CHAPTER 5

Structure Determination of Icosahedral Viruses Imaged by Cryo-electron Microscopy

ROBERT S. SINKOVITS[a,c] AND TIMOTHY S. BAKER[a,b]

[a] Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093, USA; [b] Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093, USA; [c] San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA

1 Introduction

Three-dimensional (3D) image reconstruction of icosahedral viruses by transmission electron microscopy (TEM) began with the pioneering work on negatively stained samples in Cambridge, UK.[1] This ushered in a new era for virus structure determination and helped lay a firm foundation for subsequent, near atomic resolution X-ray crystallographic studies of viruses such as tomato bushy stunt[2] and southern bean mosaic virus.[3] Crowther *et al.*'s elegant common lines formulation[4] was extremely laborious by today's standards as it required hands-on inspection and analysis of images and their Fourier transforms to identify the orientations of individual virus particles on the TEM support grid. At that time, the $\sim$30 Å resolution that could be achieved was limited not by the ability to collect and analyze adequate numbers of images but rather by effects of stain, radiation damage and various distortions to the sample.

The advent of single-particle cryo-electron microscopy (cryo-EM) subsequently revolutionized the use of microscopy for structure determination for

RSC Biomolecular Sciences No. 21
Structural Virology
Edited by Mavis Agbandje-McKenna and Robert McKenna

Published by the Royal Society of Chemistry, www.rsc.org

viruses and also a wide range of macromolecules and macromolecular complexes.[5,6] In cryo-EM, specimens are preserved in thin layers of vitrified ice, which eliminates the need for negative staining, reduces radiation damage and allows them to be imaged much closer to their native state, albeit at very low contrast. Although structures could now be solved at much higher resolutions, the extremely noisy nature of each particle image necessitated averaging the data from hundreds or even thousands of images to produce reliable 3D reconstructions. As manual processing of large amounts of data became impractical, new approaches were needed to reduce the amount of hands-on intervention. The initial focus was on developing sophisticated software for handling individual steps in the structure determination, but eventually attention turned toward automating the reconstruction from start to finish.

Viruses with a variety of morphologies have been examined by cryo-EM, but those with capsids that have icosahedral symmetry are by far the most common and also often the easiest to study in 3D.[7] They are known to infect hosts from all kingdoms of life and to package a wide range of single- and double-stranded DNA and RNA genomes. Because of the large number of human, livestock, fish and plant diseases caused by these viruses, they have been the subjects of innumerable structural studies, as exemplified throughout this book. The reason why some viruses tend to form icosahedral capsids remains an unsolved mystery, but it is widely argued that evolutionary pressures (*e.g.* genetic economy) led naturally to self-assembling systems comprised of multiple copies of one or a small number of unique subunits.[8]

The basic techniques used to generate cryo-reconstructions of single particles can be straightforwardly applied to icosahedral viruses but, by designing algorithms to exploit their high degree of symmetry, researchers have been able to reach higher resolutions. The symmetry operations that leave the icosahedron invariant result in any non-axial view having 60 equivalent views. This means that we can define an asymmetric unit (ASU) that includes just one-sixtieth of the volume and use this ASU to generate the full icosahedron. We exploit this icosahedral symmetry in two ways. First, when determining the view orientations of the virus particles in our electron micrographs, we only need to consider orientations that are in the ASU rather than the full range of orientations for a 3D object. Second, when a density map is reconstructed from the particle images, each image will make 60 contributions to the reconstruction – one from the assigned orientation in the ASU plus 59 from the symmetry-related orientations. In addition to their symmetry, we also exploit the fact that the icosahedral viruses are roughly spherical and we have developed algorithms that can very efficiently determine reasonable estimates for the orientations of the particle images in the early stages of the reconstruction. We explore all of these points in more detail in Section 3.

The full range of steps necessary for a virus structure determination project is given in Figure 1. Specimen purification has been covered in Chapter 1. Discussions in this chapter are limited to those operations required to go from electron micrographs, acquired on film or CCD camera, to a 3D structure.

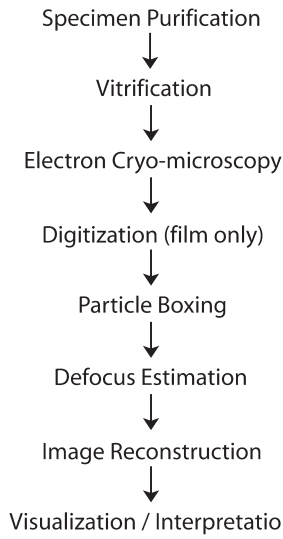

Figure 1 Primary steps involved in determining and analyzing the 3D structures of biomacromolecules using cryo-electron microscopy. This chapter covers only the steps from digitization through image reconstruction used to study icosahedral viruses.

2 Image Digitization and Preprocessing

Recording Media and Image Digitization

Micrographs are recorded on photographic film or a CCD camera and both have been used to achieve very high-resolution cryo-reconstructions.[9,10] The choice of recording method is often one of personal preference and is influenced by the tradeoff between the ease of use of CCD images and the larger field of view that can be captured with film. Indeed, CCD data are already in a format suitable for boxing particle images but film requires an additional step of digitization. This is normally performed with a flat-bed, scanning microdensitometer, to produce pixels at a step size of 6–7 μm. For example, an image recorded at a magnification of 50 000× and digitized at 7 μm intervals would generate pixels whose size correspond to 1.4 Å ($= 7\,\mu m \times 10^4$ Å/μm/50 000) in the specimen. A digitized micrograph obtained from scanning an 8×10 cm piece of film at this resolution would have dimensions of approximately 11 400 by 14 300 pixels and contain 155 megapixels. Hence, when each pixel value is represented by a four-byte, floating-point number, the entire micrograph, stored as one file, would consume ~620 megabytes (0.62 GB) of computer storage space.

Once TEM image data have been recorded and are available in digital form, several preprocessing steps are required before the image reconstruction process can be initiated. Individual virus particles must be identified in the micrographs (see particle boxing), estimates must be made of the defocus levels

in order to correct for the effects of the microscope contrast transfer function (CTF)[11] and the image transforms must be scrutinized to make sure that the images are sufficiently free of astigmatism, specimen drift or other artifacts that would degrade the quality of the reconstruction (see defocus estimation).

Particle Boxing

The next step is to identify and window out ('box') individual virus particles. Poor quality micrographs or images of 'bad' particles can always be rejected later, but care should be taken at this stage to reject particles that are overlapping with other particles or are distorted, deformed, disassembling or show other obvious defects. The size of the box should be large enough so that the boxed images contain a sufficient number of background pixels around the particles. For example, an image of a 500-Å diameter virus such as polyoma or SV40[12] recorded at 50 000× and digitized at 7 μm intervals will require a box size of at least 357 pixels. Hence, a larger box size (perhaps 395 × 395 pixels) would be used to assure that none of the virus particle is missing in the image. Another important reason for providing sufficient padding around the particles is that the use of defocus to enhance image contrast causes an otherwise perfect image to be spread out over a larger area. The density that would have been recorded at a given pixel had the image been acquired in focus is instead replaced by a Gaussian distribution that smears the intensity over neighboring pixels. As a result, the apparently featureless region surrounding the particle image actually contains information that should be used in the reconstruction. This effect is independent of the size of the particle and the extra padding that is required depends only on the value of the defocus.

Defocus Estimation

After particle boxing has been accomplished, the next step is to estimate the defocus level of each micrograph. Unlike the other parameters that are used in the calculation of the CTF correction, the defocus level is not known *a priori* to a sufficient level of accuracy and must be determined from the image data. This is typically achieved through quantitative analysis of the average (incoherent) power spectra of the particle images. A single spectrum is fairly noisy, but averaged spectra generally display a series of concentric 'Thon' rings,[13,14] whose positions are related to the value of the microscope CTF at the time the image was recorded. Various programs[15,16] are used to compute a least-squares fit between a theoretical CTF and the observed rings to determine the defocus level that best agrees with the locations of the nodes in the averaged transform (Figure 2). These programs require as input the known value for the spherical aberration of the microscope objective lens, the accelerating voltage of the electron beam, the pixel size and the averaged power spectra. While estimating the defocus levels, we also have the opportunity to identify artifacts that are not apparent from a simple visual inspection of the micrographs. High levels of

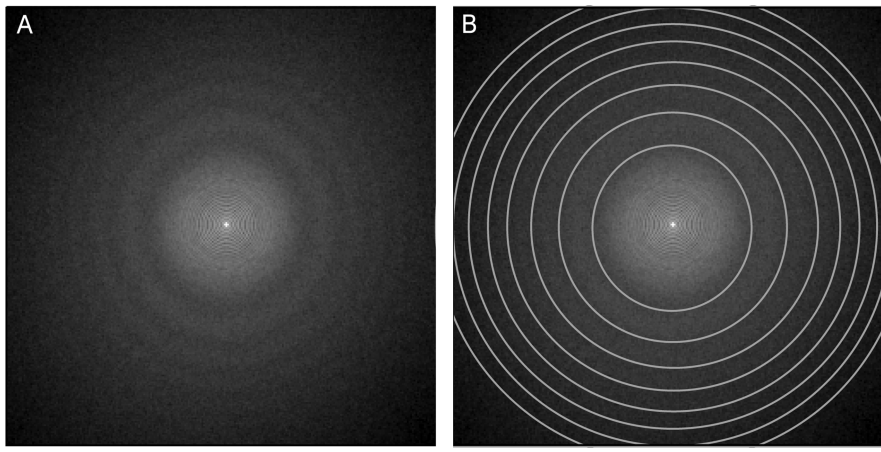

Figure 2 (A) Incoherent average of the Fourier transforms of 100 boxed images of bacteriophage P22 expanded heads. (B) Same as (A) except with nodes in the microscope contrast transfer function highlighted with gray ellipses to illustrate location and presence of slight image astigmatism. Data were acquired at an electron accelerating voltage of 200 kV and an estimated underfocus setting of 1.4 μm.

astigmatism will cause the Thon rings to become elliptical, whereas specimen drift results in loss of intensity in the rings in the direction of the drift.

3 3D Image Reconstruction

Most modern image reconstruction projects, regardless of the symmetry of the system being studied, are based on some type of iterative, model-based refinement method. In its most basic form, the strategy involves determining the origins (*i.e.* common point of reference or center of symmetry) and orientations for a set of particle images by comparing them with projections of a model, then using these aligned images to construct a new, hopefully more accurate model that serves as the starting point for the next iteration.

As expected, an actual implementation is more complex than just described, but these two elements form the core of any iterative reconstruction approach. The first complication is that the quality or resolution of the model must be estimated. There are several mathematically precise definitions for resolution, but here we take it simply to mean the level of detail that can reliably be discerned in the model. The finer detail one is able to discern in a model, the higher is the resolution of that model. For macromolecular structures, low resolution is generally considered to fall somewhere in the $\sim$20–50 Å range. Here, only gross features of the virus morphology and possibly coarse outlines of the subunits can be distinguished, whereas at very high resolutions ($<$4 Å) the tertiary structures of viral protein subunits become visible. Resolution estimates are used to gauge the quality of the model used in refinement and also

guide decisions regarding the choice of algorithms and input parameters from one iteration to the next.

Another complication arises because the origins and orientations of the particles are obtained *via* comparison of the particle images to a model, but the model itself is constructed from these same images. We require either a method to assign origins and orientations to the images in the absence of a model or a way of constructing a model without relying on aligned images. This issue is best addressed after first presenting an overview of our automated image reconstruction system.

Iterative Model-based Refinement and Automation

AUTO3DEM is an automation system that we developed to perform icosahedral reconstructions (Figure 3).[17] We specifically highlight features within AUTO3DEM, but two points should be kept in mind. First, the main steps carried out by AUTO3DEM (origin and orientation determination, resolution estimation, model construction) are generic to any iterative, model-based

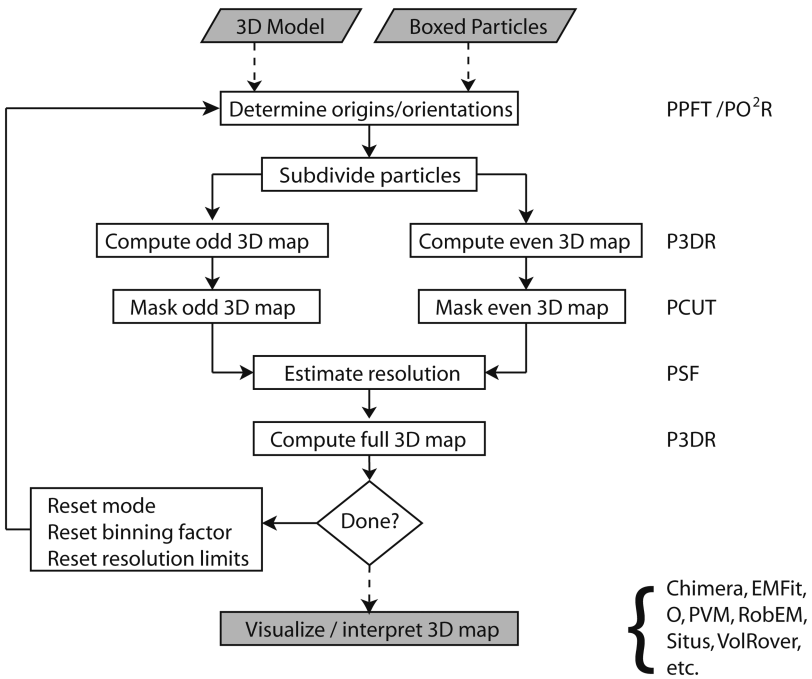

Figure 3 Simplified flow chart of automated image reconstruction process implemented by AUTO3DEM.[17] Shaded boxes represent either input data or steps performed outside of AUTO3DEM. The programs needed at each computational step are listed at the right side of the figure. Dashed and solid lines indicate one-time and iterative operations, respectively. The programs P3DR, PO^{2}R and PPFT impose or assume icosahedral symmetry.

refinement method. Second, other automated systems are available for processing images of particles with icosahedral[18] and lower or no symmetries.[19] The algorithms within AUTO3DEM have been specifically tuned for icosahedral symmetry and the underlying image reconstruction codes have been parallelized and can run on any shared or distributed memory parallel computer.

There are numerous benefits to be realized by automating the image reconstruction process. The most obvious ones are that it relieves the user of many of the repetitive and error-prone steps involved in managing large numbers of data files, setting the parameters for and running multiple programs, interpreting intermediate results and making decisions affecting the overall course of the calculations. Typographic errors may be relatively uncommon at any given step but, over the course of many iterations, even experienced users are likely to make mistakes. Another key advantage of an automation system is that it can be used to capture expert knowledge and make the software more accessible to novice users. Rather than having to understand every detail of the software, less experienced users can rely heavily on the default parameters and generate moderate resolution reconstructions with a minimum of effort. Automation can also reduce the time needed to solve a structure since the delays between the completion of one step and the initiation of the next are eliminated and computer resources are maximally utilized. Finally, the quicker turnaround time made possible through automation enables a researcher to carry out more numerical experiments and reach higher resolutions.

Starting Model/Structure

The iterative refinement process requires an initial model against which a set of particle images can be compared. This starting model does not need to be of very high resolution, but rather just have the correct size and general shape of the structure under investigation. Often a prior reconstruction obtained for a closely related virus can be used, but care must be taken since a size difference of just a few percent can cause the reconstruction process to fail. Geometric models of the proper dimension can also be used, but again are prone to the same problems. An alternative approach that we typically employ is to use the 'random model computation' (RMC)[20] to construct a starting model from a relatively small number of particle images.

A detailed understanding of the RMC is not necessary at this point since the general iterative refinement method only requires that we have a starting map and does not depend on how it was obtained. In addition, the RMC procedure closely follows that used for the general method, but with one small, yet crucial, exception as described later (see Building a Starting Model from Scratch).

Determining Particle Origins and Orientations: Global and Local Refinement

The most computationally intensive and also critical operation in the image reconstruction process is the determination of the five parameters that

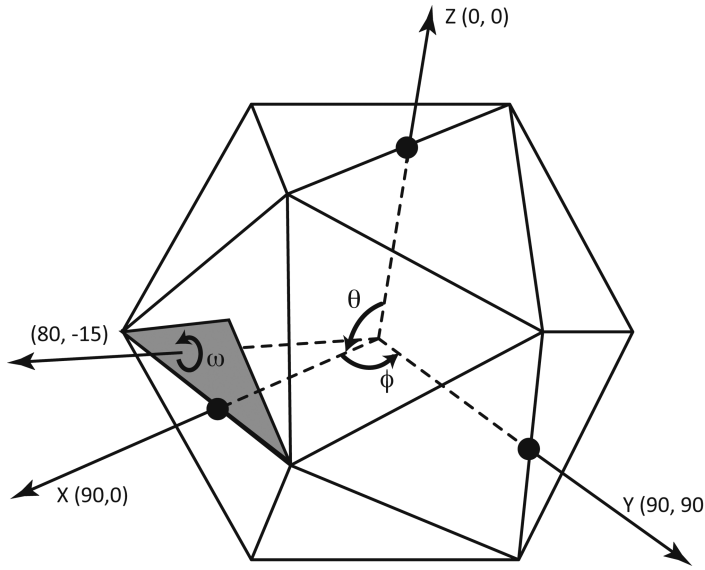

Figure 4 Schematic of icosahedron showing the three angles (θ, φ, ω) that define the view orientations of the icosahedral particles in the electron microscope. The shaded triangle denotes one of 60 equivalent asymmetric units. This unit is bounded by an adjacent pair of fivefold axes ($\theta = 90.0^\circ$, $\varphi = \pm 31.72^\circ$) and an adjacent threefold axis ($\theta = 69.09^\circ$, $\varphi = 0^\circ$). The values of θ and φ for the Cartesian axes and one general view vector appear in parentheses.

characterize each boxed particle image. These are the two coordinates (x, y) that define the origin of the particle relative to the center of the box and the three angles (θ, φ, ω) (Figure 4) that describe the orientation of the particle relative to the electron optical axis of the microscope. Two distinct algorithms are used by AUTO3DEM to carry out this step and in both cases the projections of the model are multiplied by the CTF before comparisons are made with the image data.

The first algorithm is used during the early stages of the reconstruction process, when the model is generally only accurate to and computed at a relatively low resolution (often $\gg 30$ Å) and can change significantly from one iteration to the next. To avoid a particular particle being trapped with an orientation assignment that is far from the correct one, we perform a global search of orientation space (*i.e.* all possible orientation angles) for each iteration cycle. This would normally be a very expensive 3D search over the three orientation angles, but the polar Fourier transform (PFT) algorithm[21] subdivides this brute force process into two discrete steps and greatly reduces the computation. In PFT, θ and $|\varphi|$ are estimated first, followed by ω and the sign of φ. The process is made even more efficient by taking advantage of the icosahedral symmetry and restricting the orientation search window to values of (θ, φ) that represent half of the ASU or 1/120th of the

total orientation space. (The factor of one half follows from the fact that, in the first part of the algorithm, we only need to determine the absolute value of φ.) Using the default angular step of 0.5°, the PFT program computes a set of 1430 projected views of the model that evenly cover the ASU and then each particle image is compared with every one of these model projection images. If the diameter of the particle is small (*e.g.* <400 Å), we can increase the size of the angular step to 1.0°, in which case only 370 projected views are generated.

The resolution that can be reached at this stage of the reconstruction is variable, but often falls in the range of 15–20 Å and can sometimes reach 10 Å. The time needed for this calculation depends strongly on both the number and size of the particle images and the type of computer hardware being used. We often find that this can be done in a matter of hours when run in parallel mode on a machine with eight dual core Opteron 880 processors running at a clock speed of 2.4 GHz.

In the later stages of the reconstruction process, the program PO^{2}R employs the second algorithm to perform local refinements of the particle origins and orientations.[22] The assumption is made that, at this stage of the processing, the origin and orientation of each particle are relatively close to the true values and that only a very limited region of origin and orientation space, centered about the current values of (x, y) and (θ, φ, ω), need be searched. The orientation angle search is typically done over a $9\times9\times9$ grid of points with an angular step size typically ranging between 0.1 and 1.0°. The PO^{2}R algorithm has two main advantages over the one used in PFT. First, the computational cost of PO^{2}R does not depend on the size of the angular step, but rather on the number of nearby orientations that are tested. Hence, as the resolution of the model improves, the angular step can slowly be reduced without causing an increase in run time. Second, even when the same angular step sizes are used in PPFT (parallel implementation of the PFT algorithm) and PO^{2}R, the latter generally leads to better alignments and hence higher resolutions since the comparison between the images and projections of the model is performed completely in Fourier space. Of course, both methods are needed since PO^{2}R can only be used after the orientations of the particles are correctly identified within a small region of the ASU.

Image data at full pixel resolution are generally not required during the first few iterations of a reconstruction. The PFT program has the capability of averaging together 2×2 groups of neighboring pixels in the original images to generate sampled or 'binned' images with one-quarter the total number of pixels, thereby reducing the time required for the computations. AUTO3DEM can monitor the progress of the reconstruction and automatically switch from using binned to un-binned data once it detects that the resolution of the model is no longer improving.

The PPFT and PO^{2}R programs can each more accurately determine particle origins and orientations if particles are located close to the center of the image box. For mis-centered, tightly boxed particles, portions of the particle image will be cut off and can never contribute to the 3D reconstruction.

Furthermore, images of particles collected at high defocus contain information ('Fresnel fringes') that can extend a considerable distance beyond the visible edge of the particle. Tight boxing will truncate these fringes, reduce resolution and affect origin and orientation determination. One strategy aimed at achieving higher resolutions is to re-box the particles from the original micrographs using the improved estimates for the particle origins obtained during the reconstruction.

Computing the 3D Reconstruction

The next major step in the reconstruction process is to compute a 3D model from the images of particles whose origins and orientations have been determined. AUTO3DEM uses the program P3DR (parallel 3D reconstruction)[22] to construct a density map of the virus. A density map computationally represents the object as a 3D grid of voxels, each of whose magnitude is proportional to the scattering density within a corresponding small volume of the virus.

Reconstruction of a density map from a series of 2D projection images relies on the well-known projection theorem. This states that the Fourier transform of a projection of a 3D object is equivalent to a central section of the 3D transform of the original object. The implication is that one can easily generate projections of an object from the 3D transform of the object (this operation lies at the heart of the PO^2R algorithm) and that the transforms of the particle images can be used to build a map or model of the object once the orientations and origins of the particles imaged are known. This approach is analogous to that used to generate tomographic reconstructions (tomograms), which are obtained by recording a series of images of an object that is rotated systematically ± 60–70° about an axis normal to the electron beam in a microscope.[23] In order to generate a structure that faithfully reproduces the features at all spatial frequencies, the image data must first be corrected for the CTF before performing the back projections. There are several ways of doing this, but the most common are to either correct just the phases[13] or both phases and amplitudes of the Fourier components.[24]

We took advantage of the inherent symmetry of icosahedral viruses when determining particle orientations and limited our search to orientations that fall within a single ASU. We can exploit symmetry again during the reconstruction of the density map by allowing each particle image to make 60 contributions to the reconstruction – one corresponding to the assigned orientation within the ASU and 59 from the symmetry-related orientations.

The quality of the model can be further improved by omitting 'bad' particle images from the reconstruction. The origin and orientation refinement programs, PPFT and PO^2R, both output one or more quantitative measures of how well each particle image agrees with the model. AUTO3DEM can use these scores to rank the particles and include just those that lie above a certain threshold or that have scores within a given number of standard deviations of

the average score. Particle images can also be eliminated at any time by manually editing the text files that list the particles to be processed.

Although tremendous advantages are realized by assuming icosahedral symmetry when processing images of viruses, the details of structures or substructures will be smeared or averaged out if they are not present in multiples of 60 and arranged with 532 point group symmetry. For example, the nucleocapsids of all herpesviruses are nominally icosahedral but they each possess a small portal complex at one of the vertices.[25,26] Under icosahedral averaging, this component will appear with lower density at all 12 vertices rather than the correct density at a single vertex. More seriously, the structure of such a component as deduced from an icosahedral reconstruction will likely be incorrect since fivefold symmetry would have been imposed regardless of its true symmetry. Icosahedral reconstructions are generally carried out in these cases to solve the capsid to the highest possible resolution and may be followed by lower-symmetry or asymmetric reconstructions to resolve the structure of the non-icosahedral components.[27–29]

Estimating the Resolution of the Reconstruction

An estimate of the level of resolution achieved in the reconstructed virus structure provides an objective gauge of map quality and reliability and helps guide the course of the reconstruction. The Fourier shell correlation (FSC) and phase residual are two measures commonly used to estimate resolution in single-particle reconstructions.[30] Published results often just present or report the FSC since the two tend to assess resolution very similarly. Here, we restrict our discussion to the FSC.

The FSC does not directly measure the resolution of a reconstructed density map, but rather looks at the agreement between a pair of maps. This pair, often referred to as the 'even' and 'odd' maps, is built from image data obtained by dividing the full set of images into two, mutually exclusive sets of equal size. The goal is to determine how well the two maps conform at different spatial frequencies.

An FSC curve plots the correlation coefficients (CCs) between the Fourier transforms of the maps as a function of spatial frequency. Two maps that correlate perfectly (*i.e.* are identical) in a defined band of spatial frequency have a CC = 1.0. Those that exhibit no correlation, yield a CC = 0.0. For real image reconstruction data, the FSC will typically have a value close to one at the very lowest spatial frequencies ($< 1/50\ \text{Å}^{-1}$), indicating that the 'even' and 'odd' reconstructed virus structures are similar in terms of size and overall shape and then eventually drop to zero or lower at high spatial frequencies, suggesting that details within the virus structure at that resolution are unreliable and consist of uncorrelated noise (Figure 5). The spatial frequency at which the FSC first drops below a value of 0.5 is generally considered to be a conservative estimate of the resolution, but this view is not universally accepted and some argue that a cutoff as low as 0.14 is valid.[31] In practice, the FSC curve tends to

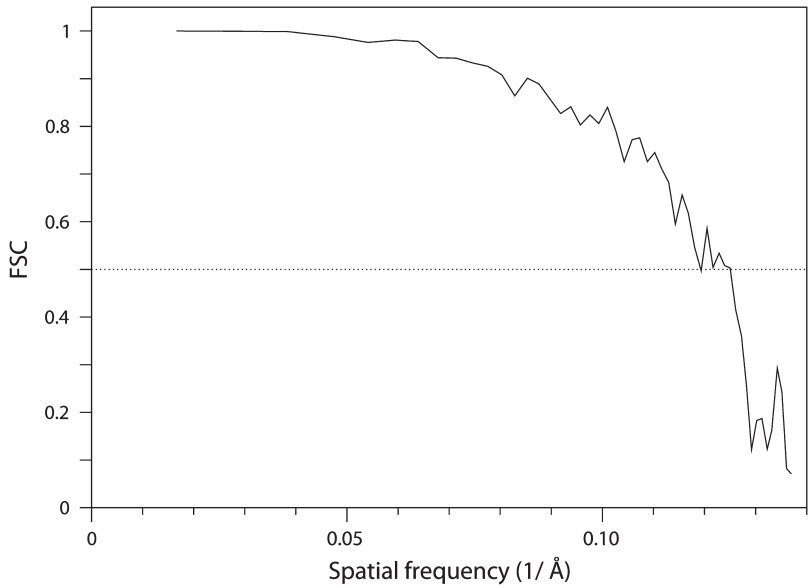

Figure 5 FSC plotted as function of spatial frequency for 3D reconstructions (even and odd maps) of PsV-F computed from a data set of 2605 boxed particle images. The spatial frequency at which the FSC drops below a value of 0.5 corresponds to a resolution of 8.1 Å.

drop very rapidly in a narrow region of spatial frequency and the difference in estimated resolution using these two cutoff values is small. Regardless, the resolution estimate arising from any FSC-type analysis is simply an estimate and not a precise determination. It merely offers the researcher some level of confidence that structural features can be reliability interpreted and correlated with other relevant data.

The disordered components of the virus, such as the genome, internal or external lipid bilayers and, in some cases, portions of the major capsid protein that contact the nucleic acid or bilayer are often solved to a much lower resolution than the icosahedral capsid. Consequently, a more accurate estimate of the resolution achieved in the 3D reconstruction is obtained when most of the disordered regions are computationally excised from the pair of maps before the FSC curve is determined. Inner and outer radii, r_1 and r_2, are selected to identify the bounds of the highly ordered region of the virus. All density values within these radial limits of the 3D maps are unmodified, but those that lie outside the limits are down-weighted smoothly to zero. That is, a Gaussian, radial falloff is applied to all densities at $r < r_1$ and $r > r_2$. Such weighting avoids sharp discontinuities in the density maps at $r = r_1$ and $r = r_2$ that would give rise to artifacts in the Fourier transforms and lead to artificially high CCs and overestimation of the resolution achieved.

For most projects, the FSC provides a reliable estimate of the resolution, but its utility begins to break down as reconstructions are attained at resolutions

beyond about 7–8 Å. In these instances, a more useful although qualitative estimate of resolution is obtained through careful examination of fine features in the map. For example, at about 6–10 Å and sometimes even lower resolution, α-helices in proteins appear as smooth, tube-like density features ('cylinders' or 'tubes') with a diameter of ~6–7 Å. When the resolution of a reconstruction improves to better than 6 Å, the density corresponding to these secondary structure elements transforms from a smooth cylinder morphology into one with more helical character. At about 4–5 Å resolution the helical pitch is unmistakable in cryo-EM density maps.[32,33] It is only at even higher resolution (4 Å or better) that the separation between adjacent β strands in sheets or bulky side chains of amino acid residues can be resolved in proteins.[9,10]

Building a Starting Model from Scratch: the Random Model Computation

The iterative, model-based refinement process requires a starting model to obtain initial estimates for the origins and orientations of the particles. Often we can start with the 3D map of a closely related virus that has been obtained *via* microscopy (single particle image reconstruction) or crystallography (X-ray diffraction of single crystals). However, when a completely new virus structure is being examined, we need some way to bootstrap the refinement process. In our reconstruction scheme, we generally use AUTO3DEM to construct an *ab initio* model using the RMC.[20]

The crux of the RMC approach is to construct a density map from a small number of particle images for which random orientations are assigned and whose origins are set to the center of the box. Although this initial 3D map will bear no resemblance to the actual structure being solved except for appropriately representing the size and symmetry of the virus, it often serves as an effective seed for the image reconstruction process. A rigorous explanation for how this technique succeeds has not been reported, but we believe that the random model contains just enough signal (features consistent with genuine structure) to jump-start the iterative process. This view is also consistent with our observation that the RMC works best when a modest number (generally <200) of images is used. Use of an entire data set of particle images tends to result in a relatively featureless, spherically symmetric model and the refinement process fails to converge and yield a reliable reconstruction.

The RMC works best for viruses whose structures include prominent features such as ridges, arches and protrusions (see, *e.g.*, Section 4), but is less reliable for particles with smoother profiles. In the latter case, multiple random models can be constructed and the one that leads to the best low-resolution model, as measured by the average value of the FSC over a fixed range of spatial frequencies, is selected as the starting point for the full reconstruction using all of the image data.

AUTO3DEM is just one of many image reconstruction packages that now employ similar, random model methods. For example, EMAN[34] and the helical

reconstruction package IHRSR[35] offer functionality similar to that embedded in AUTO3DEM.

For completeness, we point out, but with no further explanation, that there are additional ways of constructing low-resolution, starting models from a set of particle images with unassigned origins and orientations. These include the random conical tilt method and angular reconstitution.[13,14] Both of these methods have proven to be very powerful and have utility for examining lower symmetry particles and also those that are asymmetric (the ribosome being the classic example). The reader is encouraged to become familiar with these methods and with the extensive software packages in which they are implemented. These include SPIDER,[36] IMAGIC[37] and SPARX.[38]

Hand Determination

TEM records 2D projections of 3D specimens and handedness information is lost in the images. Hence there is a 50% chance that the reconstruction will be the mirror image of the correct structure. The absolute hand can often be identified from experiments in which pairs of images are collected from the same sample, tilted at two different orientations relative to the electron beam.[31,39] If tilt experiments prove to be inconclusive in distinguishing which enantiomer of a structure is correct, it may be possible to determine hand from the reconstruction itself (see below). When the hand cannot be determined by tilt experiments or direct visualization, the researcher should clearly communicate that the choice of hand is arbitrary for the structure.

For many icosahedral viruses, the subunits that form the capsid are grouped into trimers, pentamers or hexamers that are clearly distinguishable at resolutions of 30 Å or even lower. These multimeric units are further arranged on a regular lattice that can either be symmetric (*e.g.* T = 1, 3, 4, 9, 12, 16, *etc.*) or skewed (*e.g.* T = 7, 13, 19, 21, *etc.*). Since the latter possess a definite handedness, the correct hand for the reconstruction of a virus with a skewed lattice can be inferred by comparison with the lattice of related viruses for which the hand is already known.

Determination of hand is more difficult for capsids that possess a symmetric lattice, but is possible if related capsid structures of known hand have been studied and the individual subunits are clearly resolved in the reconstruction. This is often possible for reconstructions at a resolution of 15 Å or better since both the capsid subunits and the subunit oligomers (trimers, pentamers, hexamers, *etc.*) often exhibit pronounced asymmetry.

For X-ray crystal structures and cryo-reconstructions that have been solved to about 5 Å or better, the hand can be deduced directly from the appearance of secondary structural elements in the density map. For example, an α-helix-rich protein structure will exhibit helical features with a right-handed twist in the map of correct hand. Unfortunately, cryo-reconstructions at such high resolutions are still fairly difficult to achieve.

4 Image Reconstruction Example – PsV-F

A recent reconstruction of a fungal virus, PsV-F, serves to illustrate the image reconstruction process (Figure 6). A low-resolution (25–30 Å) starting model was obtained using the RMC (Section 3) with 150 particle images chosen from the micrographs (40 in total) with the highest defocus levels. AUTO3DEM was then run using the full set of 2605 images to generate a map with an estimated resolution of 8.1 Å in a completely automated manner. The hand of the resulting structure was chosen to be consistent with ScV-L-A.[40] The entire process required 70 min on a 16-processor Linux cluster, including the 8 min used to generate the starting model. Detailed statistics for the calculations (Table 1) show that, during the first three iterations, the value of the FSC never dropped below a value of 0.5 for the entire range of spatial frequencies over which the FSC was calculated. Therefore, the actual resolution should be higher than the value reported in Table 1 (highlighted with *). With image data that had been subjected to a 2×2 binning, an estimated resolution of 12.4 Å was reached in 13.8 min. Three more iterations of AUTO3DEM in search mode using unbinned image data led to a 3D reconstruction with an estimated

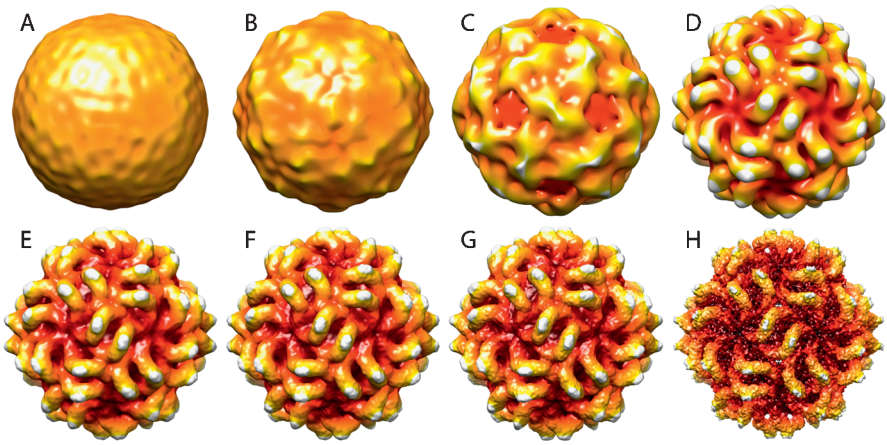

Figure 6 3D surface-shaded representations of PsV-F reconstruction viewed along a twofold axis. Radius in each 3D map is color coded from black (lowest radius) through dark red, orange, yellow, to white (highest radius). (A–D) and (E–H) represent models constructed from 150 and 2605 particle images, respectively. Timings and resolutions are given in Table 1. (A) Model constructed from images that had been assigned random orientations. (B–D) Models from iterations 1, 3 and 5 of the reconstruction process. (E–G) Iterations 9, 12 and 15 from an AUTO3DEM run that used (D) as the starting model. (E) and (F) correspond to the last iterations in search mode using binned and unbinned data, respectively. (G) Result after three further iterations in refine mode. (H) Model at estimated 6 Å resolution obtained using an 8 Å model as starting point and manually running P3DR, PO^2R, PSF and PCUT as standalone programs. An inverse temperature factor of 1/300 $Å^{-2}$ was applied to enhance high-spatial frequency terms.[41]

Table 1 AUTO3DEM statistics for 3D reconstruction of PsV-F. Iteration 0 corresponds to the generation of the initial model from images that had been assigned random orientations and hence the AUTO3DEM mode is not applicable. Iterations 0–5 constitute the RMC, whereas subsequent iterations were carried out with the full set of image data. Results marked with * indicate that the value of the FSC never dropped below 0.5 over the entire range of spatial frequencies for which it was calculated. FSC data from iterations 0 to 4 are so noisy that resolution estimates are considered to be unreliable, but suggest that the maps are at resolutions worse than 30 Å.

Iteration	*Mode*	*Images*	*CPUs*	*Binning*	*t (min)*	*Total t (min)*	*Resolution (Å)*
0	–	150	4	2×2	0.9	0.9	>30.0
1	Search	150	4	2×2	1.3	2.2	>30.0
2	Search	150	4	2×2	1.3	3.5	>30.0
3	Search	150	4	2×2	1.5	5.0	>30.0
4	Search	150	4	2×2	1.5	6.5	>30.0
5	Search	150	4	2×2	1.5	8.0	30.0*
6	Search	2605	16	2×2	3.9	11.9	25.0*
7	Search	2605	16	2×2	3.3	15.2	16.7*
8	Search	2605	16	2×2	3.3	18.5	12.5*
9	Search	2605	16	2×2	3.3	21.8	12.4
10	Search	2605	16	1×1	12.5	34.3	10.7
11	Search	2605	16	1×1	12.4	46.7	10.2
12	Search	2605	16	1×1	12.4	59.1	10.2
13	Refine	2605	16	1×1	5.0	64.1	8.9
14	Refine	2605	16	1×1	3.2	67.3	8.5
15	Refine	2605	16	1×1	2.4	69.7	8.1

resolution of 10.2 Å in an additional 37 min of wall clock time. Three iterations in refine mode further improved the resolution to 8.1 Å in just 11 min more. Starting from this 8.1 Å map, we were able to increase the resolution further to approximately 6 Å. This was accomplished by running P3DR and PO^2R as standalone programs outside the context of AUTO3DEM, manually adjusting the input parameters (inverse temperature factors,[41] method of CTF correction[13] and angular step size used in orientation determination) and visually inspecting the resultant maps. Changes to the input parameters that resulted in 'better' maps (*e.g.* secondary structural features became less ambiguous and easier to interpret) were accepted and the corresponding map was used as the starting point for the next round of origin and orientation refinement. Surface renderings at various stages of the process illustrate how the quality of the PsV-F reconstruction improved during refinement (Figure 6).

Of course, although the results with PsV-F are not atypical, not all virus image data will behave so well and structures refined to sub-nanometer resolutions may not be achieved so quickly. Nonetheless, the reconstruction of PsV-F demonstrates that it is possible in some instances to generate models to sub-nanometer resolution within 1 h after particle images have been boxed and

CTF parameters estimated. Indeed, solving larger viruses such as the 1850 Å diameter Chilo iridescent virus[42] to comparable resolution requires significantly more computation and the construction of a suitable starting model is much more difficult if the virus has a relatively smooth, featureless surface such as dengue virus.[20]

5 Summary

Recent advances in image reconstruction software have made it possible to solve the structures of icosahedral viruses to moderate resolutions in a completely automated manner. What had formerly been an extremely time-consuming step, often requiring many months of concerted effort, can now be completed in a matter of hours in the best-case scenarios once boxed image data are available.

Some of the roadblocks to structure determination have been eliminated but others remain. For example, sub-nanometer resolution cryo-reconstructions are now being achieved more frequently by users with little computational knowledge or experience with the image reconstruction process, but reaching the highest possible resolutions given the limitations of the data set has only been carried out by expert users. Our aim is to quantify and build this expert knowledge into the software so that higher resolutions can more easily and routinely be achieved. The image preprocessing steps, most notably particle boxing and defocus estimation, still require significant manual effort. Fully automating these steps would relieve the structural biologist of much tedium and pave the way to our eventual goal of performing image reconstructions directly at the microscope and providing the cryo-microscopist timely feedback on specimen quality.

Acknowledgments

This work was supported in part by grants R37 GM-033050 and R01 AI-055672 and shared instrumentation grant 1S10 RR020016-01 from the National Institutes of Health to T.S.B. The San Diego Supercomputer Center (SDSC) provided access to TeraGrid computing and support from the University of California, San Diego and the Agouron Foundation (to T.S.B.) were used to establish and to equip cryo-TEM facilities at the University of California, San Diego. R.S.S. received partial support from SDSC.

References

1. R. A. Crowther, L. A. Amos, J. T. Finch, D. J. DeRosier and A. Klug, *Nature*, 1970, **226**, 421.
2. S. C. Harrison, A. J. Olson, C. E. Schutt, F. K. Winkler and G. Bricogne, *Nature*, 1978, **276**, 368.

3. C. Abad-Zapatero, S. S. Abdel-Meguid, J. E. Johnson, A. G. W. Leslie, I. Rayment, M. G. Rossmann, D. Suck and T. Tsukihara, *Nature*, 1980, **286**, 33.
4. R. A. Crowther, D. J. DeRosier and A. Klug, *Proc. R. Soc. London, Ser. A*, 1970, **317**, 319.
5. M. Adrian, J. Dubochet, J. Lepault and A. W. McDowall, *Nature*, 1984, **308**, 32.
6. J. Dubochet, M. Adrian, J.-J. Chang, J. -C. Homo, J. Lepault, A. W. McDowall and P. Schultz, *Q. Rev. Biophys.*, 1988, **21**, 129.
7. T. S. Baker, N. H. Olson and S. D. Fuller, *Microbiol. Mol. Biol. Rev.*, 1999, **63**, 862.
8. D. L. D. Caspar and A. Klug, *Cold Spring Harbor Symp. Quant. Biol.*, 1962, **27**, 1.
9. X. Yu, L. Jin and Z. H. Zhou, *Nature*, 2008, **453**, 415.
10. X. Zhang, E. Settembre, C. Xu, P. R. Dormitzer, R. Bellamy, S. C. Harrison and N. Grigorieff, *Proc. Natl. Acad. Sci. USA*, 2008, **105**, 1867.
11. H. P. Erickson and A. Klug, *Philos. Trans. R. Soc. London*, 1971, **261**, 105.
12. T. S. Baker, J. Drak and M. Bina, *Proc. Natl. Acad. Sci. USA*, 1988, **85**, 422.
13. J. Frank, *Three-dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*, Oxford University Press, Oxford, 2006.
14. R. M. Glaeser, K. Downing, D. DeRosier, W. Chiu and J. Frank, *Electron Crystallography of Biological Macromolecules*, Oxford University Press, Oxford, 2007.
15. S. P. Mallick, B. Carragher, C. S. Potter and D. J. Kriegman, *Ultramicroscopy*, 2005, **104**, 8.
16. Z. H. Zhou, S. Hardt, B. Wang, M. B. Sherman, J. Jakana and W. Chiu, *J. Struct. Biol.*, 1996, **116**, 216.
17. X. Yan, R. S. Sinkovits and T. S. Baker, *J. Struct. Biol.*, 2007, **157**, 73.
18. W. Jiang, Z. Li, Z. Zhang, C. R. Booth, M. L. Baker and W. Chiu, *J. Struct. Biol.*, 2001, **136**, 214.
19. G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees and S. J. Ludtke, *J. Struct. Biol.*, 2007, **157**, 38.
20. X. Yan, K. A. Dryden, J. Tang and T. S. Baker, *J. Struct. Biol.*, 2007, **157**, 211.
21. T. S. Baker and R. H. Cheng, *J. Struct. Biol.*, 1996, **116**, 120.
22. Y. Ji, D. C. Marinescu, W. Zhang, X. Zhang, X. Yan and T. S. Baker, *J. Struct. Biol.*, 2006, **154**, 1.
23. V. Lucic, F. Forster and W. Baumeister, *Annu. Rev. Biochem.*, 2005, **74**, 833.
24. V. D. Bowman, E. S. Chase, A. W. Franz, P. R. Chipman, X. Zhang, K. L. Perry, T. S. Baker and T. J. Smith, *J. Virol.*, 2002, **76**, 12250.
25. G. Cardone, D. C. Winkler, B. L. Trus, N. Cheng, J. E. Heuser, W. W. Newcomb, J. C. Brown and A. C. Steven, *Virology*, 2006, **361**, 426.

26. J. T. Chang, M. F. Schmid, F. J. Rixon and W. Chiu, *J. Virol.*, 2006, **81**, 2065.
27. W. Jiang, Z. Li, Z. Zhang, M. L. Baker, P. E. Prevelige Jr and W. Chiu, *Nat. Struct. Biol.*, 2003, **10**, 131.
28. J. Chang, P. Weigele, J. King, W. Chiu and W. Jiang, *Structure*, 2006, **14**, 1073.
29. G. C. Lander, L. Tang, S. R. Casjens, E. B. Gilcrease, P. Prevelige, A. Poliakov, C. S. Potter, B. Carragher and J. E. Johnson, *Science*, 2006, **312**, 1791.
30. M. van Heel and M. Schatz, *J. Struct. Biol.*, 2005, **151**, 250.
31. P. B. Rosenthal and R. Henderson, *J. Mol. Biol.*, 2003, **333**, 721.
32. W. Jiang, M. L. Baker, J. Jakana, P. R. Weigele, J. King and W. Chiu, *Nature*, 2008, **451**, 1130.
33. S. J. Ludtke, M. L. Baker, D. H. Chen, J. L. Song, D. T. Chuang and W. Chiu, *Structure*, 2008, **16**, 441.
34. S. J. Ludtke, P. R. Baldwin and W. Chiu, *J. Struct. Biol.*, 1999, **128**, 82.
35. E. H. Egelman, *J. Struct. Biol.*, 2007, **157**, 83.
36. W. T. Baxter, A. Leith and J. Frank, *J. Struct. Biol.*, 2007, **157**, 56.
37. M. van Heel, G. Harauz and E. V. Orlova, *J. Struct. Biol.*, 1996, **116**, 17.
38. M. Hohn, G. Tang, G. Goodyear, P. R. Baldwin, Z. Huang, P. A. Penczek, C. Yang, R. M. Glaeser, P. D. Adams and S. J. Ludtke, *J. Struct. Biol.*, 2007, **157**, 47.
39. D. M. Belnap, N. H. Olson and T. S. Baker, *J. Struct. Biol.*, 1997, **120**, 44.
40. H. Naitow, J. Tang, M. Canady, R. B. Wickner and J. E. Johnson, *Nat. Struct. Biol.*, 2002, **9**, 725.
41. W. A. Havelka, R. Henderson and D. Oesterhelt, *J. Mol. Biol.*, 1995, **247**, 726.
42. X. Yan, Z. Yu, P. Zhang, A. J. Battisti, H. A. Holdaway, P. R. Chipman, C. Bajaj, M. Bergoin, M. G. Rossmann and T. S. Baker, *J. Mol. Biol.*, 2009, **385**, 1287.