IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Orientation determination in single-particle x-ray coherent diffraction imaging experiments

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2013 J. Phys. B: At. Mol. Opt. Phys. 46 164013

(http://iopscience.iop.org/0953-4075/46/16/164013)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 131.169.232.48
The article was downloaded on 13/08/2013 at 12:46

Please note that terms and conditions apply.

OPEN ACCESS
IOP Publishing

J. Phys. B: At. Mol. Opt. Phys. **46** (2013) 164013 (11pp) doi:10.1088/0953-4075/46/16/164013

Orientation determination in single-particle x-ray coherent diffraction imaging experiments

O M Yefanov[1,2,4] **and I A Vartanyants**[1,3]

[1] Deutsches Elektronen-Synchrotron (DESY), Notkestraße 85, D-22607 Hamburg, Germany
[2] Institute of Semiconductor Physics NASU, 03028 Kiev, Ukraine
[3] National Research Nuclear University 'MEPhI', 115409 Moscow, Russia

E-mail: oleksandr.yefanov@desy.de

Received 15 February 2013, in final form 9 May 2013
Published 13 August 2013
Online at stacks.iop.org/JPhysB/46/164013

Abstract
Single-particle diffraction imaging experiments at free-electron lasers (FELs) have a great potential for the structure determination of reproducible biological specimens that cannot be crystallized. One of the challenges in processing the data from such an experiment is to determine the correct orientation of each diffraction pattern from samples randomly injected in the FEL beam. We propose an algorithm (Yefanov *et al* 2010 *Photon Science—HASYLAB Annual Report*) that can solve this problem and can be applied to samples from tens of nanometres to microns in size, measured with sub-nanometre resolution in the presence of noise. This is achieved by the simultaneous analysis of a large number of diffraction patterns corresponding to different orientations of the particles. The algorithm's efficiency is demonstrated for two biological samples, an artificial protein structure without any symmetry and a virus with icosahedral symmetry. Both structures are a few tens of nanometres in size and consist of more than 100 000 non-hydrogen atoms. More than 10 000 diffraction patterns with Poisson noise were simulated and analysed for each structure. Our simulations indicate the possibility of achieving resolution of about 3.3 Å at 3 Å wavelength and incoming flux of 10^{12} photons per pulse focused to 100×100 nm^2.

(Some figures may appear in colour only in the online journal)

1. Introduction

The problem of solving the structure of individual biological specimens to high resolution is critical for many branches of modern life- and bio-science. Two widely used techniques for high-resolution structure determination are x-ray crystallography and electron microscopy. X-ray crystallography can only be used for molecules that form crystals [2], whereas transmission electron microscopy is limited to structures with a thickness well below 1 μm [3]. Therefore, the majority of samples must be sliced [4] and the minimum thickness of the slices limits the resolution of this method.

Single-particle coherent diffraction imaging [5–7] is one of the promising new techniques for the investigation of biological samples to sub-nanometre resolution. It has become possible only recently due to the development of x-ray free-electron lasers (FELs) [8–11], which produce ultra-short (10–100 fs), coherent x-ray pulses with high intensity (more than 10^{12} photons in a single pulse). Short and intense pulses are required to overcome the radiation damage of biological particles during the pulse propagation [12–15] and to produce a high number of elastically scattered photons [5, 16]. The coherence of the incident beam is important for a successful reconstruction of the electron density of the sample [17–19]. However, after the pulse propagation the particles explode, and only one projection of the sample can be measured.

[4] Present address: Center for Free-Electron Laser Science, Notkestraße 85, D-22607 Hamburg, Germany.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

© 2013 IOP Publishing Ltd Printed in the UK & the USA

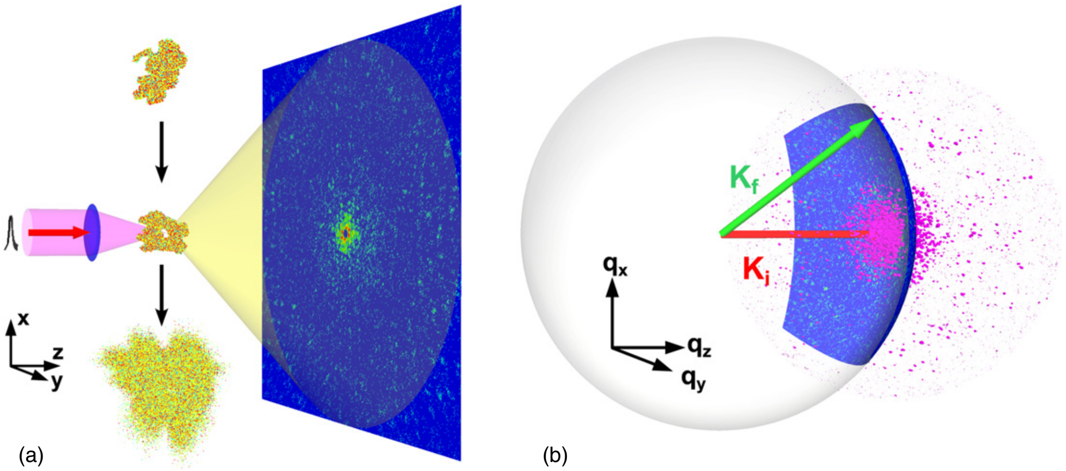

Figure 1. Schematic view of the experimental geometry. (a) In real space, the diffraction pattern from a sample in random orientation is measured by a single FEL pulse. (b) In reciprocal space, the measured diffraction pattern corresponds to a cut of the 3D intensity distribution by an Ewald sphere sector. The vectors $\mathbf{K}_i$ and $\mathbf{K}_f$ denote the incident and diffracted wave vectors, respectively.

This problem can be overcome by injecting reproducible particles one after another with random orientations and collecting a set of diffraction patterns [6]. Each measured diffraction pattern corresponds then to an unknown particle orientation. A method to determine the orientation of the particle, corresponding to each diffraction pattern, is the main subject of this paper. When the relative angular orientation of all diffraction patterns is determined the full three-dimensional (3D) intensity distribution in reciprocal space can be obtained. The structural information, or electron density of the sample, is determined then by the phase retrieval [20, 21].

During the last few years there was a significant progress in the practical implementation of these ideas at hard x-ray FELs (see, for example, [22–24]). There were few attempts to determine the 3D structure in single-particle imaging experiments [25]; however, the methods are still under development. Several approaches have been proposed so far to find an unknown particle orientation in these experiments. One is based on the common arc algorithm [26] originally developed for electron microscopy [27–29]. This algorithm exploits the fact that all two-dimensional (2D) diffraction patterns of reproducible particles in random orientations represent sections by the Ewald sphere of the 3D intensity distribution in reciprocal space. As such, all diffraction patterns have one common point, the origin of reciprocal space, and intersect along common arcs. The intensities along these arcs must be equal, and using this information the relative orientation of all diffraction patterns can be determined. The main problem of this method is its demand for a high signal-to-noise ratio, which is difficult to satisfy even with the present high-power FEL sources. It was suggested to overcome this limitation by an additional classification step [26, 30], in which diffraction patterns with similar particle orientations are averaged prior to orientation determination. This step improves the statistics of each averaged diffraction pattern, but at the same time reduces its contrast. As a result, the classification step decreases the achievable resolution and can produce artefacts in the final stage of electron density reconstruction. Another method is based on generative topographic mapping and neural networks [31, 32]. This approach works well for a low signal-to-noise ratio but scales poorly with the number of resolution elements in terms of computational time and memory. The same is valid for a method based on an expectation maximization technique [33].

Here, we propose an orientation determination method based on an improved common arc algorithm [1]. Instead of a classification step we perform a simultaneous analysis of common arcs between many diffraction patterns. To improve the quality of the orientation determination, a 3D angular refinement procedure is applied at the final step. This algorithm works well even with a low photon signal down to 0.5 photons per Shannon angle. It scales linearly with the number of resolution elements and number of measured diffraction patterns. Memory requirements are relaxed because most of the data can be processed in parts. Finally, the algorithm is highly parallelizable since most of the analysis is done between pairs of diffraction patterns.

The paper is organized in the following way. In section 2, we describe our implementation of the common arc algorithm. Section 3 describes our approach to treat poor signal-to-noise ratio data as well as the orientation refinement procedure. Tests of the proposed algorithm on simulated data from two different biological structures are presented in section 4. The paper is completed by the conclusion section. The details of the algorithm implementation are presented in appendices A–C.

2. Common arc algorithm

In a typical single-particle diffraction imaging experiment, a sample with unknown orientation is injected into the focused coherent x-ray beam of an FEL (figure 1(a)). The scattered radiation is measured in the far field by a 2D detector. This diffraction pattern can be mapped on an Ewald sphere [34] and represents a 2D cut of the 3D intensity distribution in reciprocal space (figure 1(b)). Alternatively, the diffraction pattern can be considered as a perspective projection of an Ewald sphere sector onto the 2D detector plane as viewed from the sample position (figure 2).

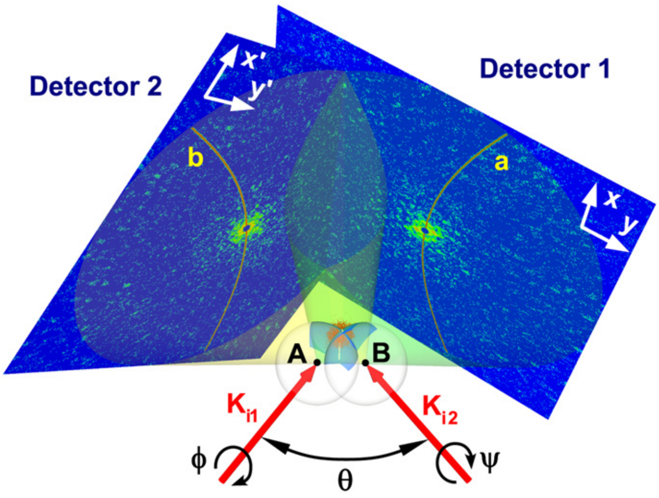

Figure 2. Measurements of two reproducible samples at random orientation can be considered as two measurements of the same sample with two different incident beam directions indicated by the vectors $\mathbf{K}_{i1}$ and $\mathbf{K}_{i2}$. The angles ϕ, θ and ψ are Euler's rotation angles. Points A and B are the centres of the corresponding Ewald's spheres. Coordinates on the first and second detector are indicated as x, y and x', y', respectively.

Our previous studies suggest [7] that, in order to increase the scattered signal, it is favourable to use longer x-ray wavelengths, since the x-ray scattering cross-sections are larger at these wavelengths. At the same time the energy of the incident x-rays should be sufficient to penetrate the sample. To achieve high resolution the detector should also cover high scattering angles. Under these conditions a large sector of an Ewald sphere is covered, which is beneficial for orientation determination.

When two independent measurements of identical particles with different orientations are considered, the orientation of the first particle can be fixed as known. The orientation of the second particle can be uniquely described relative to the first one. Alternatively, two measured diffraction patterns could be considered to originate from the same particle in different experimental geometries. In this case the particle orientation is fixed, but the direction of the incident beam and the detector orientation are different for each measurement as shown in figure 2. For the first measurement, the incident beam direction, given by its wave vector $\mathbf{K}_{i1}$, can be taken along the $\mathbf{q}_z$ axis in the reciprocal space coordinate system shown in figure 1(b). The direction of the incident beam for the second measurement is given by its wave vector $\mathbf{K}_{i2}$ (figure 2). The relative orientation of the second geometry with respect to the first one can be described by three Euler angles ϕ, θ, ψ [35]. The choice of Euler angles is convenient, since rotations around the angles ϕ and ψ in reciprocal space are equivalent to rotations by the same angles of detectors one and two in real space, respectively.

For monochromatic x-rays, the Ewald sphere has the radius $K = 2\pi/\lambda$, where λ is the wavelength of the incident radiation. The Ewald spheres corresponding to the two measurements pass through the origin of the reciprocal space coordinate system (figure 2). The origin of the first Ewald sphere (see point A in figure 2), for the incident vector $\mathbf{K}_{i1}$, is at $(0, 0, -K)$ and the origin of the second sphere (point B in figure 2), for the incident vector $\mathbf{K}_{i2}$, is at (q_{x0}, q_{y0}, q_{z0}). The coordinates q_{x0}, q_{y0} and q_{z0} are determined by a rotation of the point $(0, 0, -K)$ around the reciprocal space origin $(0, 0, 0)$ by the Euler angles ϕ and θ. The intersection of the two spheres is a common arc that also passes through the origin of reciprocal space (see figure 2). This common arc is projected on the two detectors (curves a and b in figure 2). It is clear from this construction that the intensity along these arcs must be the same at both detectors. By analysing the intensity correlations along all possible common arcs, the unique relative orientation of the two measurements can be determined.

It should be noted that a common arc can fix the relative orientation of two patterns only for experimental geometries with large scattering angles. Otherwise, the measured sector of the Ewald sphere can be considered as flat, and the common arc reduces to a straight line. This common line fixes only the angles ϕ and ψ, but not the angle θ; therefore, a simultaneous analysis of at least three diffraction patterns is needed in this case [29].

The projection of the common arc on the first detector (curve a in figure 2) can be expressed in the detector 2D coordinate system (x, y) by the following equation (see appendix A for details):

$$\begin{aligned}&\left(q_{x0}^2 - (q_{z0}+K)^2\right)x^2 + \left(q_{y0}^2 - (q_{z0}+K)^2\right)y^2 + 2q_{y0}q_{x0}xy\\&\quad + 2dq_{x0}(q_{z0}+K)x + 2dq_{y0}(q_{z0}+K)y = 0,\end{aligned} \tag{1}$$

where d is the sample–detector distance and x, y are the coordinates of the common arc projection on the first detector. Similar projection of the common arc on the second detector (curve b in figure 2) is also described by equation (1) by substituting x, y coordinates to $x', y' = -y$.

As follows from equation (1), the curvature of the common arcs a and b at the detectors one and two is determined only by the angle θ and sample–detector distance d. Practically, the coordinates of the projections of common arcs at the detector planes are obtained by solving equation (1) for each value of θ and d with the fixed angles $\phi = \psi = 0$. A set of curves corresponding to the fixed value of the angle θ and all other values of angles ϕ and ψ is determined by the rotation of the curve obtained in the previous step. This is implemented by the rotation of the coordinate system (x, y) corresponding to the first detector by angle ϕ and the coordinate system (x', y') of the second detector by an angle ψ (see figure 2 and appendix B for details).

For each set of Euler angles and fixed sample–detector distance d, the coordinates of both arcs (a and b in figure 2) are determined, and the intensities along these lines are compared by calculating the cross-correlation coefficient (CCC) $c^{ab}(\phi, \theta, \psi)$:

$$c^{ab}(\phi,\theta,\psi) = \frac{\sum_i J_i^a(\theta,\phi) J_i^b(\theta,\psi)}{\sqrt{\sum_i [J_i^a(\theta,\phi)]^2}\sqrt{\sum_i [J_i^b(\theta,\psi)]^2}}, \tag{2}$$

where $J_i^{a,b}(\theta,\phi) = \ln[I_i^{a,b}(\theta,\phi) + 1]$ are logarithms of the intensities $I_i^a(\theta,\phi)$ and $I_i^b(\theta,\phi)$ along the first (a) and second (b) arc. The correct orientation of the second measurement with respect to the first one is given by the set of angles

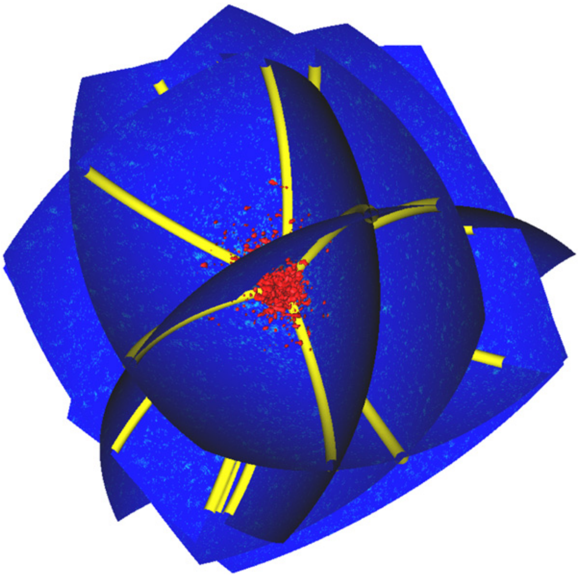

Figure 3. Few Ewald sphere sectors intersecting the 3D intensity distribution of the sample in reciprocal space. The yellow lines indicate common arcs between different patterns.

ϕ_B, θ_B and ψ_B that maximize the CCC in equation (2). To determine this orientation, all three Euler angles (ϕ, θ, ψ) are varied sequentially with some angular step and CCCs for every orientation are calculated and compared. This procedure is applied to all diffraction patterns until their orientation relative to the first one is determined (figure 3).

It should be noted here that not only orientation but also the position of a particle in space should be taken into account explicitly in the analysis. The transverse position of the sample relative to the optical axes of the incoming focused x-ray beam adds a constant phase slope to the scattered amplitude and scales its intensity. If each diffraction pattern is properly centred, then this does not cause any problems in the analysis since the phase is not recorded by the detector. The intensity of each diffraction pattern can be rescaled at the stage of composing the 3D intensity distribution in reciprocal space, as described later. At the same time, the particle–detector distance d must be taken into account explicitly, due to its strong influence on the diffraction pattern. If two measurements are performed at different sample–detector distances d_1 and d_2, equation (1) must be solved separately for both detectors taking into account the corresponding distances. This is especially important in real experimental conditions, when particles are injected in the beam, due to variations of the distance d from shot to shot. We also assume in our analysis that the particle size is much smaller than any variations of beam intensity. More details on our practical implementation of the common arc algorithm are presented in appendix B.

The common arc algorithm described in this section performs well for data sets with a high signal-to-noise ratio [26]. However, it often fails in practical applications for a low number of scattered photons. One way to overcome this problem is presented in the following section.

3. Advanced algorithm for orientation determination in the presence of noise

In the previous section, the common arcs between one diffraction pattern (that we define as a base pattern) and all other diffraction patterns were analysed. At the same time, we should note that each diffraction pattern has a common arc with other diffraction patterns (see figure 3). Therefore, common arcs between all patterns could, in principle, be analysed simultaneously. Such analysis can significantly improve the fidelity of the orientation; however, in practice it requires an increase in computational resources. A compromise can be found by implementing the following strategy. As a first step a set of base diffraction patterns N_{base} is analysed with respect to each other to determine the correct orientations of these chosen patterns. In the next step all other patterns are oriented with respect to each of these base patterns. This implementation requires N_{base} times more calculations compared to a single base pattern. In the final step all intensities are mapped to a 3D array of voxels in reciprocal space by 3D gridding and averaging procedure. The benefit of this approach is the possibility of solving the orientation problem for noisy data, as will be demonstrated in the following section (see also appendix C for a detailed discussion).

In a real experimental situation, all diffraction patterns have different intensities due to shot to shot intensity jitter of the FEL and the fact that each injected particle is hit by a different part of a focused beam. As a consequence all measured diffraction patterns have to be rescaled. This is implemented in the algorithm by utilizing the fact that each of the two patterns has a common arc and that the intensities along this arc must be equal. The scaling factor for the intensities can be determined by taking the ratio of intensities of two diffraction patterns along the common arc. Having all information about the experimental geometry, orientation and scaling factor for each pattern, the 3D intensity distribution in reciprocal space can be constructed.

The orientation determination can be significantly improved by an additional refinement that is based on the correlations between an individual pattern and the whole 3D intensity distribution. It can be implemented in the following way. First, the 3D intensity distribution is obtained from all but one selected diffraction pattern. Then the orientation of the selected pattern is varied in a small angular range and the correlation between this 2D pattern and the whole 3D intensity distribution is analysed. The orientation corresponding to the highest correlation value is considered to be the correct one. Then, the rescaled intensity of the selected pattern with the refined orientation is included in the new 3D intensity distribution in which the next diffraction pattern is excluded and the refining procedure is repeated. By applying this approach to all diffraction patterns the final 3D intensity distribution is obtained. This procedure can also be applied to identify diffraction patterns from 'wrong' particles (those that do not belong to a set of samples under investigation). Correlation coefficients of diffraction patterns originating from these particles and the whole 3D intensity distribution will be quite low, which can be used as a criterion for the rejection of these diffraction patterns from the future analysis.

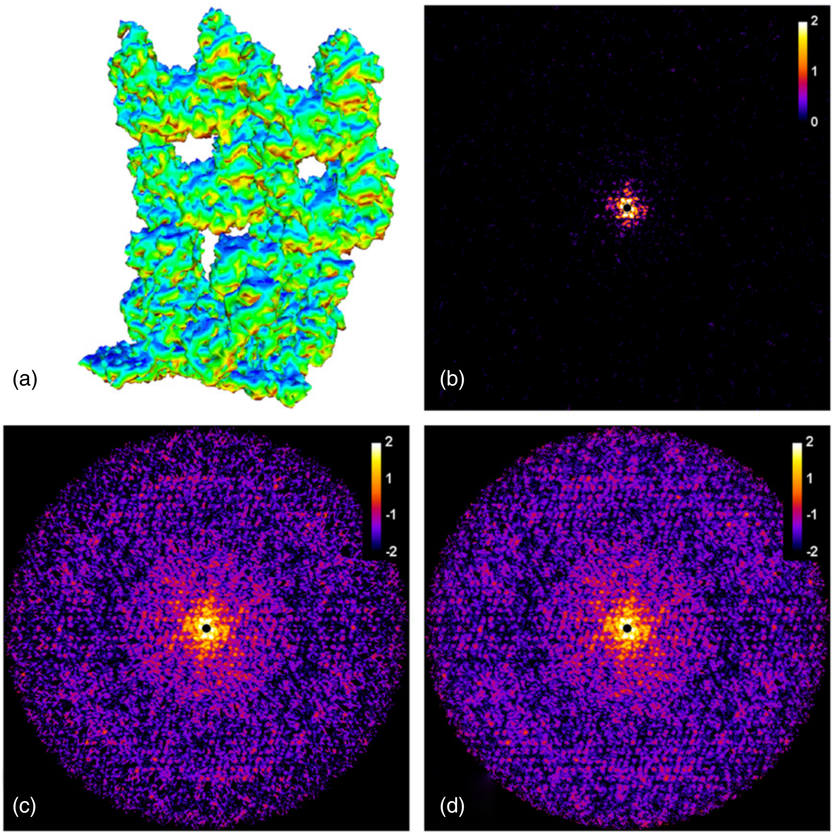

Figure 4. Artificial protein structure without any symmetry combined from the 2BTV and 8RUC macromolecular structures. (a) Iso-surface of the electron density, (b) a typical diffraction pattern (edge resolution 3.92 Å), (c), (d) 2D central cuts (edge resolution 3.3 Å) through the constructed 3D intensity distribution in reciprocal space for the patterns with a known orientation (c), and the patterns with the orientations determined using the proposed algorithm (d). All diffraction patterns are presented on a logarithmic scale.

If the structure has a known symmetry, then this can be used as an additional constraint for orientation determination [29]. Using symmetry conditions, each diffraction pattern can be oriented individually with respect to the selected symmetry axis. This is contrary to the structures without symmetry when at least two patterns are required for orientation determination. Applying symmetry conditions, it is possible to get a sufficient number of diffraction patterns for the 3D representation of the scattered intensity in reciprocal space even with a limited data set or a large area of missing data due to a big beamstop. This approach was successfully used for simulated data for a sample with icosahedral symmetry discussed in the next section as well as for experimentally measured diffraction patterns of a Mimi virus obtained in a coherent diffraction imaging experiment at FLASH [36].

It is interesting to note that the presented implementation of the common arc algorithm also allows us to determine the unknown symmetry of the object. This can be obtained by the analysis of angular orientations appearing with the highest probability. Such orientations can be found in a 3D angular map (ϕ, θ, ψ) of all possible orientations and reveal themselves as regions with high density (see appendix C for details). For example, for structures with icosahedral symmetry it will correspond to the 120 most likely orientations in reciprocal space that are related to the icosahedral symmetry transformation matrix.

A sampling rate of at least 2 in each direction in the diffraction pattern is required for a successful implementation of the algorithm described here. The same requirement is valid for the phase retrieval algorithms applied for the reconstruction of electron density of the samples. A higher sampling rate is beneficial for orientation determination because each speckle consists of more pixels. At the same time, binned experimental data with a lower sampling rate have a higher signal in each pixel, which could become important for the orientation determination of data with a low signal-to-noise ratio [37]. By testing different sampling conditions we found that an optimal sampling rate is in the region from 2 to 3. In practice, to increase the signal the experimental data could be binned for orientation determination, while the reconstruction is performed on the original unbinned data set. Applying on a final step phase retrieval algorithms [20, 21] to the 3D data set of the intensity distribution, the electron density of the sample can be obtained.

4. Numerical test of the algorithm

The algorithm was tested with two different biological structures. The first one was an artificial protein structure without any symmetry combined from the 2BTV and 8RUC macromolecular structures [38] (see figure 4(a)). It has a size of 13 × 19 × 28 nm^3 and consists of about 124 000 non-hydrogen atoms. The second one was a human adenovirus

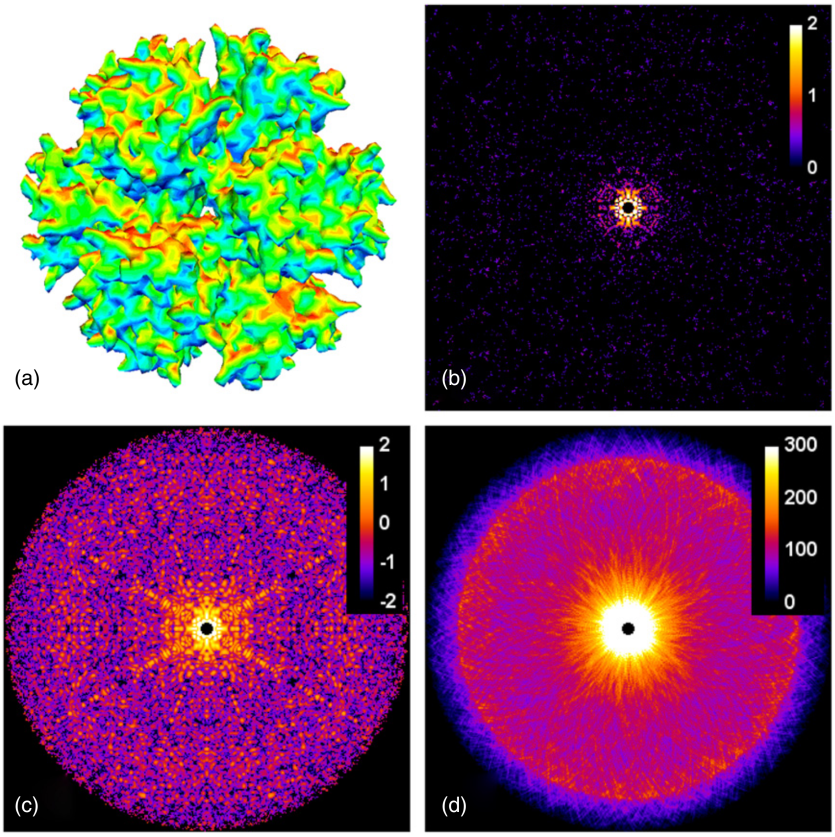

Figure 5. Human adenovirus penton base 2 12 chimera 2c6s structure with the icosahedral symmetry. (a)–(c) The same as in figure 4, (d) number of diffraction patterns contributing to each voxel of (c) (see appendix C for details). All diffraction patterns are presented on a logarithmic scale.

penton base 2 12 chimera 2c6s [38]. It has icosahedral symmetry with the diameter of 27 nm and consists of about 200 000 non-hydrogen atoms (figure 5(a)). In our simulations, we assumed completely reproducible particles; correlation analysis of particles contaminated with water molecules was discussed in [39].

Diffraction patterns for both structures where simulated at 3 Å wavelength in kinematic approximation[5]. The signal at each detector pixel was calculated as a coherent sum of the atomic form-factors from all atoms consisting the molecule. Due to a small size of the molecules considered in the simulations absorption effects were neglected. A detector size of 100 mm and a sample–detector distance of 50 mm, providing the maximum scattering angle of 45°, were assumed in our simulations. The achievable resolution in this geometry was 3.92 Å at the detector edge and 3.3 Å at its corner. The number of detector pixels was 512 × 512 for the first sample (providing a minimum sampling rate of 2.5) and 360 × 360 for the second (with a sampling rate of 2). The incoming flux was 10^{12} photons focused uniformly on 100 × 100 nm^2. At each detector pixel, noise was added according to Poisson statistics. The average flux at the edge of the detector was 0.05 and 0.15 photons per pixel corresponding to 0.45 and 0.6 photons per Shannon angle for the first and second structure, respectively. For the first structure, 36 × 36 × 18 = 23 328 patterns were simulated with a 10° increment for each Euler angle. For the second structure, 12 000 randomly oriented patterns were simulated. A beamstop with a diameter of about 2 mm covering 1.5 speckles was introduced in all simulated diffraction patterns, and this region was excluded from the calculation of correlation coefficients. Typical diffraction patterns for a single FEL pulse simulated in the experimental conditions described above are shown in figures 4(b) and 5(b) for the first and second structure, respectively.

The correct orientation of each diffraction pattern was determined using the algorithm described in the previous sections. The parameters used for the orientation determination were the following: the number of base patterns was $N_{\mathrm{base}} = 64$, the angular increment for each Euler angle was 3° in the range of angles $0 < \phi < 360°, 0 < \theta < 180°, 0 < \psi < 360°$. This allowed us to obtain the full 3D intensity distribution in reciprocal space for each sample. A central slice through this distribution constructed from the oriented diffraction patterns corresponding to the first structure is presented in figure 4(d). For comparison, the same slice through the 3D intensity distribution obtained from the known orientation of each diffraction pattern is shown in figure 4(c). It is well seen that the slice obtained as a result of orientation determination well reproduces all features of an 'ideal' intensity distribution; small deviations can be attributed to angular misalignment. This misalignment between the angles obtained from the

[5] All diffraction patterns were simulated using the computer code `moltrans`.

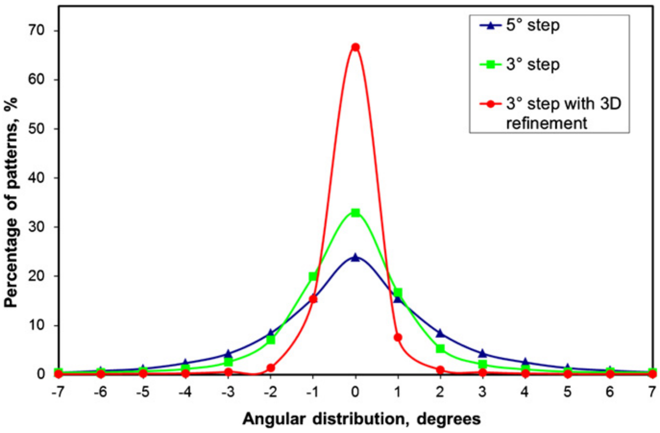

Figure 6. Distribution of the angular error of the determined orientations for the structure without symmetry. The blue line corresponds to 5° angular step, green line 3° and red line 3° after the 3D refinement (see the text for details).

common arc algorithm and the correct angles for the first structure is presented as a plot in figure 6. The accuracy of the orientation determination correlates strongly with the angular step size for the Euler angles (ϕ, θ, ψ). Clearly, a finer angular step size requires more computational time that scales as a third power of the step size. It is clearly seen in figure 6 that a 3° angular step being five times slower still gives higher accuracy in orientation determination comparing to a 5° step. An additional improvement in the angular determination is obtained by the final orientational refinement of each diffraction pattern with respect to 3D intensity distribution, as described in the previous section (see figure 6).

It is interesting to observe how the signal is increased by the number of diffraction patterns used in the analysis. In figure 5(d), a central slice through the 3D array representing the number of patterns contributing to each voxel of the constructed 3D intensity distribution is presented. It can be seen from this figure that at least 100 patterns from the analysed 12 000 contribute to each voxel inside a resolution ring of 4 Å. One more intriguing feature can be observed in this figure. Although the initial diffraction patterns were simulated up to 3.92 Å resolution at the edge of the detector, the 3D intensity distribution obtained from the algorithm has distinguishable features up to 3.3 Å resolution (see the dark outer ring in figure 5(d)). This additional signal comes from the corners of the diffraction patterns.

As was pointed above in our algorithm test, we added a round beamstop in the centre of the diffraction pattern. To make our simulations close to experiments performed at XFEL sources with the present detectors [40, 41] composed of tiles, we performed an analysis for a detector composed of two separated parts with a gap in between [36]. Our analysis has shown that due to the simultaneous analysis of a big number of diffraction patterns the algorithm works very well even for such an incomplete data set.

5. Summary

In summary, we proposed a method for the angular orientation determination in single-particle coherent imaging experiments based on the common arc algorithm. We obtained a significant improvement of this approach by introducing a simultaneous analysis of the common arcs for several diffraction patterns. This gives the possibility of applying the method to data with a low level of signal-to-noise ratio as well as to skip the classification step which can reduce achievable resolution. Additionally, we proposed an orientational refinement of diffraction patterns that can improve the quality of the final 3D intensity distribution in reciprocal space.

The algorithm proposed here has several advantages compared to other approaches [31, 33]. It scales linearly with the number of measured patterns and total number of pixels in the diffraction patterns. The algorithm is easy to parallelize, because most of the cross-correlation analysis is performed between pairs of independent diffraction patterns. It has minimum memory requirements, because the data can be processed in parts.

We foresee that this approach has the potential to be the key for the success in the analysis of single-particle diffraction imaging experiments and will allow us to reach sub-nanometre resolution in 3D imaging of biological specimens.

Acknowledgments

We are grateful to E Weckert for permanent interest and support of this project, as well as for the use of the program `moltrans` for simulation of diffraction patterns from biological structures, and to A Singer and U Lorenz for a careful reading of the manuscript. Part of this work was supported by BMBF Proposal 05K10CHG 'Coherent Diffraction Imaging and Scattering of Ultrashort Coherent Pulses with Matter' in the framework of the German–Russian collaboration 'Development and Use of Accelerator-Based Photon Sources' and the Virtual Institute VH-VI-403 of the Helmholtz association.

Appendix A. Derivation of the main equations

Equation (1) was derived using the following considerations. Both intersecting Ewald spheres (figure A1) have radii equal

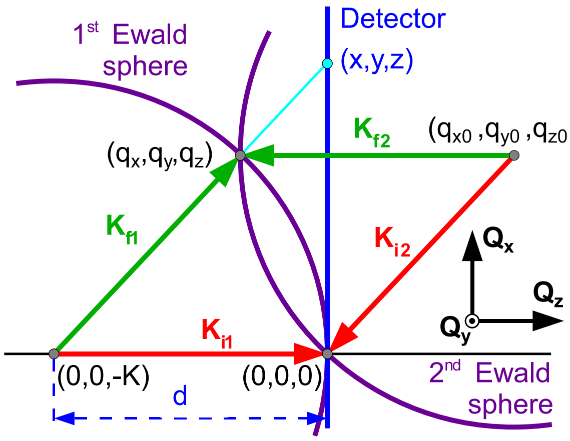

Figure A1. Schematic view of the intersection of two Ewald spheres.

to the wave vector $K = 2\pi/\lambda$ ($|\mathbf{K_i}|^2 = |\mathbf{K_f}|^2 = K^2$), where λ is the wavelength of the incident radiation. Therefore, the coordinates (q_x, q_y, q_z) of the intersection curve must satisfy the equations

$$q_x^2 + q_y^2 + (q_z + K)^2 = K^2, \tag{A.1}$$

$$(q_x - q_{x0})^2 + (q_y - q_{y0})^2 + (q_z - q_{z0})^2 = K^2. \tag{A.2}$$

The centre of the second Ewald sphere (q_{x0}, q_{y0}, q_{z0}) lies at the distance K from the centre of reciprocal space ($|\mathbf{K}_{i2}|^2 = K^2$); therefore,

$$q_{x0}^2 + q_{y0}^2 + q_{z0}^2 = K^2. \tag{A.3}$$

From equations (A.1)–(A.3), the formula describing the intersection of two Ewald spheres can be derived:

$$\left(q_{y0}^2 + q_{x0}^2\right)q_y^2 + 2q_{y0}q_{z0}(q_z - K)q_y + \left(q_{z0}^2 + q_{x0}^2\right)q_z^2 + K^2\left(q_{z0}^2 - q_{x0}^2\right) - 2Kq_{z0}^2 q_z = 0. \tag{A.4}$$

As soon as the diffracted vector $\mathbf{K}_{f1}$ (figure A1) has the same direction in both real and reciprocal spaces (angles coincide), the relation between coordinates of a pixel on the detector (x, y, z) in real space and corresponding coordinates of the end of $\mathbf{K}_f$ (q_x, q_y, q_z) in reciprocal space can be written as

$$\frac{x}{q_x} = \frac{y}{q_y} = \frac{z}{q_z}. \tag{A.5}$$

As soon as the distance from the sample to the detector (d) is fixed, $z \equiv d$. Equation (1) can be easily derived from equations (A.4) and (A.5).

Appendix B. Common arc algorithm

Due to the properties of Euler angles, the angle ϕ ($0 \leqslant \phi < 2\pi$) can be attributed to the rotation of the reciprocal space coordinate system around the incident beam ($\mathbf{K}_{i1}$), the angle θ ($0 \leqslant \theta < \pi$) corresponds to the rotation around the new position of vector $\mathbf{q}_y$ and the angle ψ ($0 \leqslant \psi < 2\pi$) is the final rotation of the coordinate system around the new position of the vector $\mathbf{q}_z$–vector $\mathbf{K}_{i2}$ in figure 2.

In practice, the curvature of the common arcs a and b in figure 2 is determined only by the angle θ and distance d. The coordinates of the projections of common arcs on detector planes can be obtained for each value of θ and d, with the fixed angles $\phi = \psi = 0$, by solving equation (1). Other curves (for all values of angles ϕ and ψ) at the fixed value of angle θ are determined by the rotation of the curve obtained in the previous step. The coordinate system (x, y), corresponding to the first detector, is rotated by an angle ϕ and the coordinate system (x', y'), corresponding to the second detector, by an angle ψ (figure 2).

The common arc algorithm described in this paper was implemented using the following scheme (figure B1). Calculation starts with fixed angles $\phi = 0$ and $\psi = 0$. Then coordinates q_{x0}, q_{y0} and q_{z0} are calculated by Euler rotation on the θ angle of the vector $\mathbf{K}_{i1}$ (point $(0, 0, -K)$). After this, equation (1) is solved and coordinates (x, y) for the common arc are found. This curve is rotated on the angle ϕ for the first pattern (each pair of (x, y) is multiplied by the corresponding rotation matrix) and on the angle ψ for the second pattern (with exchange $y \to -y$). After the full determination of the curves for both patterns, the corresponding values of intensities (along the curve) are extracted using the interpolation described below. The intensities along the curves are then correlated. This process continues for all angles ψ in the region $0 \leqslant \psi < 2\pi$ and for all angles ϕ in the region $0 \leqslant \phi < 2\pi$. Then the whole process is repeated for different θ values.

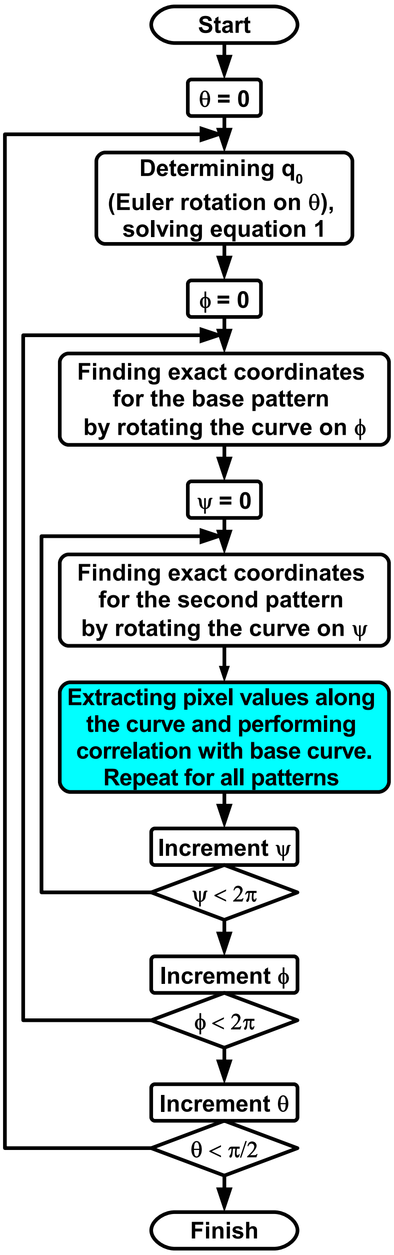
Start
θ = 0
Determining q0 (Euler rotation on θ), solving equation 1
ϕ = 0
Finding exact coordinates for the base pattern by rotating the curve on ϕ
ψ = 0
Finding exact coordinates for the second pattern by rotating the curve on ψ
Extracting pixel values along the curve and performing correlation with base curve. Repeat for all patterns
Increment ψ
ψ < 2π
Increment ϕ
ϕ < 2π
Increment θ
θ < π/2
Finish

Figure B1. Flowchart for efficient data analysis with the common arc algorithm.

The common arcs approach has difficulties when angle θ is large. In this case the intersection between two spheres reduces to a closed circle. When angle θ approaches π, this circle shrinks to a point at the origin of reciprocal space coordinates $(0, 0, 0)$. Therefore, at large θ (close to π) the projection of the common arc to the detector plane, described by equation (1), degenerates to an ellipse. Therefore, the length along an arc and a circle can be different. Moreover, the correlation coefficients found for the arcs with different curvature can hardly be compared, because the ends of such curves correspond to different q-ranges and so some curves will have a good signal at the ends and some—mostly noise. From these considerations, it is clear that it is difficult to compare curves obtained for small and big θ angles. To solve this problem we limited the range of acceptable θ angles to the range $0 \leqslant \theta \leqslant \pi/2$. To cover the range of angles $\pi/2 < \theta < \pi$ we used the fact that reciprocal space is centrosymmetric for scattering on non-absorbing objects (Friedel's law). To take this into account for angles $\pi/2 < \theta < \pi$ we invert the direction of the vector $\mathbf{K}_{i2}$ (figure 2) to its opposite $-\mathbf{K}_{i2}$, which corresponds to the following transformation: rotation of ϕ by π ($\phi \to \phi + \pi$), rotation of θ by $\theta \to \pi - \theta$ and final rotation of ψ by π ($\psi \to \psi + \pi$). To finish inversion transformation we change $y \to -y$ (equation (1)) in the detector plane. If for any reason data are not centrosymmetric (Friedel's law is violated), the range of angles θ can still be extended to $0 \leqslant \theta \leqslant 2\pi/3$. For this angular range, the approach developed for the noisy data analysis (see appendix C) may be used.

To find the intensities corresponding to each point of the curve described by equation (1), some sort of gridding must be performed. In our calculations, we used interpolation in the form of an average of four nearest neighbour pixels. We also checked other schemes of interpolation: nearest neighbour and bilinear interpolation [42]. The first one is faster but lacks accuracy, and the second is computationally much slower without noticeable increase in quality.

All common arcs for different θ values (different curvature) should cover the same q-distance in reciprocal space. Therefore the step between points on the curve should remain constant. For this reason, equation (1) was differentiated analytically and starting from the centre of detector ($x = y = 0$) each next point on the curve is calculated keeping the distance $(\mathrm{d}x^2 + \mathrm{d}y^2) = \text{const}$.

The accuracy of calculations of cross-correlation coefficient (CCC) (equation (2)) for all patterns can be increased by replacing intensities $I(q)$ with their logarithms, more precisely by $\ln(1 + I(\mathbf{q}))$. Also for noisy and high background data the following form of CCC is more beneficial:

$$c^{ab}(\phi, \theta, \psi) = \frac{\sum (J^a(\mathbf{q}_\mathrm{i}) - \langle J^a(|\mathbf{q}_\mathrm{i}|)\rangle)(J^b(\mathbf{q}_\mathrm{i}) - \langle J^b(|\mathbf{q}_\mathrm{i}|)\rangle)}{\sqrt{\sum (J^a(\mathbf{q}_\mathrm{i}) - \langle J^a(|\mathbf{q}_\mathrm{i}|)\rangle)^2}\sqrt{\sum (J^b(\mathbf{q}_\mathrm{i}) - \langle J^b(|\mathbf{q}_\mathrm{i}|)\rangle)^2}}, \qquad \text{(B.1)}$$

where $J(\mathbf{q}) = \ln(1 + I(\mathbf{q}))$ and $\langle J(|\mathbf{q}_\mathrm{i}|)\rangle$ is a radial averaged value of intensity corresponding to a ring with radius $|\mathbf{q}_\mathrm{i}|$ and a width of one pixel on a diffraction pattern.

Appendix C. Advanced algorithm for processing noisy data

The algorithm of orientation determination of noisy data is presented below. In the beginning a set of base patterns is selected. These patterns can be selected as patterns with high signal level; as an additional requirement the base patterns should have different orientation, specifically at angle θ. To select such a set, 2D cross-correlation analysis is performed between all patterns for all possible angles ϕ. The patterns with a low CCC correspond to a different angle θ. The number of the base diffraction patterns is selected according to the following considerations. If the measured signal is strong (several photons at the edge of the detector), few base patterns can be sufficient. For a weaker signal, more base diffraction patterns are required.

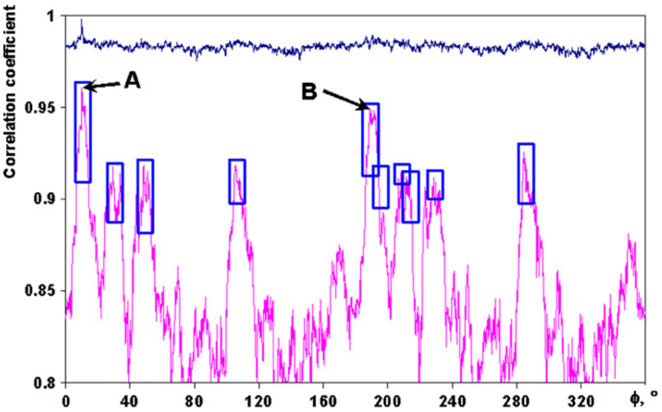

Figure C1. Correlation coefficient between two patterns for different ϕ angles. The upper curve for ideal data and the lower one for the noisy data. The blue boxes mark the ten best candidates for the correct orientation. The width of each box corresponds to $2 \times A_\mathrm{tol}$; the height is arbitrary.

Figure C1 shows typical correlations, calculated with equation (2), between two noisy diffraction patterns for different angles ϕ with fixed angles θ and ψ (bottom curve). For comparison, correlation coefficients corresponding to a perfect data set are plotted in the same figure C1 (top curve). The best correlation coefficient between two noisy patterns can correspond to completely wrong orientation (like point B in figure C1). Therefore several orientations (N_angl) corresponding to the best set of correlations must be stored. To avoid the storage of almost identical angles, some tolerance A_tol in the best orientation angle determination is necessary. Only one angle in the range $\pm A_\mathrm{tol}$ with the highest correlation coefficient is stored. This leads to the storage of only one angle per tolerance region (rectangles in figure C1). Therefore, only one point per marked rectangle in figure C1 is stored in the list of 'best' angles (for example in figure C1, $N_\mathrm{angl} = 10$ and $A_\mathrm{tol} = 5°$).

At the first step, all base patterns are correlated to each other using the algorithm described in section 2. As a result, N_angl 'best' angles between each pattern and all other $N_\mathrm{base} - 1$ base patterns are stored. Therefore, each pattern has $(N_\mathrm{base} - 1)N_\mathrm{angl}$ 'best' angles with respect to all other base patterns. Then all these angles are recalculated with respect to one selected pattern (which attributed all angles equal to 0)—we shall call it the zeroth pattern. This is done in the following way: each pattern's base angles with respect to other bases are recalculated to angles with respect to the zeroth pattern taking into account that each base has N_angl angles with respect to the zeroth. In this way $(N_\mathrm{base} - 2)N_\mathrm{angl}^2 + N_\mathrm{angl}$ angles for each pattern with respect to the zeroth are determined.

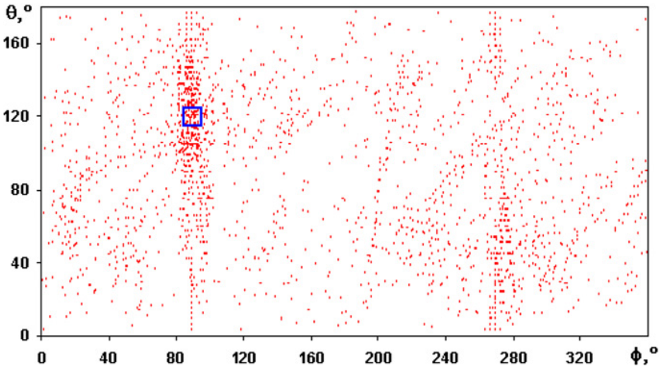

Figure C2. 2D (ϕ, θ) distribution of angles corresponding to good correlation between a pattern and all base patterns. A blue box corresponds to the region with highest density of good orientations.

Let us explain this step by an example. Consider one of the base patterns P_i. This pattern has N_{angl} angles with respect to the zeroth pattern P_0. The pattern P_i has also N_{angl} angles with respect to the first pattern P_1. But the first pattern P_1 itself has N_{angl} angles with respect to the P_0. So, pattern P_i has already $N_{\mathrm{angl}} + N^2_{\mathrm{angl}}$ angles to the P_0. Then it also has N^2_{angl} angles to the P_0 through the second pattern P_2. This process is continued for all base patterns. Finally all $(N_{\mathrm{base}} - 2)N^2_{\mathrm{angl}} + N_{\mathrm{angl}}$ angles determined for one base pattern (P_i) are plotted in 3D space of angles (ϕ, θ, ψ) and the angular region with the maximum density of points is selected as the best angle. In practice it is done in the following way: a number of points (in the tolerance region $\pm A_{\mathrm{tol}}$ for each Euler's angle) near each point is calculated. The point with the biggest number of neighbours is considered to be the best estimate. Then the correct angle is determined by averaging of all positions of the neighbours. In figure C2, an example of such operation in the 2D case (for angles θ and ϕ with fixed ψ) is presented. The best angle corresponds to the middle of the rectangle in figure C2.

As soon as all correct angles for the base patterns are determined, all other patterns can be oriented with respect to known bases. At this step only $N_{\mathrm{base}}N_{\mathrm{angl}}$ angles need to be considered for each pattern under analysis in the algorithm described above.

For better orientation determination the base patterns should be selected carefully among all those measured. All base patterns should have different orientations, more precisely different θ angles. Because initially all angles are unknown this requirement can be satisfied by the analysis of 2D correlations between different patterns. This is performed by calculating 2D correlation between pairs of patterns for different angles ϕ and the best correlation coefficient is stored. Then patterns with the worst 2D cross-correlations between each other are selected as bases. Also for experimental data analysis, base patterns should be selected according to the recorded signal quality.

The transformation to a Cartesian coordinate system in reciprocal space is performed in the following way. The whole reciprocal space is divided into an elementary set of 3D voxels with the size corresponding to the pixel size of the detector. Each voxel could contain a few values of the measured intensities that effectively increase the signal in the 3D intensity distribution (see figure 5(d)). All intensities in each voxels are averaged and the full 3D data set is obtained.

The orientation determination problem for the symmetrical structure (2c6s) without introducing the symmetry is more difficult. This is due to the fact that instead of one dense spot in figure C2 there will be 60 (for a structure with icosahedral symmetry) less dense spots in 3D angular space. So it is harder for the algorithm to select the right orientation. The accuracy (or time) of orientation determination can be greatly improved if the symmetry of the sample is known. But we want to underline the important feature of the algorithm, that it can be used even for symmetrical data with unknown symmetry and also for the structures with pseudo-symmetry.

There is one more important issue for speed and accuracy optimization while processing low flux data with noise. For the initial orientation determination there is no need to process the high-Q region of the diffraction pattern where the radially averaged photon count is less than approximately 1–2 photons per Shannon pixel (a pixel with the sampling rate equal to 1). The region with smaller photon counts just lowers the accuracy of orientation determination based on common arcs. Of course this argument cannot be applied to objects with highly anisotropic scattering in different directions, like, for example, pyramids [43]. At the same time the final 3D reciprocal space and 3D angular refinement can be performed for the full data sets; thus, the whole procedure does not lower the resolution.

Simulation parameters for the two structures analysed in this section were as follows: the number of base patterns was $N_{\mathrm{base}} = 64$, the angular step for each Euler angle was 3°, $A_{\mathrm{tol}} = 5°$ and $N_{\mathrm{angl}} = 30$. The initial orientation determination of the simulated diffraction patterns for both structures was performed for the circular low-Q region, 150 pixels in diameter. The following 3D refinement was made for diffraction patterns with the size 512 pixels (8RUC structure) and 360 pixels (2BTV structure) with the angular step of 0.5° in the range of 5° near the position obtained on the previous stage. The orientation determination of 23 328 patterns of the 8RUC structure with the parameters listed above took about one week on a single 8-core computer consuming less than 1 Gb of RAM. The refinement of the found orientation for the 512 × 512 pixel 2D diffraction patterns in the full 3D volume consisting of 512 × 512 × 512 pixels took about one day and consumed about 8 Gb of RAM due to the requirement to store the full 3D volume in memory.

References

[1] Yefanov O, Vartanyants I and Weckert E 2010 *Photon Science—HASYLAB Annual Report* (http://photon-science.desy.de/annual_report/files/2010/20101462.pdf)

[2] Drenth J 2007 *Principles of Protein X-Ray Crystallography* (Berlin: Springer)

[3] Beck M, Luccic V, Foerster F, Baumeister W and Medalia O 2007 *Nature* **449** 611

[4] Ayache J, Beaunier L, Boumendil J, Ehret G and Laub D 2010 *Sample Preparation Handbook for Transmission Electron Microscopy* (Berlin: Springer)

[5] Neutze R, Wouts R, van der Spoel D, Weckert E and Hajdu J 2000 *Nature* **406** 752

[6] Gaffney K J and Chapman H N 2007 *Science* **316** 1444

[7] Mancuso A P, Yefanov O M and Vartanyants I A 2010 *J. Biotechnol.* **149** 229

[8] Ackermann W *et al* 2007 *Nature Photon.* **1** 336

[9] Emma P *et al* 2010 *Nature Photon.* **4** 641

[10] Ishikawa T *et al* 2012 *Nature Photon.* **6** 540

[11] Altarelli M *et al* 2006 XFEL: The European x-ray Free-Electron Laser *DESY Technical Design Report DESY 2006-097* (http://xfel.desy.de/technical_information/tdr/tdr/)

[12] Howells M *et al* 2009 *J. Electron Spectrosc. Relat. Phenom.* **170** 4

[13] Quiney H M and Nugent K A 2011 *Nature Phys.* **7** 142

[14] Lorenz U, Kabachnik N M, Weckert E and Vartanyants I A 2012 *Phys. Rev.* E **86** 051911

[15] Barty A *et al* 2012 *Nature Photon.* **6** 35

[16] Bergh M, Huldt G, Timneanu N, Maia F R N C and Hajdu J 2008 *Q. Rev. Biophys.* **41** 181

[17] Vartanyants I A and Robinson I K 2001 *J. Phys.: Condens. Matter* **13** 10593

[18] Vartanyants I A and Robinson I K 2003 *J. Synchrotron Radiat.* **10** 409

[19] Williams G J, Quiney H M, Peele A G and Nugent K A 2007 *Phys. Rev.* B **75** 104102

[20] Fienup J R 1982 *Appl. Opt.* **21** 2758

[21] Marchesini S 2007 *Rev. Sci. Instrum.* **78** 011301

[22] Seibert M M *et al* 2011 *Nature* **470** 78

[23] Kassemeyer S *et al* 2012 *Opt. Express* **20** 4149

[24] Loh N D *et al* 2012 *Nature* **486** 513

[25] Loh N D *et al* 2010 *Phys. Rev. Lett.* **104** 225501

[26] Huldt G, Szoke A and Hajdu J 2003 *J. Struct. Biol.* **144** 219

[27] Vainshtein B K and Goncharov A B 1986 *Sov. Phys. Dokl.* **287** 278

[28] Vainshtein B K and Goncharov A B 1986 *Proc. 11th Int. Congr. on Electron Mircoscopy (Kyoto)* ed T Imura, S Mause and T Suzuki (Tokyo: Japan Society of Electron Microscopy) p 459

[29] van Heel M 1987 *Ultramicroscopy* **21** 111

[30] Bortel G, Faigel G and Tegze M 2009 *J. Struct. Biol.* **166** 226

[31] Fung R, Shneerson V, Saldin D and Ourmazd A 2009 *Nature Phys.* **5** 64

[32] Schwander P, Fung R, Phillips G and Ourmazd A 2010 *New J. Phys.* **12** 134

[33] Loh N-T D and Elser V 2009 *Phys. Rev.* E **80** 026705

[34] Als-Nielsen J and McMorrow D 2011 *Elements of Modern X-Ray Physics* 2nd edn (New York: Wiley)

[35] Landau L D and Lifshitz E M 1976 *Mechanics* 3rd edn (Oxford: Butterworth-Heinemann)

[36] Yefanov O, Dronyak R, Barty A, Chapman H, Hajdu J and Vartanyants I 2011 unpublished

[37] Mancuso A *et al* 2009 *Phys. Rev. Lett.* **102** 035502

[38] RCSB Protein Data Bank (www.rcsb.org)

[39] Wang F, Weckert E, Ziaja B, Larsson D S D and van der Spoel D 2011 *Phys. Rev.* E **83** 031907

[40] Strüder L *et al* 2010 *Nucl. Instrum. Methods Phys. Res.* A **614** 483

[41] Hromalik M S, Green K S, Philipp H T, Tate M W and Gruner S M 2013 *Nucl. Instrum. Methods Phys. Res.* A **701** 716

[42] Wolfram Research (www.wolfram.com)

[43] Yefanov O M, Zozulya A V, Vartanyants I A, Stangl J, Mocuta C, Metzger T H, Bauer G, Boeck T and Schmidbauer M 2009 *Appl. Phys. Lett.* **94** 123104