Single-particle reconstruction statistics: a diagnostic tool in solving biomolecular structures by cryo-EM

J. Bernard Heymann*

Laboratory for Structural Biology Research, NIAMS, National Institutes of Health, Bethesda, MD 20892, USA.

*Correspondence e-mail: heymannb@mail.nih.gov

Received 12 June 2018

Accepted 13 December 2018

Edited by A. K. Mitra, University of Auckland, New Zealand

Keywords: cryo-EM; cryo-electron microscopy; spectral signal-to-noise ratio; single-particle analysis; Fourier shell correlation; image processing.

In single-particle analysis (SPA), the aim is to obtain a 3D reconstruction of a biological molecule from 2D electron micrographs to the highest level of detail or resolution as possible. Current practice is to collect large volumes of data, hoping to reach high-resolution maps through sheer numbers. However, adding more particles from a specific data set eventually leads to diminishing improvements in resolution. Understanding what these resolution limits are and how to deal with them are important in optimization and automation of SPA. This study revisits the theory of 3D reconstruction and demonstrates how the associated statistics can provide a diagnostic tool to improve SPA. Small numbers of images already give sufficient information on micrograph quality and the amount of data required to reach high resolution. Such feedback allows the microscopist to improve sample-preparation and imaging parameters before committing to extensive data collection. Once a larger data set is available, a *B* factor can be determined describing the suppression of the signal owing to one or more causes, such as specimen movement, radiation damage, alignment inaccuracy and structural variation. Insight into the causes of signal suppression can then guide the user to consider appropriate actions to obtain better reconstructions.

1. Introduction

Electron microscopy has always had the potential to solve biomolecular structures to high resolution because of the small wavelengths of the electron beams (0.02 Å for 300 kV; Crewe *et al.*, 1970). The problem is that biological specimens are radiation-sensitive and spatial tolerances at such small scales are challenging. Nevertheless, with the introduction of field-emission guns for better beam coherence (Crewe *et al.*, 1968) and the development of the direct electron detector with greater signal-to-noise ratio (SNR; Bisogni *et al.*, 1998; Faruqi & Henderson, 2007), this potential is being realized. These hardware advances are complemented by improved computational abilities to reconstruct 3D maps from the 2D data with higher accuracy. Despite recent successes, there is still room for improvement and a better understanding of single-particle analysis (SPA). One area is in extracting more statistical information from the reconstruction process that can be used to diagnose problems and optimize the structural studies of biomolecules.

The statistics of the electron beam are those of a quantum-limited signal (Saxton & Frank, 1976; Saxberg & Saxton, 1981), equivalent to shot noise (*i.e.* following a Poisson distribution). The signal of the particles that we are interested

in is buried in this noise. The particles are assumed to be sufficiently rigid so that their projection images from different directions can be combined into a 3D reconstruction. We therefore define our signal as that part of the images that is coherent (*i.e.* that represents the same structure from different views) and the noise is the variance. The SNR is then defined as the signal power divided by the noise variance. Because the data that are collected are Poisson distributed (where the average is equal to the variance), the SNR increases with dose (although limited by radiation damage; Henderson, 1995). The SNR is further enhanced by reconstruction from many-particle images to average out noise and reveal the embedded consistent structure. This view of 3D reconstruction in single-particle analysis suggests that a simple increase in contributing particle images would lead to increasingly detailed maps.

However, several factors limit the achievable resolution that can be collectively viewed as image or signal variability. One is the interaction of the electron beam with the specimen, causing beam-induced movement (Glaeser & Hall, 2011) and radiation damage (Karuppasamy *et al.*, 2011). The conventional wisdom in electron microscopy was that radiation damage is the main limiting factor. This is based on tests on crystalline specimens, where the diffraction spot intensity decreased with dose (Hayward & Glaeser, 1979; Baker *et al.*, 2010). The advent of direct detectors (McMullan *et al.*, 2016) allowed the acquisition of series of short exposures (frames or movies) that can be aligned to compensate for movement (Li *et al.*, 2013). This indicated that before this capability, movement limited resolution more than radiation damage. Glaeser & Hall (2011) anticipated this development, suggesting that compensating for specimen movement is one of the issues that need to be addressed to improve reconstructions. Beyond these factors, errors in image alignment and structural variation may also limit the extent of detail in maps.

The principles behind 3D reconstruction have been explored in some depth in the past (Frank *et al.*, 1981; Penczek, 2002; Liao & Frank, 2010; Unser *et al.*, 1987, 1989, 2005) and incorporated into a number of software packages (https://en.wikibooks.org/wiki/Software_Tools_For_Molecular_Microscopy). Nevertheless, uncertainty regarding the best approaches and optimization remains (Heymann *et al.*, 2018; Heymann, 2018b). Here, I re-examine the theory of 3D reconstruction and demonstrate that the nearest-neighbor interpolation algorithm can be used to calculate separate signal and noise contributions. The noise power reflects Poisson statistics and is simply averaged out with increasing numbers of particles. As more data are incorporated into a reconstruction, the signal is revealed and its suppression becomes evident as the main reason for resolution limitation. The progress in reconstruction can therefore be divided into a noise-limited phase at low numbers of particles, transitioning into a signal-suppressed phase. In the latter phase it may become prohibitive to acquire more data without dealing with the factors leading to signal suppression. This study therefore aims at clarifying the key contributions to 3D reconstruction as a first step towards diagnosing problems.

2. Theory

2.1. The reconstruction signal and noise

Reconstruction can be performed in real space (Gilbert, 1972; Radermacher, 2006) or in frequency space (Crowther, Amos *et al.*, 1970; Crowther, DeRosier *et al.*, 1970). The latter has a simpler statistical basis and is therefore more suitable for analysis. The reconstruction statistics that I derive here follow established theory as developed in previous work (Frank *et al.*, 1981; Penczek, 2002; Liao & Frank, 2010; Unser *et al.*, 1987, 1989, 2005).

A single-particle image is composed of a signal (the particle of interest) and everything else, which is considered noise. The general assumption made is that the signal and noise are additive (*i.e.* uncorrelated; Penczek, 2002),

$$F = S + N, \quad (1)$$

where F , S and N are values in frequency space for the image, signal and noise, respectively. This is a simplification that ignores other important factors, such as variation owing to transfer functions and inelastic scattering. However, the fundamental statistics in the simplified model provide a useful basis for analysis.

Another way to interpret the imaging model is to define the signal as the coherent part under summation and the noise as the variable part. Thus, on averaging over c contributions, the signal remains constant, while the noise decreases,

$$\frac{1}{c} \sum_i F_i = S + \frac{1}{c} \sum_i N_i. \quad (2)$$

For the noise as independent random values, the sum of the noise power¹ is approximately equal to the residual noise power after averaging,

$$\sum_i |N_i|^2 \simeq \left| \sum_i N_i \right|^2. \quad (3)$$

We can use this relationship to calculate the noise power in the average, and thus the average noise power in the original images, as follows. The power for the average is

$$\left| \frac{1}{c} \sum_i F_i \right|^2 = |S|^2 + \frac{2}{c} \left(S \sum_i N_i^* \right)_{\text{re}} + \left| \frac{1}{c} \sum_i N_i \right|^2. \quad (4)$$

The average power in all of the contributing particle images is

$$\frac{1}{c} \sum_i |F_i|^2 = |S|^2 + \frac{2}{c} \left(S \sum_i N_i^* \right)_{\text{re}} + \frac{1}{c} \sum_i |N_i|^2. \quad (5)$$

The first and second terms in (4) and (5) are the same, and we can thus eliminate them in the difference. Further, using (3), we obtain the noise power remaining in the reconstruction,²

$$\left| \frac{1}{c} \sum_i N_i \right|^2 \simeq \frac{1}{c-1} \left(\frac{1}{c} \sum_i |F_i|^2 - \left| \frac{1}{c} \sum_i F_i \right|^2 \right). \quad (6)$$

¹ The power or intensity in frequency space is the equivalent of the variance in real space.

² Note that this is the familiar equation for unbiased variance.

If we assume the middle term (or cross-term) in (4) and (5) to be negligible (but see below and Appendix A for more), we can eliminate the noise power in a similar way using (3),³

$$|\widetilde{S}|^2 \simeq \frac{1}{c-1} \left(c \left| \frac{1}{c} \sum_i^c F_i \right|^2 - \frac{1}{c} \sum_i^c |F_i|^2 \right). \quad (7)$$

For large c , the signal power approaches the reconstruction power (*i.e.* negligible noise). Also, the sum of (6) and (7) equals the average power in the reconstruction. The SNR for the reconstruction can then be derived as

$$\alpha_r(c) = \frac{|\widetilde{S}|^2}{\left| \frac{1}{c} \sum_i^c N_i \right|^2} \simeq \frac{c \left(\left| \sum_i^c F_i \right|^2 - \sum_i^c |F_i|^2 \right)}{c \sum_i^c |F_i|^2 - \left| \sum_i^c F_i \right|^2}. \quad (8)$$

This is equivalent to the SNR for 2D images as given by Unser *et al.* (1987) and is consistent with the treatment of 3D reconstructions of Penczek (2002).

For limited data, the cross-term in (4) and (5) has significant magnitude (a similar term is noted in van Heel & Schatz, 2005). (7) includes this cross-term in estimating the signal, *i.e.*

$$|\widetilde{S}|^2 = |S|^2 + \frac{2}{c} \left(S \sum_i^c N_i^* \right)_{\text{re}}$$

I attempted to derive an analytical form for the cross-term (see Appendix A), but did not find a defensible form for it that gives an effectively zero signal at high frequency. The cross-term therefore remains enigmatic, but is apparent when calculating the signal as given in (7) (see below for results with synthetic and real data).

2.2. Particle image contributions

(8) gives the SNR for c contributions (for instance at a frequency-space pixel, although it could be a whole image or a resolution shell in frequency space). Reconstructing a 3D map from 2D images means that the contributions decrease with spatial frequency (*i.e.* the contribution of a plane, the central section, to the 3D volume in frequency space). Consider adding a central section with thickness t (typically one pixel in frequency space, but see Appendix B) to a 3D volume. Each ring of the section at a distance of k pixels covers the following fraction of the surface of the shell:

$$f(k) = \frac{2\pi kt}{4\pi k^2} = \frac{t}{2k}. \quad (9)$$

The distance k is related to the physical spatial frequency s ($\text{\AA}^{-1}$),

$$k = Ds, \quad (10)$$

where D is the particle image edge size (*i.e.* the size in pixels times the sampling or pixel size, in $\text{\AA}$). The contribution of n images at spatial frequency s is then⁴

$$c(n, s) = \frac{tn}{2Ds}. \quad (11)$$

Note that n here includes symmetry order, *i.e.* it is the number of asymmetric units. Because the signal is operationally defined as constant in (2), it is independent of the number of images. The integration of the 2D images into the 3D volume is therefore averaging of the noise as a function of the spatial frequency,

$$|N_r(n, s)|^2 = \frac{2Ds}{tn} \langle |N(s)|^2 \rangle, \quad (12)$$

where $\langle |N(s)|^2 \rangle$ is the average noise in the original particle images. The latter can therefore easily be back-calculated from the result of (6) (as also demonstrated in Rosenthal & Henderson, 2003). The particle noise is approximately white (*i.e.* constant over frequency), while the reconstruction noise increases linearly with frequency owing to the decrease in contributions.

2.3. Particle image variability suppresses the true signal

To make our analysis tractable, we implicitly assume that the true signal is a projection image of a rigid particle. However, variability in the particle images suppresses the signal and limits the achievable resolution. This variability can be attributed to different causes, including movement, radiation damage, alignment inaccuracy and structural variability (Rosenthal & Henderson, 2003). The microscopist and image processor need to understand which effect dominates in a particular data set to be able to address it. Part of this is to characterize this signal suppression.

The combined effect of the different influences on the signal is typically modeled as a Gaussian decay parametrized by a B factor (Baldwin & Penczek, 2005; Rosenthal & Henderson, 2003) or a sigma value (Penczek, 2010):

$$\varepsilon(s) = \exp\left(-\frac{B}{2}s^2\right) = \exp\left[-\frac{1}{2}\left(\frac{s}{\sigma}\right)^2\right]. \quad (13)$$

The reconstruction signal is then

$$|S_r(s)|^2 = \langle |S(s)|^2 \rangle \left[\exp\left(-\frac{B}{2}s^2\right) \right], \quad (14)$$

where $\langle |S(s)|^2 \rangle$ is the average signal (or true signal) in the original particle images at frequency s . This can be calculated from the electron-scattering cross-sections for a known structure (such as for the β -galactosidase case used here), or estimated for biomolecules in general (see below for actual calculations). The signal for most biomolecules displays a strong decay with frequency. The overall B factor described by Rosenthal & Henderson (2003) conflates this inherent signal decay with signal suppression (see Appendix C). The B factor in (14) represents resolution-limiting factors that are prohibitive to overcome by adding more data of the same quality.

³ I use the tilde here to indicate that it is not the true signal power.

⁴ $c(n, s)$ is equivalent to c in (2)–(8).

2.4. The reconstruction spectral signal-to-noise ratio

Combining the terms in (11), (8) and (13) gives an equation for the spectral signal-to-noise ratio (SSNR) for the reconstruction,

$$\alpha_r(n, s) = c(n, s) \langle \alpha(s) \rangle \varepsilon(s) = \frac{tn}{2Ds} \langle \alpha(s) \rangle \exp\left(-\frac{B}{2}\right) s^2, \quad (15)$$

where $\langle \sigma(s) \rangle$ is the average SSNR of the original particle images. This corresponds to the treatment by Rosenthal & Henderson (2003), but separating the natural signal decay with frequency from signal suppression owing to image variation.

The nature of the particle also factors into (15), with practical implications. The number n is the number of asymmetric units, and fewer images are needed for particles with higher symmetry. A larger particle means a larger box size D , which requires more data to perform a reconstruction. The box size also affects the SSNR because the signal has compact support (limited in space to the particle envelope), while the noise extends to the edge of the box. Therefore, the box size should not be excessively large to avoid decreasing the SSNR unnecessarily.

2.5. The amount of data required

(15) can be used to calculate the number of particle images required to reach a specified resolution. For an SSNR of 1, and rearranging (15), the number of images required to extend the information to a frequency s_1 is

$$n_{\alpha_r(n, s_1)=1} = \frac{2Ds_1}{t \langle \alpha(s_1) \rangle} \exp\left(\frac{B}{2} s_1^2\right). \quad (16)$$

The key parameters are therefore the average micrograph SSNR in the denominator and the B factor. With reasonable values for these parameters, the number of particles required to reach a desired resolution can be predicted to a rough extent even before collecting data.

2.6. Fourier shell correlation

The quality of a reconstruction is typically assessed using FSC analysis (Harauz & van Heel, 1986). This is calculated using two 3D maps from independent data sets ('half maps') as

$$\text{FSC}_h(s) = \frac{\sum_s (F_1 F_2^*)}{\left(\sum_s |F_1|^2 \sum_s |F_2|^2 \right)^{1/2}}. \quad (17)$$

Assuming that the cross-terms in (4) and (5) are negligible, and that noise statistics are comparable in the two maps, the approximate relationship between the SSNR in each map and the FSC is (Penczek, 2002)

$$\text{FSC}_s(s) \simeq \frac{\alpha_r(n, s)}{\alpha_r(n, s) + 1}. \quad (18)$$

(18) covers all of the input data, while (17) reflects only half. The relationship between these two measures is therefore

$$\text{FSC}_h(s) \simeq \frac{\text{FSC}_s(s)}{2 - \text{FSC}_s(s)}. \quad (19)$$

As will be shown later (also see Appendix A), this relationship becomes less accurate with limited data owing to the problematic cross-term.

3. Methods

All calculations were performed in *Bsoft* (Heymann, 2018c). Individual programs that were used are indicated in italics.

3.1. The reconstruction algorithm

The reconstruction algorithm in *breconstruct* aims at a straightforward integration in frequency space, as expressed in (2). Each image is packed into the frequency-space volume as a central section in the determined orientation. Interpolation is simple nearest neighbor with oversampling through real-space padding. The newest output from *breconstruct* reports the radial power spectrum, the signal (8), the residual noise (7), the SSNR (9) and FSC_s (18). In cases where half maps are calculated, the traditional FSC_h (17) is also displayed.

3.2. A synthetic case

To test the theory and to demonstrate how to interpret the results, I produced a synthetic map, generated projections and calculated reconstructions. I chose to examine β -galactosidase because of the availability of the data from the EMPIAR database (EMPIAR-10012 and EMPIAR-10013; Iudin *et al.*, 2016). I calculated a map with *bsf* from the coordinates of the β -galactosidase structure (PDB entry 5a1a; Bartesaghi *et al.*, 2015) using the electron-scattering cross-sections as approximated with five Gaussian curves by Peng *et al.* (1996). I generated a random set of projection images with *bproject*. This program extracts central sections of the map Fourier transform, interpolating using the method of Lanzavecchia & Bellon (1995) with a kernel width of 8 and a power of 2. For cases with noise, I added Gaussian (or white) noise to the projections with *brandom* to give an overall SNR of 0.1 (defined as the ratio of the variance of the noise-free projection to the variance of the noise). I then transferred the metadata associated with the projection images to a parameter file (STAR format) using *bmrg*. To examine alignment errors, I imposed deviations in the orientational parameters using *emgrand*. I reconstructed maps from projections (with and without noise, with and without errors, different numbers) with *D2* symmetry to a resolution of 2 Å using *breconstruct*. I calculated the radial power spectrum of the original synthetic map using *bradial*. Fourier shell correlation (Harauz & van Heel, 1986) was calculated using *bresolve* with no masking.

3.3. A real case

I processed the β -galactosidase data set from the EMPIAR database (EMPIAR-10012 and EMPIAR-10013; Iudin *et al.*, 2016) as described previously (Heymann, 2018a). I performed

various reconstructions comparable to those for the synthetic case.

4. Results

I implemented the direct calculation of reconstruction statistics in frequency space in the *Bsoft* program *breconstruct* (Heymann, 2018c). In the following sections, I present the statistics for reconstructions of β -galactosidase calculated from synthetic images derived from an atomic model and from real images (performed as part of the Map Challenge; Heymann, 2018a). Of particular importance are the spectral curves for the map, signal (equations 7 and 14) and noise (equations 6 and 12).

4.1. Reconstruction with exact particle orientations

To eliminate the possibility of signal suppression owing to alignment errors, I started with projections produced from the synthetic map of β -galactosidase, with white noise added to give an overall SNR of 0.1 (defined as the variance of the noise-free projection divided by the variance of the noise). Fig. 1(a) shows the statistics for reconstructions calculated from different numbers of images, each with a curve for the radial power spectrum, the signal power and the noise power. In each case, the noise power of the reconstruction increases with frequency (consistent with equation 12) and decreases with increasing numbers of particles (consistent with equation 6; gray curves in Fig. 1a). The radial power spectrum shows two parts: the first is the natural signal decay owing to the electron-scattering cross-section and the second is residual noise that is averaged out with increased numbers of particle images (solid-color curves in Fig. 1a). The signal (7) emerges from the noise with increasing numbers of particles (dotted curves in Fig. 1a). In the extreme of a large number of image contributions, practically all noise is averaged out and the signal power approaches the radial power spectrum of the original map [the reconstruction signal for 20 000 particles

indicated by the blue dotted curve in Fig. 1(a) is indistinguishable from the radial power spectrum of the original synthetic map].

Equations (17)–(19) relate the traditional even–odd or half-map FSC ($FSC_{h/o}$) to the FSC derived from the SSNR (FSC_s). These two measures are highly similar for the three synthetic reconstructions at low frequencies, but at higher frequencies an apparent residual correlation persists for FSC_s (Fig. 1b). This is owing to the cross-term in (4) and (5) (see Appendix A), which disappears at higher numbers of contributing images (dotted blue curve in Fig. 1b). Because this is evident from a synthetic data set with exact orientational parameters, it is simply a statistical phenomenon that is eventually averaged out with sufficient data.

4.2. Signal suppression by imposing alignment errors

One way to emulate signal suppression, as given in (14), is to impose alignment errors on otherwise exact particle orientations. I therefore generated projections from the synthetic map without any noise and imposed errors on the image orientations. I produced three, each with a combination of rotational and translational errors. The first had an average rotational error of 1° and an average translational error of 1 Å (= 1 pixel), the second 2° and 2 Å, and the third 3° and 3 Å. The signal power of the reconstructions decreases systematically with increasing error (Fig. 2a). The signal power relative to the error-free reconstruction follows an approximately Gaussian decay (Fig. 2b). While exact modeling of such decay curves is complicated and perhaps not tractable (Baldwin & Penczek, 2005), the *B* factors give a reasonable basis for describing signal suppression and, by extension, resolution-limiting effects.

4.3. Reconstruction statistics for a real case

I processed micrographs of β -galactosidase as described in Heymann (2018a). The reconstruction statistics are very

Figure 1 Reconstruction statistics for a synthetic case (β -galactosidase) with noise, calculated from 200 (red), 2000 (green) and 20 000 (blue) images, and imposing *D2* symmetry. (a) Radial power spectra for each of the maps (thick lines), and its corresponding signal (dotted colored lines, equation 7) and noise (gray dashed lines, equation 6). (b) FSC curves calculated the traditional way on half maps (equation 17, thick lines), and from (8), (18) and (19) during reconstruction (dotted lines).

similar to those of the synthetic case [Fig. 3(a) compared with Fig. 1(a)]. The differences reflect the many effects that are not modeled in the simple theoretical analysis of (1). These include inelastic scattering, shape factors, the contrast transfer function (CTF) and how it is corrected (here it is only phase-flipped). Nevertheless, the statistics of the real case have the same features as those of the synthetic case. Of particular note is the dominance of the noise at high frequencies for low numbers of images and the suppression of the signal for larger numbers of images. The FSC_s curves in Fig. 3(b) have the same behavior as for the synthetic case (Fig. 1b). A small difference is that the increased apparent correlation persists for larger numbers of images (dark blue curve in Fig. 3b). This suggests that there are additional contributions to the cross-term that are not reflected in the synthetic case.

The signal decay at larger numbers of images (20 000) is evident in Fig. 3(a). To estimate the decay magnitude as for the synthetic case (Fig. 2) the true signal is needed, but this is not known. Instead, I used the radial power spectrum of the synthetic map, scaled to fit the middle part of the radial power spectrum of the real map (Fig. 4a). The ratio of the signal to this scaled radial power spectrum gives an indication of the decay, which is approximately represented by a B factor of $\sim 70 \text{ \AA}^2$ (Fig. 4b).

4.4. The signal and noise in the particle images

The independence of the noise to alignment errors suggests that the original contribution in the particle images can be estimated from the reconstruction noise using (12) (gray

Figure 2

Signal suppression in the reconstructions of the synthetic case owing to alignment errors. (a) Radial signal power spectra of reconstructions from noiseless images with different levels of orientation errors imposed (each number indicates the imposed standard deviation in translation in pixels and in view orientation in degrees). The black curve is the radial signal power spectrum of a reconstruction without errors. (b) The ratio between the colored curves in (a) and the black curve (circles) represents the suppression of the reconstruction amplitudes owing to orientation error. Gaussian curves were fitted to these (solid lines). Corresponding σ values and B factors are shown.

Figure 3

Reconstruction statistics for a real case (β -galactosidase) calculated as in Fig. 1 from 200 (red), 2000 (green) and 20 000 (blue) images, and imposing $D2$ symmetry. (a) Radial power spectra for each of the maps (thick lines), and its corresponding signal (dotted colored lines, equation 7) and noise (gray dashed lines, equation 6). (b) Fourier shell correlation: FSC_h (thick curves) approaches zero at high frequencies, while FSC_s (dotted curves) retains a significant apparent correlation.

Figure 4

Signal suppression in the real case. (a) Comparison of the signal power spectrum of the reconstruction from 20 000 images (blue) with that of the radial power spectrum of the synthetic map (black). (b) The ratio of the signal [blue curve in (a)] to the reference [black curve in (a)] gives a profile (blue circles) that corresponds to a decay function with a B factor of $73 \text{ \AA}^2$ (dark blue curve).

curves in Fig. 5). The noise power approximates 1 because the algorithm in *breconstruct* rescales the particle images to an average of 0 and a variance of 1 before integration. The low frequencies have more noise, likely owing to inelastic scattering or solvent contributions. This indicates that the noise is not quite white (constant spectral power) as is often assumed (Rosenthal & Henderson, 2003), but it varies much less than the signal ($5\times$ versus $1000\times$). The steady decline of the signal with spatial frequency compared with the noise means that alignment is most influenced by the low frequencies.

4.5. The signal of biomolecules

The ideal reference for a biomolecule is the atomic structure, from which a map can be calculated using electron-

Figure 5

The gray curves show the noise back-calculated from the reconstruction noise for the three cases in Fig. 3 (see equation 12). For comparison, the blue curve is the signal in the map reconstructed from 20 000 images. The noise power is close to 1 because the images are normalized to a variance of 1 on input to the reconstruction algorithm.

scattering cross-sections. However, such a reference is typically not available for new specimens. Because all of these molecules have similarities (Spiegel *et al.*, 2015), there may be a reasonable spectral curve that could be used. I calculated the maps from ten atomic structures of varying shapes and sizes (Table 1) and plotted their radial power spectra (Fig. 6a). It is evident that they all follow a similar pattern, mostly just shifted up or down in power. This can now be used to approximate the radial power of a generic protein, giving a value for the average signal in (15). It also suggests a way to modify the amplitudes of a map (also called ‘sharpening’). The central sections of the three reconstructions from real data are shown in Fig. 6(b). The appearance is severely affected by the presence of high-frequency noise (the parts of the solid curves that increase at high frequency in Fig. 3a). I modified the spectral power of the maps to the average spectral power from Fig. 6(a), resulting in the maps in Fig. 6(c). In all three cases the appearance is significantly improved, with a clear distinction between foreground and background. I then low-pass filtered the maps to their estimated resolutions in Fig. 6(d), further removing high-frequency noise beyond the justifiable information in the maps.

4.6. Predicting the number of particle images that are required to reach a specified resolution

Two important pieces of information are encoded in (16): the SSNR of the original images and a B factor characterizing suppression of the true signal. The reference curve in Fig. 4(a) (the black curve) can be used as the approximate signal, with the noise close to 1 (remember that the reconstruction algorithm sets the variance of the input images to 1). Using the B factor estimated in Fig. 4(b), the predicted number of particles required to reach a defined resolution is shown in Fig. 7 (red curve). This agrees well with actual calculations from reconstructions performed by simple nearest-neighbor interpolation (Fig. 7, yellow diamonds). With a more sophisticated reconstruction in which the input images were padded in real space

to a larger size for better interpolation in frequency space, the estimated resolutions improve (Fig. 7, green triangles). In this particular case, the better interpolation required about half the number of particles to yield the same resolution. Nevertheless, where the signal suppression envelope takes effect at around 4 Å, the estimates converge. This is also the point at which the estimates of maps of β -galactosidase submitted for the Map Challenge cluster (Fig. 7, blue circles; Heymann,

Figure 6
(a) The spectral powers calculated from electron-scattering curves for a variety of biomolecules have very similar profiles. The average (black curve) can be used as a reference for sharpening maps (PDB entries 1c3a, 2abm, 2ei4, 3dyb, 5lf4, 1nqx, 2cf4, 3bwq, 4r7y and 5uf6). (b) Central slices of maps reconstructed from 200, 2000 and 20 000 images. (c) Central slices of maps corresponding to (b) after sharpening using the average profile in (a). (d) Central slices of maps corresponding to (c) after low-pass filtering to the resolution estimated at $FSC_{0.143}$: 16 Å for the 200-image map, 7 Å for the 2000-image map and 4 Å for the 20 000-image map.

Table 1
The molecular structures used to calculate the radial power spectra in Fig. 6(a).

The maps were calculated using *bsf* with a voxel size of 1 Å.

PDB code	Description	Symmetry of map	Box size (voxels)
1c3a	Flavocetin A subunit	C1	100
2abm	Aquaporin Z	C4	150
2cf4	TET1 peptidase	<i>T</i>	150
2ei4	Archaeorhodopsin-2	C1	80
3bwq	SV40 VP1	C5	120
3dyb	Proteinase K	C1	80
4r7y	MCM helicase subunit	C1	140
4ug0	80S ribosome	C1	340
5lf4	20S proteasome	Pseudo- <i>D7</i>	180
5uf6	Adeno-associated virus-DJ	<i>I</i>	340

2018b). These maps reflect slightly better resolutions because the FSC_h curves were calculated using masks, and the reconstruction algorithms feature more sophisticated interpolation than simple nearest neighbor. Of interest is that these resolution estimates occur at the inflection point in the curve. Below the inflection point, the resolution steadily improves with the number of images. The main process here is averaging the noise and revealing the signal. The slope has a third-order to fourth-order dependence on the resolution up to ~20 000 images. Beyond this the slope increases dramatically, corresponding to over a tenth-order dependence on the resolution. This is owing to resolution-limiting factors that cannot easily be overcome by using more particle images acquired in the same way. At this point the user should consider the causes of the signal suppression and how to overcome it.

Figure 7
The number of images required to reach a specified resolution (red curve) as predicted with (16) using the calculated radial power spectrum as the micrograph SSNR (black curve in Fig. 4a) and a *B* factor of 73 Å². Also shown are $FSC_{0.143}$ resolution estimates for reconstructions performed with simple (yellow diamonds) or padded (green triangles) nearest-neighbor interpolations. The blue circles indicate the resolutions calculated with respect to a reference of maps reconstructed from the same data set by participants in the Map Challenge (Heymann, 2018b). The resolution here is defined at a cutoff of 0.5 on the FSC curve, equivalent to an SSNR of 1.

5. Discussion

While SPA has achieved remarkable success recently, it is still dependent on the experience of the user (Heymann *et al.*, 2018, Heymann, 2018b) and a significant amount of trial and error. Even though high-resolution structures are now routinely solved, the effort remains considerable. The current approach is to collect very large data sets, followed by extensive processing. This can be improved by better insight into the relationship between the quality and the quantity of the data, and the practical limits to the achievable resolution. One way to gain some insight is through examining reconstruction statistics as a diagnostic tool. In Section 2 I derive separate equations for the signal and noise powers, and show how these can be used to determine the amount of data required. Ultimately, this understanding needs to be incorporated into automated workflows with feedback for resolving problematic issues.

5.1. The relationship between the number of particles and the resolution

(15) can be considered to be the ‘master’ equation for SPA. It embodies two trends: a true signal that can be recovered by

removing noise through averaging, and a suppression of the signal owing to image variability as characterized by a B factor. Given an SSNR cutoff of 1, (16) then gives an estimate of the number of particle images required to reach a given resolution. Fig. 7 demonstrates the two trends: at small numbers of particles the resolution is mainly limited by the noise in the images, while at large numbers of particles image variability dominates.

In Fig. 8 I explore the influence of different parameters on the number of particles that are required. While box size and symmetry are known parameters, the overall SNR as it is related to the SSNR is not well understood. Here I scale it to fit the results from the β -galactosidase reconstruction, assuming an overall SNR of 0.1. An increase in box size, lower symmetry and decreased SNR all appear to cause similar upward shifts in the profiles. On the other hand, a change in the B factor mostly changes the profile at large numbers of particles (Fig. 8d). What is not illustrated in this figure is the coupling between these parameters. An increase in box size also means a decrease in SNR, while a lower SNR benefits the alignability of the particles and thus a B factor reflecting alignment accuracy.

Figure 8
Effect of the magnitude of different parameters in (16) on the number of particles that are required. (a) Particle box size D . (b) Point-group symmetry (C1, asymmetric; D2, dihedral twofold; I, icosahedral). (c) Overall SNR. (d) B factor. The parameters for the green curve, unless otherwise indicated, are $\tau = 1$, $D = 200$ Å, point group C1, SNR = 0.1, $B = 100$ Å².

This complex interplay between the parameters can only be understood for a given case by analyzing the signal and noise contributions from different numbers of particles. In the analyses of Rosenthal and Henderson (see Fig. 11 in Rosenthal & Henderson, 2003) and Stagg *et al.* (2014), the signal and noise trends are not considered separately. In addition, they do not distinguish between noise-limited and signal suppression contributions as shown in Fig. 7 (see Appendix C). The result is that the natural signal decay owing to the scattering cross-section is combined with image variability described by one B factor. The value in distinguishing the different contributions is in offering a diagnostic tool for SPA. In the following sections I present a strategy to improve the efficiency and effectiveness of SPA using (16).

5.2. Using reconstruction statistics to assess resolution-limiting factors

A typical cryo-EM experiment has three stages: planning, micrograph assessment and full-scale processing (hopefully to high resolution). Even in the planning stage, (16) can be used to predict how much data we would need. In Fig. 6(a) I show that the signal spectra of diverse biomolecules are very similar. Thus, we can use a generic signal spectrum (*i.e.* the average in Fig. 6a) in (16) together with an estimate of the SSNR of a typical micrograph to anticipate the amount of data that is required. The one unresolved issue is how to scale this generic signal spectrum to obtain a representative SSNR. Here, I used the β -galactosidase data to estimate a scale, as performed for Fig. 4(a), and used it in the calculations for Fig. 8. The relationship between the SSNR and an overall SNR is complicated and requires further study.

In the next stage, the SSNR of the micrographs can be calculated from a small number of particles (as was performed for Fig. 5). At this stage it is difficult to estimate a B factor. The user can ignore it (set it to 0 in equation 16) and calculate the number of particles to reach a desired resolution (say 2 Å). Knowing that signal suppression is likely to limit the resolution, it at least gives an indication of the amount of data that is required to determine that. With experience, the user can assess micrograph quality from the SSNR and may decide to revisit issues such as better sample and grid preparation or a change in imaging parameters.

Finally, with a larger amount of data, an assessment can be made of signal suppression and a representative B factor can be calculated. This limit is reached in most SPA projects, when users typically find that adding more data does not improve their reconstructions (as assessed by the FSC). The value in estimating an associated B factor is to enable further analysis of the possible causes of signal suppression.

5.3. The causes of signal suppression

The suppression of the signal is due to image or signal variability, which is typically parameterized by a B factor. Rosenthal & Henderson (2003) defined two types of B factor: B_{image} is defined as image variability resulting from specimen movement, charging, radiation damage, inelastic scattering,

partial beam coherence and structural variability, while $B_{\text{computation}}$ is defined as image variability arising from inaccuracies in alignment and reconstruction. While all of these causes contribute to the overall B factor, the dominant effect for a specific data set can usually be ascribed to one or two of them. Reducing one dominant effect then typically results in another one becoming the leading cause. Thus, resolving resolution-limiting factors turns into a series of remediating actions, some dealing with the nature of the specimen and others with computational approaches. In the following list, I briefly comment on these causes.

(i) Alignment accuracy is a function of the SSNR of the images and the algorithm used (Baldwin & Penczek, 2005; Heymann, 2015). Increasing the micrograph SSNR will improve the alignment accuracy (Heymann, 2015), decrease the number of particles needed and boost the resolution limit.

(ii) The choice of reconstruction algorithm can affect the quality (see below).

(iii) Specimen movement is typically compensated for by dose fractionation and frame alignment (Li *et al.*, 2013). This could be further improved algorithmically by per-particle alignment in each frame. The dose-fractionation scheme itself could be optimized. It is likely that the optimal acquisition for each frame should be tuned for the expected resolution, also taking into account the typical microscope stage drift.

(iv) Charging introduces vibrations that are particularly evident in the micrographs of tilted specimens (Russo & Henderson, 2018a,b). The expectation is that signal suppression would be directional in the original images, but would be isotropic in the reconstruction owing to the random orientations of the particles.

(v) Radiation damage can be compensated for by a progressive low-pass filter towards higher dose frames (Grant & Grigorieff, 2015). A second approach is to select a set of low-dose frames with the highest resolution information (Li *et al.*, 2013).

(vi) Inelastic scattering is mainly a factor at low resolution. Particularly for thick specimens, an energy filter can be used to eliminate a significant portion of the inelastic scattering.

(vii) The electron beam has various envelopes that could limit the resolution (Frank, 1973; Wade & Frank, 1977). These are a function of the microscope design and of the microscope alignment quality.

(viii) Structural variation should be addressed by some form of classification or by addressing a continuum of conformational transitions. In both of these cases the number of images required increases with the number of states or the magnitude of conformational change.

5.4. Other reconstruction algorithms

In principle, the calculation of similar statistics should be performable for any reconstruction algorithm. One caveat is that the reconstruction algorithm must preserve the statistical information in a meaningful way. The algorithm used by *breconstruct* is a relatively simple frequency-space integration

with nearest-neighbor interpolation. Other forms of interpolation perform better in reconstruction, but change the statistics and complicate the interpretation (Penczek, 2002). These include padding the images to a larger size in real space, which is the default in *breconstruct* (Fig. 7 shows a comparison with and without padding). This algorithm appears to improve the noise averaging, but did not change the inflection point at which signal suppression becomes dominant. It remains to be seen whether interpolation by more sophisticated methods such as gridding could alleviate the effects of signal suppression (Abrishami *et al.*, 2015; Penczek *et al.*, 2004).

6. Conclusion

Reconstruction statistics offer an important diagnostic tool to assess the state of electron micrographs and resolution-limiting factors. A user can rapidly assess the quality of electron micrographs, using the feedback to optimize grid-preparation and imaging conditions. S/he can then predict the amount of data needed, making good use of limited resources. Once the initial data set has been processed, the user can determine resolution-limiting effects and identify potential causes. The most important impact of calculating the resolution statistics is therefore the ability to make better data-driven choices in SPA projects.

APPENDIX A

The signal–noise cross-term

The cross-term in (4) and (5) was highlighted by van Heel & Schatz (2005). However, the implications of their derivation suggested that the FSC curve should converge to the expected residual noise of $1/n^{1/2}$. In typical FSC curves the high frequencies oscillate around 0 and not this residual noise. A possible reason is that their equations (7) and (8) are unsubstantiated. In the calculations reported here, the cross-term has significant magnitude with limited data, but approaches 0 for a sufficient number of images (see the FSC_s curves in Fig. 1*b*).

A theoretical treatment of the cross-term remains elusive. In the following derivation I provide my best attempt, with the understanding that it was not borne out by calculations. The term $(S \sum_i^m N_i^*)_{\text{re}}$ in (4) and (5) can be viewed as the product of two independent random sets of values. The expected magnitude is the product of the individual magnitudes:

$$\frac{1}{m} \left(S \sum_i^m N_i^* \right)_{\text{re}} \simeq |S| \left| \frac{1}{m} \sum_i^m N_i \right|. \quad (20)$$

Substituting this back into (4) and (5) and solving the quadratic equation for the signal magnitude then gives

$$|S| = \left| \frac{1}{m} \sum_i^m F_i \right| - \left| \frac{1}{m} \sum_i^m N_i \right|. \quad (21)$$

This is similar to (1), but is not equivalent and cannot be reconciled with actual calculations on data.

APPENDIX B

Effective central section thickness

Any reconstruction of a 3D volume from 2D projection images is approximately equivalent to packing the central sections into the frequency-space volume. The contribution of each central section to the 3D volume is shown in (11). One parameter that is required is the effective thickness of the section, t . This is dependent on the interpolation details and is therefore a function of the algorithm. It can be determined by reconstructing maps from synthetic images without noise and calculating the coverage in 3D frequency space. The coverage for n 2D images packed into a 3D volume with random orientations is given by

$$v(s) = 1 - \left(1 - \frac{t}{2Ds} \right)^n. \quad (22)$$

The effective thickness for nearest-neighbor interpolation is about 0.5. With better interpolation schemes (such as padding the input images to larger sizes), t is closer to 1.

APPENDIX C

B factors

Rosenthal and Henderson (see Fig. 11 in Rosenthal & Henderson, 2003) used the following equation to calculate the number of particles required to reach a specific SNR in the reconstruction:

$$n = \left[\frac{\langle S \rangle}{\langle N \rangle} \frac{30\pi}{N_e \sigma_e} \right] s \exp \left(\frac{B}{2} s^2 \right). \quad (23)$$

They considered the term in square brackets to be constant (C), so that

$$\ln \frac{n}{s} = \ln C + \frac{B}{2} s^2. \quad (24)$$

This has been used in several other studies to calculate a B factor (Liu *et al.*, 2007; LeBarron *et al.*, 2008). However, the bracketed term is not constant, as shown in (16), because the signal decreases with spatial frequency. This approach thus conflates signal suppression with the natural signal decay (Fig. 6*a*), and the corresponding B factor will be much higher than that estimated here (they reported a B factor in the range 300–1000 Å²). One issue is that they collected their data on film, which is expected to give a higher B factor than for a direct detector. Therefore, I cannot make a direct comparison with the estimate in Fig. 4(*b*) using the available information.

The ‘reslog’ plots of Stagg *et al.* (2014) are similar to the right part of the curve in the plot in Fig. 7, with the resolution improving steadily with the number of images. As they state, it is a measure of the quality of the micrographs (*i.e.* the SSNR). However, the interpretation is ambiguous because it does not conform to the theory as laid out here. Furthermore, these plots do not capture the left part in Fig. 7, which is owing to other resolution-limiting factors that are not easily countered with more data.

Acknowledgements

This work was supported by the Intramural Research Program of the National Institute for Arthritis, Musculoskeletal and Skin Diseases, NIH.

References

- Abrishami, V., Bilbao-Castro, J. R., Vargas, J., Marabini, R., Carazo, J. M. & Sorzano, C. O. S. (2015). *Ultramicroscopy*, **157**, 79–87.
- Baker, L. A., Smith, E. A., Bueler, S. A. & Rubinstein, J. L. (2010). *J. Struct. Biol.* **169**, 431–437.
- Baldwin, P. R. & Penczek, P. A. (2005). *J. Struct. Biol.* **150**, 211–225.
- Bartesaghi, A., Merk, A., Banerjee, S., Matthies, D., Wu, X., Milne, J. L. & Subramaniam, S. (2015). *Science*, **348**, 1147–1151.
- Bisogni, M. G., Campbell, M., Conti, M., Delogu, P., Fantacci, M. E., Heijne, E. H. M., Maestro, P., Magistrati, G., Marzulli, V. M., Meddeler, G., Mikulec, B., Pernigotti, E., Rosso, V., Schwarz, C., Snoeys, W., Stumbo, S. & Watt, J. (1998). *Proc. SPIE*, **3445**, 298–304.
- Crewe, A. V., Eggenberger, D. N., Wall, J. & Welter, L. M. (1968). *Rev. Sci. Instrum.* **39**, 576–583.
- Crewe, A. V., Wall, J. & Langmore, J. (1970). *Science*, **168**, 1338–1340.
- Crowther, R. A., Amos, L. A., Finch, J. T., De Rosier, D. J. & Klug, A. (1970). *Nature (London)*, **226**, 421–425.
- Crowther, R. A., DeRosier, D. J. & Klug, A. (1970). *Proc. R. Soc. London Ser. A*, **317**, 319–340.
- Faruqi, A. R. & Henderson, R. (2007). *Curr. Opin. Struct. Biol.* **17**, 549–555.
- Frank, J. (1973). *Optik*, **38**, 519–536.
- Frank, J., Verschoor, A. & Boublik, M. (1981). *Science*, **214**, 1353–1355.
- Gilbert, P. F. C. (1972). *Proc. R. Soc. London B Biol. Sci.* **182**, 89–102.
- Glaeser, R. M. & Hall, R. J. (2011). *Biophys. J.* **100**, 2331–2337.
- Grant, T. & Grigorieff, N. (2015). *Elife*, **4**, e06980.
- Harauz, G. & van Heel, M. (1986). *Optik*, **73**, 146–156.
- Hayward, S. B. & Glaeser, R. M. (1979). *Ultramicroscopy*, **4**, 201–210.
- Heel, M. van & Schatz, M. (2005). *J. Struct. Biol.* **151**, 250–262.
- Henderson, R. (1995). *Q. Rev. Biophys.* **28**, 171–193.
- Heymann, J. B. (2015). *AIMS Biophys.* **2**, 21–35.
- Heymann, J. B. (2018a). *J. Struct. Biol.* **204**, 90–95.
- Heymann, J. B. (2018b). *J. Struct. Biol.* **204**, 360–367.
- Heymann, J. B. (2018c). *Protein Sci.* **27**, 159–171.
- Heymann, J. B., Marabini, R., Kazemi, M., Sorzano, C. O. S., Holmdahl, M., Mendez, J. H., Stagg, S. M., Jonic, S., Palovcak, E., Armache, J. P., Zhao, J., Cheng, Y., Pintilie, G., Chiu, W., Patwardhan, A. & Carazo, J. M. (2018). *J. Struct. Biol.* **204**, 291–300.
- Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. (2016). *Nature Methods*, **13**, 387–388.
- Karuppasamy, M., Karimi Nejadasl, F., Vulovic, M., Koster, A. J. & Ravelli, R. B. G. (2011). *J. Synchrotron Rad.* **18**, 398–412.
- Lanzavecchia, S. & Bellon, P. L. (1995). *J. Vis. Commun. Image Represent.* **6**, 122–131.
- LeBarron, J., Grassucci, R. A., Shaikh, T. R., Baxter, W. T., Sengupta, J. & Frank, J. (2008). *J. Struct. Biol.* **164**, 24–32.
- Li, X., Mooney, P., Zheng, S., Booth, C. R., Braumfeld, M. B., Gubbens, S., Agard, D. A. & Cheng, Y. (2013). *Nature Methods*, **10**, 584–590.
- Liao, H. Y. & Frank, J. (2010). *Structure*, **18**, 768–775.
- Liu, X., Jiang, W., Jakana, J. & Chiu, W. (2007). *J. Struct. Biol.* **160**, 11–27.
- McMullan, G., Faruqi, A. R. & Henderson, R. (2016). *Methods Enzymol.* **579**, 1–17.
- Penczek, P. A. (2002). *J. Struct. Biol.* **138**, 34–46.
- Penczek, P. A. (2010). *Methods Enzymol.* **482**, 35–72.
- Penczek, P. A., Renka, R. & Schomberg, H. (2004). *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **21**, 499–509.
- Peng, L.-M., Ren, G., Dudarev, S. L. & Whelan, M. J. (1996). *Acta Cryst.* **A52**, 257–276.
- Radermacher, M. (2006). *Electron Tomography: Methods for Three-Dimensional Visualization of Structures in the Cell*, edited by J. Frank, pp. 245–273. New York: Springer.
- Rosenthal, P. B. & Henderson, R. (2003). *J. Mol. Biol.* **333**, 721–745.
- Russo, C. J. & Henderson, R. (2018a). *Ultramicroscopy*, **187**, 43–49.
- Russo, C. J. & Henderson, R. (2018b). *Ultramicroscopy*, **187**, 56–63.
- Saxberg, B. E. H. & Saxton, W. O. (1981). *Ultramicroscopy*, **6**, 85–89.
- Saxton, W. O. & Frank, J. (1976). *Ultramicroscopy*, **2**, 219–227.
- Spiegel, M., Duraisamy, A. K. & Schröder, G. F. (2015). *J. Struct. Biol.* **191**, 207–213.
- Stagg, S. M., Noble, A. J., Spilman, M. & Chapman, M. S. (2014). *J. Struct. Biol.* **185**, 418–426.
- Unser, M., Sorzano, C. O., Thévenaz, P., Jonić, S., El-Bez, C., De Carlo, S., Conway, J. F. & Trus, B. L. (2005). *J. Struct. Biol.* **149**, 243–255.
- Unser, M., Trus, B. L., Frank, J. & Steven, A. C. (1989). *Ultramicroscopy*, **30**, 429–434.
- Unser, M., Trus, B. L. & Steven, A. C. (1987). *Ultramicroscopy*, **23**, 39–51.
- Wade, R. H. & Frank, J. (1977). *Optik*, **49**, 81–92.