Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Structural Biology 144 (2003) 219–227

Journal of Structural Biology

www.elsevier.com/locate/yjsbi

Diffraction imaging of single particles and biomolecules

G. Huldt,[a] A. Szőke,[a,b] and J. Hajdu[a,*]

[a] ICM Molecular Biophysics, Biomedical Center, Uppsala University, Box 596, SE 751 24 Uppsala, Sweden
[b] Lawrence Livermore National Laboratory, Livermore, CA 94551, USA

Received 2 June 2003, and in revised form 6 August 2003

Abstract

Theory predicts that with a very short and very intense X-ray pulse, the image of a single diffraction pattern may be recorded from a large macromolecule, a virus, or a nanocluster of proteins without the need for a crystal. A three-dimensional data set can be assembled from such images when many copies of the molecule are exposed to the beam one by one in random orientations. We outline a method for structure reconstruction from such a data set in which no independent information is available about the orientation of the images. The basic requirement for reconstruction and/or signal averaging is the ability to tell whether two noisy diffraction patterns represent the same view of the sample or two different views. With this knowledge, averaging techniques can be used to enhance the signal and extend the resolution in a redundant data set. Based on statistical properties of the diffraction pattern, we present an analytical solution to the classification problem. The solution connects the number of incident X-ray photons with the particle size and the achievable resolution. The results are surprising in that they show that classification can be done with less than one photon per pixel in the limiting resolution shell, assuming Poisson-type photon noise in the image. The results can also be used to provide criteria for improvements in other image classification procedures, e.g., those used in electron tomography or diffraction.

Keywords: Single molecule diffraction; Single particle diffraction; Image classification; Averaging; Orientation; 3D reconstruction

1. Introduction

Emerging radiation sources offer exciting new possibilities in biomolecular imaging. X-ray free-electron lasers will provide femtosecond X-ray pulses with a peak brilliance more than 10 orders of magnitude higher than that currently available from synchrotrons. Such light sources may permit non-crystalline biological samples to be imaged with X-rays, and could thus remove a current bottle-neck in structure determination (Neutze et al., 2000). Unfortunately, the intense radiation pulse emitted by the laser will destroy any biological sample in a single shot, precluding the collection of multiple diffraction patterns from a single particle or molecule. We therefore assume that the sample is reproducible, and that single-shot diffraction images can be collected from individual sample particles exposed to the beam one-by-one in unknown orientations. The mathematical treatment of this problem is not unique to planned experiments with X-ray free-electron lasers, but can be extended to diffraction studies with electrons, neutrons, and other types of scattering probes. As a consequence, this paper has a broder scope than simply anticipating experiments with X-ray lasers.

The diffraction pattern of an object is proportional to the squared modulus of the molecular transform (the three-dimensional Fourier transform of the electron density). The coordinates of the diffraction space, usually called reciprocal space, are those of the scattering vector (or momentum transfer vector) between the incident and scattered X-rays. In order to reconstruct the electron density, reciprocal space must be sampled with sufficient density and the diffracted intensities must be known with an acceptable accuracy. In view of this, there are two reasons why a large number of diffraction patterns need to be collected. As a single diffraction image samples only a spherical slice through the origin of reciprocal space (see Fig. 1), so the sample must be imaged in multiple orientations for the space to be adequately covered. In addition, the signal-to-noise ratio of raw diffraction images will probably be insufficient for

[*] Corresponding author. Fax: +46-18-51-17-55.
E-mail address: janos@xray.bmc.uu.se (J. Hajdu).

1047-8477/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2003.09.025

a high-resolution reconstruction, and it will be necessary to obtain a redundant data set so that the signal can be enhanced by averaging.

When the orientation of the samples is unknown, the images must be classified according to the view of the sample that they present before signal averaging is possible. Methods to sort and average images have been developed for single-particle electron microscopy (Mueller et al., 2000; Saxton and Frank, 1977; van Heel, 1987; van Heel et al., 1996; van Heel et al., 1997), and have produced substantially increased resolution even for irregular objects like the ribosome (Mueller et al., 2000). With particles displaying high symmetry, the resolution can be extended further by exploiting the symmetry of the structure (Bottcher et al., 1997; Stowell et al., 1998).

There are important differences between the task of classifying tomographic images of electron microscopy (micrographs) and diffraction patterns of single molecules. Some of these stem from differences between planar (tomography) and spherical sectioning (diffraction), while others reflect differences in the way the images are formed, which also affects their statistical properties. Perhaps the most prominent difference is that the diffraction pattern has a known center, whereas in the micrograph, the molecular image has to be located and centered. Equally significant are the differences in background: in the micrograph the molecular image and the background are separate (although the background contributes to the noise in the image), but in a diffraction pattern there is no obvious way to distinguish the background from the diffraction pattern. Also important is that the micrograph has to be corrected for imperfections of the microscope (the contrast transfer function), whereas diffraction images are perfect in that sense and need no correction. We note that diffraction patterns can also be obtained in electron microscopes, with similar advantages and disadvantages as discussed here.

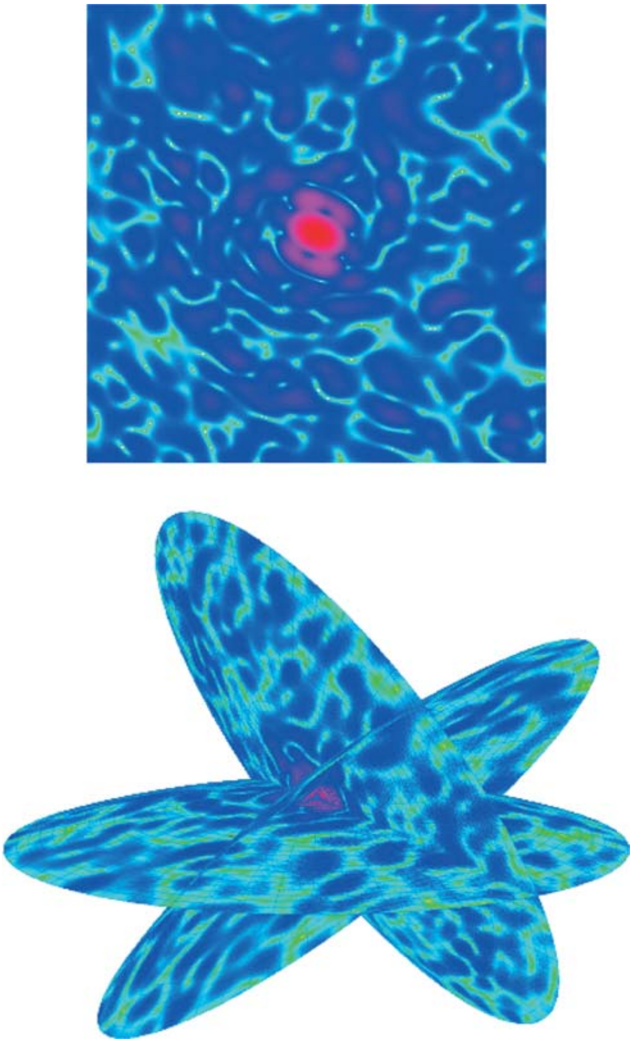

Fig. 1. Intersection of images in diffraction space. The top figure shows a diffraction image of lysozyme. The image is a projection of a spherical section of the molecular transform onto a plane. The bottom figure shows three diffraction images that intersect in diffraction space.

1.1. Classification and averaging of diffraction images

Averaging techniques are based on the assumption that the data set is redundant. The images can thus be sorted into classes that correspond to a distinct view (orientation) of the sample. Images within each class are then averaged; if the classification is correct, the signal adds constructively but the noise does not. We note that, according to the sampling theorem (Jerri, 1977), a finite set of views of the sample is sufficient for full reconstruction; thus, it is sufficient if the data set is redundant with respect to a number of views satisfying this condition. Errors in classification as well as heterogeneity in the samples degrade the signal to noise ratio and the intrinsic resolution of the class averages. It is

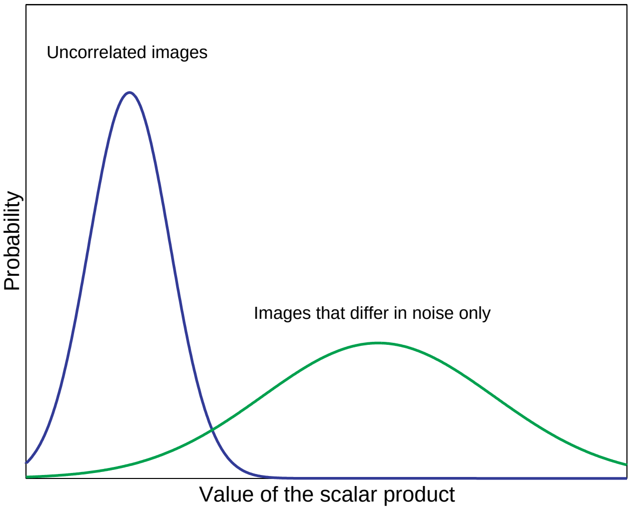

Fig. 2. Overlapping distributions: The area of the overlapping sections gives the probability of misalignment.

important, therefore, that the number of classes be adapted to the signal-to-noise ratio of the raw images. Methods to accurately classify diffraction images with extremely low signal-to-noise ratio, as well as methods to identify wrongly classified images, need to be developed.

Once a complete set of averaged images is obtained, they can be used as reference images to check and correct the original classification of each noisy diffraction image in an iterative process. Images that are significantly different from any of the class averages can be removed at this stage. Also, the procedure must at some point include a search for images that present the same view of the sample, but which are rotated with respect to one another around the axis of the beam.

1.2. Construction of a three-dimensional data set

After classification and averaging, the mutual three-dimensional orientation of the class averaged images must be determined in order to assemble a

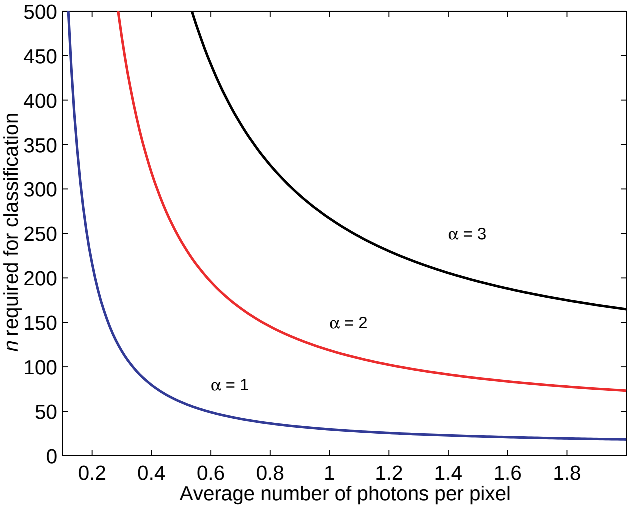

Fig. 3. The number of pixels n needed for classification as a function of the average number of photons scattered into a pixel. Curves are drawn for $\alpha = 3$ (corresponding to a 99% certainty of the classification), $\alpha = 2$ (95% certainty) and $\alpha = 1$ (68%). This is independent of the size of the molecule.

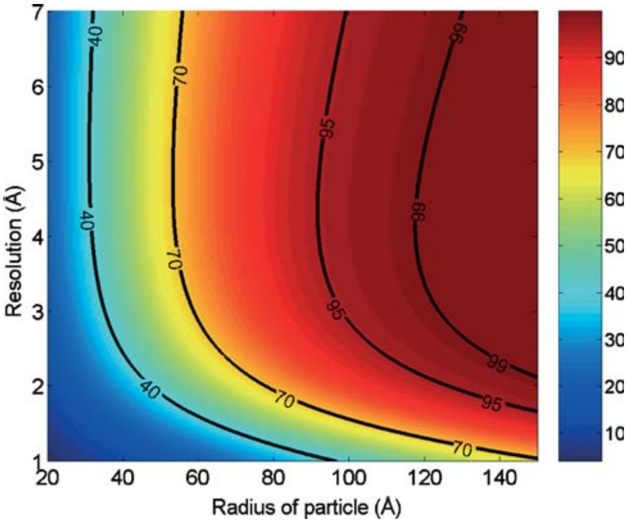

Fig. 5. Certainty of classification as a function of radius and resolution. Contours corresponding to 40, 70, 95, and 99% certainty of classification ($\alpha \approx 0.5$, 1, 2, and 3) are indicated, as well as a colorbar with values in percent. For the calculations we assumed that 3×10^{12} photons were focused into a 100 nm spot. The density of the particle was taken as 1/15 atoms/Å^3, corresponding to an electron density of 0.4 electrons/Å^3.

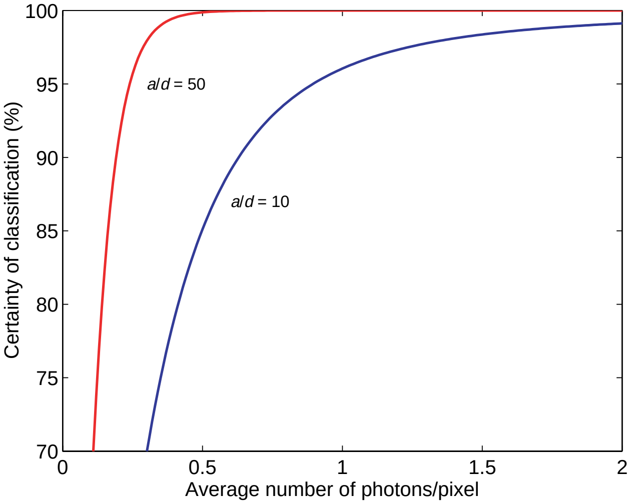

Fig. 4. Certainty of classification as a function of the average number of photons per pixel. The results depend on the ratio between the radius of the particle a and the resolution d.

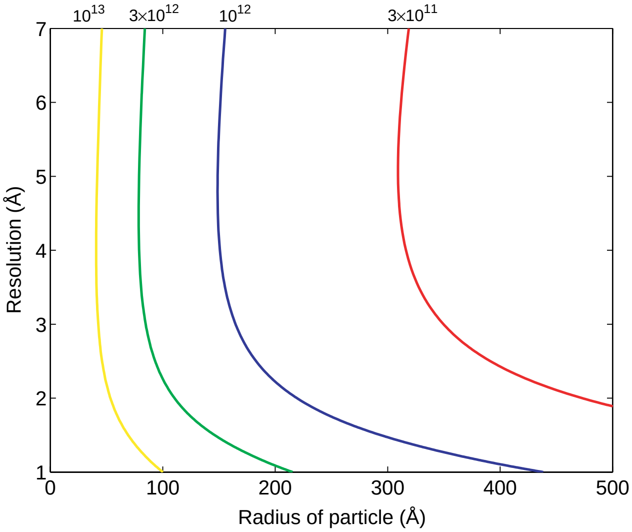

Fig. 6. Classification with 90% certainty at different radii, resolutions and incident number of photons. The density of the particle is 1/15 atoms/Å^3.

three-dimensional data set. This may be possible through the method of common lines (see e.g. Frank, 1996; van Heel et al., 2000), a technique widely used in electron microscopy, where the micrographs represent planar sections through the center of the molecular transform. Diffraction images are different and represent spherical sections. Each pair of images will intersect in an arc that also passes through the origin of the molecular transform (Fig. 1). If the signal (after averaging) is strong enough for the line of intersection to be found in two averaged images, it will then be possible to establish the relative orientation of these images. We note that due to the curvature of the sections, the common arc will provide a three-dimensional fix rather than a hinge-axis. Moreover, the centric symmetry of the modulus of the molecular transform ensures that we obtain 2×2 independent repeats of the common lines in the two images. This feature provides redundancy for determining sample orientation, and is unique to diffraction images.

1.3. Reconstruction of the electron density

The molecular transform is related to the electron density simply and directly by a three-dimensional Fourier transform. Unfortunately, the formation of diffraction images is associated with a loss of information: the molecular transform is a complex, continuous function, whereas the diffraction data are real, discrete, and irregularly spaced in reciprocal space. This leads to a reconstruction problem where the data contain less information then the solution. Such a problem is ill-posed, and as a consequence a very broad set of solutions may fit the data within experimental error. To cure the ill-posedness, we need to include additional information about the sample that constrains the solutions to those that are physically acceptable, and thus allows us to discriminate between spurious solutions and those that are realistic. Classical crystallography has a similar problem.

It was surmised by Sayre (Sayre, 1980) that if the amplitudes of the molecular transform could be oversampled, there would be enough information to replace the lost phases and reconstruct the electron density. The idea has its basis in sampling theory, which states that a band-limited function, such as the molecular transform of a finite-size molecule, can be fully represented by a set of discrete equidistant samples (Jerri, 1977). By sampling the amplitudes more finely than the sampling theorem requires, it may be possible to compensate for the missing phases.

In a recent publication (Szőke, 1999), Szőke has shown that the electron density can, indeed, be reconstructed from a simulated, oversampled continuous diffraction pattern, obtained from a crystal that is made of two similar molecules. He used an approach based on principles used in holography (encoded in the EDEN package, Szőke 1997) and some a priori information. Miao, Hodgson and Sayre used the iterative Gerchberg–Saxton–Fienup algorithm (Gerchberg and Saxton, 1972) to successfully reconstruct electron densities from both simulated (Miao et al., 2001) and real (Miao et al., 2002) diffraction images from non-crystalline samples. A third demonstration, by Oszlányi and Faigel (unpublished) used a maximum likelihood optimizer for this purpose. These approaches represent major developments in phasing, and could be applied to obtain three-dimensional structures from oversampled diffraction images like those of single particles and molecules.

In classical crystallography, the set of Bragg reflections constitute a uniform three-dimensional grid/lattice in reciprocal space. (Actually, both Szőke (1999) and Miao et al. (2001, 2002) used diffraction intensities measured on a three-dimensional regular grid.) Diffraction data sets derived from samples without translational symmetry, on the other hand, yield a highly non-uniform sampling of the molecular transform with a decreasing sampling density at higher resolutions. This is also true for tomograms. One could limit the analysis to those samples that lie on a regular grid, but this would be a very inefficient use of data and seems incompatible with the idea of a highly oversampled diffraction pattern. Interpolating onto a regular grid does not improve the situation; it moves the problem of ill-posedness from real space to reciprocal space, but does not change its nature. Reconstruction algorithms will have to deal intelligently with the above problems.

Fortunately, there are extensive mathematical treatments of matrix inversion (Golub and Loan, 1996), image processing (Bertero and Boccaci, 1998), and of reconstruction in computed tomography (Natterer, 1986; Natterer and Wübbeling, 2001). One can state with some confidence that those inverse problems have similar difficulties, but are "easier." Therefore, reconstruction algorithms for single particle diffraction will be a subset of those that work well for matrix inversion or tomography. We have recently extended EDEN, the holographic method for reconstructing the electron density in crystals, to deal with diffraction patterns from single particles (Hau-Riege et al., in preparation), and expect to be able to find the optimum electron density under conditions of incomplete, noisy measurements on an irregular set of points in reciprocal space.

2. Classification of diffraction images

The first step in the reconstruction process is to classify the diffraction images according to the view of the sample that they present. The images within each class can then be averaged to produce the set of high-quality views of the sample that is required for an atomic-resolution reconstruction of the structure. It is the precision and noise-tolerance of the classification procedure, rather than that of the reconstruction method, that sets the lower limit on the quality of the raw diffraction images. In the following sections, we estimate the minimum number of photons that must be scattered into a diffraction image in order for a set of images to be accurately classified.

The essence of the classification problem is whether two images present similar views of the sample or not. In order to determine this, we divide each image into resolution elements, or pixels, where we simply add the diffracted intensity. More precisely, we think of an image as a vector $\boldsymbol{g} = \{g_0, g_1, \dots, g_{n-1}\}$, where n is the number of pixels used in the representation and g_k is the total photon count of the pixel. Note that with this definition, a pixel does not necessarily correspond to a physical resolution element of the detector but can be of any size or shape. We then correlate two images $\boldsymbol{g}$ and $\boldsymbol{h}$ through their scalar product, or the discrete cross-correlation function at zero lag (displacement)

$$(\boldsymbol{g}, \boldsymbol{h}) = \sum_{i=0}^{n-1} g_i h_i. \tag{1}$$

To determine whether a set of images can be classified by this similarity measure, we derive the probability distribution of the scalar product for two limiting cases: (i) images that present different views of the sample and (ii) images that present the same view of the sample but differ in the distribution of noise. Classification will be regarded as possible if the two distributions can be distinguished.

2.1. The statistics of the diffraction image

2.1.1. The instantaneous intensity scattered by a molecule

The instantaneous intensity elastically scattered within the differential solid angle $\mathrm{d}\Omega$, centered on the scattering vector $\boldsymbol{k}$, by a particle with electron density ρ is proportional to the squared modulus of the molecular transform $F = \mathcal{F}[\rho]$ (Shmueli, 1996), where $\mathcal{F}$ stands for the Fourier transform

$$I(\boldsymbol{k}, t)\mathrm{d}\Omega = |F(\boldsymbol{k}, t)|^2 I_\mathrm{T}(\boldsymbol{k}, t)\mathrm{d}\Omega. \tag{2}$$

Here I_T is the intensity per unit solid angle scattered from a single free electron (Thomson scattering),

$$I_\mathrm{T}(\boldsymbol{k}, t) = r_\mathrm{e}^2 \mathfrak{P}(\boldsymbol{k}) I_\mathrm{in}(t), \tag{3}$$

where r_e^2 is the classical electron radius, $\mathfrak{P}$ a factor that depends on the polarization of the incident radiation, and I_in is the intensity of the incident electric field in photons per unit area and unit time.

2.1.2. The integrated intensity

The integrated intensity measured during time $t_2 - t_1$ within a pixel that spans solid angle $\Omega_\mathcal{P}$ is given by

$$W(\boldsymbol{k}) = \int_{\Omega_\mathcal{P}} \mathrm{d}\Omega \int_{t_1}^{t_2} I(\boldsymbol{k}, t)\mathrm{d}t. \tag{4}$$

In the present analysis we assume that the electron density stays approximately unchanged for the duration of the pulse, so that only the incident intensity varies in time. We can therefore write the integration over time as

$$\int_{t_1}^{t_2} I(\boldsymbol{k}, t)\mathrm{d}t = |F(\boldsymbol{k}, t)|^2 W_\mathrm{T}(\boldsymbol{k}), \tag{5}$$

where

$$W_\mathrm{T}(\boldsymbol{k}) = r_\mathrm{e}^2 \mathfrak{P}(\boldsymbol{k}) \int_{t_1}^{t_2} I_\mathrm{in}(t) = r_\mathrm{e}^2 \mathfrak{P}(\boldsymbol{k}) W_\mathrm{in}, \tag{6}$$

W_in being the total number of photons incident on the sample within the time interval. We also assume that the scattered intensity is approximately constant over the solid angle $\Omega_\mathcal{P}$, so that Eq. (4) can be written as

$$W(\boldsymbol{k}) = \int_{\Omega_\mathcal{P}} |F(\boldsymbol{k}, t)|^2 W_\mathrm{T}(\boldsymbol{k})\mathrm{d}\Omega = |F(\boldsymbol{k}, t)|^2 W_\mathrm{T}(\boldsymbol{k})\Omega_\mathcal{P}. \tag{7}$$

2.1.3. The statistics of the intensity

Biological macromolecules can be represented as a collection of atoms distributed in space. To a first approximation, this distribution can be regarded as random. The probability distribution of diffraction intensities from a crystal exposed to polarized radiation were derived by Wilson (1949) and the derivation for a non-crystalline asymmetric particle is analogous. It shows that the squared modulus of the molecular transform follows negative exponential statistics

$$p(|F|^2) = \frac{1}{\langle |F|^2 \rangle} \mathrm{e}^{-|F|^2/\langle |F|^2 \rangle}. \tag{8}$$

The approximation holds well at reasonably high resolution (say higher than 3 Å). It follows from Eq. (7) that under the assumptions we have made, the same statistics hold for the integrated intensity W

$$p(W) = \frac{1}{\langle W \rangle} \mathrm{e}^{-W/\langle W \rangle}. \tag{9}$$

We will take $\langle W \rangle$ to be the average of Eq. (7) taken over all angles at a constant k, thus making it a function of k only

$$\langle W \rangle(k) = \frac{\int W \mathrm{d}\Omega}{\int \mathrm{d}\Omega} = \frac{\Omega_\mathcal{P}}{4\pi} \int |F(\boldsymbol{k})|^2 W_\mathrm{T}(\boldsymbol{k})\,\mathrm{d}\Omega = \Omega_\mathcal{P} \langle |F|^2 W_\mathrm{T} \rangle(k). \tag{10}$$

2.1.4. The statistics of the photon count

Due to the fundamentally stochastic nature of the interaction between radiation and matter, the number of photons K actually recorded by a detector will deviate from the classical value W. These deviations, which we will refer to as *photon noise*, follow a Poisson distribution (Goodman, 2000)

$$p(K|W) = \frac{W^K}{K!}\mathrm{e}^{-W}. \tag{11}$$

Combining Eqs. (9) and (11) and integrating to calculate the total probability (Papoulis, 1991) of the photon count, we arrive at the Bose–Einstein distribution (geometrical distribution) of the photon count (Goodman, 2000)

$$p(K) = \frac{1}{1+\langle W\rangle}\left(\frac{\langle W\rangle}{1+\langle W\rangle}\right)^K. \tag{12}$$

2.2. The distribution of the scalar product

We derive the distribution of the scalar product for the special case where the image vectors are constructed from a single annulus of high-resolution pixels. We will also assume that the pixels are all independent samples of the molecular transform in the sense of the sampling theorem (Jerri, 1977). Under these conditions, all pixels are independent and identically distributed, and the distribution is given by Eq. (12).

Let the two image vectors $\boldsymbol{g} = \{g_0, g_1, \ldots, g_{n-1}\}$ and $\boldsymbol{h} = \{h_0, h_1, \ldots, h_{n-1}\}$ be realizations of the random vectors $\boldsymbol{G}$ and $\boldsymbol{H}$. Likewise, let the corresponding noise-free images vectors $\boldsymbol{g}^{(0)}$ and $\boldsymbol{h}^{(0)}$ be realizations of the random vectors $\boldsymbol{G}^{(\mathbf{0})}$ and $\boldsymbol{H}^{(\mathbf{0})}$. As the scalar product is a sum of independent random variables, the central limit theorem tells us that it will be asymptotically normally distributed (the distribution is essentially normal for n as low as 30) (Papoulis, 1991). As $\boldsymbol{G}$ and $\boldsymbol{H}$ are also identically distributed we can write

$$(\boldsymbol{G}, \boldsymbol{H}) \in \mathrm{AsN}(n\mu, \sqrt{n}\sigma), \tag{13}$$

where μ and σ denote the common mean and average of the products $G_k H_k$. They are computed from the joint probability density function $p(g_k, h_k)$ of an arbitrary pair of discrete random variables G_k and H_k

$$\begin{aligned} \mu &= \sum_{i,j} ijp(i,j), \\ \sigma^2 &= \sum_{i,j}(ij-\mu)^2 p(i,j), \end{aligned} \tag{14}$$

where the sums are over all integers i and j. Through the theorem of marginal distributions (Papoulis, 1991) we can express the joint probability density function as

$$\begin{aligned} p(g_k,h_k) &= \int\int p\left(g_k,h_k,g_k^{(0)},h_k^{(0)}\right)\mathrm{d}g_k^{(0)}\,\mathrm{d}h_k^{(0)} \\ &= \int\int p\left(g_k|g_k^{(0)},h_k|h_k^{(0)}\right)p\left(g_k^{(0)},h_k^{(0)}\right)\mathrm{d}g_k^{(0)}\,\mathrm{d}h_k^{(0)} \\ &= \int\int p\left(g_k|g_k^{(0)}\right)p\left(h_k|h_k^{(0)}\right)\mathrm{p}\left(g_k^{(0)},h_k^{(0)}\right)\mathrm{d}g_k^{(0)}\,\mathrm{d}h_k^{(0)}. \end{aligned} \tag{15}$$

In the second equality we have used Bayes' theorem of conditional probability and in the third the fact that the noise in the two images is independent.

The first two factors in the last integral are given by Eq. (11), while the third can be obtained from Eq. (9). If the noise-free images $\boldsymbol{g}^{(0)}$ and $\boldsymbol{h}^{(0)}$ correspond to independent views of the sample, then their joint probability function factorizes

$$p\left(g_k^{(0)}, h_k^{(0)}\right) = p\left(g_k^{(0)}\right)p\left(h_k^{(0)}\right). \tag{16}$$

If the images display the same view of the sample, then the noise-free images are identical.

$$p\left(g_k^{(0)}, h_k^{(0)}\right) = p\left(g_k^{(0)}\right)\delta\left(h_k^{(0)} - g_k^{(0)}\right). \tag{17}$$

In both cases we have expressed the integrand of Eq. (15) in terms of known functions, and the resulting integrals can be solved analytically. By inserting the resulting distributions into (14), it is possible to calculate μ and σ^2 for the two cases in terms of the mean classical (photon-noise-free) photon count $\langle W\rangle$ scattered to the pixel. The results are presented in Table 1.

2.3. Classification criterion

The mean and variance of the scalar product takes on higher values if the images present the same view of the sample than if they present independent views. We will consider the two distributions to be distinct if their overlap is smaller then a given fraction of their total area (Fig. 2).

Let the mean of the scalar product be denoted $n\mu_{gg'}$ for images presenting the same view and $n\mu_{gh}$ for inde-

Table 1
Expectation value and variance of the product of the photon count in two pixels at the same resolution

	Noise-free	With Poisson noise
μ_g	$\langle W\rangle$	$\langle W\rangle$
σ_g^2	$\langle W\rangle^2$	$\langle W\rangle^2 + \langle W\rangle$
μ_{gh}	$\langle W\rangle^2$	$\langle W\rangle^2$
σ_{gh}^2	$3\langle W\rangle^4$	$3\langle W\rangle^4 + 4\langle W\rangle^3 + \langle W\rangle^2$
$\mu_{gg'}$	$2\langle W\rangle^2$	$2\langle W\rangle^2$
$\sigma_{gg'}^2$	$20\langle W\rangle^4$	$20\langle W\rangle^4 + 32\langle W\rangle^3 + 13\langle W\rangle^2$

The subscript gg' indicates two images presenting the same view, gh two images presenting different views. $\langle W\rangle$ is the average (classical) intensity at that resolution, and it is expressed in number of photons per pixel.

pendent images, where n is the number of pixels used in the representation. The difference between the means of the two distributions is then

$$n\delta = n\left(\mu_{gg'} - \mu_{gh}\right). \tag{18}$$

Analogously, the sum of the standard deviations is

$$\sqrt{n}S = \sqrt{n}\left(\sigma_{gg'} + \sigma_{gh}\right). \tag{19}$$

The two distributions will be regarded as distinct if the difference between their means is a factor $\alpha/2$ larger than the sum of their standard deviations

$$n\delta > \frac{\alpha}{2}\sqrt{n}S \tag{20}$$

or equivalently, if

$$n > \left[\frac{\alpha}{2}\left(\frac{S}{\delta}\right)\right]^2. \tag{21}$$

The parameter α determines the area of overlap between the distributions and thus the probability of a correct classification. It is analogous to the standard deviation of a normal distribution, i.e., $\alpha = 2$ corresponds to a probability of correct classification of approximately 95%. Eq. (21) connects the number of independent pixels (n) to the significance of the classification (through α) and to the average number of photons scattered into a pixel (through S, δ and Table 1). In Fig. 3, this relationship is plotted for a few values of α.

2.4. Interpretation in terms of particle size and achievable resolution

2.4.1. The number of independent pixels

In this analysis we have assumed that each pixel measures data that is independent of the data in the adjoining pixel. The number of such pixels that can be extracted from a diffraction pattern depends on the band limit of the molecular transform, or equivalently, on the support of the electron density. For a particle that can be inscribed in a cube with side a, points that are separated by a distance of $1/2a$ in each dimension are guaranteed to be independent by the sampling theorem. An estimate of the number of independent sample points within an annulus of radius k in the image is then

$$n(a,d) = \frac{2\pi k}{1/2a} = 4\pi ak = 4\pi\frac{a}{d}. \tag{22}$$

Note that $d = 1/k$ is what we call the resolution of the annulus (we use this term because d is related to the resolution of the reconstructed electron density). This estimate is low, mainly because the cube is a crude estimate of the molecular shape. However, we see that the number of independent pixels along the circle depends on the quotient a/d.

By considering the limiting case when the inequality (21) becomes an equality, solving for α and using Eq. (22) for n,

$$\alpha = 2\sqrt{n}\frac{\delta}{S} = 4\sqrt{\pi\frac{a}{d}}\frac{\delta}{S}, \tag{23}$$

we obtain a relation for the probability of correct classification that depends on the average photon count per pixel and the quotient a/d. In Fig. 4 we have plotted the probability as function of average photon count for two different values of a/d, using Eq. (23) and Table 1 with Poisson noise.

2.4.2. The average photon count per independent pixel

The next step is to express the average photon count per pixel as a function of resolution and particle size. The average photon count per pixel is given by Eq. (10) and restated here

$$\langle W\rangle(k) = \Omega_{\mathcal{P}}\langle|F|^2 W_{\mathrm{T}}\rangle(k).$$

To estimate the solid angle $\Omega_{\mathcal{P}}$ spanned by the pixel, we note again that the boundary of the particle constitutes the band limit of the molecular transform. The distance between independent sample points of the molecular transform (in a cartesian sampling scheme) is then $1/2a$, and the area of an independent unit of the transform will be taken to be $(1/2a)^2$. Based on the Ewald construction (Ewald, 1921), we can say that an independent pixel collects all those photons whose wave vectors fall within an area of $(1/2a)^2$. Since the length of the wave vector is defined as $1/\lambda$, the solid angle $\Omega_{\mathcal{P}}$ spanned by an independent pixel is

$$\Omega_{\mathcal{P}} = \left(\frac{\lambda}{2a}\right)^2. \tag{24}$$

To calculate $\langle|F|^2 W_{\mathrm{T}}\rangle(k)$ we will assume that the incident radiation is unpolarized, so that

$$\langle W_{\mathrm{T}}\rangle(k) = r_{\mathrm{e}}^2 W_{\mathrm{in}}\langle\mathfrak{P}\rangle(k), \tag{25}$$

$$\langle\mathfrak{P}\rangle(k) = \frac{1}{8}(8 + k^4\lambda^4 - 4k^2\lambda^2), \tag{26}$$

which allows us to write

$$\langle|F|^2 W_{\mathrm{T}}\rangle(k) = \langle|F|^2\rangle(k)\langle W_{\mathrm{T}}\rangle(k). \tag{27}$$

Note that there is a discrepancy here with the calculation of the statistics, in which we assumed that the incident radiation is polarized.

The angular average of the squared modulus of the molecular transform remains to be calculated. To do this we assume that the particle consists of N_{C} carbon equivalent atoms. Under the isolated-atom approximation (Shmueli, 1996) we can write

$$\langle|F|^2\rangle(k) = N_{\mathrm{C}}f_{\mathrm{C}}^2(k), \tag{28}$$

where f_C is the Fourier transform of the atomic electron density (the atomic scattering factor for carbon).

Up till now, all we have assumed about the molecules shape is that it can be inscribed in a cube with side $2a$. To relate the number of atoms N_C to the size, we will assume that the particle is spherical with radius a and density ρ_C. The number of carbon atoms is then related to the volume by

$$N_C = \rho_C \frac{4\pi a^3}{3}. \tag{29}$$

We arrive at the following expression for the average number of photons scattered into an independent pixel,

$$\langle W \rangle(a, k) = \left[\frac{\pi}{3}\lambda^2 \rho_C r_e^2\right] a f_C^2(k) \langle \mathfrak{P} \rangle(k) W_{\text{in}}. \tag{30}$$

Eq. (30) in conjunction with Table 1 gives us expressions for S and δ in Eq. (21). For our quantitative conclusions, we assume an incident X-ray pulse which is focused into a spot with diameter 100 nm. The density of the carbon cloud is 1/15 atoms/Å^3 and the scattering factor is calculated trough the analytical approximation given in Wilson (1995). In Fig. 5 we plot the probability of correct classification as a function of particle radius and resolution when the number of incident photons W_{in} is 3×10^{12}. In Fig. 6 we plot the maximum resolution at which classification is possible, as a function of particle radius and for several different intensities of the X-ray pulse.

The decrease in accuracy of classification at low resolution—despite a higher average photon count—reflects the fact that fewer independent data points are available within a low resolution annulus than within a high resolution annulus. The double-valued curves in Figs. 5 and 6, while puzzling at first sight, reflect this feature.

3. Conclusions

We have presented a simple but realistic statistical model for the classification of diffraction images. Our quantitative conclusions are presented in Eqs. (21), (22), and (30) and Figs. 5 and 6, which connect the number of incident X-ray photons, the particle size and the achievable resolution. We have shown that less then one photon per independent pixel can be enough for classification, even in the presence of a Poisson-type photon noise. As expected, the larger the particle and the larger the incident X-ray fluence, the higher the resolution and the higher the significance level of the classification.

The classification scheme that we have analyzed can be expanded and improved upon in various ways. It does not take low-resolution data into account, and those will obviously be included in any practical solution. Low-resolution pixels have higher intensities, so they have better statistics and could help "homing in" early on in the classification. Nevertheless, their angular resolution is lower, and the quality of the averaged images will ultimately be determined by our ability to align the high-resolution data. Moreover, the classification criterion is based on the statistical properties of a generic diffraction image; a more sensitive decision could be made by adapting the criterion to each specific pair of images, at a computational cost. The accuracy of the classification should also be improved by iteration; i.e., if, after a number of classes are established by a first round of classification, their class averages are used as classifiers. We should then compare each noisy image with every one of the class averages, and decide whether the image has been classified properly or belongs to a different class. Such procedures are used successfully in electron microscopy (van Heel et al., 2000) .

In this initial analysis, we did not consider molecular motion and the heterogeneity of individual molecules. The spatial variance of the diffraction pattern is the signal that we use for classification. Molecular motion will cause this variance to decrease, so that the diffraction intensities within each resolution shell approach the average value given by the sum of the scattering factors. Differences between individual sample molecules will have the same effect on the averaged images. Radiation damage will further increase the molecular motion during the pulse, substantially if the pulse is long. We intend to approach these effects, and the limits that they impose on classification, in forthcoming work. We also intend to discuss the improvements possible through the iterated classification outlined above.

This paper deals with general problems in classifying diffraction patterns from reproducible particles/molecules/structures exposed to a wave front in random and unknown orientations. The importance of our treatment is that it establishes clear statistical (mathematical) criteria for the achievable resolution of diffraction images. Such criteria can establish standards of achievement of classification algorithms as well as guidance for research into new methods of signal processing.

Acknowledgments

We are grateful to Marin van Heel for discussions. This work was supported by the Swedish Research Councils STINT and VR.

References

Bertero, M., Boccaci, P., 1998. Inverse Problems in Imaging. IOP Publishing, Bristol.

Bottcher, B., Wynne, S.A., Crowther, R.A., 1997. Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. Nature 386, 88–91.

Ewald, P.P., 1921. Kristallogr. Z., 129–156.

Frank, J., 1996. Three-Dimensional Electron Microscopy of Macromolecular Assemblies. Academic Press, San Diego.

Gerchberg, R.W., Saxton, W.O., 1972. A practical algorithm for the determination of phase from image and diffraction plane pictures. Optik 35, 237–246.

Golub, G.H., van Loan, C.F., 1996. Matrix Computations, third ed. The Jason Hopkins Univ. Pr., Baltimore.

Goodman, J.W., 2000. Statistical Optics, John Wiley & Sons, Inc., reprint.

Jerri, A.J., 1977. The Shannon sampling theorem—its various extensions and applications: a tutorial review. Proc. IEEE 65 (11), 1292–1304.

Miao, J., Hodgson, K., Sayre, D., 2001. An approach to three-dimensional structures of biomolecules by using single-molecule diffraction images. Proc. Natl. Acad. Sci. USA 98 (12), 6641–6645.

Miao, J., Ishikawa, T., Johnson, B., Anderson, E.H., Lai, B., Hodgson, K.O., 2002. High resolution 3D X-ray diffraction microscopy. Phys. Rev. 89, 088303.

Mueller, F., Sommer, I., Baranov, P., Matadeen, R., Stoldt, M., Wohnert, J., Gorlach, M., van Heel, M., Brimacombe, R., 2000. The 3D arrangement of the 23 S and 5 S rRNA in the *Escherichia coli* 50 S ribosomal subunit based on a cryo-electron microscopic reconstruction at 7.5 Å resolution. J. Mol. Biol. 298, 35–39.

Natterer, F., 1986. The Mathematics of Computerized Tomography. Wiley, New York.

Natterer, F., Wübbeling, F., 2001. Mathematical Methods in Image Reconstruction. Society for Industrial and Applied Mathematics, Philadelphia.

Neutze, R., Wouts, R., van der Spoel, D., Weckert, E., Hajdu, J., 2000. Potential for femtosecond imaging of biomolecules with X-rays. Nature 406, 752–757.

Papoulis, A., 1991. Probability, Random Variables and Stochastic Processes, third ed. Electrical & Electronic Engineering Series. McGraw-Hill.

Saxton, W.O., Frank, J., 1977. Motif detection in quantum noise-limited electron micrographs by cross-correlation. Ultramicroscopy 2, 219–227.

Sayre, D., 1980. In: Schlenker, M. et al. (Eds.), Imaging Processes and Coherence in Physics. Springer Lecture Notes in Physics, vol. 112. Springer-Verlag, Berlin, pp. 229–235.

Shmueli, U. (Ed.), 1996. International Tables for Crystallography. Reciprocal Space, vol. B. Kluwer Academic Publishers, Dordrecht.

Stowell, M.H.B., Miyazawa, A., Unwin, N., 1998. Macromolecular structure determination by electron microscopy: new advances and recent results. Curr. Opin. Struct. Biol. 8, 595–600.

Szőke, A., 1997. Holography with a complicated reference. J. Imag. Sci. Technol. 41, 332–341.

Szőke, A., 1999. Time-resolved holographic diffraction at atomic resolution. Chem. Phys. Lett. 313, 777–788.

van Heel, M., 1987. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy 21, 111–124.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., Schatz, M., 1996. A new generation of the IMAGIC image processing system. J. Struct. Biol. 116, 17–24.

van Heel, M., Orlova, E.V., Harauz, G., Stark, H., Dube, P., Zemlin, F., Schatz, M., 1997. Angular reconstitution in three-dimensional electron microscopy. Historical and theoretical aspects. Scanning Microsc. 11, 195–210.

van Heel, M., Gowen, B., Matadeen, R., Orlova, E.V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M., Patwardhan, A., 2000. Single-particle electron cryo-microscopy: towards atomic resolution. Quart. Rev. Biophys. 33, 269–307.

Wilson, A.J.C., 1949. The probability distribution of X-ray intensities. Acta Cryst. 2, 318–321.

Wilson, A.J.C. (Ed.), 1995. International Tables for Crystallography. Mathematical, Physical and Chemical Tables, vol. C. Kluwer Academic Publishers, Dordrecht.