Journal of Structural Biology 167 (2009) 185–199

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Ab initio maximum likelihood reconstruction from cryo electron microscopy images of an infectious virion of the tailed bacteriophage P22 and maximum likelihood versions of Fourier Shell Correlation appropriate for measuring resolution of spherical or cylindrical objects

Cory J. Prust [a,1], Peter C. Doerschuk [b,*,2], Gabriel C. Lander [c,3], John E. Johnson [c,3]

[a] Department of Electrical Engineering and Computer Science, Milwaukee School of Engineering, 1025 N. Broadway, Milwaukee, WI 53202-3109, USA
[b] Department of Biomedical Engineering and School of Electrical and Computer Engineering, Cornell University, 305 Phillips Hall, Ithaca, NY 14853-5401, USA
[c] Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037, USA

ARTICLE INFO

Article history:
Received 30 September 2008
Received in revised form 4 March 2009
Accepted 28 April 2009
Available online 18 May 2009

Keywords:
Bacteriophage P22
Infectious particle
3-D reconstruction
Cryo electron microscopy

ABSTRACT

A maximum likelihood reconstruction method for an asymmetric reconstruction of the infectious P22 bacteriophage virion is described and demonstrated on a subset of the images used in [Lander, G.C., Tang, L., Casjens, S.R., Gilcrease, E.B., Prevelige, P., Poliakov, A., Potter, C.S., Carragher, B., Johnson, J.E., 2006. The structure of an infectious P22 virion shows the signal for headful DNA packaging. Science 312(5781), 1791–1795]. The method makes no assumptions at any stage regarding the structure of the phage tail or the relative rotational orientation of the phage tail and capsid but rather the structure and the rotation angle are determined as a part of the analysis. A statistical method for determining resolution consistent with maximum likelihood principles based on ideas for cylinders analogous to the ideas for spheres that are embedded in the Fourier Shell Correlation method is described and demonstrated on the P22 reconstruction. With a correlation threshold of .95, the resolution in the tail measured radially is greater than 0.0301 $Å^{-1}$ (33.3 Å) and measured axially is greater than 0.0142 $Å^{-1}$ (70.6 Å) both with probability $p = 0.02$.

1. Introduction

Motivated by recent reconstructions from cryo electron microscopy (cryo EM) images of tailed bacteriophages epsilon15 (Jiang et al., 2006) and P22 (Lander et al., 2006), alternative maximum likelihood reconstruction and resolution calculation methodologies are described and demonstrated on the same P22 images used in Lander et al. (2006). Maximum likelihood dates back to the early 1900s (Lehmann and Casella, 1998, Section 10.1, p. 515) but continues as an important method for deriving statistical estimators in structural biology (e.g., Blanc et al. (2004) and McCoy et al. (2005) in crystallography and Scheres et al. (2007) and Singh et al. (2004) in cryo EM). In comparison with the reconstruction method described in Lander et al. (2006), the chief advantage of the reconstruction method described in the present paper is its *ab initio* character which manifests itself in two main differences: First, it is not necessary to have a 3-D structure of the tail machine before determining the 3-D structure of the tailed phage. Second, although the 6-fold symmetric tail machine is attached at a 5-fold symmetry axis of the capsid, it is not necessary to specify the rotational position of the tail machine relative to the symmetry axes of the capsid; Instead, all possible rotations, a range of 12 degrees (please see Supplemental material, Section L), are considered by the reconstruction method and the best is selected. Similar to the results of Lander et al. (2006), the portal end of the tail shows approximate 12-fold rotational symmetry even though no such symmetry was imposed.

The approach of this paper can be applied to any tailed bacteriophage for which an icosahedrally symmetric structure can be determined. If the tail is long and flexible, only the proximal part of the tail will be resolved in the 3-D reconstruction. More generally, the approach can probably be applied to viruses where the infection process results in a distinguished attachment site on the surface of the virus which replaces the role of the tail. Such problems are of current interest, e.g., in the case of polio virus, Bubeck et al. (2005) and Zhang et al. (2008). Finally, the methods

* Corresponding author. Fax: +1 607 254 3508.
E-mail addresses: cprust@ll.mit.edu (C.J. Prust), pd83@cornell.edu (P.C. Doerschuk), glander@scripps.edu (G.C. Lander), jackj@scripps.edu (J.E. Johnson).
[1] C.J.P. was with Purdue University, Department of Electrical Engineering and Computer Science, then with Lincoln Laboratory, Massachusetts Institute of Technology, and now with Milwaukee School of Engineering, Department of Electrical and Computer Engineering, 1025 N. Broadway, Milwaukee, WI 53202-3109, USA.
[2] P.C.D. was with Purdue University, School of Electrical and Computer Engineering; and is now with Department of Biomedical Engineering and School of Electrical and Computer Engineering, Cornell University, 305 Phillips Hall, Ithaca, NY 14853-5401, USA.
[3] G.C.L. and J.E.J. are with Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037, USA.

1047-8477/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2009.04.013

used in this paper show how maximum likelihood approaches can be used for complicated structures by assembling the structure out of parts and estimating parameters that determine the structure of each part and the relative locations and orientations of the parts. The computations required in this approach are moderately extensive, e.g., best done on a set of tens of PCs. However, the computations are much less extensive than would be required for a statistical *ab initio* 3-D reconstruction of the infectious bacteriophage (i.e., capsid, tail, and genome) since such a reconstruction would be of a particle without any symmetry. Because the approach of this paper is based on combining an icosahedrally symmetric reconstruction with an ξ-fold symmetric tail reconstruction, not all possible distortions of the capsid to accommodate the tail can be represented. However, to the extent that the distortions of the capsid which allow the joining of the tail are ξ-fold symmetric, then the proximal part of the tail reconstruction will include those distortions so that the sum of the icosahedrally symmetric reconstruction and the ξ-fold symmetric tail reconstruction will accurately reflect the structure of the tailed bacteriophage.

Relative to standard Fourier Shell Correlation (FSC) calculations (please see van Heel and Schatz (2005) for new ideas on setting FSC thresholds and an extensive bibliography of FSC investigations containing 26 entries dating back to initial papers such as Frank et al. (1981) and Saxton et al. (1982)), the resolution methodology described in the present paper has several potentially attractive features: (1) It is linked to the maximum likelihood criteria used to determine the 3-D reconstruction algorithm. (2) It is possible to measure resolution independent of orientation, as is appropriate for spherical objects, or with respect to translations along and rotations around a specific axis, as may be natural for a cylindrical object such as the tail of a phage. (3) It is not necessary to perform two reconstruction calculations each with half of the entire data set. (4) It provides a probability of correctness, i.e., the answer is of the form that resolution is greater than a particular number with a certain probability.

The reconstruction method has three phases: (1) Use a standard cryo EM reconstruction algorithm to compute an icosahedrally symmetric reconstruction of the tailed phage. (2) Use the reconstruction of Phase (1) to determine origin location and projection orientation for each image by quadratic correlation. The projection orientation is only determined up to one of the 60 rotations of the icosahedral group, since the reconstruction of Phase (1) has icosahedral symmetry. (3) Use a mathematical description of the tail, the capsid reconstruction of Phase (1), the origin locations and projection orientations (up to a rotation from the icosahedral group) of Phase (2), and the maximum likelihood criteria to determine a 3-D reconstruction of the entire tailed phage. While the details of Phases (1) and (2) are described in the numerical results (Section 3), the major portion of the reconstruction part of the present paper concerns Phase (3) (Section 2).

The resolution method has two phases: (1) Compute the Hessian of the log likelihood at the maximum likelihood parameter estimates, i.e., compute the matrix of mixed second-order partial derivatives of the log likelihood with respect to the parameters evaluated at the particular vector of parameters that maximizes the likelihood. As is described in Section 4.1, the parameter estimation error, i.e., the difference between the true value of the parameter vector and the value determined by the maximum likelihood estimator, is approximately Gaussian in distribution and the negative of the inverse of this Hessian is approximately the parameter estimation error covariance matrix. (2) As is described in Section 4.4, use a Monte Carlo procedure to compute many FSC curves where the structures compared by FSC are drawn at random from the multivariate Gaussian probability density function (pdf) determined in Phase (1). From this ensemble of FSC curves, it is possible to compute a histogram which approximates the pdf for the resolution at which FSC first falls below any threshold, where the threshold might depend on the magnitude of the reciprocal space position, as is described in van Heel and Schatz (2005). From this histogram it is possible to compute the probability that the resolution exceeds a particular value. For cylindrical objects, two alternatives to FSC are described which are appropriate for measuring axial and rotational resolution, respectively.

2. Reconstruction method

For further details of the reconstruction method, please see Prust (2006).

2.1. Mathematical model for the phage capsid and tail

Real space coordinates are denoted by $\mathbf{x}$ with rectangular, cylindrical, and spherical coordinates denoted by $(x, y, z)^T$, (r, ϕ, z), and $(|\mathbf{x}|, \theta, \phi)$, respectively. Correspondingly, reciprocal space coordinates are denoted by $\mathbf{k}$, (k_x, k_y, k_z), (k_r, ϕ', k_z), and (k, θ', ϕ'). The capsid electron scattering intensity, denoted by $\rho_c(\mathbf{x})$, is described in a coordinate system in which the rotational symmetry axes intersect at the origin of the coordinate system, the z axis is a 5-fold symmetry axis, and the quadrant of the x–z plane that has $x > 0$ and $z > 0$ includes one of the five 2-fold symmetry axes closest to the positive z axis (Yin et al., 2003; Altmann, 1957; Laporte, 1948). The tail electron scattering intensity, denoted by $\rho_t(\mathbf{x})$, is described in a coordinate system in which the long axis of the tail is the z axis of the coordinate system and the end of the tail that attaches to the capsid is the end at more positive z value. The entire particle is described in the same coordinate system as the capsid. The electron scattering intensity of the complete particle, denoted by $\rho(\mathbf{x})$, is therefore

$$\rho(\mathbf{x}) = \rho_c(\mathbf{x}) + \rho_t(\mathbf{x} - \boldsymbol{\delta}) \tag{1}$$

where

$$\boldsymbol{\delta} = (0, 0, z_t)^T. \tag{2}$$

It is not necessary to consider rotation of the tail when attaching the tail to the capsid because the mathematics to be used in the sequel can represent the tail in any rotation around the z axis. It will be convenient to assume that the tail is length-centered in the tail coordinates, i.e., $\rho_t(\mathbf{x})$ is nonzero only in a symmetric region $-z_0/2 \leqslant z \leqslant +z_0/2$ $(z_0 > 0)$ in which case $z_t \leqslant 0$.

Because of the cylindrical shape of the tail and the rotational symmetry of the tail around its long axis, a cylindrical coordinate system is used. Because the tail is by definition periodic with period 2π in ϕ, $\rho_t(\mathbf{x})$ can be expressed as a Fourier series with radially- and axially-dependent weights denoted by $c_l(r, z)$. Assuming that the tail has a known maximum radius, denoted by R_+, it follows that the radial dependence of the weights can be expanded in a sum of weighted Bessel functions in r. The axial dependence of the weights can be expanded as a sum of weighted complex exponentials in z. Assume that the tail has rotational symmetry of order ξ ($\xi = 1$ is permitted). It follows from Eqs. 185, 174, 180, and 92 in the Supplemental material that

$$\rho_t(r, \phi, z) = \sum_{l=-\infty}^{+\infty} \sum_{p=1}^{\infty} \sum_{n=-\infty}^{+\infty} c_{l,p,n} q(z) f_n(z) h_{l,p}(r) \exp(il\xi\phi) \tag{3}$$

where $c_{l,p,n}$ are unknown complex-valued coefficients,

$$q(z) = \begin{cases} 1, & |z| \leqslant z_0/2 \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

$$f_n(z) = \exp\left(i(2\pi/z_0)nz\right), \tag{5}$$

$$h_{l,p}(r) = \begin{cases} 0, & r \geqslant R_+ \\ J_{|l\xi|}(\gamma_{|l\xi|,p} r/R_+), & r < R_+ \end{cases}, \tag{6}$$

$J_l(x)$ is the lth Bessel function of the first type, and $\gamma_{l,p}$ is the pth zero of $J_l(x)$. Since $\rho(\mathbf{x})$ is real, it follows by Hermitian symmetry (see Eq. 179 in the Supplemental material) that $c_{l,p,n}$ has the symmetry

$$c_{-l,p,-n} = c^*_{l,p,n}. \tag{7}$$

Note that Eq. (7) implies that

$$\Im[c_{0,p,0}] = 0 \tag{8}$$

where $\Im$ indicates the imaginary part. The reciprocal-space representation of the electron scattering intensity $\rho_t(\mathbf{x})$ is denoted by $P_t(\mathbf{k})$ and is the 3-D Fourier transform of $\rho_t(\mathbf{x})$ which is defined by

$$P_t(\mathbf{k}) = \int \rho_t(\mathbf{x}) \exp(-i2\pi\mathbf{k}^T\mathbf{x})\mathrm{d}\mathbf{x}. \tag{9}$$

It then follows from Eqs. 202, 203, 198, and 118 in the Supplemental material that

$$P_t(\mathbf{k}) = \sum_{l=-\infty}^{+\infty}\sum_{p=1}^{\infty}\sum_{n=-\infty}^{+\infty} L_{t(k_r,\phi',k_z),(l,p,n)} c_{l,p,n} \tag{10}$$

where

$$L_{t(k_r,\phi',k_z),(l,p,n)} = Q(k_z - n/z_0)\exp(il\xi(\phi' - \pi/2))H_{l,p}(k), \tag{11}$$

$$Q(k_z) = z_0 \mathrm{sinc}(k_z z_0), \tag{12}$$

$$H_{l,p}(k_r) = \frac{R_+^2 \gamma_{|l\xi|,p} J_{l\xi-1}(\gamma_{|l\xi|,p}) J_{|l\xi|}(2\pi k_r R_+)}{(2\pi k_r R_+)^2 - \gamma^2_{|l\xi|,p}}, \tag{13}$$

and $\mathrm{sinc}(z) = \sin(\pi z)/(\pi z)$.

The reciprocal-space representation of the electron scattering intensity of the capsid (complete particle), i.e., of $\rho_c(\mathbf{x})$ $[\rho(\mathbf{x})]$, is denoted by $P_c(\mathbf{k})$ $[P(\mathbf{k})]$ and is the 3-D Fourier transform of $\rho_c(\mathbf{x})$ $[\rho(\mathbf{x})]$ where the 3-D Fourier transform is defined by Eq. (9). Eq. (1) implies that

$$P(\mathbf{k}) = P_c(\mathbf{k}) + \exp(-i2\pi\mathbf{k}^T\boldsymbol{\delta})P_t(\mathbf{k}). \tag{14}$$

Results related to the icosahedral average of the complete particle (capsid plus tail) are necessary in the approach of this paper. The icosahedral group has $N_g = 60$ operations each of which is a rotation that can be expressed (for a specific coordinate system) as a 3×3 matrix. For $\beta \in \{0,\ldots,N_g-1\}$, let $S_\beta \in \mathbb{R}^{3\times 3}$ be the matrices which, since they are rotation matrices, satisfy $S_\beta^{-1} = S_\beta^T$ and $\det S_\beta = +1$. If a function f is rotated to yield a function f' and the rotation is described by the rotation matrix R then the definition used in this paper is that $f'(\mathbf{x}) = f(R^{-1}\mathbf{x})$. With these preliminary results, the icosahedral average of the complete particle, denoted by $\bar{\rho}(\mathbf{x})$, is

$$\bar{\rho}(\mathbf{x}) \doteq \frac{1}{N_g}\sum_{\beta=0}^{N_g-1}\rho(S_\beta^{-1}\mathbf{x}) \tag{15}$$

$$= \frac{1}{N_g}\sum_{\beta=0}^{N_g-1}\left[\rho_c(S_\beta^{-1}\mathbf{x}) + \rho_t(S_\beta^{-1}\mathbf{x} - \boldsymbol{\delta})\right] \tag{16}$$

$$= \rho_c(\mathbf{x}) + \frac{1}{N_g}\sum_{\beta=0}^{N_g-1}\rho_t(S_\beta^{-1}\mathbf{x} - \boldsymbol{\delta}) \tag{17}$$

since $\rho_c(\mathbf{x})$ has icosahedral symmetry, i.e., $\rho_c(S_\beta^{-1}\mathbf{x}) = \rho_c(\mathbf{x})$ for $\beta \in \{0,\ldots,N_g-1\}$. The icosahedral average, $\bar{\rho}(\mathbf{x})$, also has icosahedral symmetry, i.e., $\bar{\rho}(S_\beta^{-1}\mathbf{x}) = \bar{\rho}(\mathbf{x})$ for $\beta \in \{0,\ldots,N_g-1\}$. Let $\overline{P}(\mathbf{k})$ be the reciprocal space representation of $\bar{\rho}(\mathbf{x})$. Eq. (17) implies that

$$\overline{P}(\mathbf{k}) = P_c(\mathbf{k}) + \frac{1}{N_g}\sum_{\beta=0}^{N_g-1}\exp(-i2\pi\mathbf{k}^T S_\beta\boldsymbol{\delta})P_t(S_\beta^{-1}\mathbf{k}). \tag{18}$$

As is described in Supplemental material Section B.6, the mathematics used in this paper does not uniquely represent the tail since if $c_{l,p,n}$ represents the tail $\rho_t(\mathbf{x})$ then $\exp(-il\phi_0)c_{l,p,n}$ represents the same tail rotated around the z axis by the angle ϕ_0 where $\phi_0 \in [0, 2\pi)$ is arbitrary. However, when the tail is attached to the capsid, only 5 of these rotations are equivalent for the combination of capsid and tail since only under the 5 rotations $\phi_0 \in \{n2\pi/5 : n \in \{0,\ldots,4\}\}$ is the capsid unaltered since the z axis is a 5-fold symmetry axis of the capsid.

2.2. Mathematical model for the image formation process and the difference image

A standard image formation equation is used. Let $\sigma_i(\boldsymbol{\chi})$ be the ith real-space image and $\Sigma_i(\boldsymbol{\kappa})$ be the corresponding reciprocal space image which is its 2-D Fourier transform defined analogously to Eq. (9) where $\boldsymbol{\chi} \in \mathbb{R}^2$ and $\boldsymbol{\kappa} \in \mathbb{R}^2$ are the 2-D coordinate vectors in real and reciprocal space, respectively, where $\kappa \doteq |\boldsymbol{\kappa}|$ and $\chi \doteq |\boldsymbol{\chi}|$. Let $\boldsymbol{\chi}_{0,i}$ be the offset between the location of the particle's center in the ith image and the center of the ith image. Let $G(\kappa)$ be the contrast transfer function (CTF) (Baker et al., 1999, p. 873; Scherzer, 1949). Let R_i be the rotation matrix that describes the orientation of the particle in the microscope or, equivalently, the projection orientation. Then (Erickson, 1973, Eq. 11c; Yin et al., 2003, Eq. 10),

$$\Sigma_i(\boldsymbol{\kappa}) = \exp(-i\boldsymbol{\kappa}^T\boldsymbol{\chi}_{0,i})G(\kappa)P(R_i^{-1}(\boldsymbol{\kappa}^T,0)^T). \tag{19}$$

As is described in Section 1, the approach of this paper is based on difference images where the difference is between the experimental image and the predicted image where the prediction is based on an icosahedrally symmetric reconstruction of the complete particle. As a part of the reconstruction, estimates are made of the origin offset for each image, the particle orientation for each image, and the icosahedrally symmetric electron scattering intensity. In this paper, the following assumptions are made:

1. The origin offset estimate, denoted by $\hat{\boldsymbol{\chi}}_{0,i}$, is exact, i.e., $\hat{\boldsymbol{\chi}}_{0,i} = \boldsymbol{\chi}_{0,i}$.
2. The estimate of the rotation matrix describing the particle orientation, denoted by $\widehat{R}_i$, is exact up to a rotation from the icosahedral group, i.e.,
$$\widehat{R}_i = R_i S_{\beta_i}. \tag{20}$$
3. The estimate of the icosahedrally averaged electron scattering intensity of the complete particle, denoted by $\hat{\rho}(\mathbf{x})$, is exact, i.e., $\hat{\rho}(\mathbf{x}) = \bar{\rho}(\mathbf{x})$.

A predicted image is needed in order to form the difference image and the natural predicted image, denoted by $\widehat{\Sigma}_i(\mathbf{k})$, is

$$\widehat{\Sigma}_i(\boldsymbol{\kappa}) = \exp(-i\boldsymbol{\kappa}^T\hat{\boldsymbol{\chi}}_{0,i})G(\kappa)\widehat{P}(\widehat{R}_i^{-1}(\boldsymbol{\kappa}^T,0)^T) \tag{21}$$

$$= \exp(-i\boldsymbol{\kappa}^T\boldsymbol{\chi}_{0,i})G(\kappa)\overline{P}\left((R_iS_{\beta_i})^{-1}(\boldsymbol{\kappa}^T,0)^T\right) \tag{22}$$

$$= \exp(-i\boldsymbol{\kappa}^T\boldsymbol{\chi}_{0,i})G(\kappa)\overline{P}\left(S_{\beta_i}^{-1}\left(R_i^{-1}(\boldsymbol{\kappa}^T,0)^T\right)\right) \tag{23}$$

$$= \exp(-i\boldsymbol{\kappa}^T\boldsymbol{\chi}_{0,i})G(\kappa)\overline{P}\left(R_i^{-1}(\boldsymbol{\kappa}^T,0)^T\right) \tag{24}$$

since $\overline{P}(\mathbf{k})$ has icosahedral symmetry.

The ith difference image, denoted by $\Delta_i(\boldsymbol{\kappa};R_i)$, is

$$\Delta_i(\boldsymbol{\kappa};R_i) = \Sigma_i(\boldsymbol{\kappa}) - \widehat{\Sigma}_i(\boldsymbol{\kappa}) \tag{25}$$

$$= \exp(-i2\pi\boldsymbol{\kappa}^T\boldsymbol{\chi}_{0,i})G(\kappa)\left[P(R_i^{-1}(\boldsymbol{\kappa}^T,0)^T) - \widehat{P}(R_i^{-1}(\boldsymbol{\kappa}^T,0)^T)\right] \tag{26}$$

$$= \exp(-i2\pi\boldsymbol{\kappa}^T\boldsymbol{\chi}_{0,i})G(\kappa)\Bigg[\exp(-i2\pi(\boldsymbol{\kappa}^T,0)R_i\boldsymbol{\delta})P_t(R_i^{-1}(\boldsymbol{\kappa}^T,0)^T) - \frac{1}{N_g}\sum_{\beta=0}^{N_g-1}\exp(-i2\pi(\boldsymbol{\kappa}^T,0)R_iS_\beta\boldsymbol{\delta})P_t((R_iS_\beta)^{-1}(\boldsymbol{\kappa}^T,0)^T)\Bigg]. \tag{27}$$

Since the origin offset is known by Assumption 1, it is natural to assume that the boxed images are shifted such that the origin offset in

the shifted image is zero. In that case the factor $\exp(-i2\pi\boldsymbol{\kappa}^T\boldsymbol{\chi}_{0,i})$ has value 1. The icosahedral averaging that creates $\overline{P}(\mathbf{k})$ causes the single tail to be replicated $N_g = 60$ times in 12 groups of 5 with one group at each 5-fold symmetry axis where each replication is at 1/60 the scattering intensity of the true tail. We refer to these replicated low-scattering-intensity tails as "ghosts". Let $\Sigma_g(\boldsymbol{\kappa}; R_i)$ denote the ghost tail reciprocal-space image with definition

$$\Sigma_g(\boldsymbol{\kappa}; R_i) = \frac{1}{N_g}\sum_{\beta=0}^{N_g-1} \exp(-i2\pi(\boldsymbol{\kappa}^T, 0)R_iS_\beta\boldsymbol{\delta})P_t((R_iS_\beta)^{-1}(\boldsymbol{\kappa}^T, 0)^T). \quad (28)$$

Note that

$$\Sigma_g(\boldsymbol{\kappa}; R_iS_\beta) = \Sigma_g(\boldsymbol{\kappa}; R_i) \quad (29)$$

for all $\beta \in \{0, \ldots, N_g - 1\}$ because the $\{S_\beta\}$ form a multiplicative group. Using this definition, the difference image can be written as

$$\Delta_i(\boldsymbol{\kappa}; R_i) = \exp(-i2\pi\boldsymbol{\kappa}^T\boldsymbol{\chi}_{0,i})G(\kappa)\Big[\exp(-i2\pi(\boldsymbol{\kappa}^T, 0)R_i\boldsymbol{\delta})P_t(R_i^{-1}(\boldsymbol{\kappa}^T, 0)^T) - \Sigma_g(\boldsymbol{\kappa}; R_i)\Big] \quad (30)$$

which describes the difference image as the superposition of the true tail and the ghost tails.

2.3. Statistical model for the noisy images

The statistical model falls within the general class of models described in Doerschuk and Johnson (2000) and Yin et al. (2003) and only the main characteristics are briefly summarized here. The central feature, and the key difference from the numerical examples described in Doerschuk and Johnson (2000) and Yin et al. (2003), is that the orientations of the particles are still independent random variables but the probability density functions (pdfs) for these random variables are not identical. In particular, each particle has its pdf concentrated on the $N_g = 60$ icosahedrally related orientations that are the outcome of orienting the image of a nonsymmetrical particle with predicted images of an icosahedrally symmetric particle. As is described in Section 2.2, it is assumed that the difference boxed images are shifted if necessary so that the center of the capsid is projected to the center of the image. Therefore, there is no uncertainty in the location of the center of the capsid in the image and so, in the notation of Doerschuk and Johnson (2000) and Yin et al. (2003), $\boldsymbol{\chi}_{i,j} = (0,0)^T$. In the calculations described in this paper, we assume that there is only one class of capsid and one class of tail. Therefore, in the notation of Doerschuk and Johnson (2000) and Yin et al. (2003), $N_\eta = 1$. This restriction could be removed. The reciprocal space image is assumed to be corrupted by additive zero-mean white Gaussian noise with known variance σ^2. The variance is, in fact, estimated from the images in a preliminary calculation.

Removing the one class restriction would require a precise definition of how multiple classes occur. For instance, if the capsid has only one class but the tail has multiple classes, then the following generalization would be natural: (1) Merge all of the data to compute an icosahedrally symmetric reconstruction. (2) Use the icosahedrally symmetric reconstruction to compute difference images. (3) Use a multiclass generalization of the algorithm described in this paper, exactly following the multiclass algorithm of Doerschuk and Johnson (2000), to reconstruct multiple tail structures. Alternatively, if the capsid has multiple classes but only one possible tail exists, then the following generalization would be natural: (1′) Use the multiclass algorithm of Doerschuk and Johnson (2000), to reconstruct multiple icosahedrally symmetric structures and label each image with its estimated class. (2′) Based on the estimated class label, compute difference images using the appropriate icosahedrally symmetric reconstruction. (3′) Apply the algorithm described in this paper to the difference images to determine a reconstruction of the single class of tail. Finally, if there are multiple classes of capsid and multiple classes of tail and all possible mixtures of capsid and tail are present in the data, then a combination of these two generalizations would be necessary, in particular, (1′), (2′), and (3).

Let the ith difference image be arrayed in a vector denoted by y_i. Let the unknown coefficients, $c_{l,p,n}$, be arrayed in a vector denoted by c. Let the additive pixel noise for the ith difference image be arrayed in a vector denoted by w_i which is, therefore, Gaussian with mean 0 and covariance $Q_i = \sigma^2 I_{N_y}$ where N_y is the number of pixels in the image. From Eq. (30), the ith difference image depends linearly on $P_t(\mathbf{k})$. From Eq. (10), $P_t(\mathbf{k})$ depends linearly on the unknown coefficients $c_{l,p,n}$ (which are the elements of the vector c). Therefore, there is a matrix, which is denoted by L_Δ, that relates the ith difference image to the unknown coefficients $c_{l,p,n}$ (which are the elements of the vector c). (In Eq. (11) the elements of the simpler matrix relating $P_t(\mathbf{k})$ to c, where c has elements $c_{l,p,n}$, is shown explicitly). The matrix depends on the identity of the image, i.e., on i, because it depends on the orientation of the particle, i.e., on R_i. The matrix also depends on random variables, such as which of the $N_g = 60$ icosahedrally related orientations is present, and the collection of such random variables for the ith image is denoted by z_i. In the calculations of this paper, the only random variables on which the matrix depends are the orientations and therefore, the integrals in Eqs. (35)–(37) are actually discrete sums though in more general problems the matrix could depend on additional variables such as translations in which case the integrals would not reduce to discrete sums. The conclusion of this paragraph is that there is an equation,

$$y_i = L_\Delta(i, z_i)c + w_i, \quad (31)$$

analogous to Doerschuk and Johnson (2000, p. 1718) and Yin et al. (2003, Eq. 19, p. 31), which describes the entire imaging system.

2.4. Maximum likelihood reconstruction method and expectation-maximization algorithm

Use of the maximum likelihood estimation ideas and formulas of Doerschuk and Johnson (2000) and Yin et al. (2003) allows a reconstruction of the tail by estimating values for the unknown $c_{l,p,n}$ coefficients which in turn specify the tail through Eq. (3). With the value of the matrix L_Δ and the pdfs for the random variables on which L_Δ depends, both new in this paper for this application, the algorithm is as follows. First, pre-compute the following quantities:

$$a_i(y_i; z_i) = \sum_{j=1}^{N_T(i)} \left\{ \ln\left[(2\pi)^{N_y/2}\sqrt{\det Q_{ij}(z_i)}\right] + \frac{1}{2}y_{ij}^T Q_{ij}^{-1}(z_i)y_{ij} \right\} \quad (32)$$

$$b_i(y_i; z_i) = \sum_{j=1}^{N_T(i)} L_\Delta^T(i, j, z_i)Q_{ij}^{-1}(z_i)y_{ij} \quad (33)$$

$$D_i(z_i) = \sum_{j=1}^{N_T(i)} L_\Delta^T(i, j, z_i)Q_{ij}^{-1}(z_i)L_\Delta(i, j, z_i) \quad (34)$$

where, for the ith virion, these equations allow $N_T(i)$ tilt images, denoted by y_{ij}, to be processed and allow the pixel noise covariance Q_{ij} and L_Δ to depend on both the virion index i and the tilt series index j. These quantities allow rapid evaluation of a Gaussian pdf which is the product of $N_T(i)$ independent Gaussian pdfs with means $L_\Delta c$ the covariances Q_{ij} which is the pdf needed in the expectation of the expectation-maximization algorithm (z_i are the so-called nuisance parameters of the algorithm). Second, determine an initial condition for the values of the $c_{l,p,n}$ coefficients. In the numerical results presented in Section 3, that initial condition is random as is described in Section 3.1. Third, starting at this initial condition, iterate the following actions until the values of the $c_{l,p,n}$ coefficients have converged.

1. Compute

$$\gamma_i(c_0, y_i) = \int_{z_i} p(y_i|z_i, c_0)p(z_i)dz_i \tag{35}$$

$$\beta_i(c_0, y_i) = \int_{z_i} b_i(y_i; z_i)p(y_i|z_i, c_0)p(z_i)\mathrm{d}z_i \tag{36}$$

$$\Delta_i(c_0, y_i) = \int_{z_i} D_i(z_i)p(y_i|z_i, c_0)p(z_i)dz_i \tag{37}$$

where c_0 is the current value of the vector constructed from the $c_{l,p,n}$ coefficients.

2. Combine these results to compute

$$g = \sum_{i=1}^{N_v} \frac{1}{\gamma_i(c_0, y_i)} \beta_i(c_0, y_i) \tag{38}$$

$$F = \sum_{i=1}^{N_v} \frac{1}{\gamma_i(c_0, y_i)} \Delta_i(c_0, y_i). \tag{39}$$

where N_v is the number of virions, i.e., the number of images if each virion has a tilt series containing only one image.

3. Solve the linear system

$$Fc = g \tag{40}$$

for the vector c which is the new value of the vector constructed from the $c_{l,p,n}$ coefficients.

Action 1 evaluates the expectations of the expectation-maximization algorithm while Actions 2 and 3 evaluate the maximization of the expectation-maximization algorithm where the function that must be maximized turns out to be a quadratic form in c so the location of the maximum can be computed by solving a linear system, i.e., Eq. (40). These calculations generalize immediately to the case where each virion is from one of a finite number of different classes and the class label for the virion shown in a particular image is not known (Doerschuk and Johnson, 2000; Yin et al., 2003). Because the approach of this paper is based on difference images (Eq. 30), multiclass algorithms require care that the image and the prediction of the image used to compute the difference image come from the same class and three multiclass situations and algorithms are described in the second paragraph of Section 2.3.

2.5. Parallel computation methods

The calculations implied by the algorithm described in this paper are large and so efficiency and parallel computing on a cluster of commodity PCs has been critical. Prust (2006, Section 2.2.6, pp. 19–21) provides methods based on Eq. (7) that allow fast computation of L_t and therefore of L_Δ. The parallel software is based on modifications of the software described in Zheng (2002). The modifications are the new L_Δ and the new pdfs for the random variables z_i on which L_Δ depends. (In the calculations reported in this paper, the z_i are the orientation parameters for each image and take 1 of 60 values for each image but the possible values differ from image to image. For each image, the pdf on the z_i is uniform over the possible values for that image). With regard to the pdfs, the key modification, a generalization, is to provide a different pdf for each difference image. This has major implications for the storage footprint of the software, specifically the storage required for the D matrices (Eq. 34), which will be returned to in Section 5.

3. Numerical results 1: The reconstruction of P22

For further details concerning the reconstruction of P22, please see Prust (2006).

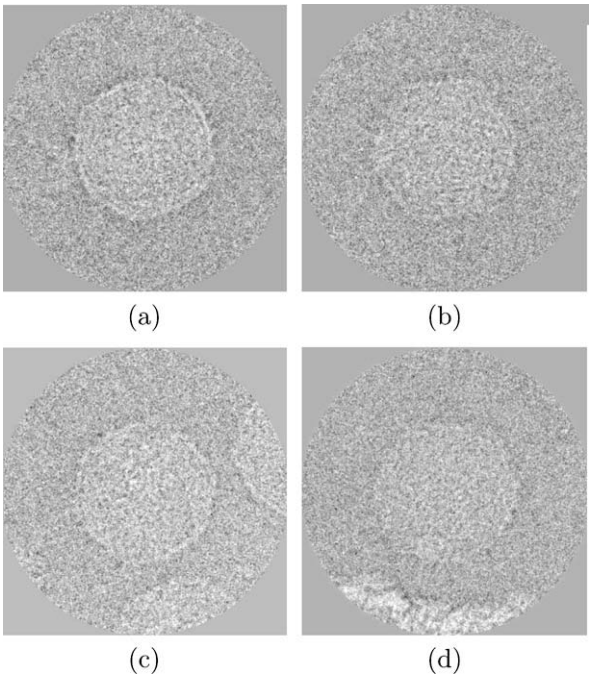


Fig. 1. Examples of accepted and rejected P22 images. (a and b) Show a pair of P22 images that were included in the 3-D reconstruction while (c and d) show images that were rejected.

3.1. Practical issues

The boxed images were hand selected based on absence of adjacent particles, broken particles, or junk in the image. No preference was given to images based on the visibility of the tail. Fig. 1[4] shows samples of both accepted images and discarded images. No masking of the images was performed because of the difficulty of designing a procedure that did not mask side-pointing tails.

The selected images were oriented, modulo a rotation from the icosahedral group, and centered by quadratic correlation. A library of 5000 reference images with projection directions uniformly distributed through the asymmetric unit of the icosahedral group was computed by Spider (Frank et al. (1996)) using command `PJ3Q` from a high-resolution icosahedrally symmetric reconstruction of the capsid of P22 (Lander et al. (2006)). The rest of the processing was performed in Matlab.[5] For each boxed image, estimates of the orientation (modulo a rotation from the icosahedral group) and origin offset are computed by maximizing the normalized quadratic correlation between the boxed image at a variety of shifts and the reference images each with a different orientation. Let $\sigma_i^{\mathrm{ref}}(\boldsymbol{\chi})$ denote the ith reference image and hence the ith orientation and let $\boldsymbol{\chi}_0$ denote the origin offset. Let $\sigma(\boldsymbol{\chi})$ denote one of the boxed images. The estimates for that boxed image are

$$\hat{i}, \hat{\boldsymbol{\chi}}_0 = \underset{i\in\{1,\ldots,5000\}, \boldsymbol{\chi}_0\in\{(m_1\Delta, m_2\Delta)^T : m_1, m_2\in\{-m_{\max},\ldots,+m_{\max}\}\}}{\arg\max} J_1(\sigma, \sigma_i^{\mathrm{ref}}, \boldsymbol{\chi}_0) \tag{41}$$

where

$$J_1(\sigma, \sigma_i^{\mathrm{ref}}, \boldsymbol{\chi}_0) = \frac{\sum_{\boldsymbol{\chi}} \sigma(\boldsymbol{\chi} - \boldsymbol{\chi}_0)\sigma_i^{\mathrm{ref}}(\boldsymbol{\chi})}{\sqrt{\left[\sum_{\boldsymbol{\chi}} \sigma^2(\boldsymbol{\chi})\right]\left[\sum_{\boldsymbol{\chi}} \left[\sigma_i^{\mathrm{ref}}(\boldsymbol{\chi})\right]^2\right]}}, \tag{42}$$

[4] All 2-D images (boxed images, cross sections, *etc.*) in this paper were made with Matlab (http://www.mathworks.com/) and all 3-D surface plots were made with the Spider/Web system (Frank et al., 1996).

[5] http://www.mathworks.com/.

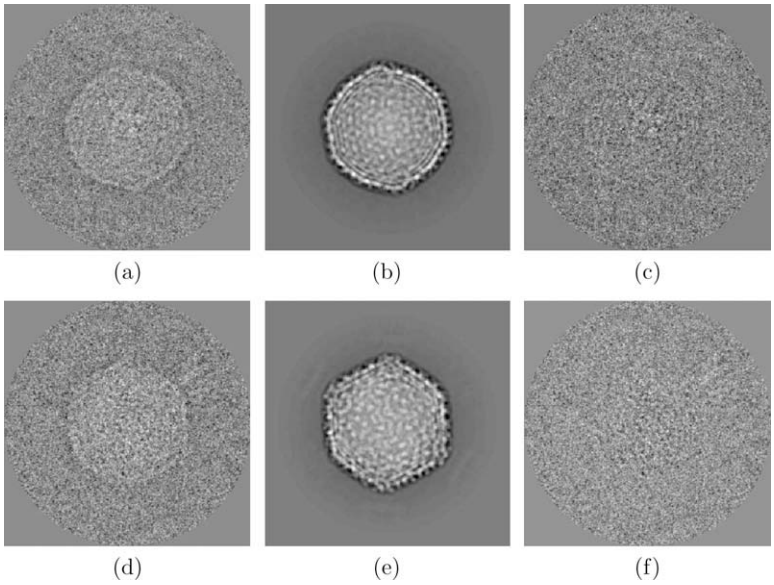


Fig. 2. Generation of difference images for the P22 reconstruction. (a and d) Show a pair of raw P22 images. (b and e) Show the corresponding reference image from the 5000 image library. (c and f) Show the resulting difference images using the optimal gains.

Δ is the image sampling interval, and $m_{\max} = 2$.

Difference images were computed by subtracting the reference image from the shifted boxed image after computing, by least squares, an optimal gain to apply to the reference image in order to compensate for the unknown scaling of the boxed image. The optimal gain for a particular boxed image, denoted by $\hat{g}$, is the gain that minimizes a cost function, denoted by J_2:

$$\hat{g} = \arg\min_{g} J_2(g, \sigma, \sigma_i^{\text{ref}}, \hat{\boldsymbol{\chi}}_0) \tag{43}$$

where

$$J_2(g, \sigma, \sigma_i^{\text{ref}}, \hat{\boldsymbol{\chi}}_0) = \sum_{\boldsymbol{\chi}} \left[\sigma(\boldsymbol{\chi} - \hat{\boldsymbol{\chi}}_0) - g\sigma_i^{\text{ref}}(\boldsymbol{\chi})\right]^2. \tag{44}$$

The calculation of $\hat{g}$ can be done explicitly in terms of $\sigma, \sigma_i^{\text{ref}}$, and $\hat{\boldsymbol{\chi}}_0$. Fig. 2 shows sample difference images for the P22 reconstruction.

Expectation-maximization is an iterative algorithm that requires an initial condition. As is described in more detail in Supplemental material Section C, the calculations described in this paper used random initial conditions computed in two steps. First, compute pseudo random variables $c'_{l,p,n} \in \mathbb{R}$ from a pdf which is uniform on the interval $[-\omega, +\omega]$ subject to the energy constraint that

$$\omega^2/4 \leqslant \sum_{l=-\infty}^{+\infty} \sum_{p=1}^{\infty} \sum_{n=-\infty}^{+\infty} [c'_{l,p,n}]^2 \leqslant 3\omega^2/2. \tag{45}$$

Second, set $c_{l,p,n}$ equal to $c'_{l,p,n}$ for those values of (l, p, n) where $c_{l,p,n} \in \mathbb{R}$ (please see Eq. 8) and set $c_{l,p,n}$ equal to $c'_{l,p,n} \exp(i\xi_{l,p,n})$ for the remaining values of (l, p, n) where $\xi_{l,p,n}$ are pseudo random variables from a pdf which is uniform on the interval $[0, 2\pi]$. The value of ω was set to $\sqrt{0.00002}$ which yielded satisfactory performance. Because expectation-maximization is guaranteed only to converge to a local maximum of the likelihood function, a multi-start optimization was performed in which 99 initial conditions were tested and the best answer, i.e., the answer with highest likelihood, was retained.

Similar to most cryo EM and X-ray crystallography reconstruction algorithms, the resolution of the reconstruction is increased in a series of steps. Resolution of the model is controlled by truncating the l, p, and n sums in Eq. (3) or equivalently Eq. (10) to $-l_{\max} \leqslant l \leqslant l_{\max}, 1 \leqslant p \leqslant p_{\max}$, and $-n_{\max} \leqslant n \leqslant n_{\max}$. Table 1 lists the values of $l_{\max}$, $p_{\max}$, and $n_{\max}$ for each step and the corresponding number of $c_{l,n,p}$ coefficients, denoted by N_c. The multi-start optimization described in the previous paragraph was used for Step 1 and the answer from Step 1 was the best of the multi-start answers. In Steps 2–6, the single initial condition was always the answer from the previous step augmented with additional $c_{l,p,n}$ coefficients with value 0.

Table 1
Parameters at each step of the reconstruction algorithm. $\pm l_{\max}, p_{\max}$, and $\pm n_{\max}$ are the truncation limits for the infinite sums in the tail model (Eq. 3 or equivalently Eq. 10). N_c is the total number of $c_{l,p,n}$ coefficients.

Step	$l_{\max}$	$p_{\max}$	$n_{\max}$	N_c
1	1	3	1	27
2	1	3	3	63
3	1	5	3	105
4	1	5	5	165
5	1	7	5	231
6	2	7	5	385

In Steps 1 and 2 the difference image is predicted by Eq. (30) which includes both the true tail and the ghost tails. However, in Steps 3–6, in order to save computation, the ghost tails are omitted, i.e., $\Sigma_g(\boldsymbol{\kappa}; R_i)$ is deleted from Eq. (30).

The parameters for the reconstruction were as follows: A total of $N_v = 276$ images were used. Each image had a sampling interval of $\Delta = 4.04$Å/pixel and measured 288×288 pixels. The CTF was unity. The cylinder containing the tail had radius $R_+ = 130$Å and length $z_0 = 380$Å. The symmetry of the tail was $\xi = 6$ fold rotational symmetry. The tail coordinate system was displaced by $z_t = -380$Å from the capsid coordinate system which is the same as the whole-particle coordinate system. (Recall that the tail is nonzero in the tail coordinate system for the region $-z_0/2 \leqslant z \leqslant +z_0/2$ and that the capsid, tail, and whole-particle coordinate systems are described in Eqs. (1) and (2)).

3.2. The 3-D cube

Fig. 3 shows the final answer (i.e., Step 6) of the reconstruction at a single viewing angle but at various contour levels. Fig. 4 shows cross sections through the final answer (i.e., Step 6). Fig. 5 shows two views of the assembled P22 virus structure. Supplemental

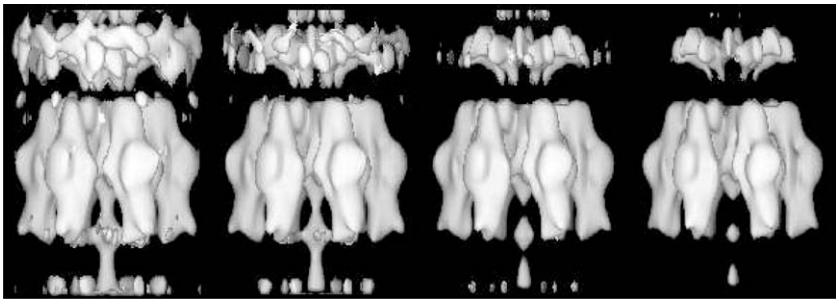


Fig. 3. Final P22 tail reconstruction (i.e., Step 6) rendered at increasing contour levels from left to right.

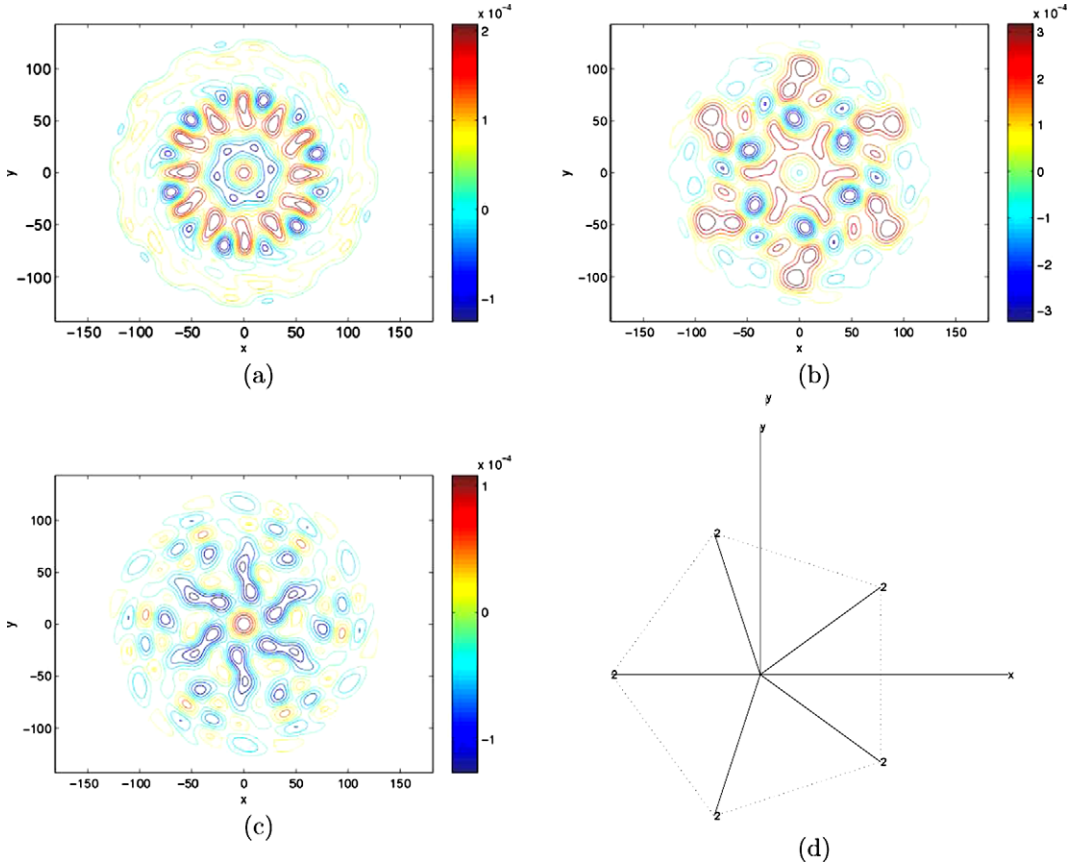


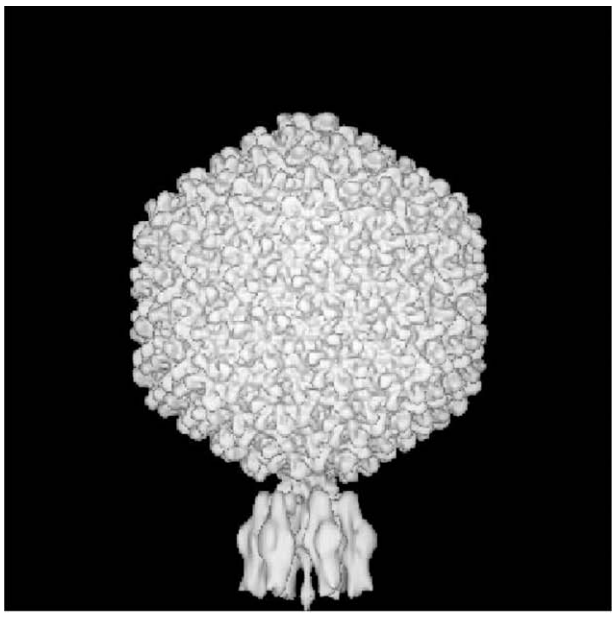
Fig. 4. Cross sectional plots in the x–y plane of the final P22 tail reconstruction (i.e., Step 6). The cross sections are at distances 240, 380, and 530 Å from the center of the capsid in (a–c), respectively, which implies that (a–c) show cross sections of the tail near the portal end of the tail, the midpoint of the tail, and 40 Å proximal to the free end of the tail, respectively. (d) Shows the coordinate system which is described in Section 2.1. In summary, the rotational symmetry axes of the icosahedral symmetry intersect at the origin of the coordinate system, the z axis is a 5-fold symmetry axis, the quadrant of the $x-z$ plane that has $x>0$ and $z>0$ includes one of the five 2-fold symmetry axes closest to the positive z axis (Yin et al., 2003; Altmann, 1957; Laporte, 1948) and the tail extends in the negative z direction. The lines marked "2" are the projections onto the x–y plane of the icosahedral 2-fold symmetry axes that are closest to the negative z axis where the tail is located.

material Section D contains additional figures showing the 3-D real-space reconstruction of the tail. Supplement Fig. 12 shows the resulting structure at each step of the reconstruction process (see Table 1). Supplement Fig. 13 shows the final answer (i.e., Step 6) from various viewing angles. Note that there is no ambiguity in the relative location and orientation of the tail and capsid structures since the tail coordinate system is locked to the capsid coordinate system and because the tail reconstruction algorithm can reconstruct the tail in any rotation as implied by the data. The fact that the tail is not rotationally blurred demonstrates that the tail structure attaches to the capsid in a specific way (i.e., the 6-fold symmetric tail specifically connects to the 5-fold symmetric capsid). The resolution of the reconstruction is presented in Section 5.

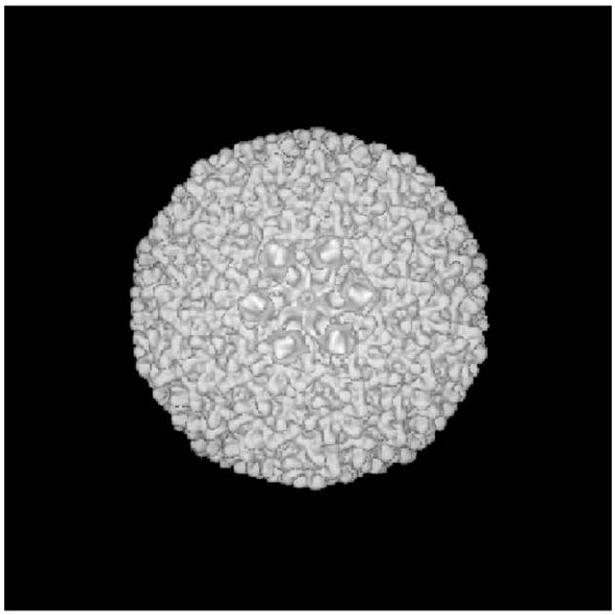
3.3. The angle between a tail spike and the icosahedral symmetry axes

The mathematics used in this paper can represent a ξ-symmetric tail independent of the rotation of the tail around the z axis relative to the capsid. Therefore, the rotational relationship between the protein molecules in the tail and the icosahedral symmetry of the capsid can be determined free of any initial assumptions.

One method to visualize the rotational relationship is to select a region of z values and integrate $\rho(\mathbf{x})$ over this region to create a 2-D averaged cross sectional real-space image of the tail. Assume that the region of z is specified in the tail coordinate system (which is displaced by z_t from the complete particle coordinate system as is shown in Eqs. (1) and (2)) by $-z_0/2 \leqslant z_- \leqslant z \leqslant z_+ \leqslant z_0/2$. The



(a)



(b)

Fig. 5. 3-D reconstruction of the complete P22 virus structure. Side and end-on views are shown in (a and b), respectively. Note that the rotational uncertainty in assembling the capsid and tail structures was uniquely determined by the tail reconstruction algorithm.

integral can be done analytically because the only factor of Eq. (3) that must be integrated is $f_n(z)$ and that integral has the value (please see Supplemental material Section G, Eq. 209)

$$\int_{z=z_-}^{z_+} f_n(z)\mathrm{d}z = \frac{z_0}{\pi n}\exp\left(i\pi n(z_+ + z_-)/z_0\right)\sin\left(\pi n(z_+ - z_-)/z_0\right). \quad (46)$$

Alternatively, a 3-D real-space cube, such as is visualized in Section 3.2, can be summed in the z direction over the appropriate subset of planes to create an averaged 2-D cross sectional real-space image of the tail.

Taking the second approach, Fig. 6 shows averaged cross sections of the real-space 3-D cube visualized in Fig. 3 and Fig. 12(f) (Supplementary material) and Fig. 13 in which the sampling interval is 2 Å in all three coordinates. Specifically, Fig. 6(a) shows the sum of planes between $z_- = 120$Å and $z_+ = 150$Å in the tail coordinate system (-260Å and -230Å in the total particle coordinate) and Fig. 6(b) shows the sum of planes between $z_- = -100$Å and $z_+ = 50$Å in the tail coordinate system (-480Å and -330Å in the total particle coordinate system) where the center of the capsid is at the origin in the total-particle coordinate system. Fig. 6(a and b) correspond to the portal-end and mid-tail portion of the tail structure, respectively. The image shown in Fig. 6(b) shows the 6-fold symmetric locations of the protein molecules that make up the tail as present in the mid-tail portion of the structure. The center of mass of the molecule closest to the x axis in the positive rotational direction is 33.74° from the x axis. Fig. 6(a) shows similar information for the portal end of the structure. Here, in addition to the exact 6-fold symmetry, there is an approximate 12-fold symmetry, also seen in Lander et al. (2006), that was not imposed on the structure. The center of mass of the molecule closest to the x axis in the positive rotational direction is 0.83° from the x axis. As discussed in the final paragraph of Section 2.1, the structure described here is in one of 5 equivalent coordinate systems where the 5 systems are related by rotations by $2\pi n/5$ ($n \in \{0,\ldots,4\}$) around the z axis. If the coordinate system is chosen in order to make these angles as small as possible, which is a way to make the angles unique, then 33.74° becomes 9.74° and 0.83° is unaltered.

The diagrams shown in Fig. 6(c and d) display the relationships between the molecules at the portal-end or the mid-tail portion of the tail structure and the icosahedral symmetries of the capsid structure. Specifically, the diagrams show the x and y axes, the centers of mass of the six (Fig. 6c) or 12 (Fig. 6d) molecules, and the x–y projection of the five 2-fold rotational symmetries of the capsid icosahedral symmetry that lie closest to the negative z axis, i.e., lie closest to the attachment point of the tail. Essentially the same information is provided in Fig. 4B of Lander et al. (2006). Given the small range of angles that are relevant (as is shown in Supplemental material Section L, the range of angles is 6° for 12-fold and 12° for 6-fold) and the inaccuracies of the structures, these angles appear to support those of Lander et al. (2006). Unlike the method of Lander et al. (2006), in which a 3-D structure for the tail machine is attached at a specific arbitrary rotation angle relative to the capsid symmetry axes and then the combined capsid-tail structure is refined, in the approach described in this paper no assumption about this angle is ever made at any stage of the algorithm and so this is an *ab initio* determination of the value of the angle. The reason that no value for this angle is ever assumed is that the mathematics used to represent the tail can represent the tail in any rotational position as is described in the final paragraph of Section 2.1.

At any cross sectional level, by using the maximum likelihood estimation error ideas of Section 4.1 and Monte Carlo ideas analogous to those of Section 4.4 to compute many structures and therefore many averaged cross sections of the tail, it would be possible to determine statistical information about the location of the molecules in the cross section of the tail. For instance, such statistical information might be the 2×2 covariance matrix describing the uncertainty in the center of mass location or a histogram describing the uncertainty in the angle. Given the level of precision with which this angle can be determined, these calculations would not provide substantial additional insight.

4. Resolution methods

A standard definition of resolution is based on the Fourier Shell Correlation (FSC) function (van Heel, 1987, Eq. 2; Harauz and van Heel, 1986, Eq. 17; Baker et al., 1999, p. 879) which

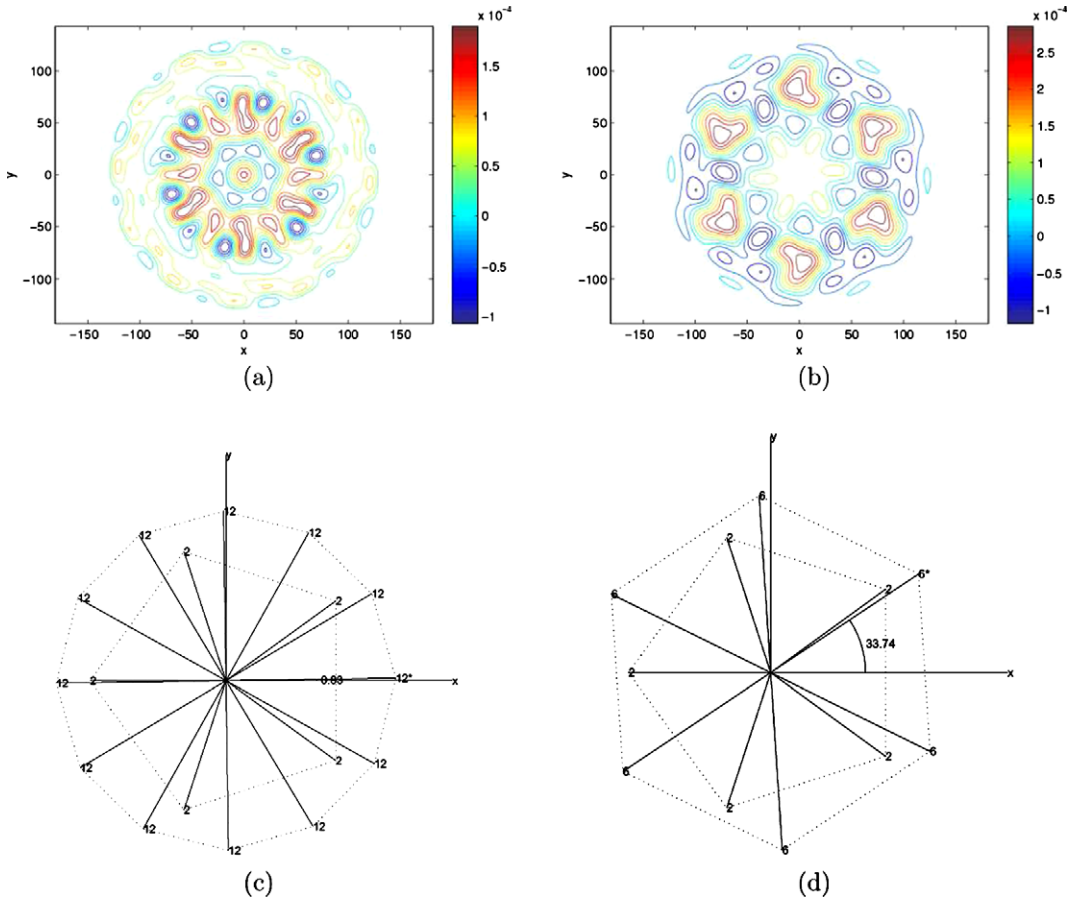


Fig. 6. Averaged cross sections of the tail (a and b) and the relationships between the tail molecules and the icosahedral capsid symmetries (c and d). (a): Averaged cross section near the portal end of the tail, specifically, averaged over distances of 230–260 Å from the center of the capsid. The 6-fold symmetry is exactly present. An approximate 12-fold symmetry is also present. (b): Averaged cross section near the midpoint of the tail, specifically, averaged over distances of 330–480 Å from the center of the capsid. The 6-fold symmetry is exactly present. No approximate 12-fold symmetry is present. (c): An abstraction of (a) showing the centers of mass of the twelve molecules (marked "12") and the projection onto the x–y plane of the five icosahedral 2-fold symmetry axes that are closest to the negative z axis where the tail is located. The angle from the x axis to the closest center of mass in the positive angular direction (marked "12^*") is 0.83°. (d): An abstraction of (b) showing the centers of mass of the six molecules (marked "6") and the projection onto the x–y plane of the five icosahedral 2-fold symmetry axes that are closest to the negative z axis where the tail is located. The angle from the x axis to the closest center of mass in the positive angular direction (marked "6^*") is 33.74°. The coordinate system is described in Section 2.1 and summarized in Fig. 4.

compares the reciprocal-space scattering densities of two structures. The purpose of FSC, as described for instance in Saxton et al. (1982) (starting on p. 131 line 17), is to determine how the image noise effects the 3-D reconstruction where the determination is done in 3-D reciprocal space as a function of the magnitude of the spatial frequency vector. In the standard approach, this determination is made by comparing two 3-D reconstructions made from nonoverlapping subsets of the available images. In the approach described here, this determination is made based on the shape of the likelihood function at the maximum where the shape is measured as the matrix of mixed second-order partial derivatives of the log likelihood function with respect to the parameters evaluated at the parameter values that maximize the likelihood function. In the approach described here, the goal is to compute the probability density functions (pdfs) of certain random variables. We are not able to perform these calculations symbolically. Therefore, we use Monte Carlo methods to compute histograms which approximate the pdfs. So, while the Monte Carlo methods are important to implement our approach, they are not intrinsic to our approach.

Let $P^{\mathrm{a}}(\mathbf{k})$ and $P^{\mathrm{b}}(\mathbf{k})$ be the two reciprocal-space scattering densities to be compared. The FSC function [denoted by $p_{\mathrm{FSC}}(k)$] is a function of the magnitude of the reciprocal-space frequency vector (k) and is defined by

$$p_{\mathrm{FSC}}(k) \doteq \frac{\int P^{\mathrm{a}}(\mathbf{k})\left[P^{\mathrm{b}}(\mathbf{k})\right]^{*} \mathrm{d}\Omega'}{\sqrt{\int |P^{\mathrm{a}}(\mathbf{k})|^2 \mathrm{d}\Omega' \int |P^{\mathrm{b}}(\mathbf{k})|^2 \mathrm{d}\Omega'}} \tag{47}$$

$$= \frac{S^{\Omega'}_{P^{\mathrm{a}},P^{\mathrm{b}}}(k)}{\sqrt{S^{\Omega'}_{P^{\mathrm{a}},P^{\mathrm{a}}}(k) S^{\Omega'}_{P^{\mathrm{b}},P^{\mathrm{b}}}(k)}} \tag{48}$$

where $\mathrm{d}\Omega'$ is integration over the angles of spherical coordinates (i.e., $\int \mathrm{d}\Omega' = \int_{\phi'=0}^{2\pi} \int_{\theta'=0}^{\pi} \sin(\theta') \mathrm{d}\theta' \mathrm{d}\phi'$ where θ' and ϕ' are the angles of spherical coordinates in reciprocal space) and

$$S^{\Omega'}_{P^{\mathrm{a}},P^{\mathrm{b}}}(k) \doteq \int P^{\mathrm{a}}(\mathbf{k})\left[P^{\mathrm{b}}(\mathbf{k})\right]^{*} \mathrm{d}\Omega' \tag{49}$$

for any pair of reciprocal-space functions $P^{\mathrm{a}}(\mathbf{k})$ and $P^{\mathrm{b}}(\mathbf{k})$. Note that $p_{\mathrm{FSC}}(k)$ is real valued because $\rho(\mathbf{x})$ is real valued and that $|p_{\mathrm{FSC}}(k)| \leqslant 1$ by the Cauchy–Schwarz inequality. The two structures, $P^{\mathrm{a}}(\mathbf{k})$ and $P^{\mathrm{b}}(\mathbf{k})$, are often the reconstructions based on even and odd numbered images, respectively. Once the FSC has been computed, the resolution is defined as the smallest value of k such that $p_{\mathrm{FSC}}(k)$ is less than a threshold which may depend on k (van Heel and Schatz, 2005).

Our interest is a resolution measure with the following properties:

1. The resolution measure distinguishes between axial and radial directions because that distinction is intrinsic in the mathematics used in this paper, e.g., the cutoff for the n sum versus the l and p sums in Eq. (3) influence axial versus radial directions.
2. The resolution measure is related to the maximum likelihood criteria used to compute the reconstruction.
3. The resolution measure attaches a probability to its results, analogous to the probability included in tests like t tests.
4. The resolution measure provides the resolution of the reconstruction based on the full set of images rather than by comparing two probably lower resolution reconstructions based on even and odd numbered images, respectively.

In the remainder of this section a statistical version of FSC is described, suitable for separate axial and radial resolution measures, that is based on the general theory of errors in maximum likelihood estimation. As described above for standard FSC, it is still necessary to have a threshold and the ideas described do not include new ideas about the specification of the threshold. However, all current threshold ideas of which we are aware, including k dependent thresholds such as those advocated by van Heel and Schatz (2005), can be used.

4.1. General theory of estimation error covariance for maximum likelihood estimators

The standard theory (Efron and Hinkley, 1978) for the estimator error covariance of a maximum likelihood estimator is described in this section. Let y be the vector of data and c be the vector of unknown parameters. Let the estimate of c, which is a function of y, be denoted by $\hat{c}(y)$. Let the Hessian of the log likelihood function, the matrix of mixed second-order partial derivatives of the log likelihood function, be denoted by $H(c)$ with i,jth element defined by $\partial^2 \ln p(y|c)/\partial c_i \partial c_j$ where $p(y|c)$ is the conditional probability density function on the data y given the unknown parameters c. Let c_* be the true value of the parameters. The key result (Efron and Hinkley, 1978) is that the estimation error, $\hat{c}(y) - c_*$, is approximately Gaussian distributed with mean vector $\mathbf{0}$ and covariance matrix $-[H(\hat{c}(y))]^{-1}$.

4.2. A simpler example

In this section the general theory of Section 4.1 is demonstrated on the simpler example of Yin et al. (2003, Section 2.1) both to provide intuition and to demonstrate the accuracy of the general theory. Demonstrating the accuracy in almost any nonlinear problem requires Monte Carlo simulation and hence can only be done on simpler problems.

The simple example is to assume that each experiment produces a data point, denoted by y_v, which is the sum of a random variable, denoted by L_v, times the unknown quantity, denoted by c, plus a noise denoted by n_v:

$$y_v = L_v c + n_v. \tag{50}$$

The integer $v \in \{1, \ldots, N_v\}$ is an index describing independent realizations of the experiment. The quantities y_v, L_v, c, and n_v are all real numbers. The goal is to estimate the value of c from all of the y_v data. Note the similarity with the cryo EM situation (Eq. 31). Assume that the sets of random variables $\{L_v : v \in \{1, \ldots, N_v\}\}$ and $\{n_v : v \in \{1, \ldots, N_v\}\}$ are independent of each other, that the sequence of random variables L_v is independent and identically distributed according to a Gaussian pdf with mean m_z and variance σ_z^2, and that the sequence of random variables n_v is independent and identically distributed according to a Gaussian pdf with mean 0 and variance σ^2. Then, by direct computation (Yin et al., 2003, Eq. C14), the log likelihood function is

$$\ln p(y|c) = -\frac{N_v}{2}\ln\left(c^2\sigma_z^2 + \sigma^2\right) - \frac{N_v}{2}\ln(2\pi) - \frac{1}{2}\frac{1}{c^2\sigma_z^2 + \sigma^2} \times \sum_{v=1}^{N_v}(y_v - cm_z)^2 \tag{51}$$

and it is possible to show (Yin et al., 2003, Eqs. 5–7)that the maximum likelihood estimate of c, denoted by $\hat{c}$, is one of the three roots of the polynomial

$$0 = -c^3\sigma_z^4 - c^2 m_z \sigma_z^2 \bar{y} + c\left[\sigma_z^2 r_y - \sigma^2(\sigma_z^2 + m_z^2)\right] + m_z\sigma^2\bar{y} \tag{52}$$

where

$$\bar{y} = \frac{1}{N_v}\sum_{v=1}^{N_v} y_v \tag{53}$$

$$r_y = \frac{1}{N_v}\sum_{v=1}^{N_v} y_v^2. \tag{54}$$

The Hessian required by the general theory is just the negative of the inverse of the second derivative with respect to the unknown c because c is a single real number. Starting from Yin et al. (2003, Eq. C15), the second derivative can be computed directly for arbitrary value of c with the result that

$$\frac{\partial^2 \ln p(y|c)}{\partial c^2} = \frac{-N_v}{\left(c^2\sigma_z^2 + \sigma^2\right)^3}\begin{bmatrix} -\sigma_z^6 c^4 + \sigma_z^2\sigma^4 + 3\sigma_z^4 r_y c^2 - \sigma_z^2 r_y \sigma^2 - 2\sigma_z^4 c^3 m_z \bar{y} \\ +6\sigma_z^2 c m_z \bar{y}\sigma^2 - 3\sigma_z^2 c^2 m_z^2\sigma^2 + m_z^2\sigma^4 \end{bmatrix} \tag{55}$$

Each iteration of Monte Carlo (indexed by the integer t) includes the following computations:

1. Compute N_v pairs of pseudo-random variables L_v and n_v drawn from the pdfs described previously.
2. Compute N_v measurements y_v from Eq. (50).
3. Compute $\bar{y}$ and r_y from Eqs. (53) and (54), respectively.
4. Compute the coefficients and then the three roots of the polynomial described by Eq. (52).
5. Among the real roots of Step 4, the estimate, denoted by $\hat{c}(t)$, is the root that maximizes the log likelihood given by Eq. (51). Since the polynomial is of third order, at least one real root is guaranteed.
6. Using Eq. (55), compute the Hessian at the estimate $\hat{c}(t)$ and denote the result by $h(t)$. The value $-1/h(t)$ is the estimate of the estimation error covariance.
7. Compute the true error, denoted by $\delta(t)$ and defined by $\delta(t) = c - \hat{c}(t)$.

The Monte Carlo estimates of the mean and variance of the error after T Monte Carlo iterations are the sample mean and variance of the $\delta(t)$ values, specifically,

$$\bar{\delta} \doteq \frac{1}{T}\sum_{t=1}^{T}\delta(t) \tag{56}$$

$$s^2 \doteq \frac{1}{T}\sum_{t=1}^{T}\left[\delta(t) - \bar{\delta}\right]^2, \tag{57}$$

respectively.

Consider a case where the true value of c is 5.0, $N_v = 100, m_z = 2.0, \sigma_z = 0.1$, and $\sigma = 0.3$. Based on $T = 10^6$ Monte Carlo iterations, the Monte Carlo estimate for the mean and variance of the estimation error are $\bar{\delta} = 0.000120$ and $s^2 = 0.000849$. In addition, the histogram of the T different estimation error variance results from the general theory of Section 4.1, i.e., the T different values of the covariance estimate $-1/h(t)$ computed from the Hessian $h(t)$, are shown in Fig. 7. The fact that the histogram is

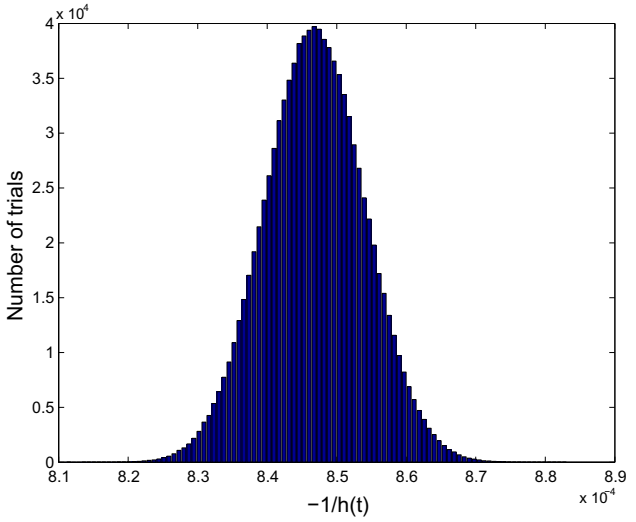


Fig. 7. The histogram of results from the general theory of Section 4.1, i.e., the histogram of different values of $-1/h(t)$ where $h(t)$ is the Hessian. The Monte Carlo estimate for the mean and variance of the estimation error are $\bar{\delta} = 0.000120$ and $s^2 = 0.000849$. The histogram is accurately centered around s^2 (the sample mean of the histogram is 0.000847) and is narrow (the square root of the sample variance of the histogram is 7.253×10^{-6}) demonstrating the high quality of the general theory in this specific example.

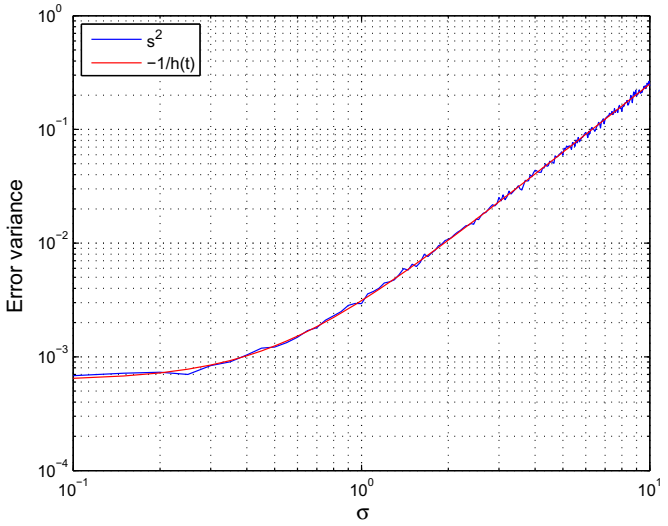


Fig. 8. The variance of the estimation error, i.e., s^2, computed by Monte Carlo using 10^3 trials and the approximate variance based on the general theory for maximum likelihood estimators, i.e., $-1/h(t)$, both as a function of standard deviation of the noise corrupting the data, i.e., σ. Note that $-1/h(t)$ accurately tracks s^2 and both have a strong dependence on the noise in the original data, which is described by the standard deviation σ.

narrow and is centered around $s^2 = 0.000849$ demonstrates the accuracy of the general theory of Section 4.1 on a problem that resembles a scalar version of the cryo EM problem.

The calculation described in the previous paragraph and Fig. 7 does not make clear how the measure of performance proposed in this paper depends on the SNR of the original data. Therefore, in Fig. 8 are plotted the Monte Carlo variance of the estimation error, i.e., s^2, (based on 10^3 Monte Carlo trials) and the result of the general theory for maximum likelihood estimators, i.e., $-1/h(t)$, as a function of the SNR of the original data. The parameters are the same as in the previous paragraph except that σ varies in order to vary the SNR of the original data. The fact that $-1/h(t)$ tracks s^2 accurately as the SNR changes and the fact that both change dramatically as the SNR improves illustrate the high quality of the general theory for this specific example and the fact that the general theory is really about how SNR of the data influences SNR of the estimates.

4.3. Fourier Axial Correlation (FAC) and Fourier Radial Correlation (FRaC) for cylindrical objects

As shown in Eq. (47), the standard FSC quantity is an integration over the two angles of spherical coordinates. For a cylindrical object, especially when the resolution can be independently controlled in the axial and radial directions, it is natural to consider integrations over various combinations of cylindrical coordinates. If the integration is taken over cylindrical shells of reciprocal space, i.e., ϕ' and k_z, then the criteria is called Fourier Radial Correlation (FRaC)[6] and is denoted by $p_{\text{FRaC}}(k_r)$ and the criteria measures radial resolution. Alternatively, if the integration is taken over cross sectional planes of reciprocal space, i.e., ϕ' and k_r, then the criteria is called Fourier Axial Correlation (FAC) and is denoted by $p_{\text{FAC}}(k_z)$ and the criteria measures axial resolution. Define the integrals over cylindrical shells and over cross sectional planes, both in reciprocal space, by

$$S^{\phi',k_z}_{P^{\text{a}},P^{\text{b}}}(k_r) \doteq \int_{k_z=-\infty}^{+\infty}\int_{\phi'=0}^{2\pi} P^{\text{a}}(\mathbf{k})\left[P^{\text{b}}(\mathbf{k})\right]^{*} \mathrm{d}\phi'\mathrm{d}k_z \tag{58}$$

$$S^{\phi',k_r}_{P^{\text{a}},P^{\text{b}}}(k_z) \doteq \int_{k_r=0}^{\infty}\int_{\phi'=0}^{2\pi} P^{\text{a}}(\mathbf{k})\left[P^{\text{b}}(\mathbf{k})\right]^{*} \mathrm{d}\phi' k_r \mathrm{d}k_r \tag{59}$$

for any pair of reciprocal-space functions $P^{\text{a}}(\mathbf{k})$ and $P^{\text{b}}(\mathbf{k})$. Then FRaC and FAC are defined by

$$p_{\text{FRaC}}(k_r) \doteq \frac{S^{\phi',k_z}_{P^{\text{a}},P^{\text{b}}}(k_r)}{\sqrt{S^{\phi',k_z}_{P^{\text{a}},P^{\text{a}}}(k) S^{\phi',k_z}_{P^{\text{b}},P^{\text{b}}}(k_r)}} \tag{60}$$

$$p_{\text{FAC}}(k_z) \doteq \frac{S^{\phi',k_r}_{P^{\text{a}},P^{\text{b}}}(k_z) + S^{\phi',k_r}_{P^{\text{a}},P^{\text{b}}}(-k_z)}{\sqrt{\left[S^{\phi',k_r}_{P^{\text{a}},P^{\text{a}}}(k_z) + S^{\phi',k_r}_{P^{\text{a}},P^{\text{a}}}(-k_z)\right]\left[S^{\phi',k_r}_{P^{\text{b}},P^{\text{b}}}(k_z) + S^{\phi',k_r}_{P^{\text{b}},P^{\text{b}}}(-k_z)\right]}} \tag{61}$$

$$= \frac{\Re\left[S^{\phi',k_r}_{P^{\text{a}},P^{\text{b}}}(k_z)\right]}{\sqrt{S^{\phi',k_r}_{P^{\text{a}},P^{\text{a}}}(k_z) S^{\phi',k_r}_{P^{\text{b}},P^{\text{b}}}(k_z)}} \tag{62}$$

where $\Re$ indicates the real part. As is shown in Supplemental material Section K, both $p_{\text{FRaC}}(k_r)$ and $p_{\text{FAC}}(k_z)$ are real valued because $\rho(\mathbf{x})$ is real valued. In addition, by the Cauchy-Schwarz inequality, $|p_{\text{FRaC}}(k)| \leqslant 1$. Finally, starting with Eq. (62) and using $|\Re[S^{\phi',k_r}_{P^{\text{a}},P^{\text{b}}}(k_z)]| \leqslant |S^{\phi',k_r}_{P^{\text{a}},P^{\text{b}}}(k_z)|$ followed by the Cauchy-Schwarz inequality, it follows that $|p_{\text{FAC}}(k)| \leqslant 1$. Defining $p_{\text{FAC}}(k_z)$ as the sum of terms at k_z and $-k_z$ allows it to combine the positive and negative frequencies which have the same interpretation with respect to resolution. In addition, summing the terms makes $p_{\text{FAC}}(k_z)$ a function of $|k_z|$ which, like k_r, ranges from 0 to ∞ while, without the sum of terms, it would be a function of k_z, which ranges from $-\infty$ to $+\infty$. Finally, it is only with the sum of terms that $\rho(\mathbf{x})$ real implies that $p_{\text{FAC}}(k_z)$ is also real (Supplemental material Section K).

For the tail model of Eq. (3), $S^{\phi',k_z}_{P^{\text{a}},P^{\text{b}}}(k_r)$ and $S^{\phi',k_r}_{P^{\text{a}},P^{\text{b}}}(k_z)$ can be computed symbolically in terms of the $c^{\text{a}}_{l,p,n}$ and $c^{\text{b}}_{l,p,n}$ for the two structures $P^{\text{a}}(\mathbf{k})$ and $P^{\text{b}}(\mathbf{k})$ under the assumption that both structures share the same values of z_0 (Eqs. 4 and 5) and R_+ (Eq. 6). The reason that the calculations can be done symbolically is the orthogonality of the factors of $L_{t(k_r,\phi',k_z),(l,p,n)}$ (Eq. 11):

[6] "FRaC" is used in order not to conflict with Fourier Ring Correlation which is abbreviated FRC.

$$\int_{\phi'=0}^{2\pi} \exp(il\xi\phi')\exp(-il'\xi\phi')\mathrm{d}\phi' = 2\pi\delta_{l,l'} \tag{63}$$

$$\int_{k_r=0}^{\infty} H_{l,p}(k_r)H^*_{l,p'}(k_r)k_r\mathrm{d}k_r = \frac{R_+^2}{2}\left[J_{|l|+1}(\gamma_{|l|,p})\right]^2\delta_{p,p'} \tag{64}$$

$$\int_{k_z=-\infty}^{\infty} Q(k_z - n/z_0)Q^*(k_z - n'/z_0)\mathrm{d}k_z = z_0\delta_{n,n'} \tag{65}$$

where Eq. (63) is an elementary integral, Eq. (64) follows from Supplemental material Eq. 127 since $H_{l,p}(k_r) \in \mathbb{R}$, and Eq. (65) is Supplemental material Eq. 225. Using these results,

$$S^{\phi',k_r}_{P^a,P^b}(k_r) = 2\pi z_0 \sum_{l=-\infty}^{+\infty}\sum_{p=1}^{\infty}\sum_{n=-\infty}^{+\infty}\sum_{p'=1}^{\infty} c^a_{l,p,n}(c^b_{l,p',n})^* H_{l,p}(k_r)H_{l,p'}(k_r) \tag{66}$$

$$S^{\phi',k_r}_{P^a,P^b}(k_z) = 2\pi \sum_{l=-\infty}^{+\infty}\sum_{p=1}^{\infty}\sum_{n=-\infty}^{+\infty}\sum_{n'=-\infty}^{+\infty} c^a_{l,p,n}(c^b_{l,p,n'})^* \times \frac{R_+^2}{2}\left[J_{|l|+1}(\gamma_{|l|,p})\right]^2 Q(k_z - n/z_0)Q(k_z - n'/z_0). \tag{67}$$

As is shown in Supplemental material Section J, Eq. (67) implies that Eq. (62) can be simplified and that the simplified equation is rational in $k_z z_0$. Taking the limit as $k_z z_0$ grows large leads to the result (Eq. 225) that

$$\lim_{k_z\to\infty} p_{\mathrm{FAC}}(k_z) = \frac{\sum_l\sum_p j_{l,p}\sum_n\sum_{n'}\Re\left[c^a_{l,p,n}c^{b*}_{l,p,n'}\right](-1)^{n+n'}}{\sqrt{\left[\sum_l\sum_p j_{l,p}\sum_n\sum_{n'}|c^a_{l,p,n}|^2(-1)^{n+n'}\right]\left[\sum_l\sum_p j_{l,p}\sum_n\sum_{n'}|c^b_{l,p,n}|^2(-1)^{n+n'}\right]}} \tag{68}$$

where

$$j_{l,p} \doteq \frac{R_+^2}{2}\left[J_{|l|+1}(\gamma_{|l|,p})\right]^2. \tag{69}$$

4.4. The connection between estimation error covariance and FAC, FRaC, and FSC

Let c_* denote the vector containing the true $c_{l,p,n}$ coefficients, let $\hat{c}(y)$ denote the vector containing the maximum likelihood estimates of the $c_{l,p,n}$ coefficients from the image data y, and let $H(\hat{c}(y))$ be the Hessian of the log likelihood evaluated at the maximum likelihood estimate of the $c_{l,p,n}$ coefficients. The Hessian can be computed starting from Doerschuk and Johnson (2000, Eq. 15, p. 1721) by taking partial derivatives of the log likelihood with respect to two components of the vector containing the $c_{l,p,n}$ coefficients and using the fact that the first derivative of the log likelihood when evaluated at the maximum likelihood estimate is zero because the estimate is the maximum. The resulting formula is Prust (2006, Section 3.2, pp. 35–36)

$$H(\hat{c}(y)) = \sum_{i=1}^{N_v}\left(\frac{-1}{\gamma_i(\hat{c},y_i)}\right)\Delta_i(\hat{c},y_i,k) \tag{70}$$

where it is important to note that $H(\hat{c}(y))$ can be computed in terms of the γ and Δ variables of Eqs. (35) and (37) so its computation does not add to the memory footprint or computational cost of the algorithm. Similar to the comment at the end of Section 2.4, these calculations can be generalized to the case where each virion is from one of a finite number of different classes and the class label for the virion shown in a particular image is not known (Doerschuk and Johnson, 2000; Yin et al., 2003) with the interesting result that the Hessian matrix has a block diagonal structure where each class corresponds to a different block. From the general theory of Section 4.1, the estimation error $\hat{c}(y) - c_*$ is Gaussian distributed with mean 0 and covariance $-[H(\hat{c}(y))]^{-1}$. Therefore, if the entire experiment and computation were repeated many times, the estimates $\hat{c}(y(n))$ would be Gaussian distributed with mean c_* and covariance $-[H(\hat{c}(y))]^{-1}$ where n indexes the repetitions. If the true c_* is approximated by the maximum likelihood estimate $\hat{c}(y)$, then the distribution of $\hat{c}(y(n))$ is fully specified and sampling from this distribution is a generalization of the two reconstructions traditionally computed by using even versus odd numbered images.

Suppose there are two structures that are either computed from different images or are different samples from the distribution of the previous paragraph. Denote the estimated $c_{l,p,n}$ coefficients by $\hat{c}^a$ and $\hat{c}^b$. From Eqs. (60) and (66) (Eqs. 62 and 67), FRaC (FAC) can be computed for any value of k_r (k_z) from $\hat{c}^a$ and $\hat{c}^b$. As described in Section 4, resolution in FSC is measured as the smallest value of k for which $p_{\mathrm{FSC}}(k)$ is below the threshold. Adopting the same definition for $p_{\mathrm{FRaC}}(k_r)$ and $p_{\mathrm{FAC}}(k_z)$ implies that the resolution, denoted by k_{r*} and k_{z*}, respectively, is defined to be

$$k_{r*} = \min_{k_r\geqslant 0}\{k_r : p_{\mathrm{FRaC}}(k_r) < t_{\mathrm{FRaC}}(k_r)\} \tag{71}$$

$$k_{z*} = \min_{k_z\geqslant 0}\{k_z : p_{\mathrm{FAC}}(k_z) < t_{\mathrm{FAC}}(k_z)\} \tag{72}$$

where $t_{\mathrm{FRaC}}(k_r)$ [$t_{\mathrm{FAC}}(k_z)$] is the radial (axial) threshold which is possibly k_r (k_z) dependent. Though it is not shown in the notation, k_{r*} (k_{z*}) is a function of $p_{\mathrm{FRaC}}(k_r)$ [$p_{\mathrm{FAC}}(k_z)$] which is a function of $\hat{c}^a$ and $\hat{c}^b$. Therefore, k_{r*} and k_{z*} are derived random variables, derived from $\hat{c}^a$ and $\hat{c}^b$.

Because k_{r*} and k_{z*} are derived random variables, their pdfs can be computed by Monte Carlo. Before starting the iterative Monte Carlo calculation, it is most efficient to compute the Cholesky factorization of $H(\hat{c}(y))$, which is denoted by $H^{1/2}$ and which has the property that $H(\hat{c}(y)) = H^{1/2}(H^{1/2})^T$. Each iteration of Monte Carlo (indexed by the integer t) includes the following computations:

1. Compute $v^a, v^b \in \mathbb{R}^{N_c}$ whose components are independent and identically distributed Gaussian pseudo random variables with mean 0 and variance 1 where N_c is the number of $c_{l,p,n}$ coefficients.
2. Compute $c^a = H^{1/2}v^a + \hat{c}(y)$ and $c^b = H^{1/2}v^b + \hat{c}(y)$ which are samples from the Gaussian distribution for the estimates. c^a and c^b play the role of the structures computed from even versus odd numbered images.
3. From Eqs. (60, 66, and 71), compute $k_{r*}(t)$.
4. From Eqs. (62, 67, and 72), compute $k_{z*}(t)$.

After T Monte Carlo iterations, an estimate of the pdf for k_{r*} (k_{z*}), denoted by $p_{k_{r*}}(k_r)$ [$p_{k_{z*}}(k_z)$], can be determined by computing the histogram for the set $k_{r*}(t)$ [$k_{z*}(t)$] for $t \in \{1,\ldots,T\}$. Once a probability p is chosen, e.g., $p = .01$, the estimates for k_{r*} and k_{z*}, denoted by $\hat{k}_r$ and $\hat{k}_z$, respectively, can be determined as the values such that

$$\int_{k_r=0}^{\hat{k}_r} p_{k_{r*}}(k_r)\mathrm{d}k_r = p \tag{73}$$

$$\int_{k_z=0}^{\hat{k}_z} p_{k_{z*}}(k_z)\mathrm{d}k_z = p. \tag{74}$$

Alternatively, the same pdfs can be used to compute confidence intervals. Let $\bar{k}_r$ ($\bar{k}_z$) be the sample mean of $k_{r*}(t)$ [$k_{z*}(t)$], i.e.,

$$\bar{k}_r = \frac{1}{T}\sum_{t=0}^{T} k_{r*}(t) \tag{75}$$

$$\bar{k}_z = \frac{1}{T}\sum_{t=0}^{T} k_{z*}(t). \tag{76}$$

Then the symmetric 100q% confidence intervals are $[\bar{k}_r - \delta_r, \bar{k}_r + \delta_r]$ and $[\bar{k}_z - \delta_z, \bar{k}_z + \delta_z]$ where δ_r and δ_z are defined by

$$\int_{k_r=\bar{k}_r-\delta_r}^{\bar{k}_r+\delta_r} p_{k_{r*}}(k_r)\mathrm{d}k_r = q \tag{77}$$

$$\int_{k_z=\bar{k}_z-\delta_z}^{\bar{k}_z+\delta_z} p_{k_{z*}}(k_z)\mathrm{d}k_z = q, \tag{78}$$

respectively.

While the equations are not shown here, these ideas can also be applied to the FSC based on Yin et al. (2003, Eqs. 23 and 25).

5. Numerical results 2: The resolution of the reconstruction of P22

As described in Section 2.5, the storage footprint of the current software system is large. Each D matrix requires $N_c(N_c+1)/2$ locations (where N_c is the number of $c_{l,p,n}$ coefficients used) and the number of D matrices is the number of abscissas, which is $N_g = 60$ since the orientation uncertainty is due to the uncertainty in which icosahedrally related orientation is present, times the number of images (denoted by N_v) since each image has a different set of $N_g = 60$ possible orientations. Fitting the D matrices into memory constrains the number of $c_{l,p,n}$ coefficients N_c and/or the number of images N_v and the largest calculations reported here have $N_c = 385$ (implied by $l_{\max} = 2, p_{\max} = 7$, and $n_{\max} = 5$) and $N_v = 276$. With so few images, the traditional method of performing reconstructions with even and with odd numbered images and then comparing the two 3-D cubes by FSC is not attractive.

Independent of image quality and image number, N_c sets an upper bound on the resolution that can be achieved. The N_c $c_{l,p,n}$ coefficients describe a function that is nonzero in a cylinder of length z_0 and radius R_+. Since the function is constrained to have ξ-fold rotational symmetry, the function is uniquely defined by its values on $1/\xi$ of the cylinder's volume. Alternatively, the same volume might be represented by N_c voxels where each voxel measures $T \times T \times T$. Equating the volume measured in voxels and the volume of $1/\xi$ of the cylinder gives $N_c T^3 = \pi R_+^2 z_0/\xi$ which implies that

$$T = \left[\frac{\pi R_+^2 z_0}{\xi N_c}\right]^{1/3}. \tag{79}$$

The resolution that a model with N_c coefficients is able to represent when averaged in all directions, independent of the number and quality of the images, is unlikely to exceed about $2T$ unless the basis functions used in Eq. (3) much more efficiently represent the electron scattering intensity function in comparison with voxel basis functions (i.e., basis functions which are 1 in a voxel and 0 outside of the voxel). For $\xi = 6, N_c = 385, R_+ = 130$Å, and $z_0 = 380$Å, the value of T is $T = 20.59$Å. Therefore the achieved resolution may be limited by the value of N_c rather than the number and quality of the images.

A second measure of the upper bound on resolution that is set by N_c independent of the image quality and image number can be determined by considering special 3-D structures composed of just those $c_{l,p,n}$ and corresponding basis functions with the highest spatial frequencies. In the truncated system used for numerical calculations, that special structure has $c_{l,p,n}$ coefficients defined by

$$c_{l,p,n}^{\mathrm{HF}} = \begin{cases} 1, & l \in \{\pm l_{\max}\}, p = p_{\max}, n \in \{\pm n_{\max}\} \\ 0, & \text{otherwise} \end{cases} \tag{80}$$

where "HF" stands for "high frequency". Let $P^{\mathrm{HF}}(\mathbf{k})$ be the corresponding reciprocal space function. The averaged distribution in reciprocal space of the energy in the special structure is determined by $S^{\phi',k_z}_{P^{\mathrm{HF}},P^{\mathrm{HF}}}(k_r)$ (averaged over cylindrical shells) and $S^{\phi',k_r}_{P^{\mathrm{HF}},P^{\mathrm{HF}}}(k_z)$ (averaged over cross sectional planes) and these functions are plotted in Fig. 9. While the curves oscillate, the curve has permanently decreased to below 10% of its maximum value by 0.0495Å^{-1} for $S^{\phi',k_z}_{P^{\mathrm{HF}},P^{\mathrm{HF}}}(k_r)$ and by 0.0150Å^{-1} for $S^{\phi',k_r}_{P^{\mathrm{HF}},P^{\mathrm{HF}}}(k_z)$ and to below 1% of its maximum value by 0.0568Å^{-1} for $S^{\phi',k_z}_{P^{\mathrm{HF}},P^{\mathrm{HF}}}(k_r)$ and by 0.0226Å^{-1} for $S^{\phi',k_r}_{P^{\mathrm{HF}},P^{\mathrm{HF}}}(k_z)$. With $N_c = 385$ (implied by $l_{\max} = 2, p_{\max} = 7$, and $n_{\max} = 5$) the mathematical model cannot achieve higher spatial resolution than somewhere between the 10% and 1% values of k_r and k_z independent of the number and quality of the images used to determine the values of the $c_{l,p,n}$ coefficients.

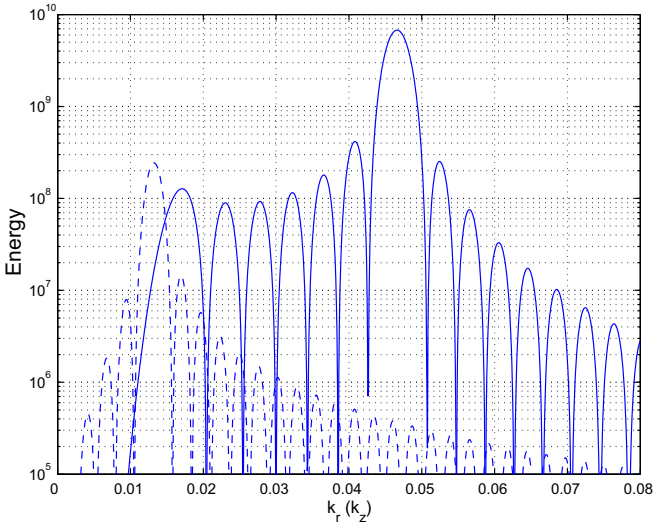


Fig. 9. The distribution of energy in reciprocal space ($0 \leqslant k_r, k_z \leqslant 0.08$Å^{-1}) for the highest frequency $c_{l,p,n}$ and corresponding basis function. Both $S^{\phi',k_z}_{P^{\mathrm{HF}},P^{\mathrm{HF}}}(k_r)$, showing energy averaged over cylindrical shells as a function of the radius of the shell (solid curve), and $S^{\phi',k_r}_{P^{\mathrm{HF}},P^{\mathrm{HF}}}(k_z)$, showing energy averaged over cross sectional planes as a function of the axial coordinate (dashed curve), are shown.

Due to the considerations of the previous two paragraphs, the resolution cannot be above $k_{\mathrm{FRaC}} \doteq 0.06$Å^{-1} or $k_{\mathrm{FAC}} \doteq 0.023$Å^{-1} for FRaC or FAC, respectively, and plots are stopped at $k_* \doteq \max(k_{\mathrm{FRaC}}, k_{\mathrm{FAC}}) = 0.06$Å^{-1}. Fig. 10 shows 4 realizations of $p_{\mathrm{FRaC}}(k_r)$ (Eq. 60) and $p_{\mathrm{FAC}}(k_z)$ (Eq. 62). The plots of $p_{\mathrm{FAC}}(k_z)$ clearly show the approach of $p_{\mathrm{FAC}}(k_z)$ to the asymptotic value as $k_z z_0$ ($z_0 = 380$Å) grows large as is expected from Eq. (68). It is apparent that resolution is not limited by the number or quality of the images, since the curves remain high over the entire range of 0 to k_{FRaC} or 0 to k_{FAC} for FRaC or FAC, respectively, but rather is limited by the number of $c_{l,p,n}$ coefficients (i.e., by N_c) that our current software can accommodate. Therefore, in order to compute resolutions less than k_{FRaC} and k_{FAC}, it is necessary to choose a strict threshold in Eqs. (71) and (72). Using $T = 1000$ Monte Carlo iterations and threshold functions $t_{\mathrm{FRaC}}(k_r) = 0.95$ (Eq. 71) and $t_{\mathrm{FAC}}(k_z) = 0.95$ (Eq. 72) the histograms of k_{r*} (Eq. 71) and k_{z*} (Eq. 72) values are shown in Fig. 11. The reason that the k_{r*} histogram of Fig. 11(a) is bimodal is that the k_r value at which $p_{\mathrm{FRaC}}(k_r)$ first drops below the threshold typically occurs in one of two successive oscillations of $p_{\mathrm{FRaC}}(k_r)$ where the oscillations are due to the $H_{l,p}(k_r)$ functions (defined in Eq. 13) via Eq. (66) in Eq. (60). Choosing $p = 0.02$ in Eqs. (73) and (74) leads to a radial resolution of $\hat{k}_r = 0.0301$Å^{-1} (33.3 Å) and an axial resolution of $\hat{k}_z = 0.0142$Å^{-1} (70.6 Å) where both are with probability $p = 0.02$, that is, the probability that the resolution is actually less than $\hat{k}_r = 0.0301$Å^{-1} (33.3 Å) and $\hat{k}_z = 0.0142$Å^{-1} (70.6 Å) is $p = 0.02$.

6. Discussion

Two connected methodological contributions are presented in this paper. The first is a maximum likelihood reconstruction

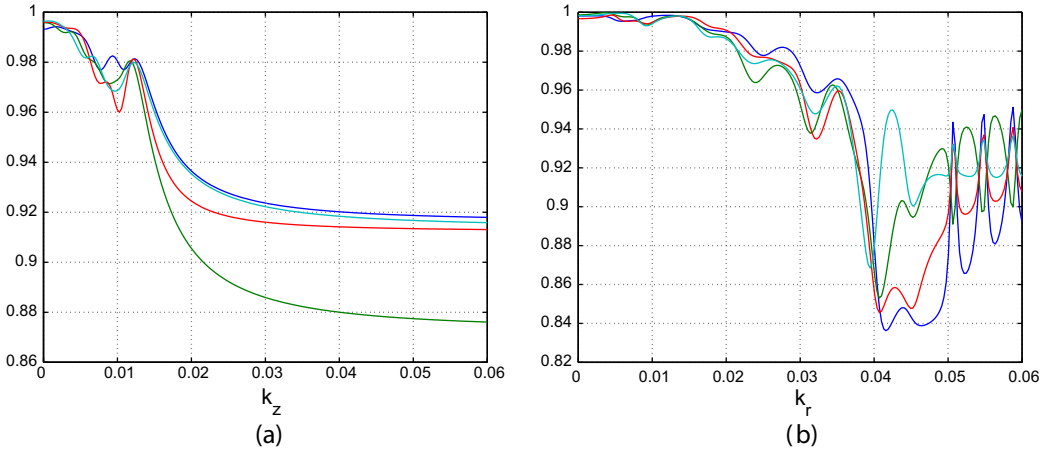


Fig. 10. Four example realizations of $p_{\mathrm{FAC}}(k_z)$ (a) and $p_{\mathrm{FRaC}}(k_r)$ (b) defined in Eqs. (60) and (62), respectively, and plotted on the range from 0 to $k_* = 0.06\text{Å}^{-1}$. A different color is used for each realization so that the curve of an individual realization can be more easily tracked.

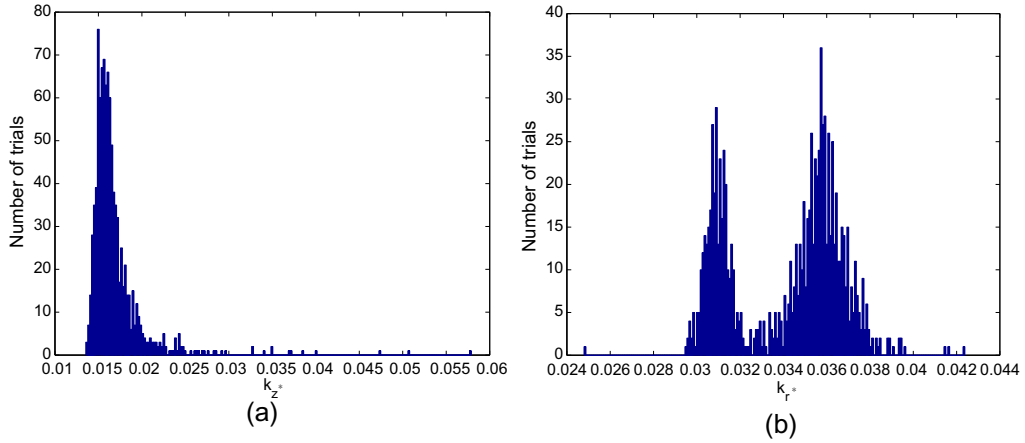


Fig. 11. Histograms of k_{z*} (a, defined in Eq. (72), range $0.01 \leqslant k_{z*} \leqslant 0.06\text{Å}^{-1}$) and k_{r*} (b, defined in Eq. (71), range $0.024 \leqslant k_{r*} \leqslant 0.044\text{Å}^{-1}$).

method for an asymmetric reconstruction of an object of the form "sphere plus cylinder" where the sphere part has icosahedral symmetry, the cylinder part has ξ-fold rotation symmetry around the axis of the cylinder, and the axis of the cylinder and a 5-fold axis of the icosahedron are coincidental. While an icosahedrally symmetrized reconstruction of the object is used, in particular, the method described in this paper is really based on images that are the difference of the experimental image and the predicted icosahedrally symmetric image, no previous structure for the cylinder is required. In addition, the rotation angle between the cylindrical and spherical objects is determined without any assumptions directly from the data. The second methodological contribution is a statistical method of measuring resolution that combines standard Fourier Shell Correlation (FSC) ideas with standard statistical maximum likelihood ideas and can measure resolution axially and radially in a cylindrical object as well as radially in a spherical object as is done by FSC.

The reconstruction and the resolution methods are used to study the infectious P22 bacteriophage virion using a subset of the images used in Lander et al. (2006). Without making any assumptions at any stage of the reconstruction algorithm, the rotational positioning of the components of the tail are determined relative to the icosahedral symmetry axes of the capsid as is described in Fig. 6. The combination of the reconstruction and resolution calculations is important because limitations of the reconstruction software make a resolution calculation based on reconstructions from even versus from odd numbered images unattractive and the methods described here can be applied to reconstructions from the complete set of images. With a correlation threshold of 0.95, the resolution in the tail measured radially is greater than 0.0301Å^{-1} (33.3Å) and measured axially is greater than 0.0142Å^{-1} (70.6Å) both with probability $p = 0.02$.

Acknowledgments

C.J.P. and P.C.D. gratefully acknowledge the support provided by the National Institutes of Health under Grant 1R01EB000432-01 and the National Science Foundation under Grants CCR-0098156 and CCF-0735297 and the computations were carried out by C.J.P. and P.C.D. at Purdue University using facilities supported by the Army Research Office under Contract DAAD19-99-1-0015 and the National Science Foundation under Grant EIA-0130538. Electron microscopic imaging was conducted at the National Resource for Automated Molecular Microscopy, which is supported by the NIH through the National Center for Research Resources' P41 program (RR17573).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jsb.2009.04.013.

References

Altmann, S.L., 1957. On the symmetries of spherical harmonics. Proc. Camb. Phil. Soc. 53, 343–367.

Baker, T.S., Olson, N.H., Fuller, S.D., 1999. Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. Microbiol. Mol. Biol. Rev. 63 (4), 862–922.

Blanc, E., Roversi, P., Vonrhein, C., Flensburg, C., Lea, S.M., Bricogne, G., 2004. Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. Acta Crystallogr. D60 (Part 12(1)), 2210–2221.

Bubeck, D., Filman, D.J., Cheng, N., Steven, A.C., Hogle, J.M., Belnap, D.M., 2005. The structure of the poliovirus 135S cell entry intermediate at 10-Angstrom resolution reveals the location of an externalized polypeptide that binds to membranes. J. Virol. 79 (2), 7745–7755.

Doerschuk, P.C., Johnson, J.E., 2000. Ab initio reconstruction and experimental design for cryo electron microscopy. IEEE Trans. Info. Theory 46 (5), 1714–1729.

Efron, B., Hinkley, D.V., 1978. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. Biometrika 65 (3), 457–487.

Erickson, H.P., 1973. The Fourier transform of an electron micrograph—first order and second order theory of image formation. In: Barer, R., Cosslett, V.E. (Eds.), Advances in Optical and Electron Microscopy, vol. 5. Academic Press, London and New York, pp. 163–199.

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M., Leith, A., 1996. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J. Struct. Biol. 116, 190–199. Available from: <http://www.wadsworth.org/resnres/frank.htm>.

Frank, J., Verschoor, A., Boublik, M., 1981. Computer averaging of electron micrographs of 40S ribosomal subunits. Science 214 (4527), 1353–1355.

Harauz, G., van Heel, M., 1986. Exact filters for general geometry three dimensional reconstruction. Optik 73 (4), 146–156.

Jiang, W., Chang, J., Jakana, J., Weigele, P., King, J., Chiu, W., 2006. Structure of epsilon15 bacteriophage reveals genome organization and DNA packaging/injection apparatus. Nature 439 (7076), 612–616.

Lander, G.C., Tang, L., Casjens, S.R., Gilcrease, E.B., Prevelige, P., Poliakov, A., Potter, C.S., Carragher, B., Johnson, J.E., 2006. The structure of an infectious P22 virion shows the signal for headful DNA packaging. Science 312 (5781), 1791–1795.

Laporte, O., 1948. Polyhedral harmonics. Z. Naturforschg. 3a, 447–456.

Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation, 2nd edition. Springer-Verlag, New York.

McCoy, A.J., Grosse-Kunstleve, R.W., Storoni, L.C., Read, R.J., 2005. Likelihood-enhanced fast translation functions. Acta Crystallogr. D61 (Part 4), 458–464.

Prust, C.J. 2006. Model-based statistical inference problems concerning non-linear 3-D tomography with applications to the structural biology of asymmetric virus particles. Ph.D. thesis, School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA.

Saxton, W.O., Baumeister, W., 1982. The correlation averaging of a regularly arranged bacterial cell envelope protein. J. Microsc. 127 (2), 127–138.

Scheres, S.H.W., Gao, H., Valle, M., Herman, G.T., Eggermont, P.P.B., Frank, J., Carazo, J.-M., 2007. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nat. Methods 4 (1), 27–29.

Scherzer, O., 1949. The theoretical resolution limit of the electron microscope. J. Appl. Phys. 20, 20–29.

Singh, V., Marinescu, D.C., Baker, T.S., 2004. Image segmentation for automatic particle identification in electron micrographs based on hidden markov random field models and expectation maximization. J. Struct. Biol. 145 (1–2), 123–141.

van Heel, M., 1987. Similarity measures between images. Ultramicroscopy 21, 95–100.

van Heel, M., Schatz, M., 2005. Fourier shell correlation threshold criteria. J. Struct. Biol. 151, 250–262.

Yin, Z., Zheng, Y., Doerschuk, P.C., Natarajan, P., Johnson, J.E., 2003. A statistical approach to computer processing of cryo electron microscope images: virion classification and 3-D reconstruction. J. Struct. Biol. 144 (1/2), 24–50.

Zhang, P., Mueller, S., Morais, M.C., Bator, C.M., Bowman, S., nad Hafenstein, Valorie D., Wimmer, E., Rossmann, M.G., 2008. Crystal structure of CD155 and electron microscopic studies of its complexes with polioviruses. PNAS 105 (47), 18284–18289.

Zheng, Y. 2002. Parallel implementations of 3-D reconstruction algorithms for cryo electron microscopy: a comparative study. Master's thesis, School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA.