Common arc method for diffraction pattern orientation

Gábor Bortel* and Miklós Tegze

Research Institute for Solid State Physics and Optics of the Hungarian Academy of Sciences, 1525 Budapest, PO Box 49, Hungary. Correspondence e-mail: gb@szfki.hu

Very short pulses of X-ray free-electron lasers opened the way to obtaining diffraction signal from single particles beyond the radiation dose limit. For three-dimensional structure reconstruction many patterns are recorded in the object's unknown orientation. A method is described for the orientation of continuous diffraction patterns of non-periodic objects, utilizing intensity correlations in the curved intersections of the corresponding Ewald spheres, and hence named the common arc orientation method. The present implementation of the algorithm optionally takes into account Friedel's law, handles missing data and is capable of determining the point group of symmetric objects. Its performance is demonstrated on simulated diffraction data sets and verification of the results indicates a high orientation accuracy even at low signal levels. The common arc method fills a gap in the wide palette of orientation methods.

1. Introduction

The idea of single-particle structure determination by means of diffraction beyond the radiation dose limit using very short pulses of X-ray free-electron lasers (XFEL) emerged more than a decade ago (Neutze *et al.*, 2000; Hajdu, 2000). Although early papers envisaged the three-dimensional structure of single biomolecules (Miao *et al.*, 2001, 2004; Webster & Hilgenfeld, 2002; Huldt *et al.*, 2003), recent experiments have studied particles of a few hundred nm in size at most with a resolution of a few nm (Chapman *et al.*, 2006, 2011; Barty *et al.*, 2008; Song *et al.*, 2008; Mancuso *et al.*, 2009, 2010; Bogan *et al.*, 2010; Loh *et al.*, 2010; Seibert *et al.*, 2011). It is true, however, that the principle is proven and the progress continuous. One type of such measurements records a large number of continuous diffraction patterns of replicas of the non-periodic object in unknown random orientations and the three-dimensional scattering density is determined through elaborate evaluation processes. The original concept of data processing consists of three separate steps: improving the statistics of low-intensity patterns by grouping and averaging, orientation by finding the intersection of the patterns, and finally real-space structure reconstruction by iterative phase retrieval. Here we focus on how the unknown random orientation of the individual scattering patterns can be determined.

The method proposed first for orientation originates from the field of cryo-electron microscopy (DeRosier & Klug, 1968; Hart, 1968; Crowther, 1971; van Heel, 1987; van Heel *et al.*, 2000; Frank, 1996; Penczek *et al.*, 1996) where planar central sections of the three-dimensional Fourier transform of the object are derived from the measured tomographic projections (based on the central section theorem); they are oriented by identifying their straight intersection lines through the origin, the *common lines*. The case of diffraction is different: the measured diffraction patterns define randomly oriented Ewald spheres in the reciprocal space, so their intersections are circles instead of straight lines. They can be oriented *via* identification of these *common arcs*; however, there are significant differences because of the geometry. While the concept of common arcs has been known for a long time (Huldt *et al.*, 2003), the method has never been applied or analysed in detail. There is a single case where the tangent to the common arc – the common line – was utilized to orient diffraction patterns (Shneerson *et al.*, 2008). This is an approximate solution, which is valid when the curvature of the Ewald sphere is negligible, typically in low-resolution, short-wavelength measurements. That geometry also necessitates the inspection of triplets of patterns when determining the orientation. Other methods have also been proposed for orienting low-intensity diffraction patterns based on generative topographic mapping (Fung *et al.*, 2009) and expansion maximization compression (Loh & Elser, 2009) which are able to utilize the similarity of patterns of close orientations and do not require preliminary classification. These methods were shown to be closely related (Moths & Ourmazd, 2010) and, in principle, are able to operate at the lowest intensities close to the limit where the differences due to counting noise or different orientation can be separated (Elser, 2009). A method for obtaining the reciprocal-space intensity distribution based

research papers

on a different principle was also proposed (Saldin *et al.*, 2009), and the two approaches were critically compared (Elser, 2010).

In this paper we supply the missing orientation method based on the *common arcs* of continuous diffraction patterns. The method is described in detail, its operation is demonstrated on the simulated data sets of two scattering objects, and the results are analysed carefully. Various questions that have arisen during experiments, *e.g.* the question of counting noise, validity of Friedel's law, symmetric objects and missing data, are all addressed by the common arc orientation method. We obtain rather precise orientations at significantly lower intensities than anticipated (Shneerson *et al.*, 2008). Also, the complete set of symmetry operations can be determined in the case of symmetric scattering objects. The excellent performance is attributed to the exact handling of the curvature of the common arcs and the high redundancy in the complete set of the pairwise determined relative orientations. The elements of the common arc orientation method are described in §2 and its application is demonstrated in §3.

2. Methods

2.1. Principle of the common arc method

The *Ewald* spheres are two-dimensional spherical slices of the reciprocal space through its origin that represent the achievable region for a single diffraction measurement at constant energy. The common arc method determines the unknown orientation of scattering objects by finding the intersection arcs of the corresponding Ewald spheres. The operation of this orientation method can be divided into two main steps:

(i) *Determine relative orientation of all pattern pairs by searching for matching common arcs.* The correlation of intensity distribution along intersection arcs as a function of relative orientation is calculated for each pair of diffraction patterns. The best relative orientation is determined by finding the largest correlation. This yields $N(N-1)/2$ unique relative orientations, representing the maximum pairwise obtainable information, N being the number of patterns to orient. However, ultimately we need only $N-1$ relative orientations, as the whole data set can be arbitrarily oriented. This means $N/2$-fold redundancy, representing the fact that each pattern

intersects all other patterns (and *vice versa*) providing orientation information. This opens up the space for a consistency check and averaging of orientations that are performed in the next step.

(ii) *Determine consistent absolute orientation of each pattern by selection and averaging.* The orientation of one arbitrarily selected pattern is fixed and all other pattern orientations are determined relative to this in several ways utilizing all available relative orientation information. The resulting set of absolute orientations for each pattern is checked for consistency; the reliable ones, which fall close to each other, are selected and averaged. Then the obtained absolute orientation of each pattern can be used to construct the three-dimensional reciprocal-space data for phase retrieval and three-dimensional structure reconstruction.

In the following sections we discuss the details and the formalism of the common arc method.

2.2. Relative orientation

Initially, all the scattering patterns are given in the same 'laboratory' Cartesian coordinate system, where typically the *xy* 'detector plane' is perpendicular to the *z* 'beam axis'. These data must be located in the reciprocal space by appropriate orientation and projection onto the Ewald sphere. However, without knowing the actual orientation of the sample, we cannot properly orient the Ewald sphere. Therefore, temporarily all patterns are projected on the same sphere in a standard, laboratory setting. To bring two such projected but not yet oriented patterns (P_a and P_b) into a correct relative orientation, where their intersection along the common arc becomes obvious, one has to rotate at least one of them. To describe this rotation, *i.e.* relative orientation, we use three *Euler* angles, Φ, Θ, Ψ (in the same convention as shown in Fig. 1 of Shneerson *et al.*, 2008). The three rotations described by $\mathrm{rot}_z(\Phi) \times \mathrm{rot}_x(\Theta) \times \mathrm{rot}_z(-\Psi)$ are illustrated in Fig. 1 and can be explained as follows: one of the patterns in its standard setting (*e.g.* P_b) is rotated about the beam axis by $-\Psi$ azimuthal angle, which brings the tangent to the common arc through the origin (*i.e.* the common line) of this pattern to the *x* axis. Then this pattern is tilted about the *x* axis by the Θ angle. We call this Θ angle the hinge angle, as this defines the 'inclination' of the two intersecting Ewald spheres and ultimately determines the curvature of the common arc. Finally, the tilted pattern is rotated again about the beam axis by the Φ azimuthal angle to bring the common arc of P_b into an exact overlap with that of P_a which is still in its standard setting. In another view, the inverse of this last rotation could be equivalently applied to the P_a pattern. Although the Euler angles are not always the best choice for orientation/rotation parameterization (*versus e.g.* quaternion representation), in this case they exactly correspond to the beam axial rotation of the two patterns and their actual inclination. Therefore we find them the most suitable to the problem.

After performing the above rotations we have set the two scattering patterns into correct relative orientation as illustrated in the upper panel of Fig. 2. The two spheres are the
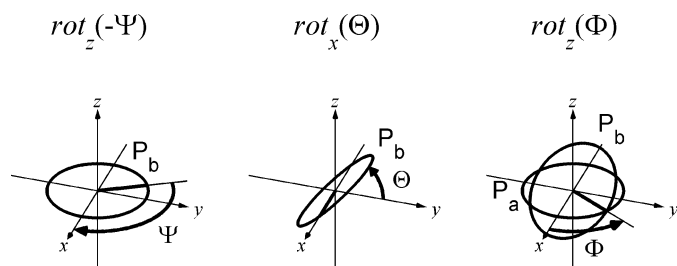
$rot_z(\text{-}\Psi)$ $rot_x(\Theta)$ $rot_z(\Phi)$

Figure 1
Illustration of the three Euler rotations describing the relative orientation of two patterns, P_a and P_b. Circles indicate either the tangent planes to the Ewald spheres at the origin of the reciprocal space or equally the detector planes in the real space. For further explanation, see text.

intersecting Ewald spheres; the areas with a polar grid illustrate the measured regions of the two scattering patterns. Their intersection, the common arc (CA), is plotted in red and the tangent to the common arc through the origin (O) is the approximating common line (CL), shown in green. It can be seen how limited the scattering range is, where the common line well approximates the common arc.

2.3. Common arc

In order to be able to correlate intensities along the common arc we have to determine its equation. It follows from the parameterization of relative orientations by the Euler angles that the Φ and Ψ rotations define only the azimuthal position of the common arc in the P_a and P_b patterns. The curvature of the common arc is defined solely by the Θ hinge angle. This geometry suggests polar gridding of the patterns on the Ewald sphere and polar parameterization for the equation
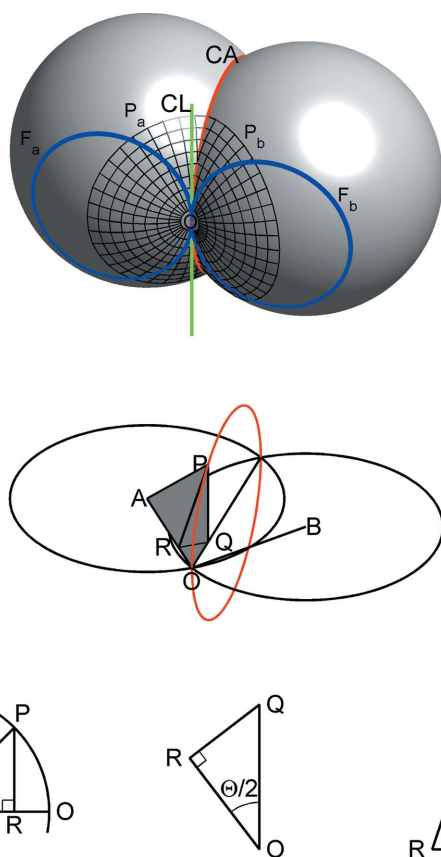

Figure 2
Upper panel: intersection of two Ewald spheres showing the geometry of two diffraction patterns (P_a, P_b), their common arc in red (CA), its tangent at the origin, the approximating common line in green (CL) and the common arc of a pattern with the other pattern's Friedel pair in blue (F_a, F_b). The Friedel pairs of the Ewald spheres, *i.e.* their inversion across the origin (O), are not shown for clarity. Middle panel: the same view of the construction made transparent. The spheres are shown by their main circles only, P is an arbitrary point of the common arc, Q and R are its projections onto the OAB plane and the OA line, respectively. Lower panel: extract of the three highlighted right-angled triangles used in the derivation of the equation of common arc. For further explanation, see text.

of the common arc. To avoid confusion with the Euler angles or the Bragg angle, let μ denote the *polar* and ν the *azimuthal* angular coordinate. μ runs from the minimal to the maximal scattering angle, its zero corresponds to the forward scattering direction and ν optimally covers a whole circle. The middle panel of Fig. 2 repeats the upper panel in a transparent view to aid the derivation of the equation. Here O marks the origin of the reciprocal space, A and B are the origins of the two Ewald spheres (indicated by their main circles in the OAB plane), P is an arbitrary point of the common arc, Q is defined as the projection of P onto the OAB plane and R as the projection of P onto the OA line. Further, $PAR\angle = \mu$ by definition, $QPR\angle = \nu$ if ν is measured from the normal of the OAB plane and $AOB\angle = \Theta$, the hinge angle. Now let us consider the three highlighted right-angled triangles that are also shown in the last row of Fig. 2. Temporarily assuming unit radius for the spheres, from the first triangle $\overline{OR} = 1 - \cos\mu$ and $\overline{PR} = \sin\mu$. Then in the second triangle $\overline{QR} = \overline{OR}\tan(\Theta/2) = (1 - \cos\mu)\tan(\theta/2)$. Finally, in the third triangle $\sin(\nu) = \overline{QR}/\overline{PR}$ with proper substitution yields the following base equation for the common arc in the polar coordinates:

$$\sin\nu = \frac{1 - \cos\mu}{\sin\mu}\tan\frac{\Theta}{2}. \qquad (1)$$

It is plotted in red in Fig. 3 for several Θ angles to illustrate how the common arc bends starting from a line, becomes a closed circle and finally shrinks to a point as Θ runs from 0 to π. The formula has the following properties: the origin $\mu = 0$, $\nu = 0$ is always an asymptotic solution to this equation; this is a fix point to all common arcs. Apart from the $\Theta = 0$ and $\Theta = \pi$ singular cases, there are two symmetric branches starting from the origin: ν and $\pi - \nu$ representing the two halves of the common arc. Depending on Θ, these two branches meet again at $\mu = \pi - \Theta$ polar and $\nu = \pi/2$ azimuthal angle, and the common arc becomes a full common circle. If $\Theta = 0$ (the two Ewald spheres exactly overlap), or if $\Theta = \pi$ (the two Ewald spheres face each other), the common arc becomes degenerate (extends to the whole sphere or shrinks to a point). These singular cases are not described by the above equation. It is more important, however, that this equation defines the arcs in the two investigated patterns along which we will have to find similar intensity distributions. It is done with an exhaustive three-dimensional search of the Θ, Φ and Ψ angles that yield the maximum correlation. Although the above base equation depends on the Θ hinge angle only, it is assumed that the Φ and Ψ azimuthal rotations have already been applied to the two intersecting patterns. These rotations by definition simply just shift the ν azimuthal coordinates, so the polar coordinates of the common arc will be (μ, $\Phi - \nu$) and (μ, $\Psi + \nu$) in the P_a and P_b patterns, respectively. Note the opposite signs of ν due to opposite curvature of arcs in the two patterns.

To easily obtain the data points along the common arc, it is expedient to resample the measured patterns to a polar grid by some kind of interpolation and/or averaging during preprocessing. (At least one transformation of the data is likely to be unavoidable for any data evaluation, as the typical

Cartesian two-dimensional detector pixel arrangements result in a hard-to-use distorted grid on the Ewald sphere, which is inconsistent with our ultimately desired three-dimensional Cartesian grid in the reciprocal space required by iterative reconstruction methods.) This way the advantageously chosen parameterization and polar gridding of the patterns together make possible a fast implementation of the three-dimensional maximum search procedure.

2.4. Friedel's law

If the scattering process involves no phase shift, it can be described by real scattering factors and Friedel's law applies. The intensities in the reciprocal space possess inherent inversion symmetry, since the corresponding scattering amplitudes are complex conjugates. For light elements, Friedel's law is applicable in most of the X-ray scattering cases, if

Figure 3
Common intersection arcs of two Ewald sphere projected scattering patterns (P_a, P_b) extending to 45° scattering angle for several Θ hinge angles. It is assumed that the Φ and Ψ azimuthal rotations already applied bring the tangent common line to the same (vertical) direction. The red curves show the two branches of the 'direct' intersection arcs (D_1, D_2), and the blue ones show the 'Friedel' intersection arcs (F_1, F_2). Note the flipped branches of the Friedel arcs.

we avoid the anomalous scattering near absorption edges. However, it becomes definitely invalid if the energy is lowered to the soft-X-ray region, for example to utilize transparency in the water window.

If the conditions of the measurement make Friedel's law applicable, that will advantageously affect our common arc orientation method as follows. Because of the inversion symmetry, any two opposite points in the reciprocal space will have the same intensity. Consequently, any measured scattering pattern determines the intensity distribution in the reciprocal space on two Ewald spheres: on one sphere directly and on the opposite sphere through the inverted pattern. We can exploit this when searching for the common arc of P_a and P_b patterns by using their opposite patterns $\overline{P}_a$ and $\overline{P}_b$, and simultaneously comparing intensities along the intersection arc of P_a and $\overline{P}_b$, for example (other combinations would not yield independent information). This effectively doubles the average length of the correlated arcs, making the correlation factor more reliable, less sensitive to noise.

The 'Friedel' common arcs are shown in both the upper panel of Fig. 2 and in Fig. 3 in blue. Their equation can also be calculated using equation (1) above, except $\tan(\Theta/2)$ must be replaced by $-\cot(\Theta/2)$ owing to the inversion of one of the spheres. For the same reason, the Friedel common arcs are flipped. Fig. 3 illustrates the complementary role of the two sets of arcs.

The option to include the Friedel common arcs in the relative orientation search and forcing the inversion symmetry when the three-dimensional reciprocal-space intensity is constructed from the oriented individual scattering patterns depends only on the physical applicability of Friedel's law and in no other way influences the process of orientation.

2.5. Correlation

When searching for the Φ, Θ, Ψ Euler angles determining the best matching relative orientation of two patterns along their common arcs, a weighted *Pearson* correlation factor is calculated as a measure of the similarity. This takes the following form:

$$\chi = \frac{\sum_{l=1}^{L} w_l (a_l - \bar{a})(b_l - \bar{b})}{\left[\sum_{l=1}^{L} w_l (a_l - \bar{a})^2\right]^{1/2} \left[\sum_{l=1}^{L} w_l (b_l - \bar{b})^2\right]^{1/2}},$$

$$\bar{a} = \frac{\sum_{l=1}^{L} w_l a_l}{\sum_{l=1}^{L} w_l}, \quad \bar{b} = \frac{\sum_{l=1}^{L} w_l b_l}{\sum_{l=1}^{L} w_l}. \quad (2)$$

Here all sums run through the L pixels along the common arcs (either including the Friedel common arcs or not), a_l and b_l denote the interpolated intensities from the two compared patterns, and w_l implements an optional weighting. These are obtained as follows. The scattering patterns already given on the polar grid described above are pre-normalized with the average scattering-angle dependence of all patterns. This
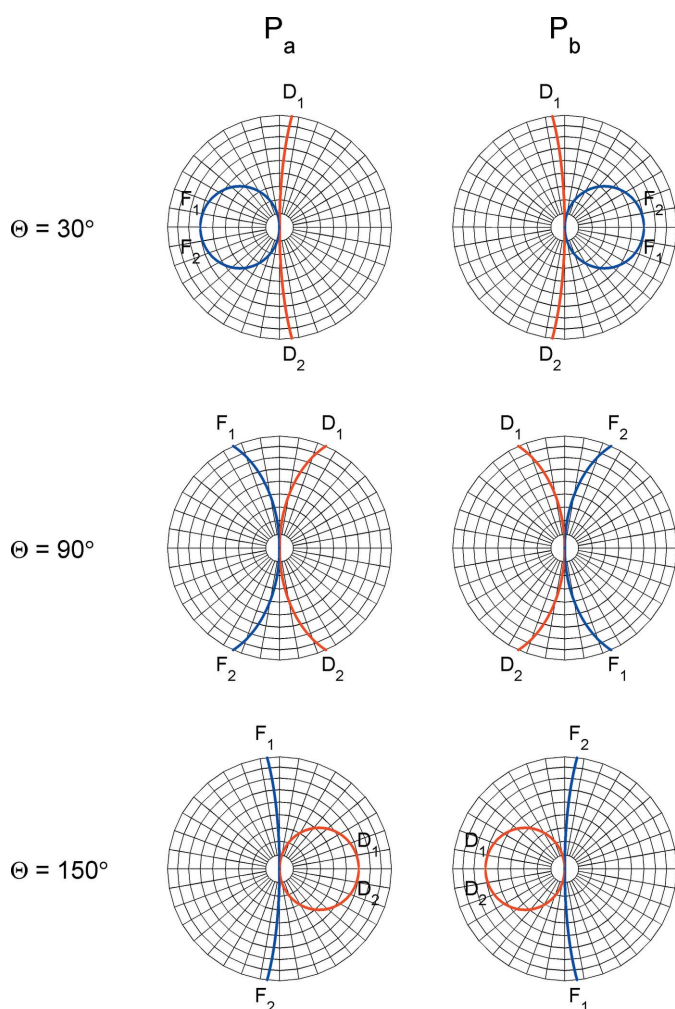
prevents the domination of often orders of magnitude higher intensities at low scattering angles. At the same time, their higher reliability due to counting statistics can be taken into account by properly chosen w_l weights. These weights can also express the higher sensitivity of the outer pixels to the orientation. The exact weighting scheme is a parameter of the method, which typically can be a function of intensity, statistics or even the scattering angle. However, regardless of the scale of the correlated values and the applied weighting, the correlation factor falls in the $[-1, +1]$ range, -1 representing anticorrelation, 0 no correlation and $+1$ the perfect correlation. This puts the $\chi(\Theta, \Phi, \Psi)$ correlation maps on an absolute scale, when searching for the maximal χ value on a discrete grid of the Φ, Θ, Ψ Euler angles. This is useful in the judgment of the best matching common arc of each pattern pair.

2.6. Absolute orientation

Once the relative orientation of all pattern pairs described by the rotation matrix R_{ij} is found, an absolute orientation matrix, O_i, has to be determined for each pattern. For one selected pattern, e.g. O_1, it can be arbitrary, e.g. the eye matrix, as the orientation of the whole reciprocal space is irrelevant. Then, the orientation of all other patterns is obtained by $O_i^{(1)} = R_{i1} \times O_1$. In principle this step solves the orientation problem, as all orientations are determined. However, one should go further to exploit the high degree of redundancy in the complete set of relative orientations. This is achieved by successive application of two relative rotations: $O_i^{(j)} = R_{ij} \times R_{j1} \times O_1$, to utilize all intermediate rotations. The above equations yield $N - 1$ candidate orientations for each pattern. In an ideal case these would be identical, but because of the sampled, interpolated and noisy data and the discrete grid used to search the relative orientations, they scatter, or even can be completely erroneous, i.e. identify a false common arc.
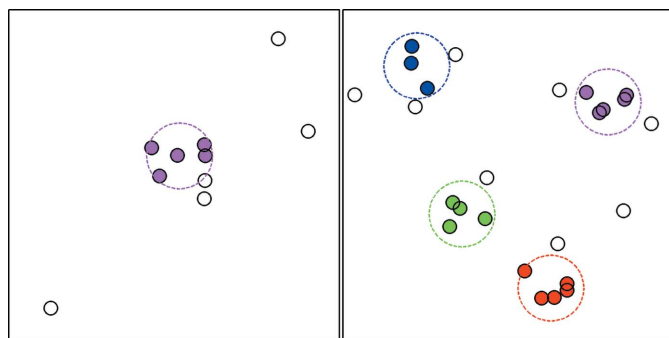

Figure 4
Simple illustration of selection and averaging procedure in two dimensions. Points on the plane represent orientations; their distance corresponds to the misorientation angle. A group quality threshold, i.e. a radius, is given as a parameter, and the neighbours within this radius are determined for each element. The most populated neighbourhood is selected and averaged. Left panel: only one group is determined (solid circles), the remaining elements are treated as outliers (empty circles). Right panel: the procedure is applied repeatedly to obtain a predefined number of groups (various colours), in between excluding the already clustered elements.

Therefore, a 'selection and averaging' procedure is performed on the $N - 1$ candidate orientations to improve the reliability and precision of the final absolute orientation assigned to the given pattern. This selection and averaging is illustrated in the left panel of Fig. 4, and is realized as follows. (i) For each pattern, the distances of candidate orientations are calculated as $\delta_{jj'} = \arccos(\{\text{trace}[O_i^{(j)} \times O_i^{(j')}] - 1\}/2)$. The misorientation angle given by this formula is a natural metric in the rotational group (Morawiec, 2004). (ii) For each candidate, $O_i^{(j)}$, the number of other candidates, $O_i^{(j')}$, falling within a distance range given by a tolerance angle, $\delta_{\lim}$, is determined. (iii) The most populated neighbourhood is selected; in case of a tie, the smaller average misorientation is preferred. (iv) The orientations within this group of candidate orientations are averaged and assigned to the given pattern as its most probable absolute orientation, O_i. The selection within a range eliminates outliers and the averaging improves the precision. The quality of the result is ensured by the proper choice of the tolerance angle parameter, which should allow averaging of naturally scattered orientations but should exclude erroneous ones.

2.7. Symmetric objects

The structure of single particles one wishes to determine may be symmetric. One can think of simple geometric body-shaped nanoparticles (Chapman et al., 2011), icosahedral viruses (Seibert et al., 2011) or the quaternary structure of multi-chain proteins. The real-space symmetry of the scattering density given by the point group leads to at least the same symmetry of the Fourier intensities in the reciprocal space. An inversion is possibly added (if not present before), due to Friedel's law, as discussed above. From the point of view of orientation problem, symmetry means that a given pattern's orientation cannot be unambiguously determined. The same measured pattern can be located in several equivalent settings in the reciprocal space, related by unknown symmetry operations. Note that we use the word 'setting', instead of orientation, as they may be related by other operations than a simple rotation (see below).

The common arc orientation method was extended not just to handle the scattering patterns of such symmetric objects, but also to take advantage of them. With this addition it is possible to identify the symmetry elements and also their orientation (the direction of rotation axes and normal vectors of mirror planes) in the reciprocal space. The idea is as follows: equivalent settings of the same pattern will yield several equally high peaks in the $\chi(\Theta, \Phi, \Psi)$ correlation map, when searching for the matching common arcs as a function of the relative orientation of two patterns, P_a and P_b. Each of these peaks indicates a highly correlating common arc and its position in the map provides the appropriate relative rotation. As a complication, equivalent settings can be related by operations other than a proper rotation. These are the inversion, a reflection or an improper rotation. However, these all can be separated into an inversion and a proper rotation. The latter component can be found by doubling the correlation map, i.e. calculating $\overline{\chi}(\Theta, \Phi, \Psi)$ using one of the inverted

patterns, $\overline{P}_b$, for example. If reflective symmetry is present, this map contains the same number of peaks with similar quality as the first map; otherwise it contains no peaks at all. The total number of peaks found on the two maps (χ and $\overline{\chi}$) should be equal to the number of equivalent settings, *i.e.* the *order* or *multiplicity, M*, of the point-group symmetry of the reciprocal space. By defining appropriate criteria on the local maxima on the maps this number can be determined automatically. However, in the present implementation of the algorithm, the expected multiplicity is given as an input parameter and that number of peaks is identified on the correlation maps. This approach does not exclude the possibility of determining the unknown multiplicity by testing all possible values of M (being a small integer number), and selecting the value that yields a consistent point-group symmetry (see below).

Once the M relative operations (rotations, or possibly rotations combined with inversion), ${}^s R_{ij}$, are determined for all pattern pairs from the correlation maps, we derive the absolute settings of the patterns based on similar equations that we used in the absence of symmetry: ${}^s O_i^{(1)} = {}^s R_{i1} \times O_1$ and ${}^{s,s'} O_i^{(j)} = {}^s R_{ij} \times {}^{s'} R_{j1} \times O_1$. The s superscript refers to one of the M symmetry equivalents. The only difference is that some of these equivalent settings may involve inversion and also that we will obtain a total $M + M^2 \times (N - 2)$ of them. This set of candidates contains all the M-equivalent absolute settings mixed together, with a redundancy of the order of approximately $M \times N$-fold. Therefore, a kind of clustering must sort them out and the redundancy is exploited by averaging. We can apply again the above-described selection and averaging procedure M times repeatedly, and remove the already averaged elements between the runs. The procedure is illustrated in the right panel of Fig. 4. This approach resembles the *quality threshold clustering* algorithm (Heyer *et al.*, 1999) in the sense that it limits the extent of clusters, but here the number of clusters is also limited. Consequently, some elements will not belong to any of the clusters; these are treated as outliers. By obtaining M-averaged absolute settings for each pattern the orientation problem of a symmetric object is solved, but we can go further and determine the symmetry operations themselves.

Any operation that transforms one setting of a pattern to an equivalent one is a *symmetry operation*. These can be written with the help of the above-determined absolute orientations as ${}^s O_i \times ({}^{s'} O_i)^{-1}$. This set contains $M^2 \times N$ symmetry operations derived from the equivalent settings of all patterns. However, many of them should be close to others, as there are only M common symmetry operations for all patterns, representing again an $M \times N$-fold redundancy. Their mixture is clustered into M distinct, quality-assured groups again by the repeatedly applied selection and averaging procedure to provide the M-averaged symmetry operations, *i.e.* the elements of the point group of the reciprocal space. The special directions of these operations (axes of rotations and normal of reflections) reveal the high-symmetry directions of the oriented patterns, which could be very useful during a visual inspection. Also, and perhaps more importantly, the relation of the determined symmetry operations can be

analysed to verify whether they form a valid point group. If not, possibly there may be a clue as to whether one should cancel or include symmetry elements and accordingly modify M, the expected multiplicity. Moreover, it is also possible to test all possible values of M (up to a reasonable limit), and pick the one that yields a consistent point group. This approach enables determination of the symmetry without its *a priori* knowledge.

One may argue against the implicit treatment of Friedel's inversion symmetry as described earlier, stating that inversion is just a symmetry element and could be treated the same way as any other reflective symmetry element described in this section. However, there is a significant difference, which makes inversion unique: it is the only symmetry element that can be applied where there is a lack of orientation information. It requires only fixing of the origin, in contrast to other types of point-group symmetries that involve some unknown orientation of the symmetry element (mirror plane or rotation axis). Therefore, the latter symmetry elements can only be treated as described here. While the inversion could be treated both ways, it is preferable to handle it implicitly during the correlation mapping as it improves the reliability of the determined relative orientations.

2.8. Missing data

An ideal single-particle diffraction pattern would cover the whole annulus between the coaxial cones defined by the beamstop and the crystallographic resolution. However, the patterns obtained experimentally often have unmeasured, or unusable regions even within this area. These missing regions are typically due to gaps between elements of a multidetector system, saturation at low angles, bad pixels or any other experimental deficiency. In general, an arbitrary mask on the data must be handled.

Missing data can be treated when calculating the above-defined correlation factor, simply by using only the pixels along the common arc that are present in both patterns ($l \in P_a \cap P_b$). In principle, that is all that has to be done; however, this makes the calculation more demanding as the number of common pixels and the normalization vary with the relative orientation. Yet to be able to pre-normalize the patterns, as an *approximation*, we have slightly modified the formula for the correlation written as the ratio of the covariance and the product of standard deviations:

$$\chi \simeq \frac{\sum\limits_{l \in P_a \cap P_b} w_l (a_l - \bar{a})(b_l - \bar{b}) / \sum\limits_{l \in P_a \cap P_b} w_l}{\left[\sum\limits_{l \in P_a} w_l (a_l - \bar{a})^2 / \sum\limits_{l \in P_a} w_l \right]^{1/2} \left[\sum\limits_{l \in P_b} w_l (b_l - \bar{b})^2 / \sum\limits_{l \in P_b} w_l \right]^{1/2}},$$

$$\bar{a} = \frac{\sum\limits_{l \in P_a} w_l a_l}{\sum\limits_{l \in P_a} w_l}, \quad \bar{b} = \frac{\sum\limits_{l \in P_b} w_l b_l}{\sum\limits_{l \in P_b} w_l}. \tag{3}$$

Here, the weighted covariance (the nominator) is calculated for the pixels of the common arc present in both patterns, but

Table 1
Parameters of the two scattering objects used in preparation of the simulated single-particle diffraction experiment.

	Virus model	Protein molecule
Size of the object	~5000 Å	~150 Å
Symmetry of object	Pseudo fivefold axis	Exact 422 point group
Wavelength/energy of radiation	50 Å/248 eV	1.24 Å/10 keV
Friedel's law	Invalid	Valid
Pulse fluence	$0.03 \times 10^{12}/(10\ \mu m)^2$	$30 \times 10^{12}/(100\ nm)^2$
Crystallographic resolution	100 Å	2.4 Å
Maximal scattering angle	29°	30°
Counts in outer Shannon–Nyquist pixels	0.7	0.2
Fraction of missing pixels	0.22%	0.15%
Number of patterns	100	100

the weighted mean and standard deviations are calculated for the pixels present in the given pattern (allowing its preliminary calculation). This approximation assumes that the weighted mean and standard deviation are not affected very much by the elements present only in one, but not in the other pattern. It is acceptable if a moderate fraction of data points is missing. The formula returns the exact value if there are no missing data.

3. Results and discussion

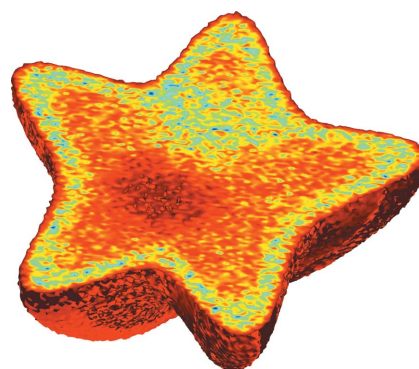## 3.1. Preparation of synthetic data

Beyond a description of the common arc orientation method, the secondary objective of our work is to demonstrate its capabilities in finding the correct orientation of single-particle diffraction patterns. To be able to *validate* the results the operation is demonstrated on synthetic, but as realistic as possible, data created under total control of parameters. This is achieved in three separate groups of tasks. (i) Calculate numerous diffraction patterns of selected test objects in random orientation. Make them more realistic by including counting noise and introducing missing regions in the patterns. (ii) Transform the patterns from the Cartesian detector grid to a polar grid suitable for the orientation algorithm. Perform the common arc method: map the correlations with relative orientation, determine the absolute orientation of each pattern and the symmetry operations, if applicable. Also combine the oriented two-dimensional patterns to the three-dimensional reciprocal-space distribution ready for structure reconstruction. (iii) Verify the obtained orientations and the constructed three-dimensional distributions against the ones used at the preparation phase and analyse the errors that characterize the method.

To demonstrate and test all aspects of the common arc method we have chosen *two scattering objects*. One of them was inspired by recent experiments performed at the already operational X-ray free-electron lasers (Seibert *et al.*, 2011). It is a model of a large virus with internal structure and pseudo-

symmetry, investigated with non-atomic resolution using longer-wavelength radiation and Friedel's law is assumed to be invalid. The other one is an example of objects targeted by the ultimate single-particle diffraction experiments planned at future X-ray free-electron laser sources (Miao *et al.*, 2001). It is a symmetric protein structure investigated with atomic resolution using shorter-wavelength radiation and Friedel's law is assumed to be valid. The details of the pattern-generating procedure are described below and the important parameters of these two objects and the simulated experiment are listed in Table 1.

(i) *Virus model with internal structure and pseudo-symmetry.* The overall shape of the model object was created with the help of Gielis curves (Gielis, 2003) extended to two dimensions, also called the spherical product of two superformulas. These functions define simple two-dimensional surfaces with the help of a few control parameters resembling the shape of various bodies found in nature. This boundary was smeared out with the help of a Fermi–Dirac function yielding the primary three-dimensional scattering density. With the linear combination of several such primary densities it is possible to create some lower- or higher-density regions within the particle or even features such as a shell with different density. Then a 20% random fluctuation of this combined scattering density was introduced to imitate a fine local structure with relatively low contrast. Finally, an imaginary part was added to the density in order to yield scattering patterns not obeying Friedel's law. Its amplitude was set to 10% of the real density according to the typical f_2/f_1 ratio of the atomic scattering factor corrections of C, N and O elements at the given radiation energy. The absolute value of the complex scattering density is illustrated in Fig. 5.

(ii) *Symmetric protein molecule at atomic resolution.* Our choice was the ribulose 1,5-bisphosphate carboxylase/oxygenase protein assembly, RuBisCo (Wildman, 2002; Portis & Parry, 2007), partially because of its abundance in the related literature (Miao *et al.*, 2001) including our study on grouping of its low-intensity patterns (Bortel & Faigel, 2007; Bortel *et al.*, 2009) and partially because of its highly symmetric tertiary

Figure 5
Absolute value of the three-dimensional complex scattering density of the virus model shown on a combined isosurface and cut-slice plot. The object has a pseudo-symmetric shape, and internal structure on various length scales with small contrast. An imaginary part was added to make Friedel's symmetry invalid.

structure. Structural data were taken from the Protein Data Bank (PDB) ID 1ej7 (Duff *et al.*, 2000), and the biologically active unit of eight long and short chains was generated using the 422 point-group symmetries. The weight of the approximately spherical assembly is about 0.54 MDa. According to the high energy of the radiation, Friedel's law was assumed to be valid.

Calculation of the scattering amplitudes was somewhat different for the two objects. For the three-dimensional electron-density model the scattering amplitude in the reciprocal space was calculated by three-dimensional discrete Fourier transform. The simulated scattering patterns on the planar Cartesian detector were obtained by interpolation to the corresponding two-dimensional projected grid of the randomly oriented Ewald spheres. For the atomistic protein model the scattering amplitudes at each scattering vector of the two-dimensional pattern were calculated directly by summing up the spherical wave contributions of each atom, taking into account the atom's form factor and its phase due to its position. In the case of both objects the scattering amplitudes were squared and *scaled* to photon counts that enabled us to include counting statistics in the simulated scattering patterns. For this we assumed a realistic 0.4 e $\text{\AA}^{-3}$ average electron density in the case of the virus model object, the experimentally achievable photon fluence (number of incoming photons per focal spot area) and detector pixel size corresponding to the Shannon–Nyquist sampling. The solid angle of this pixel was taken as $(\lambda/2D)^2$, λ being the wavelength and D the object size (Huldt *et al.*, 2003; Shneerson *et al.*, 2008).

As the last step of synthetic data creation, a random 0.1% of the pixels and the ones having counts larger than 1% of the forward scattering value were masked out, representing some bad and saturated pixels, respectively. One of the synthetic scattering patterns of the RuBisCo enzyme is shown in Fig. 6. Note that, in spite of valid Friedel's law, the individual scattering patterns are not centrosymmetric because of the
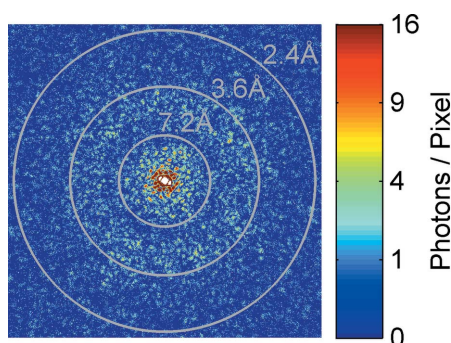
curvature of the Ewald sphere. Nevertheless, an approximate symmetry is observable in the central region, where the tangent plane well approximates the sphere. A total of 100 patterns were generated for both objects and passed to the orientation procedure, keeping the orientation information for verification purposes. We have to note that these simulated patterns could also be interpreted as the result of a classification and averaging procedure of a larger set of weaker diffraction patterns.

3.2. Orientation by the common arc method

The scattering patterns created on a Cartesian detector grid are first transformed onto a polar grid suitable for the common arc search procedures. This extra step was intentionally included as a preparation to accept *real data*. The different geometry of the two grids, specifically the mismatch of the point densities, necessitated a combined interpolation and averaging procedure. To utilize most of the information, interpolation is the better choice where the targeted polar grid is denser than the original Cartesian grid, and averaging is preferable in the opposite case. The step size of the polar grid was chosen as $0.5°$ in the polar and $1°$ in the azimuthal direction. Consequently, the steps of the three Euler angles during the orientation search were also $1°$. In the correlation-factor calculations all pixel values were normalized with their corresponding solid angle, *i.e.* the intensity was used rather than the pixel-size-dependent counts. Furthermore, all patterns were pre-normalized by the average scattering-angle dependence to prevent domination of the low-scattering-angle, high-intensity regions. The weights used in the correlation factor were chosen proportionally to the scattering angle representing a higher sensitivity of the outer pixels to the orientation. The validity of Friedel's law and an eightfold symmetry multiplicity was assumed in the case of RuBisCo. The common arc orientation of 100 patterns with the above parameters takes several hours on a typical desktop computer. The most time-consuming task is to calculate the correlation
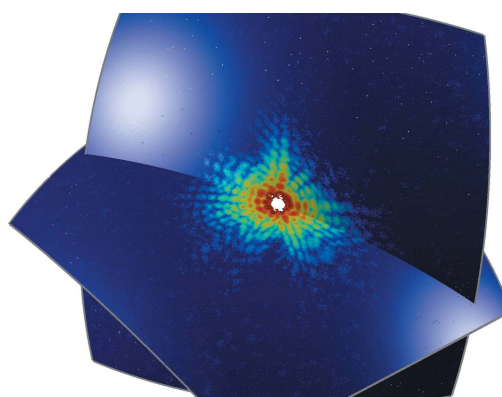

Figure 6
One of the 100 calculated scattering patterns of the RuBisCo protein. The square-root intensity scale represents photon counts in the Shannon–Nyquist sampling pixel size, having $(\lambda/2D)^2$ solid angle. To make low counts visible, only the 0–16 counts range is shown, larger values are saturated to the same colour. The average count in the outermost indicated resolution shell of 2.4 Å is about 0.24. The mask of the missing and saturated pixels is visible in the centre as a white area that is handled by the orientation algorithm.

Figure 7
Zoomed intersection of two oriented scattering patterns of the virus model showing the perfect correlation of intensities along the common arc. The hole in the centre is due to the saturated and missing pixels that were made transparent.

map of the relative orientations. It could be speeded up by parallelization.

The success of the common arc orientation process can be assessed from several *indicators* already available during its progress. The maximal correlation values of the $\chi(\Theta, \Phi, \Psi)$ maps indicating common arcs are strongly affected by the signal-to-noise ratio in the corresponding patterns. While in the high-intensity, low-noise case the peak correlation approaches 1 and the peak clearly stands out from the map, in the opposite extreme they ultimately diminish in the average correlations of non-common arcs. Therefore it is worth analysing the amplitude, shape and width of the peaks and also visualizing the corresponding relative orientation. Fig. 7 shows a perfectly matching common arc of two patterns of the virus model object found by the algorithm. Here the correlation factor of the intensities along the arc is 0.95, but the distribution of the local maxima of all pattern pairs falls between 0.5 and 0.95. During verification of the resulting orientations in the next section we will see that such low correlations already allow one to correctly determine the orientations.

When determining the absolute orientation by the selection and averaging procedure, the *fraction* of selected and averaged orientations serves as another useful indicator of the quality of the result. Its higher value expresses the consistency of the candidate orientations and the reliability of the average. A value of 100% would indicate that all common arcs have been consistently found and contribute to the average. Nevertheless, our experience shows that a value as low as 10% is already enough to correctly orient all patterns with low error (see the verification below). The relevant parameter, the averaging range angle, δ_{lim}, was chosen as $3°$, a few times the grid step size, which was confirmed by the typical peak widths found on the orientation map and present in the misorientation-angle distributions. This shows the importance and effectiveness of this step in filtering the consistent orientations.
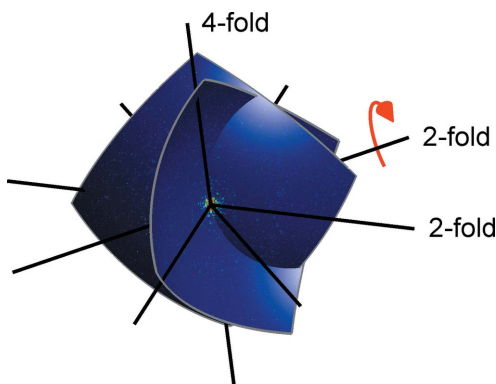
In the case of symmetric objects, such as RuBisCo, the symmetry operations determined by the common arc method also give the opportunity of a self-consistency check. Fig. 8 illustrates how the symmetry operations are obtained. Two equivalent settings of one selected example pattern defined by multiple peaks of the $\chi(\Theta, \Phi, \Psi)$ maps reveal the underlying symmetry operation, in this case a $180°$ rotation indicated by the red arrow. If these operations are found to be common for the equivalent setting of all patterns, then ultimately they define the corresponding twofold symmetry axis with high reliability *via* selection and averaging. Other symmetry operations can be derived similarly, a total of eight, the expected multiplicity. The axes of these operations are also shown in Fig. 8. Their folds were determined from the corresponding rotation angles that happened to be $0°$, $90°$ and $180°$, which is already a good sign. Furthermore, the symmetry axes make $45°$, $90°$ or $135°$ angles. It can be verified that these symmetry operations do form the complete 422 point group. In general, finding symmetry elements that define a valid point group is a strong indication of the correctly identified symmetry group, merely from its multiplicity. To demonstrate the feasibility of determining the multiplicity itself, we attempted to perform the common arc orientation by assuming a multiplicity of 7 or 9. The result was either lack or a surplus of symmetry operations and the required completeness of the point group revealed the correct symmetry. The high accuracy of the symmetry operations determined from the equivalent settings of all patterns – less than $0.01°$ error in the rotation angles and the direction of the axes – is attributed to the high degree of redundancy.

3.3. Verification of the results

The above observations characterize the common arc orientation method in itself and can also be examined in the case of real data. Now, taking advantage of synthetic data, we can judge the results based on the original orientation information kept from the pattern-preparation phase. This allows
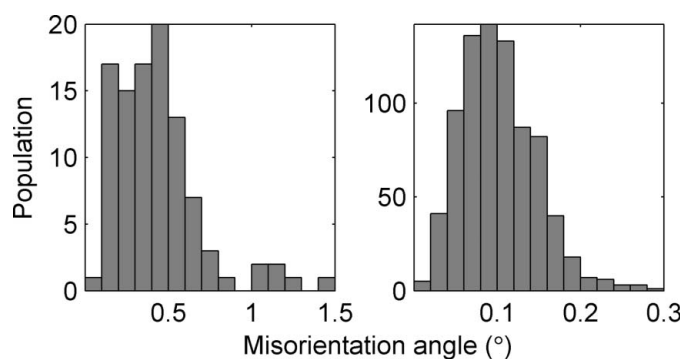

Figure 8
Two possible settings of a pattern and the four- and twofold axes of symmetry operations as found by the common arc orientation method. The two settings are found to be related by a $180°$ rotation (indicated by the red arrow) that defines one of the twofold axes. All equivalent settings of all patterns define the symmetry operations with high precision. These form the RuBisCo's complete 422 point group. Plotting of all eight equivalent and also the eight Friedel inverted settings of the oriented pattern would make the figure too busy.

Figure 9
Distribution of the misorientation angles of the oriented patterns for the virus model object (left panel) and the RuBisCo protein assembly (right panel). The values verify the orientations determined by the common arc method against the original ones kept from the data preparation. Low average and maximum misorientation angles prove perfect orientation.

us to quantitatively verify the results and specify the operation limits and the effectiveness of the method.

Fig. 9 shows the distribution of *misorientation* angles of the original and the oriented patterns for both scattering objects. The $0.4°$ and $0.1°$ mean values convert to 1.5 and 0.4 pixel positional error at the perimeter (the highest-resolution ring) of the individual scattering patterns. This means that in both cases all patterns have been oriented by the common arc method precisely enough for structure determination to high resolution. It has to be noted that more patterns or better statistics would further decrease the orientation errors. There is also good reason to expect that the method works for other objects and X-ray pulses with similar parameters as listed in Table 1.

This success naturally raises the question of the intensity that is required for successful orientation. The parameters listed in Table 1 represent a safe operation regime for our two scattering objects. Values leading to lower counts in the diffraction patterns already caused significant orientation errors for a few percent of the patterns, but a large fraction of the patterns was still correctly oriented. In the case of the virus model object the stated fluence is well above the parameters of the already operational FLASH and LCLS XFEL sources – it is only 3% of 10^{12} photons focused to $10 \times 10 \, \mu m^2$. However, assuming 10^{12} photons focused to $100 \times 100 \, nm^2$ from the European XFEL source, successful orientation of the RuBisCo patterns requires preliminary grouping and averaging, the so-called classification of patterns (Huldt *et al.*, 2003; Bortel & Faigel, 2007; Bortel *et al.*, 2009), with an average 30

Figure 10
Three main slices of the three-dimensional intensity distribution in the reciprocal space as constructed from the original two-dimensional scattering patterns of RuBisCo oriented by the common arc method. The fuzzy surface of the roughly spherical volume is due to the voxels that do not have a contribution from any of the oriented patterns. Voxels further in are defined by more and more patterns with higher reliability. These data have a correlation of 0.98 with the original intensity distribution in the reciprocal space.

patterns per group. Its feasibility was demonstrated in our previous work (Bortel *et al.*, 2009). Here we note that this preliminary classification and averaging of the two-dimensional patterns is advisable only if necessary; otherwise, as in the case of the virus model object, it is preferable to directly orient the measured patterns.

For easy comparison to the work of Shneerson *et al.* (2008) on the orientation by the common lines, the mean photon count in a Shannon–Nyquist pixel at the highest-resolution ring is also given in Table 1. The sufficient 0.7 and 0.2 counts for our two scattering objects indicate that the common arc method is able to orient at significantly *lower intensities* with *higher precision* than was anticipated from the performance of the approximate method. The cited work states 10 counts per pixel and $4°$ average misorientation error and draws a rather pessimistic conclusion on the feasibility of such experiments. We attribute the origin of the improvement to the precise handling of the curvature of the common arcs and to exploiting the high degree of redundancy of all relative orientations which in a noisy case improves the tolerance of the method. This signal level also gets close to that of other algorithms (Loh & Elser, 2009; Fung *et al.*, 2009).

The iterative phase-retrieval algorithms – the last step in determining the actual density of the scattering object – typically require the reciprocal-space intensities on a three-dimensional Cartesian grid, suitable for fast Fourier transform cycles. These data were prepared by binning all original two-dimensional patterns (not the ones transformed to the polar grid) to the voxels of the three-dimensional grid utilizing the orientation information provided by the common arc method. Contributions to each voxel were then averaged. Three main slices of these volume data through the origin of the reciprocal space are shown in Fig. 10. The outer regions with no pattern contribution yield a fuzzy surface of the roughly spherical volume that is defined by the original diffraction patterns. The intensity of the voxels further in, with contributions from many patterns, possesses very high reliability. Although this volume contains all pixels of all patterns, it obviously has to be trimmed to make it usable by the reconstruction process. The voxel–voxel correlation of these data and the original intensity distribution in the reciprocal space were calculated as a quantitative measure of the accuracy. The high correlation-factor value of 0.98 indicates that the common arc orientation method including the necessary data transformation from the two-dimensional detectors to the three-dimensional volume grid results in a faithful reproduction of the diffraction intensity.

For successful real-space structure reconstruction the noise level of the three-dimensional diffraction intensity should be sufficiently low. Various studies (*e.g.* Miao & Sayre, 2000) demonstrate the noise tolerance of the iterative phase-retrieval algorithms, even in the case of complex valued three-dimensional densities. As the common arc method is able to orient the two-dimensional noisy patterns with high precision, the signal-to-noise ratio of the three-dimensional reciprocal-space data improves by increasing the number of measured patterns and the signal-to-noise-level requirement of the
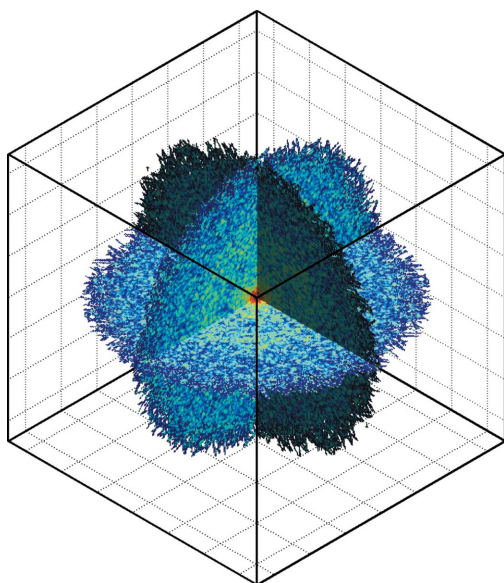
reconstruction procedure can be met. This way the success of structure reconstruction depends on the quality and quantity of the measured diffraction data.

4. Conclusion

The common arc method for the orientation of continuous diffraction patterns of non-periodic single particles has been described in detail. The method relies on the curved intersection of Ewald spheres in finding the relative orientation of all pairs of diffraction patterns, and then determines a consistent orientation for each of them *via* selection and averaging. The approach is significantly new compared to recent application of the approximate method of common lines (Shneerson *et al.*, 2008) adapted from the field of cryo-electron microscopy. Taking simulated diffraction patterns of various scattering objects, the common arc orientation method is shown to be able to operate at significantly lower intensity levels (less than a count per pixel) and yield orientations with higher precision (within a few tenths of a degree) than was anticipated earlier. Also it was extended to handle missing regions in the diffraction patterns and to take advantage of Friedel's symmetry, if it exists. In the case of symmetric scattering objects the symmetry operations are automatically obtained from the equivalent settings of the patterns and ultimately the complete point-group symmetry can be determined. It is a non-iterative method providing deterministic results, with various indicators on the reliability and consistency of the orientations; therefore it can also aid iterative orientation algorithms with a biased state in emerging from their random initial states. It is also shown that the common arc orientation process introduces no artefacts into the three-dimensional reciprocal-space intensity distribution and therefore can contribute to the success of structure reconstruction.

References

Barty, A. *et al.* (2008). *Nature Phot.* **2**, 415–419.
Bogan, M. J. *et al.* (2010). *Aerosol Sci. Technol.* **44**, i–vi.
Bortel, G. & Faigel, G. (2007). *J. Struct. Biol.* **158**, 10–18.
Bortel, G., Faigel, G. & Tegze, M. (2009). *J. Struct. Biol.* **166**, 226–233.
Chapman, H. N. *et al.* (2006). *Nature Phys.* **2**, 839–843.
Chapman, H. N. *et al.* (2011). *Nature (London)*, **470**, 73–77.
Crowther, R. A. (1971). *Philos. Trans. R. Soc. London Ser. B*, **261**, 221–230.
DeRosier, D. & Klug, A. (1968). *Nature (London)*, **217**, 130–134.
Duff, A. P., Andrews, T. J. & Curmi, P. M. (2000). *J. Mol. Biol.* **298**, 903–916.
Elser, V. (2009). *IEEE Trans. Inform. Theory*, **55**, 4715–4722.
Elser, V. (2010). http://arxiv.org/abs/1007.3777v1.
Frank, J. (1996). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. San Diego: Academic Press.
Fung, R., Shneerson, V., Saldin, D. K. & Ourmazd, A. (2009). *Nature Phys.* **5**, 64–67.
Gielis, J. (2003). *Am. J. Bot.* **90**, 333–338.
Hajdu, J. (2000). *Curr. Opin. Struct. Biol.* **10**, 569–573.
Hart, R. G. (1968). *Science*, **159**, 1464–1467.
Heel, M. van (1987). *Ultramicroscopy*, **21**, 111–124.
Heel, M. van, Gowen, B., Matadeen, R., Orlova, E. V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M. & Patwardham, A. (2000). *Q. Rev. Biophys.* **33**, 307–369.
Heyer, L. J., Kruglyak, S. & Yooseph, S. (1999). *Genome Res.* **9**, 1106–1115.
Huldt, G., Szoke, A. & Hajdu, J. (2003). *J. Struct. Biol.* **144**, 219–227.
Loh, N. D. *et al.* (2010). *Phys. Rev. Lett.* **104**, 225501.
Loh, N. D. & Elser, V. (2009). *Phys. Rev. E*, **80**, 026705.
Mancuso, A. P. *et al.* (2009). *Phys. Rev. Lett.* **102**, 035502.
Mancuso, A. P., Yefanov, O. M. & Vartanyants, I. A. (2010). *J. Biotechnol.* **149**, 229–237.
Miao, J., Chapman, H. N., Kirz, J., Sayre, D. & Hodgson, K. O. (2004). *Annu. Rev. Biophys. Biomol. Struct.* **33**, 157–176.
Miao, J., Hodgson, K. O. & Sayre, D. P. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 6641–6645.
Miao, J. & Sayre, D. (2000). *Acta Cryst.* A**56**, 596–605.
Morawiec, A. (2004). *Orientations and Rotations, Computations in Crystallographic Textures*. USA: Springer-Verlag.
Moths, B. & Ourmazd, A. (2010). http://arxiv.org/abs/1005.0640v1.
Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. (2000). *Nature (London)*, **406**, 752–757.
Penczek, P. A., Zhu, J. & Frank, J. (1996). *Ultramicroscopy*, **63**, 205–218.
Portis, A. R. & Parry, M. A. (2007). *Photosynth. Res.* **94**, 121–143.
Saldin, D. K., Shneerson, V. L., Fung, R. & Ourmazd, A. (2009). *J. Phys. Condens. Matter*, **21**, 134014.
Seibert, M. M. *et al.* (2011). *Nature (London)*, **470**, 78–81.
Shneerson, V. L., Ourmazd, A. & Saldin, D. K. (2008). *Acta Cryst.* A**64**, 303–315.
Song, C., Jiang, H., Mancuso, A., Amirbekian, B., Peng, L., Sun, R., Shah, S. S., Zhou, Z. H., Ishikawa, T. & Miao, J. (2008). *Phys. Rev. Lett.* **101**, 158101.
Webster, G. & Hilgenfeld, R. (2002). *Single Mol.* **3**, 63–68.
Wildman, S. G. (2002). *Photosynth. Res.* **73**, 243–250.