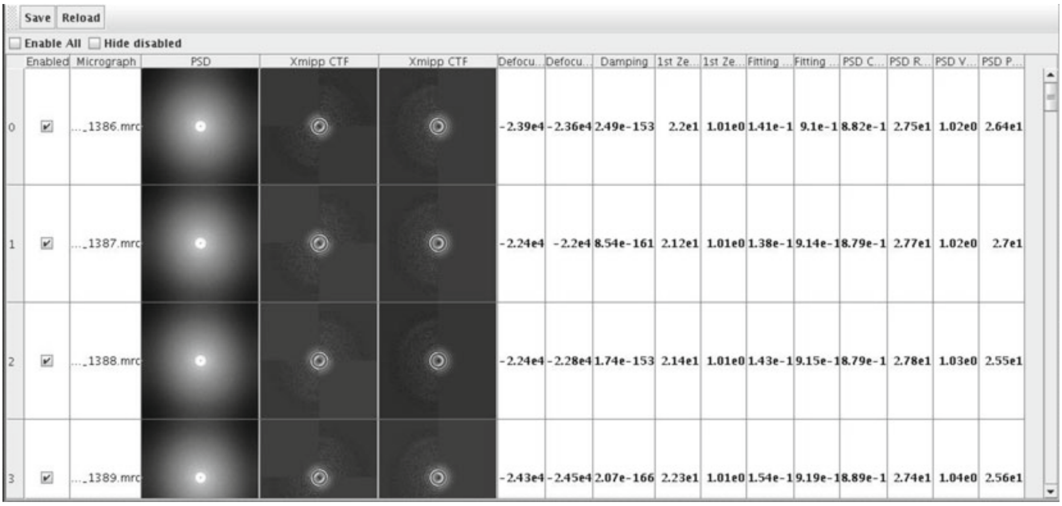
We present the protocols results as applied to the Bovine Papilloma virus images published in (28) and kindly provided by Drs. Wolf and Grigorieff for this book chapter. The dataset consists of 49 micrographs of an approximate size of 10,000 × 10,000 pixels. The sampling rate is 1.237 Å/pixel, the microscope voltage 300 kV, and the nominal magnification ×56,588. At this magnification, the projection of a single virus fits in an image of size 512 × 512. The execution time for the different protocol steps are reported for the parallel run of Xmipp on Intel Xeon 2.666 GHz processors. The processors belong to a cluster with 28 nodes, 8 processors, and 16 GB of RAM per node.

3.6.1. Micrograph Screening

Estimating the CTF and evaluating the quality of all micrographs took 15 min for the BPV dataset in 32 processors. The estimated micrograph defocus ranged between 1.8 and 2.3 μm and no micrograph was removed for having an astigmatic or drifted CTF (Fig. 4 shows the typical output of the Xmipp protocols for this step). The maximum resolution present in the micrographs is about 6 Å (the CTF envelope drops below 1% its maximum value, which implies a loss in image power by a factor 0.01^2).

3.6.2. Particle Picking

Semiautomatic particle picking was used. We manually trained the algorithm with the first 4 micrographs. 238 virus projections were manually picked among the 4 micrographs (the automatic picking



	Enabled	Micrograph	PSD	Xmipp CTF	Xmipp CTF	Defocu...	Defocu...	Damping	1st Ze...	1st Ze...	Fitting ...	Fitting ...	PSD C...	PSD R...	PSD V...	PSD P...
0	☑	..._1386.mrc				-2.39e4	-2.36e4	2.49e-153	2.2e1	1.01e0	1.41e-1	9.1e-1	8.82e-1	2.75e1	1.02e0	2.64e1
1	☑	..._1387.mrc				-2.24e4	-2.2e4	8.54e-161	2.12e1	1.01e0	1.38e-1	9.14e-1	8.79e-1	2.77e1	1.02e0	2.7e1
2	☑	..._1388.mrc				-2.24e4	-2.28e4	1.74e-153	2.14e1	1.01e0	1.43e-1	9.15e-1	8.79e-1	2.78e1	1.03e0	2.55e1
3	☑	..._1389.mrc				-2.43e4	-2.45e4	2.07e-166	2.23e1	1.01e0	1.54e-1	9.19e-1	8.89e-1	2.74e1	1.04e0	2.56e1

Fig. 4. Display of the micrograph screening results. Each row shows the PSD, fitted CTF, and different quality criteria for a different micrograph. The table can be sorted by any of the quality criteria and bad micrographs can be manually deselected so that they are no longer considered in the analysis.

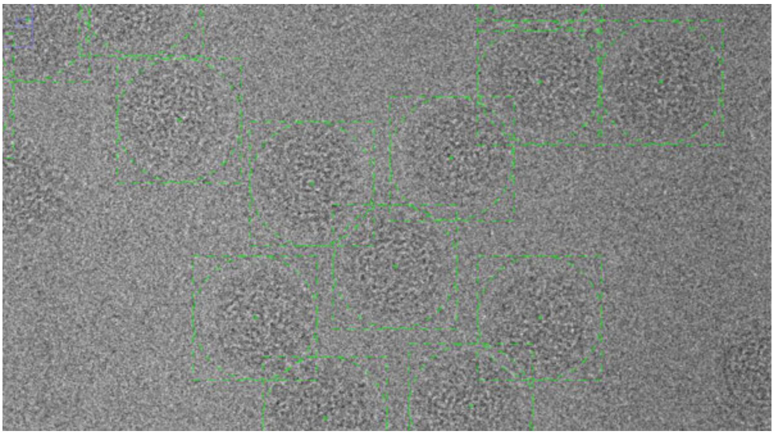


Fig. 5. Sample region of a micrograph with virus particles automatically identified.

algorithm was used after the first 153 manually picked particles). 6,475 particles were automatically selected in 50 min from 45 micrographs (an average of 130 particles/min) (see Fig. 5). Among the 6,475 particles, 1,158 were false-positives (mostly they were true virus particles on the carbon film instead of the grid holes), which amounts to 17.9% of the selected particles. This kind of false-positives was easily removed during the revision of the protocol output. A total of 5,317 (=6,475–1,158) particles were correctly identified. 410 additional particles were manually selected (they corresponded to false-negatives). In total 5,727 (=410 + 5,317) particles were selected during the automatic phase, 92.8% of them were automatically selected. Thus, we see an advantage in the manual workload by using a semisupervised approach to particle picking. Adding the 238 particles manually selected during the training phase then gave a total of 5,965 particles in the dataset.

3.6.3. Particle Extraction and Screening

The selected particles were corrected for the CTF phase flip and screened in search of outliers using the multivariate procedure described in Subheading 3.3. The process took 3 min in a single processor. All particles whose PCA distance was larger than 7 were discarded (245 particles out of 5,965, 4.1%; the visual appearance of these particles was not particularly different from the other images in the dataset, but for some reason they did not follow the general trend). Discarded particles mostly corresponded to a couple of micrographs that had a relatively lower correlation between the first and third zero of the CTF. For some reason the CTF in these micrographs was more difficult to estimate, and they also resulted in particles whose projections into the PCA space were outliers. It is worth noting that these two quality measures are completely independent so that both seem to indicate some strange behavior of the discarded micrographs. After the screening, the dataset was composed of 5,720 particle projections.

3.6.4. 2D Analysis

In order to explore the acquired projections, we reduced the image size from 512 × 512 to 128 × 128 (3 min in 8 processors). We then applied a CL2D analysis with 64 classes (6 h in 32 processors). Figure 6 shows the first 24 class averages. The CL2D iterations converged rather clearly, with only 1.7% of the images changing from one class to the other after 15 iterations (this is normally a sign of finding good classes). In the set of class averages, we could not identify any class that did not correspond to true virus projections. Therefore, we assume that the set of experimental projections corresponds to a single, homogenous population.

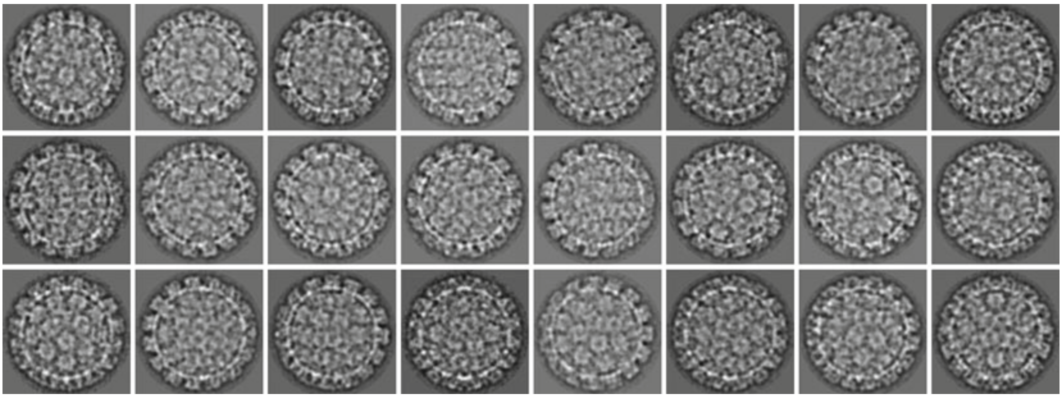


Fig. 6. First 24 classes of the 64 classes calculated by CL2D.

3.6.5. Initial Volume

To construct the initial volume, we downloaded the three-dimensional structure of the Bluetongue virus capsid from the Electron Microscopy Data Bank (http://www.ebi.ac.uk/pdbe/emdb, entry code: 5147, Zhang, Jin (29)). This virus is totally unrelated to the Bovine Papilloma virus, and they only share the fact that both have an icosahedral capsid. The volume at the EMDB has an alternate 222 symmetry (twofold symmetry axes on X, Y and Z). We filtered the Bluetongue capsid to 100 Å, and scaled the capsid to fit in a box of size 128 × 128 × 128. We used this lowpass filtered volume as a reference for the alignment and 3D reconstruction of the class averages found by CL2D. We used seven iterations of the 3D Model Refinement pipeline described in Subheading 3.5. The angular sampling was 10° for the first four iterations, 5° in the next two, and 3° in the last one. Images were allowed to freely move in the projection sphere in the first four iterations, and then their Euler angles were restricted to a maximum change of 10° in the next two iterations, and 6° in the last iteration. The first four iterations performed a fully 5D parameter search, while the rest used a 3D + 2D search. CTF was not corrected. The whole process took 3 min in 8 processors, at the end of which we had already an initial volume that we rescaled (30 s in a single processor) to a size of 512 × 512 × 512. Figure 7 shows the Bluetongue initial structure and its refinement using the CL2D classes. This refined volume will serve as starting volume for the 3D data analysis of the Bovine Papilloma virus.

3.6.6. 3D Model Refinement

At this point we entered into a 3D Model refinement process starting from the initial volume constructed in the previous step. We performed 16 iterations of the process. In the first four iterations we performed a fully 5D parameter search allowing the angles to take any value in the projection sphere. The rest of iterations were

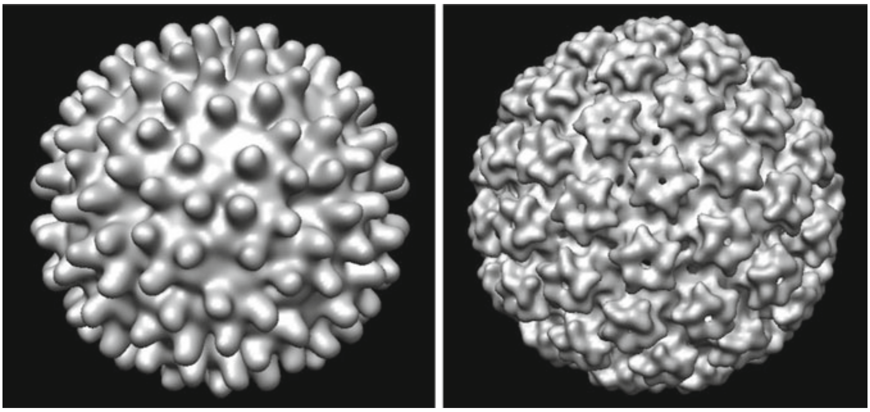


Fig. 7. *Left*: Bluetongue capsid filtered at 100 Å. *Right*: Volume obtained after refining the Bluetongue capsid using the CL2D classes. This example shows the robustness of the 3D reconstruction process in this specific case to the initial volume.

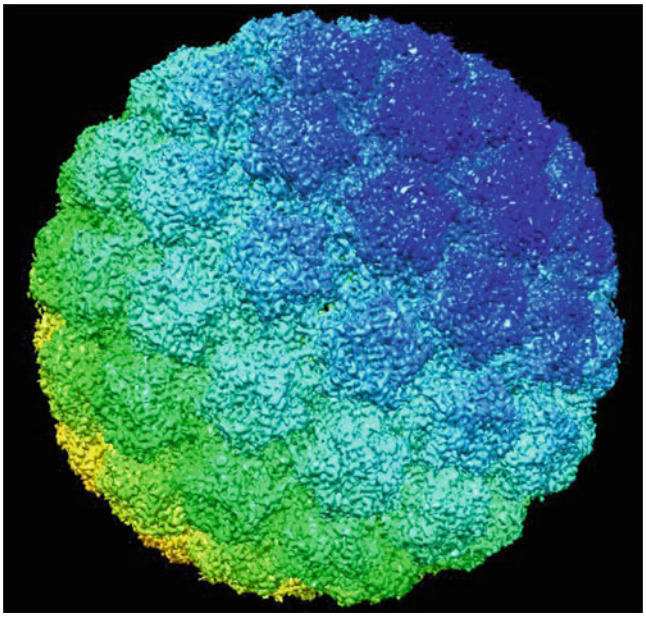


Fig. 8. Final reconstruction of the Bovine Papilloma virus at 5.75 Å.

performed with a 3D + 2D strategy. The angular sampling was 10° in the first four iterations, 5° in the next two, 3° in the following two, 2° in the following two, 1° in the next three, and 0.5° in the last three iterations. In the first four iterations the projection directions could take any value in the projection sphere, in the next two they could move up to a maximum of 30°, in the next two up to a maximum of 15°, in the next two up to a maximum of 10°, and the rest of iterations up to a maximum of 6°.

Starting from iteration 12, we started correcting for the CTF amplitude effects with a Wiener filter (since projections were phase corrected when extracted, the reconstruction up to iteration 11 is phase corrected; the angular step at that level was 2°). The final result is shown in Fig. 8. The resolution achieved in iteration 16 is 5.75 Å (at FSC = 0.5) which is in agreement with the fact that the maximum visible frequency in the micrographs is about 6 Å, and it confirms our hypothesis that the set of images belonged to a single structural population (or at least structural differences are in details finer than those recorded by the microscope). Wolf, Garcea (28) report a resolution of 4.9 Å which is further improved to 3.6 Å by averaging the subunits. The difference from 5.75 to 4.9 Å could be explained by a regularization internal to the 3D reconstruction algorithm used by Wolf, Garcea (28), overfitting or any other reason that increases the consistency of the reconstruction between the two halves of the dataset.

Alternatively we could have performed the amplitude correction by taking the phase corrected volume (Iteration 11), and

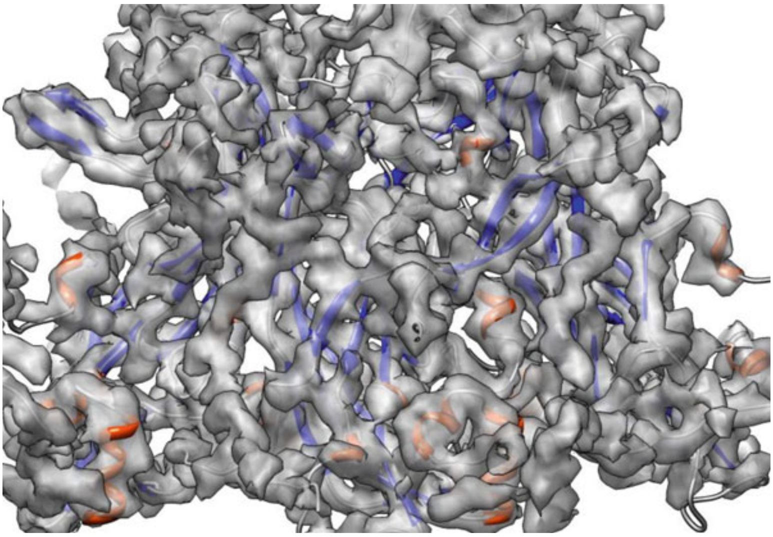


Fig. 9. Overview of the fitting between the EM reconstruction and its atomic model.

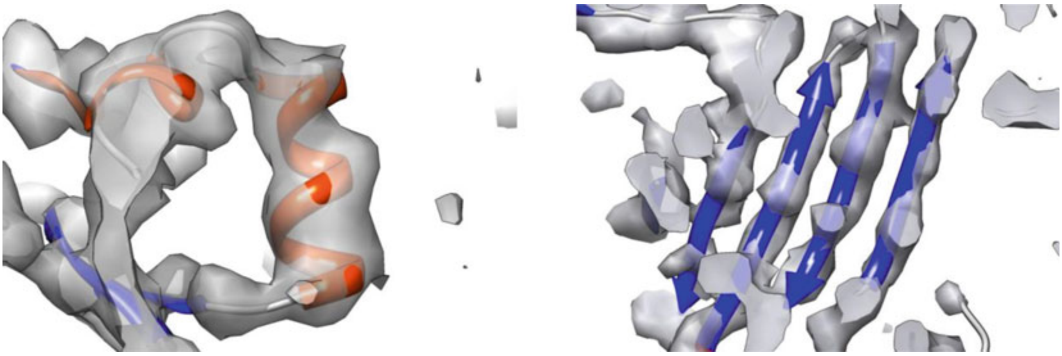


Fig. 10. *Left*: Detail of an α-helix. *Right*: Detail of a β-strand.

applying a B-factor correction (25). This takes 2 min in a single processor, and in this particular case it produced a superior result to the Wiener filter (α-helices were better defined). The resolution of the amplitude corrected volume with the B-factor cannot be measured through the FSC since this measure is insensitive to multiplicative factors.

Wolf, Garcea (28) created an atomic model that was fitted to their EM volume. The atomic model is at the Protein Data Bank (http://www.pdb.org) under the access code 3IYJ. Figures 9, 10, and 11 show some detail of that atomic model fitted to our EM reconstruction. The secondary structure elements (α-helices and β-sheets) can be clearly seen as well as some side chains.

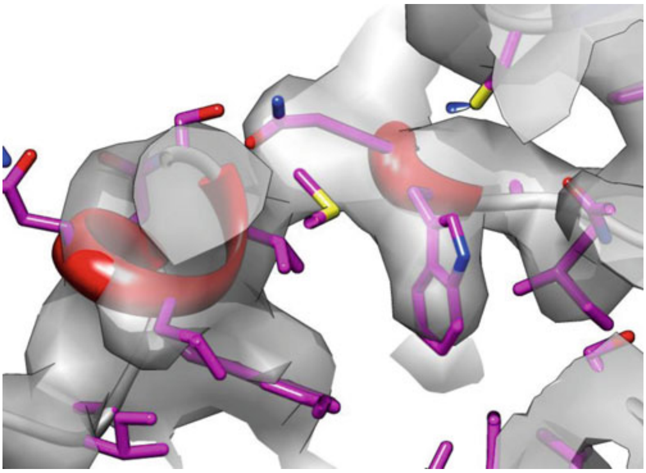


Fig. 11. Detail of the fitting of some side chains.

Table 1
Computing times for each process using a variable number of processors (see text). Each processor is core of an Intel Xeon 2.666 GHz

Process	Time
Micrograph screening	15 min
Semi-automatic particle picking	50 min (+1 h manual training and revision)
Particle screening	3 min
Particle downscaling	3 min
CL2D	6 h
Initial volume construction	3 min
3D Model refinement Iterations 1–11 (phase corrected)	13 h
Amplitude correction via B-factor	2 min
3D Model refinement Iterations 12–16 (amplitude corrected)	18 h

3.6.7. Execution Time

Table 1 shows the time spent at each of the different steps. In 9 h, we screened the micrographs, selected the particles, screened them, checked that there were no contaminants, and constructed an initial model for the 3D Model refinement. The next 13 h were used to construct a phase corrected volume. Finally, 18 h were needed to construct an amplitude corrected volume with a very fine angular step in the projection matching. In parallel to this last step, a second amplitude corrected volume was computed by B-factor correction in only 2 min. These times were obtained by using at most 32 processors of a cluster, which is a reasonable number of processors available in standard clusters.

3.7. Conclusions

In this chapter we have introduced the basic protocols used in 3D Electron Microscopy for Single Particles and illustrated them as they are applied in Xmipp. These protocols cover all the way from the micrographs to the final 3D reconstructed volume. We have applied these protocols to images of the Bovine Papilloma virus capsid. This dataset was rather homogeneous and no heterogeneity analysis was needed. Obtaining a 3D map with a resolution that allows the identification of α-helices (below 5.75 Å) took less than 24 h in a cluster with 32 processors available. The presence of heterogeneity in the sample complicates the 3D analysis, specially its validation. However, as shown by this example, 3DEM is advancing towards high-throughput, high-resolution protocols.

4. Note

1. All packages have been written with a command-line interface (i.e., the user issue commands from a shell, each one performing a small task). The whole image processing task consists of the concatenation of many small tasks that start at the micrograph level and finish with the 3D reconstruction. However, all packages now offer a graphical interface that provides a more user-friendly interface. In some cases, the different programs have been bundled in a few protocols (30). Protocols reflect the typical image processing pipeline and are suited for most users, especially novice ones. Despite the simplification brought by the protocols, packages still offer their full functionalities so that experienced users may diverge from the most typical image processing path.

Acknowledgements

C.O.S. Sorzano is a recipient of a Ramón y Cajal fellowship awarded by the Spanish Ministry of Science and innovation (Micinn). J. Vargas and J. Otón are Juan de la Cierva fellows (Micinn). This work has been partly funded by grants CSD2006-0002 (Spanish Ministry of Education), ACI2009-10220 (Micinn), BIO2010-16566 (Micinn), and PIF08-020-2 (CSIC).

References

1. Frank J (2006) Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state. Oxford University Press, New York
2. Sorzano COS, Jonic S, Cottevieille M et al (2007) 3D electron microscopy of biological nanomachines: principles and applications. Eur Biophys J 36:995–1013
3. Wang L, Sigworth FJ (2006) Cryo-EM and single particles. Physiology 21:13–18
4. Frank J, Radermacher M, Penczek P et al (1996) SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J Struct Biol 116: 190–199
5. Ludtke SJ, Baldwin PR, Chiu W (1999) EMAN: semiautomated software for high-resolution single-particle reconstructions. J Struct Biol 128:82–97
6. van Heel M, Harauz G, Orlova EV et al (1996) A new generation of the IMAGIC image processing system. J Struct Biol 116:17–24
7. Sorzano COS, Marabini R, Velázquez-Muriel J et al (2004) XMIPP: a new generation of an open-source image processing package for electron microscopy. J Struct Biol 148:194–204
8. Sorzano COS, Jonic S, Núñez-Ramírez R et al (2007) Fast, robust and accurate determination of transmission electron microscopy contrast transfer function. J Struct Biol 160:249–262
9. Sorzano COS, Otero A, Olmos EM, Carazo JM (2009) Error analysis in the determination of the electron microscopical contrast transfer function parameters from experimental power spectra. BMC Struct Biol 9:18
10. Oppenheim AV, Schafer RW, Buck JR (1999) Discrete-time signal processing. Prentice-Hall, Upper Saddle River
11. Sorzano COS, Iriarte-Ruiz A, Marabini R, Carazo JM (2009) Effects of the downsampling scheme on three-dimensional electron microscopy of single particles. In: Proc. IEEE workshop Intell Signal Proc. Budapest, Hungary
12. Jonic S, Sorzano COS, Cottevieille M et al (2007) A novel method for improvement of visualization of power spectra for sorting cryo-electron micrographs and their local areas. J Struct Biol 157:156–167
13. Mindell JA, Grigorieff N (2003) Accurate determination of local defocus and specimen tilt in electron microscopy. J Struct Biol 142:334–347
14. Akarun L, Yardunci Y, Cetin AE (1997) Adaptive methods for dithering color images. IEEE Trans Image Process 6:950–955
15. Bracewell RN (2006) Fourier analysis and imaging. Springer, New York
16. Sorzano COS, Recarte E, Alcorlo M et al (2009) Automatic particle selection from electron micrographs using machine learning techniques. J Struct Biol 167:252–260
17. Penczek PA, Zhu J, Frank J (1996) A common-lines based method for determining orientations for N > 3 particle projections simultaneously. Ultramicroscopy 63:205–218
18. van Heel M (1987) Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy 21:111–124
19. Sorzano COS, Bilbao-Castro JR, Shkolnisky Y et al (2010) A clustering approach to multireference alignment of single-particle projections in electron microscopy. J Struct Biol 171:197–206
20. Liu W, Pokharel PP, Príncipe JC (2007) Correntropy: properties and applications in non-Gaussian signal processing. IEEE Trans Signal Process 55:5286–5298
21. Scheres SHW, Valle M, Núñez R et al (2005) Maximum-likelihood multi-reference refinement for electron microscopy images. J Mol Biol 348:139–149
22. Pascual-Montano A, Donate LE, Valle M et al (2001) A novel neural network tecnique for analysis and classification of EM single-particle images. J Struct Biol 133:233–245
23. Crowther RA, Amos LA (1971) Harmonic analysis of electron microscope images with rotational symmetry. J Mol Biol 60:123–130
24. Bárcena M, San Martin MC, Weise F, Ayora S, Alonso JC, Carazo JM (1998) Polymorphic quaternary organization of the Bacillus subtilis bacteriophage SPP1 replicative helicase (G40P). J Mol Biol 283:809–819
25. Fernández JJ, Luque D, Castón JR, Carrascosa JL (2008) Sharpening high resolution information in single particle electron cryomicroscopy. J Struct Biol 164:170–175
26. Frank J, Penczek P (1995) On the correction of the contrast transfer function in biological electron microscopy. Optik 98:125–129
27. Scheres SHW, Gao H, Valle M et al (2007) Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nat Methods 4:27–29
28. Wolf M, Garcea RL, Grigorieff N, Harrison SC (2010) Subunit interactions in bovine papillomavirus. Proc Natl Acad Sci USA 107:6298–6303
29. Zhang X, Jin L, Fang Q et al (2010) A cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry. Cell 141:472–482
30. Scheres SHW, Núñez-Ramírez R, Sorzano COS et al (2008) Image processing for electron microscopy single-particle analysis using XMIPP. Nat Protoc 3:977–990