CellPress

Structure
Resource

Autofocused 3D Classification of Cryoelectron Subtomograms

Yuxiang Chen,[1,2] Stefan Pfeffer,[1] José Jesús Fernández,[3] Carlos Oscar S. Sorzano,[3] and Friedrich Förster[1,*]
[1]Department of Molecular Structural Biology, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
[2]Computer Aided Medical Procedures (CAMP), Technische Universität München, Boltzmannstrasse 3, 85748 Garching, Germany
[3]National Centre for Biotechnology, National Research Council (CNB-CSIC), Campus UAM, Darwin 3, Cantoblanco, 28049 Madrid, Spain
*Correspondence: foerster@biochem.mpg.de
http://dx.doi.org/10.1016/j.str.2014.08.007

SUMMARY

Classification of subtomograms obtained by cryoelectron tomography (cryo-ET) is a powerful approach to study the conformational landscapes of macromolecular complexes in situ. Major challenges in subtomogram classification are the low signal-to-noise ratio (SNR) of cryo-tomograms, their incomplete angular sampling, the unknown number of classes and the typically unbalanced abundances of structurally distinct complexes. Here, we propose a clustering algorithm named AC3D that is based on a similarity measure, which automatically focuses on the areas of major structural discrepancy between respective subtomogram class averages. Furthermore, we incorporate a spherical-harmonics-based fast subtomogram alignment algorithm, which provides a significant speedup. Assessment of our approach on simulated data sets indicates substantially increased classification accuracy of the presented method compared to two state-of-the-art approaches. Application to experimental subtomograms depicting endoplasmic-reticulum-associated ribosomal particles shows that AC3D is well suited to deconvolute the compositional heterogeneity of macromolecular complexes in situ.

INTRODUCTION

Cryoelectron tomography (cryo-ET) is a 3D imaging technique to visualize macromolecular complexes in their physiological environment (Lučić et al., 2005). In cryo-ET, the 3D density map (tomogram) of a frozen-hydrated sample is reconstructed from 2D projections, which are acquired from different tilt angles using a transmission electron microscope (TEM). The applicable electron dose limits the spatial resolution of the tomograms typically to approximately 5–10 nm (Grünewald et al., 2003). If multiple copies of the macromolecule of interest are present, aligning them to a common coordinate system and averaging them enhance the signal and, hence, increase the resolution (subtomogram averaging) (Bartesaghi and Subramaniam, 2009; Förster and Hegerl, 2007). Resolutions in the subnanometer regime have been obtained from more than 100,000 subtomograms (Schur et al., 2013). More commonly, the number of subtomograms is in the range of few thousands, yielding resolutions of up to 15–20 Å (Briggs et al., 2009; Chen et al., 2013; Eibauer et al., 2012). Due to the advance of data acquisition and image processing, cryo-ET is becoming an increasingly important tool for structural studies of macromolecules in situ, e.g., complexes associated with their native membranes (Bartesaghi and Subramaniam, 2009; Briggs, 2013; Förster and Hegerl, 2007; Pfeffer et al., 2012, 2014).

Subtomogram averaging normally comprises the following steps. (1) Localize the different copies of the macromolecule of interest. This can, for example, be accomplished by a six-dimensional exhaustive cross-correlation search with a structural template of the molecule under scrutiny, commonly referred to as template matching (Förster et al., 2010; Frangakis et al., 2002). (2) Classify the obtained candidates/subtomograms to ensure the homogeneity of the data set. Heterogeneity can be due to false-positives but also to conformational differences of the particles depicted by the subtomograms. (3) Align and average the subtomograms to obtain higher resolution structures. Several software packages have been developed for this purpose, including AV3 (Förster and Hegerl, 2007), Protomo (Winkler, 2007), EMAN2 (Tang et al., 2007), PEET (Heumann et al., 2011), Dynamo (Castaño-Díez et al., 2012), and PyTom (Chen et al., 2013; Hrabe et al., 2012).

In this paper, we focus on the second step, subtomogram classification, which is particularly challenging for several reasons. (1) The signal-to-noise ratio (SNR) of cryoelectron tomograms is poor (typically in the range of 0.1–0.01). (2) The tilt range for data acquisition is limited, typically from −60° to 60°, which results in an incomplete sampling in Fourier space (missing wedge problem). (3) The number of classes is typically unknown beforehand. (4) The classes of subtomograms can be unbalanced (strongly differing populations). (5) The structural differences between the class averages can be subtle.

Several approaches have been introduced for classification of cryoelectron subtomograms. We and others have previously introduced the constrained principal-component analysis (CPCA); the constrained correlation coefficient (*CCC*), in which two volumes are correlated only in their overlapping regions in Fourier space, is used as the similarity score of the correlation matrix, which is then analyzed by principal-component analysis (PCA) and *k*-means clustering (Bartesaghi et al., 2008; Förster et al., 2008). Alternative PCA-based classification approaches are probabilistic PCA with expectation maximization

Box 1. Algorithm 1: AC3D

Input:
SS: A set of input subtomograms
k: Number of classes.

Output:
CS: Class-labeled subtomograms

Begin
01 Prealign *SS*
02 Initialize *k* class centers $SV = \{V_1, \cdots, V_k\}$
03 while #class changes > 0.5% **do**
04 Align *SS* to *SV* and obtain the corresponding scores *SCS*
05 Determine the noise class so that $SS = SS' \cup SS_{noise}$
06 Calculate the focused scores (*FSS*) of SS' with respect to *SV*
07 Determine the class labels according to *FSS*, which results in *CS*
08 Update the alignment of *CS* according to the class assignment
09 Average classes in *CS* to get the new class centers *SV*
10 end while
end

(Yu et al., 2010) and wedge-masked differences-corrected PCA (Heumann et al., 2011). Scheres et al. (2009), as well as Stölken et al. (2011), formulated the classification problem statistically and developed a maximum likelihood (ML) approach for simultaneous alignment and classification. There are also other approaches that conduct simultaneous subtomogram alignment and classification: Winkler (2007) and Hrabe et al. (2012) extended real space subtomogram averaging protocols to multireference procedures. Xu et al. (2012) proposed a fast rotational matching (FRM) method for subtomogram alignment and a local feature enhancement strategy for classification. Kuybeda et al. (2013) used a nuclear norm-based, collaborative similarity measure for subtomogram alignment. Despite their successes when applied to respective data sets, the performances of all these methods tend to be limited, in particular for unbalanced classes and subtle structural differences.

Here, we propose an unsupervised learning approach named AC3D that automatically focuses the classification on the most variable parts of 3D structures. This similarity metric can capture subtle differences and does not involve any human intervention, thus alleviating bias. Based on this metric, we introduce an iterative multireference clustering scheme that makes use of a fast subtomogram alignment algorithm to achieve a substantial speedup. Moreover, we adapt *k*-means++ as the initialization strategy for the clustering procedure to avoid being trapped in local optima and to accelerate the convergence. Comparisons of AC3D against the CPCA approach (Förster et al., 2008) and the ML approach MLTOMO (Scheres et al., 2009) on a simulated data set show significant improvements of classification accuracy. Application of AC3D on experimental cryo-tomograms of ER-associated ribosomes yields clearly distinct conformations, including established ribosome states without any human intervention or prior knowledge.

RESULTS

Overall Classification Workflow

The overall workflow is first briefly described in Algorithm 1, and some important components are explained in the following sections (Box 1).

The iterative optimization procedure of AC3D is a multireference scheme, which is closely related to *k*-means clustering. However, we use a more efficient initialization (see "Initialization of Class Assignment").

The basic workflow goes as follows: First, the subtomograms are prealigned using our single-reference fast alignment algorithm described by Chen et al. (2013). After initialization, the class centers (the subtomogram averages) are computed. During each iteration, subtomograms are aligned and assigned to the "closest" class center. All the class centers get updated subsequently, using the assigned class members and their respective alignments. The whole procedure iterates until it converges or the maximal number of iterations is reached.

There are a few challenges when implementing this algorithm for cryo-ET. First, an appropriate similarity metric is required to measure the "distance" of each subtomogram to the class average. We make use of the *CCC*, which constrains the correlation to the commonly sampled region in Fourier space (Förster et al., 2005). However, computing the *CCC*s is time consuming because each subtomogram has to be optimally aligned to the class centers prior to computing the *CCC*. The alignment is a problem of 6 degrees of freedom (DoF): 3 for translation and 3 for rotation. We tackle this problem by a fast 6 DoF alignment algorithm we introduced earlier (Chen et al., 2013), which is briefly explained later. Second, the SNR of cryoelectron tomograms is relatively low, making it difficult to identify outliers/noise that may deteriorate the clustering performance. This problem is explicitly handled here using the score distribution functions. Third, it is difficult to classify subtle structural differences in cryo-ET data. The *CCC* quantifies the similarity between two volumes globally or within a subjectively chosen real-space mask of interest (Förster et al., 2008). An objective and robust way has to be found to define the mask where significant differences are located, because the noise may otherwise deteriorate the classification performance. Here, we present an algorithm to automatically focus the clustering on the variable parts of the macromolecule of interest and calculate the so-called focused score as the similarity measure. These three features are discussed in detail in the following sections.

Initialization of Class Assignment

k-means clustering normally starts with a random initialization of the class assignment. Nevertheless, it is known that the performance of *k*-means strongly depends on the starting condition. There is no guarantee that the global optimum can be achieved. Moreover, a bad initialization decelerates the convergence of *k*-means. A common strategy is to run *k*-means multiple times with different seeds and then to choose the result with the best score as the final output. However, this strategy is effectively not applicable here, because each iteration is computationally intensive.

Arthur and Vassilvitskii (2007) proposed an algorithm named *k*-means++ to improve the initialization step. The basic idea is to choose *k* cluster centers successively, each of which is

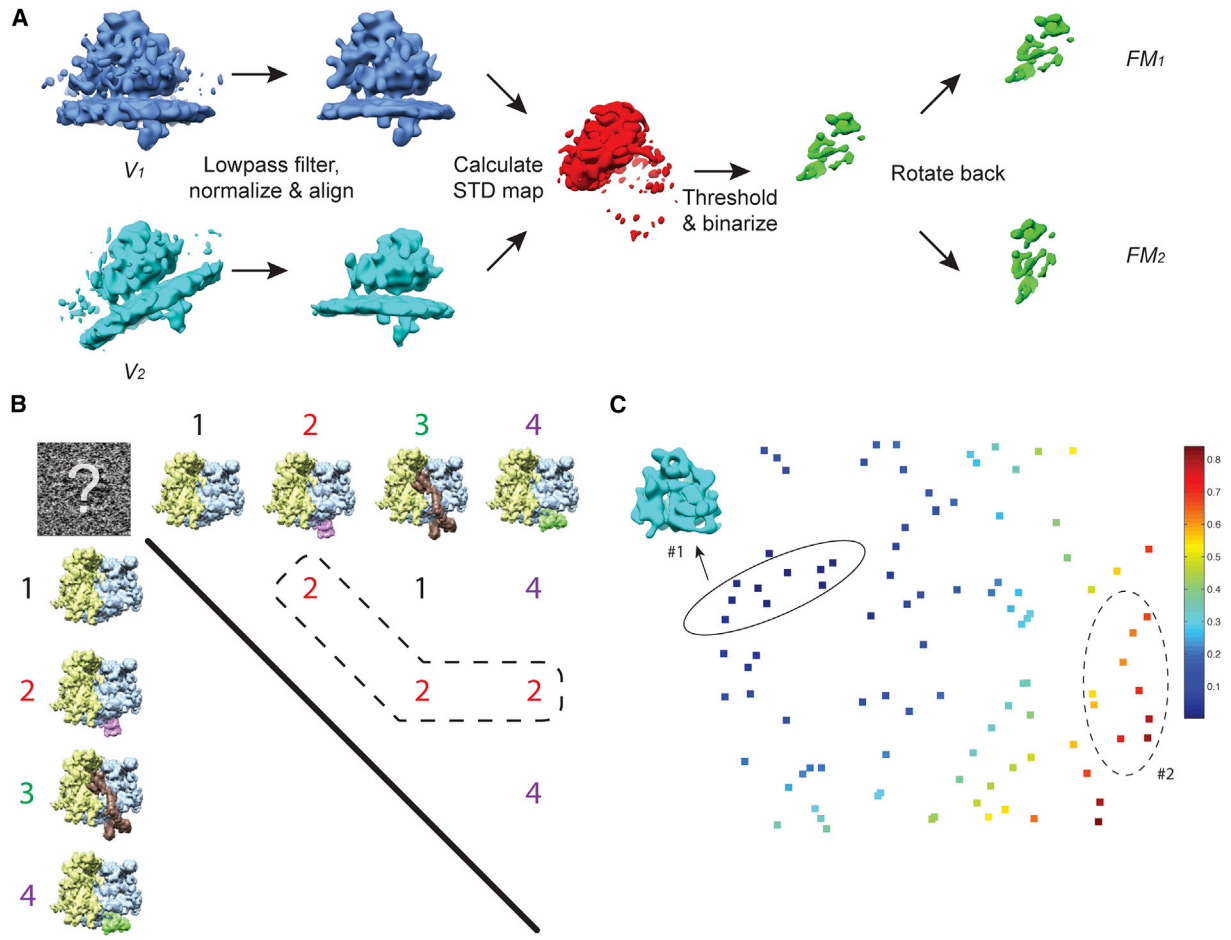

Figure 1. Main Methodological Features of AC3D

(A) Calculation of the focus masks FM_1 and FM_2 of two class centers V_1 and V_2.

(B) A voting strategy is used for multiclass label determination. The subtomogram under investigation with unknown class label (top left) will be assigned to the class with the most votes from pairwise comparisons, i.e., class 2 in this case.

(C) *k*-means++ is adapted as the initialization strategy. In the 2D simplification, each square represents a subtomogram. Assuming a subset of subtomograms (upper left, outlined with a solid line) is already chosen yielding the first class center (#1), the next class center (#2) is then the subtomogram average of a new subset (e.g., bottom right, outlined with a dashed line), in which each subtomogram is randomly picked with a probability proportional to the squared distance function (indicated by the colors of the squares and the scale bar).

randomly picked with a probability proportional to its squared distance from the closest existing center. It is shown that *k*-means++ converges faster than *k*-means with random initialization and guarantees that it is $O(\log k)$ competitive with the optimal clustering. In contrast, the performance of *k*-means with random initialization can be arbitrarily worse than the optimum (Kanungo et al., 2004).

Here, we implement *k*-means++ with a few important modifications for application to cryo-ET (Figure 1C): (1) the class center is not a single subtomogram but rather an average of a certain portion of the whole data set containing *N* subtomograms. The reason is that one single subtomogram has low SNR and is affected by the missing wedge. (2) The first class center is the average of the aligned subtomograms with top $\lfloor N/k \rfloor$ scores, which are obtained by the *CCC*s from the prealignment. This class is, hence, similar to the average of the whole data set. (3) The subsequent class centers are the averages of $\lfloor N/k \rfloor$ subtomograms from the whole data set. These subtomograms are chosen at random, with probabilities proportional to the squared distance functions. (4) The distance function, *D*, used here is the normalized Euclidean distance, which can be derived from the *CCC*. Mathematically, given a set of class centers $SV = \{V_1,\cdots,V_k\}$ and a subtomogram *S*, *D* can be calculated as:

$$D(S,SV) = \min_{V \in SV} \sqrt{2 - 2 \cdot CCC(V,S)}. \qquad \text{(Equation 1)}$$

The final initialization algorithm is presented in Algorithm 2. We emphasize that the computational cost of this step is marginal compared to the others in Algorithm 1, and the whole clustering procedure normally converges faster with this strategy (Box 2).

Fast Alignment of Subtomograms

The most time-consuming task in AC3D is the alignment of each subtomogram against the class centers. The computational time

Box 2. Algorithm 2: Initialization of AC3D

Input:
SS: A set of input aligned subtomograms
k: Number of desired classes

Output:
SV: A set of initial class centers

Begin
01 $n = \lfloor N/k \rfloor$
02 Sort *SS* according to the scores and average the top *n* subtomograms to get V_1
03 $SV = \{V_1\}$
04 for $i = 2{:}k$ **do**
05 $SS' = \{\}$
06 for $j = 1{:}n$ **do**
07 $\forall S \in SS$, calculate $P \propto D^2(S,SV)$
08 Pick $S_j \in SS$ without replacement at random with probability P_j
09 $SS' \leftarrow SS' \cup \{S_j\}$
10 end for
11 Average SS' to get V_i and $SV \leftarrow SV \cup \{V_i\}$
12 end for
end

grows linearly with the number of subtomograms times the number of classes. The speed of subtomogram alignment is the bottleneck of the entire procedure and thus limits its practical use. Recently, we proposed a fast alignment algorithm based on spherical harmonics (Chen et al., 2013), which can be applied here to address this issue. Here, we briefly recapitulate this algorithm.

The fast subtomogram alignment consists of two major components: fast translational matching (FTM) and FRM, which are then combined into an integrated framework using expectation maximization, i.e., the original 6 DoF problem is divided into two 3 DoF problems (translation and rotation) and solved by FTM and FRM iteratively. FTM is well known: the two volumes to be aligned are first constrained to common areas in Fourier space (Frangakis et al., 2002), and their cross-correlation function can be efficiently computed using fast Fourier transform (Roseman, 2003).

However, FRM for cryo-ET is not trivial, and we proposed to solve it using spherical harmonics analysis in Fourier space (Chen et al., 2013). Mathematically, FRM evaluates the *CCC* as a function of rotation $\boldsymbol{R}$ of two 3D volumes, V_1 and V_2, efficiently. We define $\widehat{V_1}$ and $\widehat{V_2}$ as the Fourier transforms of V_1 and V_2, and two spherical mask functions m_1 and m_2, indicating their respective missing wedges in Fourier space. We first convert the Fourier transforms of volumes to spherical coordinates: $\widehat{V}(k_x,k_y,k_z) = \widehat{V}(k,\ \theta,\ \phi)$. Then, *CCC* can be calculated as follows (Chen et al., 2013):

$$CCC(\boldsymbol{R}) = \frac{\sum_{k=1}^{k_{max}} SCC_{12}(\boldsymbol{R};k)\cdot k^2}{\sqrt{\sum_{k=1}^{k_{max}} SCC_{11}(\boldsymbol{R};k)\cdot k^2}\cdot\sqrt{\sum_{k=1}^{k_{max}} SCC_{22}(\boldsymbol{R};k)\cdot k^2}},$$

$$SCC_{12}(\boldsymbol{R};k) = \widehat{V_1}(k,\theta,\ \phi)m_1 \star \widehat{V_2}(k,\theta,\ \phi)m_2,$$

$$SCC_{11}(\boldsymbol{R};k) = \left|\widehat{V_1}(k,\theta,\ \phi)\right|^2 m_1 \star m_2, \text{ and}$$

$$SCC_{22}(\boldsymbol{R};k) = m_1 \star \left|\widehat{V_2}(k,\theta,\ \phi)\right|^2 m_2. \qquad \text{(Equation 2)}$$

Here, k_{max} is the maximal frequency band involved, and $\star$ is the spherical correlation operator, which can be efficiently computed by the SO(3) Fourier transform (SOFT) (Kostelec, 2008) and spherical Fourier transform (SFT) (Healy et al., 2003). The calculation of SOFT and SFT involves spherical harmonics functions. This is the generalized convolution theorem of spherical functions. The peak of *CCC* then indicates the best scoring rotation.

Noise Class Handling

The subtomograms under investigation often include outliers, typically false positives from the automated or manual detection or subtomograms that are too noisy to be aligned accurately. These outliers tend to degrade the clustering performance. To ensure the robustness of the classification with respect to such outliers, we assign a certain percentage of all the subtomograms to a "noise class" during each iteration. This step is conducted before the class label determination step. If a subtomogram is assigned to the noise class, it will be excluded from the remaining steps of that iteration. Importantly, the subtomogram will be included again in the subsequent iterations and may be assigned to a different class.

To decide which subtomogram belongs to the noise class, we calculate the probabilities using the score distributions. Given a set of subtomograms $SS = \{S_1,\cdots,S_N\}$ and a set of class centers $SV = \{V_1,\cdots,V_k\}$, we first align *SS* to *SV* using the fast subtomogram alignment algorithm. For each $V_j \in SV$, we will have a set of similarity scores $SCS_j = \{SC_1,\cdots,SC_N\}$. Assuming that the noise class has a low score and it is statistically independent of all class centers *SV*, the probability of a subtomogram $S_i \in SS$ not being class V_j is then $P_{i,j}\{SC_i < SC\},\ \forall SC \in SCS_j$. Finally, the overall probability of S_i being noise can be calculated by $P_i = \prod_{j=1}^{k} P_{i,j}$. Sorting the probabilities and setting a threshold of the list will then yield the noise class.

Focus Mask and Focused Score

Another critical step is the automatic calculation of the focus mask, *FM*, and the corresponding focused score, *FS*. Given two volumes (class centers) V_1 and V_2, we calculate *FM* as follows (Figure 1A). (1) Low-pass filter V_1 and V_2 (according to the corresponding resolution) to reduce noise influence and normalize them (mean = 0, standard deviation [SD] = 1) so that they have approximately the same intensity scale. (2) Align V_1 and V_2 to make sure they have the highest correlation. (3) Multiply the aligned V_1 and V_2 with a mask *M*, if provided, to enforce *FM* to be computed inside *M*. Note that this step is optional and that *M* is only used for explicitly constraining the classification, e.g., filtering out hypervariable areas. (4) Calculate the SD map STD of the aligned V_1 and V_2. In this case, the STD of two volumes is essentially the same as their absolute difference map. (5) Threshold STD (e.g., top 10% of the intensity values) and binarize it by setting the areas above

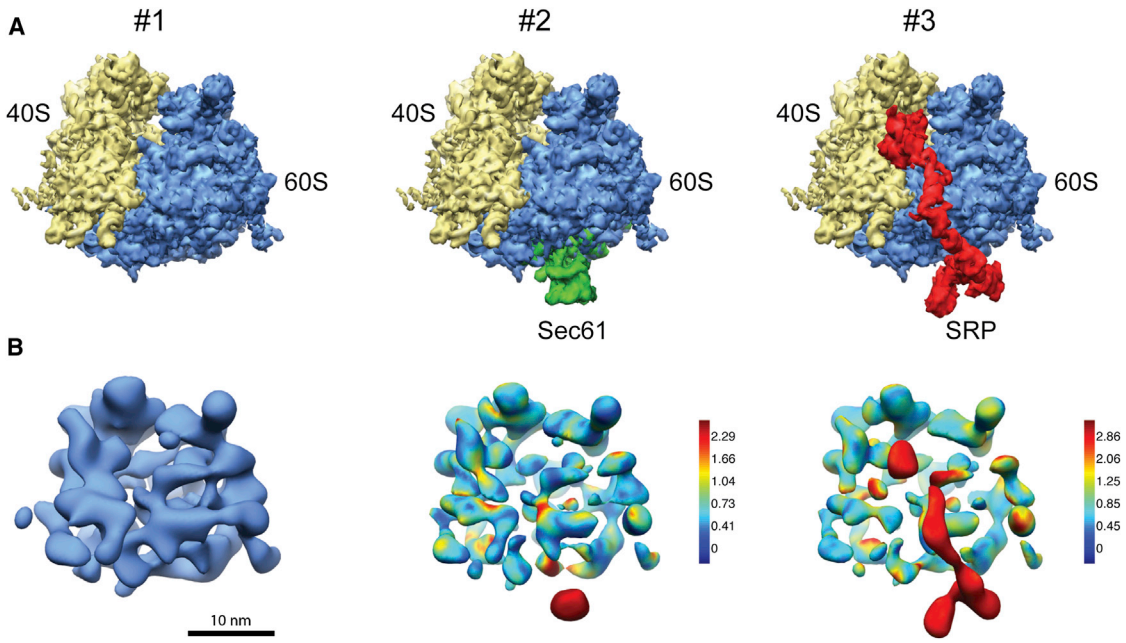

Figure 2. Classification Result on Simulated Data Set of 80S Ribosomes

(A) Densities simulated from atomic models of three ribosome states. From left to right: *S. cerevisiae* 80S ribosome (model #1), 80S ribosome bound to the Sec61 translocon (model #2), and 80S ribosome bound to the SRP (model #3).

(B) Classification result. From left to right: subtomogram average of class #1, subtomogram average of class #2 colored by the STD against #1, and subtomogram average of class #3 colored by the STD against #1.

the threshold to 1 and those below it to 0, resulting in *FM*. (6) Transform *FM* back to the respective orientations and positions of V_1 and V_2, which results in a pair FM_1 and FM_2. Note that, for each pair (V_1, V_2), their focus masks are also a pair (FM_1, FM_2).

Finally, $FS_{j,i}$ of a subtomogram S_i and V_j can be obtained by first aligning S_i to V_j and then calculating the local *CCC* (Förster et al., 2008):

$$FS_{j,i} = \sum_{x,y,z} S_i'(x,y,z) \cdot V_j'(x,y,z),$$

$$S_i' = \frac{FM_j(x,y,z) \cdot \left(FT^{-1}\left(\widehat{S}_i \cdot \omega \right) - \overline{S_i'} \right)}{\sqrt{\sum_{x,y,z} \left(FM_j(x,y,z) \cdot \left(FT^{-1}\left(\widehat{S}_i \cdot \omega \right) - \overline{S_i'} \right) \right)^2}},$$

$$\overline{S_i'} = \frac{1}{\sum_{x,y,z} FM_j(x,y,z)} \sum_{x,y,z} FT^{-1}\left(\widehat{S}_i \cdot \omega \right). \qquad \text{(Equation 3)}$$

Herein, $\widehat{S}_i$ is the Fourier transform of S_i, FT^{-1} is the inverse Fourier transform, and ω is the corresponding sampling region in Fourier space. V_j' can be computed analogously. Note that if FM_j is a unit volume, $FS_{j,i}$ is identical to $CCC_{j,i}$.

Multiclass Label Determination

Binary class label determination is straightforward. Given a subtomogram S_i and two class centers (V_1, V_2), we first calculate $FM_{1,i}$ and $FM_{2,i}$ and their corresponding $FS_{1,i}$ and $FS_{2,i}$. The class label of S_i will then correspond to the class average, with the larger value between $FS_{1,i}$ and $FS_{2,i}$.

Multiclass label determination, i.e., class assignment with more than two classes, is not trivial because *FM* is defined pairwise. Focus masks that incorporate the structural discrepancies of more than two volumes are less discriminative than those pinpointing pairwise differences, because the focus mask of multiple volumes will involve more voxels than any pairwise *FM*. In order to use the pairwise *FM* for classification, we use a voting strategy for the multiclass label assignment (Figure 1B). *FS* is defined with respect to a pair of class centers for each subtomogram. *FS* can be considered as a binary classifier, which generates a vote to one of the classes from the pair analyzed. For each comparison of a subtomogram S_i with a pair of class centers (V_k,V_l), the binary class label is determined according to the vote. The final class label of S_i is determined by a voting of all the pairwise comparisons.

Classification of Simulated Ribosome Subtomograms

We first assessed our algorithm on a simulated data set of *Saccharomyces cerevisiae* 80S ribosomes bound to different cofactors involved in signal-recognition-particle (SRP)-mediated protein translocation into the endoplasmic reticulum (ER) (Figure 2A): the 80S ribosome alone, the 80S ribosome bound to the Sec61 translocon, and the 80S ribosome bound to the SRP. For convenience, we name the 80S ribosome as class #1, the 80S ribosome bound to the Sec61 channel as class #2, the 80S ribosome bound to the SRP as class #3, and noise particles as class #0. For comparison, this data set was also classified into four classes using CPCA in combination with *k*-means clustering

Table 1. Results of Compared Classification Approaches for Simulated Ribosome Data Set

CPCA

		Predicted			
		#0	#1	#2	#3
Actual	#0	100	0	0	0
	#1	15	76	59	0
	#2	8	56	36	0
	#3	7	0	0	43

	% TPR	% FPR
#0	100	10
#1	50.7	22.4
#2	36	19.7
#3	86	0

MLTOMO

		Predicted			
		#0	#1	#2	#3
Actual	#0	100	0	0	0
	#1	2	106	42	0
	#2	1	68	31	0
	#3	0	28	15	7

	% TPR	% FPR
#0	100	1
#1	70.7	38.4
#2	31	19
#3	14	0

AC3D

		Predicted			
		#0	#1	#2	#3
Actual	#0	93	7	0	0
	#1	4	125	21	0
	#2	2	0	98	0
	#3	1	0	0	49

	% TPR	% FPR
#0	93	2.3
#1	83.3	2.8
#2	98	7
#3	98	0

Classes #1–#3 are shown in Figure 2, and class #0 corresponds to the noise class. From the class assignments, the TPR and FPR were computed.

(Förster et al., 2008) and the ML approach MLTOMO implemented in Xmipp (Scheres et al., 2009). For CPCA, five eigenvectors were retained for *k*-means; and for MLTOMO, 20 iterations were executed with reg0 = 5, regF = 0, and reg_steps = 5.

The confusion matrices are shown in Table 1, in which also the true positive rates (TPR) and false positive rates (FPR) are listed. Table 1 indicates a significantly better performance of AC3D compared to CPCA and MLTOMO in terms of both TPR and FPR. Moreover, the classification results of AC3D (class centers) are shown in Figure 2B, in which the 3D densities are colored by the STD map (prior to threshold) to illustrate the autofocus ability of AC3D.

To demonstrate the benefits of two key components of our approach, i.e., the advanced initialization (*k*-means++) and the focused score, we evaluated the classification results of AC3D with each of these two features turned off (Table 2). When the random class assignment was used in the initialization step, the obtained accuracies were essentially identical in this case, but the convergence was slower (two more iterations) compared to AC3D with *k*-means++. Thus, *k*-means++ increases the classification speed. When the conventional *CCC* was used as the similarity metric in AC3D, the classification accuracy degraded dramatically. Thus, the superior classification performance of AC3D compared to CPCA and MLTOMO can be almost exclusively attributed to the focused score.

Classification of ER-Associated Ribosomes

We further tested AC3D on an experimental data set of mammalian ribosomes bound to the ER protein translocon. In previous studies of the same sample, we could resolve the membrane-bound 80S ribosome and two complexes with prominent lumenal domains: the translocon-associated protein complex (TRAP) and the oligosaccharyl-transferase complex (OST) (Pfeffer et al., 2012, 2014). The acquired subtomograms depict ribosomes bound to ER-derived microsomes. Because of the highly variable diameters of the microsomes, the curvature of the membrane would dominate the classification; to prevent classification according to membrane curvature, we constrained the classification on the ribosome and the ER lumenal region.

The whole data set was first classified into four classes, and the resulting four classes are depicted in Figure 3A: class #1 clearly captures 80S ribosomes bound to a translocon population with only TRAP; class #2 80S ribosomes bound to a translocon population with TRAP and OST; class #3 60S large ribosomal subunits with only TRAP; and class #4 60S ribosomal subunits associated with TRAP- and OST-containing translocons. The populations of the four classes are 564 (21.8%), 970 (37.5%), 737 (28.5%), and 313 (12.1%) particles, respectively.

We compared the obtained subtomogram assignments with our results in (Pfeffer et al., 2014), where the foci for classification were chosen based on biological prior knowledge. In detail, we conducted CPCA classification on the same data set, first with a sphere mask focusing on the entire ribosome and then with another sphere mask covering the ER-lumenal region. The resulting class averages of the knowledge-based approach are essentially the same as those derived by AC3D (Figure 3A). The confusion matrix of the classification results from CPCA and AC3D is shown in Table 3. Both measures indicate good agreement between knowledge-based CPCA and AC3D.

Moreover, we conducted a further classification round of the particles included in classes #1 and #2, focusing on the 80S ribosome part only. The number of classes was set to three, and we obtained the class averages shown in Figure 3B. Consistent with previous studies using cryoelectron microscopy single-particle analysis (Frank and Gonzalez, 2010; Melnikov et al., 2012; Wilson and Doudna Cate, 2012), we observe a highly flexible ribosomal L1 stalk (Figure 3B, right panel). Furthermore, we find a nonribosomal density of approximately 100 kDa bound to the ribosomal stalk base in classes C1 and C2, but not C3 (Figure 3B), which likely corresponds to canonical translation elongation or termination factors. The number of subtomograms assigned to class C1 was 637 (41.5%); class C2, 507 (33%); and class C3, 390 (25.4%). The classification result is furthermore quantitatively assessed by the Fourier shell correlation (FSC) curves. Three types of FSC curves were calculated for each class: intraclass FSC, interclass FSC, and FSC of a random, same-sized portion of subtomograms (Figure 3C), from which we can see that the intraclass FSCs are generally better or similar than the random FSCs. Since the FSC measures the global similarity, which is dominated by the structurally invariant core ribosome, the superiority of intraclass FSCs is more obvious when compared to interclass FSCs, which indicate the level of

Table 2. Influence of AC3D's Initialization and Focused Score on Classification Accuracy

AC3D with Random Initialization

Actual \ Predicted	#0	#1	#2	#3
#0	91	6	2	1
#1	6	123	21	0
#2	3	0	97	0
#3	0	0	0	50

	% TPR	% FPR
#0	91	3
#1	82	2.4
#2	97	7.7
#3	100	0.3

Convergence: eight iterations

AC3D without Focused Score

Actual \ Predicted	#0	#1	#2	#3
#0	93	1	0	6
#1	2	74	54	20
#2	3	49	37	11
#3	2	8	33	7

	% TPR	% FPR
#0	93	2.3
#1	49.3	23.2
#2	37	29
#3	14	10.6

Convergence: seven iterations

AC3D

Actual \ Predicted	#0	#1	#2	#3
#0	93	7	0	0
#1	4	125	21	0
#2	2	0	98	0
#3	1	0	0	49

	% TPR	% FPR
#0	93	2.3
#1	83.3	2.8
#2	98	7
#3	98	0

Convergence: six iterations

Classification of the simulated ribosome data set was performed by AC3D with random initialization and with a uniform *FM* for comparison with the AC3D enabling all the features.

similarity between the different classes. Taken together, these classification results suggest that AC3D is capable of separating different conformations of ER-associated ribosomes, which all agree with previous studies relying on much larger data sets.

DISCUSSION

Here, we presented a multireference clustering algorithm (AC3D) for subtomogram classification and simultaneous alignment. For large data sets, AC3D, like other multireference approach, is computationally more efficient than clustering approaches requiring pairwise correlations of all subtomograms, such as PCA-based approaches (Bartesaghi et al., 2008; Förster et al., 2008). The main distinguishing feature of AC3D among multireference approaches is the ability to automatically focus the similarity measurement to regions of significant structural discrepancies. This autofocus ability does not require any prior knowledge or human intervention, which avoids hypothesis-driven bias of classification results. Moreover, we adapted *k*-means++ for the initialization of the iterative clustering algorithm, which improves the convergence speed and makes the procedure less vulnerable to local optima. Last, but not least, the integration of a fast, spherical harmonics-based subtomogram alignment algorithm makes AC3D computationally highly efficient compared to other state-of-the-art approaches without compromising on accuracy.

A problem that AC3D shares with essentially all multireference classification approaches is that the user must specify the number of classes, *k*, which is not straightforward. A common guideline is to oversample *k* properly, because it is safer for the small classes to be discovered and the clustering result will become more stable. In a subsequent step, the classes can be either manually examined and aggregated or automatically merged using hierarchical clustering of the class averages (Hrabe et al., 2012).

AC3D is open-source software and is available to the whole community at http://www.biochem.mpg.de/foerster. Some features of AC3D can also be incorporated into other approaches. For example, the efficient subtomogram alignment algorithm can be integrated into the ML approach, which will make the processing of large cryo-ET data sets feasible. Whereas AC3D performs excellently for figuring out whether cofactors are present or absent in complexes as shown here, further studies need to be conducted to find out which approach is a better choice when structural variations are not confined to relatively small areas of subtomograms.

When evaluated on a realistic simulated data set of 80S ribosomes bound to different cofactors, AC3D achieved a nearly perfect classification of the different states, while two other tested state-of-the-art classification approaches, CPCA and MLTOMO, yielded significantly less accurate class assignments. The data set was designed so that it encapsulated three challenges of subtomogram classification. (1) Particularly between two classes, the bare 80S ribosome and the 80S ribosome bound to the Sec61 channel, the structural difference arose from only an ~60 kDa density, indicating that AC3D can identify highly subtle structural heterogeneity in low SNR data. (2) The populations of different classes were unbalanced by a factor of up to three. (3) A considerable amount of outliers was present. It is highly encouraging that AC3D yielded a near-perfect classification result under these challenging conditions, which often occur in experimental data from physiological samples.

We then applied AC3D to an experimental data set of ER-associated ribosomes. For the ER-lumenal part of the complex, we retrieved essentially the same classes that we previously obtained using biological knowledge-based classifications (Pfeffer et al., 2014): the OST complex was present in the translocon holocomplex in substoichiometric amounts. The most prominent classes for the cytosolic ribosomal density were assembled 80S ribosomes and 60S ribosomal subunits. Thus, the smallest structural difference detected in the initial classification was the presence or absence of the ~250 kDa lumenal OST density. The significant enrichment of OST in translocon complexes bound to fully assembled 80S ribosomes (62.8% occupancy) compared to 60S ribosomal subunits (29.8% occupancy) suggests that OST have a higher affinity to translocon complexes engaged in cotranslational translocation of a nascent peptide across the ER membrane. This affinity variation would imply

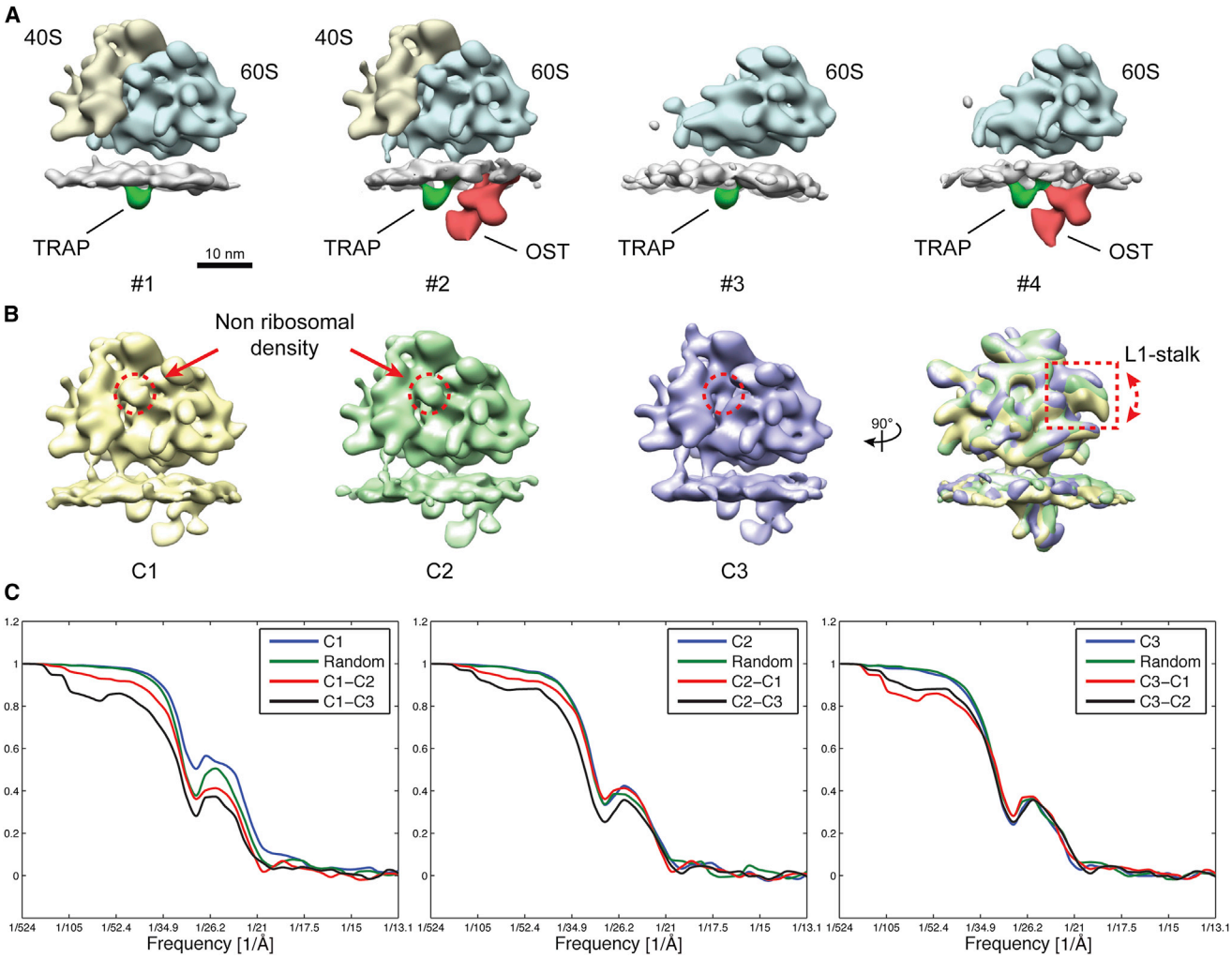

Figure 3. Classification Result for Mammalian Ribosomes Bound to the Native ER Protein Translocon

(A) The whole data set was first classified into 4 classes that apparently corresponded to the following assemblies: 80S ribosomes bound to a translocon population with only TRAP (class #1), 80S ribosomes bound to a translocon population with TRAP and OST (class #2), 60S ribosomes with only TRAP (class #3), and 60S ribosomes with TRAP and OST (class #4).

(B) Classes #1 and #2 were merged and further classified into three classes (C1, C2, and C3) with the focus on the ribosome density. The dotted circles mark the presence/absence of a nonribosomal density bound to the ribosomal stalk base, which likely corresponds to canonical translation elongation or termination factors. The three class averages are overlaid on the rightmost side to show the high flexibility of the ribosomal L1 stalk (outlined with a dotted rectangle).

(C) The FSC curves of the class averages in (B). For each class, three types of FSC curves are plotted: the intraclass FSC, the interclass FSC, and the FSC of a random portion with the same number of subtomograms.

that the ER protein translocon is not a temporally invariant complex but rather undergoes compositional dynamics according to the translational state of the associated ribosome.

More subtle structural differences were detected when we classified 80S ribosomal densities, revealing well-established flexibility of the L1 stalk and cofactor binding to the ribosomal stalk base. The approximate mass of 100 kDa of the cofactor would be consistent for example with the 95 kDa eukaryotic elongation factor 2. Previously, different conformational states of the ribosome during translation could only be observed in cryo-electron microscopy single-particle data of purified ribosome particles. The classification results presented here for ribosomes in their native membrane suggest that cryo-ET in conjunction with subtomogram classification by AC3D will become a powerful method to study the mechanics of large macromolecular machines in their physiological environment.

EXPERIMENTAL PROCEDURES

Simulation of Ribosome Subtomograms

Three different states of ribosomes were simulated using atomic models from the Protein Data Bank (PDB) (Figure 2A): the *Saccharomyces cerevisiae* 80S ribosome (IDs: 3IZB, 3IZE, 3IZF, and 3IZS), the *S. cerevisiae* 80S ribosome bound to the Sec61 translocon (ID: 2WWB), and the *S. cerevisiae* 80S ribosome bound to the SRP (ID: 1RY1). The simulations were conducted as described by Chen et al. (2013) for SNR = 0.01. For testing the performance on an unbalanced data set, the number of particles for each class was 150, 100, and 50, respectively. Furthermore, 100 noise particles were added into the data set to test the robustness. They were spheres with diameters ranging

Table 3. Confusion Matrix of Classification Results from Knowledge-Based CPCA and Unbiased AC3D on ER-Associated Ribosomes

	AC3D			
CPCA	#1	#2	#3	#4
#1	299	381	250	68
#2	253	575	21	14
#3	9	8	327	49
#4	3	6	139	182

from 15 to 30 nm. They had similar mean values as the 80S ribosome and the same SNR. In total, 400 subtomograms of 100^3 voxels were simulated with a defocus of 4 μm and a voxel size of 0.47 nm. The tilt angles ranged from −60° to 60°, with 3° as the angular increment. The tomograms were randomly translated with respect to the center within the range of 10 voxels and randomly rotated.

Experimental Data Set of ER-Associated Ribosomes

Rough microsomes were prepared from dog pancreas and vitrified on lacey carbon molybdenum electron microscopy grids (Ted Pella) as described by Pfeffer et al. (2012). Tilt series were acquired using an FEI Titan Krios TEM equipped with a Gatan K2 Summit direct electron detector, operated in frame mode with five to seven frames per projection image. The TEM was operated at an acceleration voltage of 300 kV. Single-axis tilt series were recorded from −60° to 60°, with an angular increment of 2° at a nominal defocus of 4 μm and an object pixel size of 2.62 Å using the Serial EM acquisition software (Mastronarde, 2005). The cumulative electron dose did not exceed 60 electrons per square angstrom.

Frames from the Gatan K2 Summit direct electron detector were aligned using quasi-expectation maximization implemented in the MATLAB toolbox AV3 (Förster et al., 2005). Phase correction of single projections was performed using the MATLAB scripts described by Eibauer et al. (2012) rather than the slightly more accurate but computationally more demanding Wiener filtering (Chen et al., 2013). Tomogram reconstruction (object pixel: 2.1 nm) and template matching were accomplished using PyTom (Hrabe et al., 2012) as described by Pfeffer et al. (2012), followed by extraction of ribosome candidates. A preliminary classification (Förster et al., 2008) was carried out to remove most of the false-positives, e.g., gold markers, ER membranes, or carbon edges. Finally, 2,584 subtomograms (200^3 voxels, object pixel: 0.262 nm) were retained and reconstructed for further processing.

ACKNOWLEDGMENTS

Canine pancreatic microsomes were a kind gift from the Zimmermann lab (Saarland University, Saarbrucken, Germany). This work was supported by funding from the Deutsche Forschungsgemeinschaft (FO 716/3-1).

Received: April 28, 2014
Revised: July 3, 2014
Accepted: August 8, 2014
Published: September 18, 2014

REFERENCES

Arthur, D., and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, H. Gabow, ed. (Philadelphia, PA: Society for Industrial and Applied Mathematics), pp. 1027–1035.

Bartesaghi, A., and Subramaniam, S. (2009). Membrane protein structure determination using cryo-electron tomography and 3D image averaging. Curr. Opin. Struct. Biol. *19*, 402–407.

Bartesaghi, A., Sprechmann, P., Liu, J., Randall, G., Sapiro, G., and Subramaniam, S. (2008). Classification and 3D averaging with missing wedge correction in biological electron tomography. J. Struct. Biol. *162*, 436–450.

Briggs, J.A.G. (2013). Structural biology in situ—the potential of subtomogram averaging. Curr. Opin. Struct. Biol. *23*, 261–267.

Briggs, J.A.G., Riches, J.D., Glass, B., Bartonova, V., Zanetti, G., and Kräusslich, H.-G. (2009). Structure and assembly of immature HIV. Proc. Natl. Acad. Sci. USA *106*, 11090–11095.

Castaño-Díez, D., Kudryashev, M., Arheit, M., and Stahlberg, H. (2012). Dynamo: a flexible, user-friendly development tool for subtomogram averaging of cryo-EM data in high-performance computing environments. J. Struct. Biol. *178*, 139–151.

Chen, Y., Pfeffer, S., Hrabe, T., Schuller, J.M., and Förster, F. (2013). Fast and accurate reference-free alignment of subtomograms. J. Struct. Biol. *182*, 235–245.

Eibauer, M., Hoffmann, C., Plitzko, J.M., Baumeister, W., Nickell, S., and Engelhardt, H. (2012). Unraveling the structure of membrane proteins in situ by transfer function corrected cryo-electron tomography. J. Struct. Biol. *180*, 488–496.

Förster, F., and Hegerl, R. (2007). Structure determination in situ by averaging of tomograms. Methods Cell Biol. *79*, 741–767.

Förster, F., Medalia, O., Zauberman, N., Baumeister, W., and Fass, D. (2005). Retrovirus envelope protein complex structure in situ studied by cryo-electron tomography. Proc. Natl. Acad. Sci. USA *102*, 4729–4734.

Förster, F., Pruggnaller, S., Seybert, A., and Frangakis, A.S. (2008). Classification of cryo-electron sub-tomograms using constrained correlation. J. Struct. Biol. *161*, 276–286.

Förster, F., Han, B.-G., and Beck, M. (2010). Visual proteomics. Methods Enzymol. *483*, 215–243.

Frangakis, A.S., Böhm, J., Förster, F., Nickell, S., Nicastro, D., Typke, D., Hegerl, R., and Baumeister, W. (2002). Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. Proc. Natl. Acad. Sci. USA *99*, 14153–14158.

Frank, J., and Gonzalez, R.L., Jr. (2010). Structure and dynamics of a processive Brownian motor: the translating ribosome. Annu. Rev. Biochem. *79*, 381–412.

Grünewald, K., Desai, P., Winkler, D.C., Heymann, J.B., Belnap, D.M., Baumeister, W., and Steven, A.C. (2003). Three-dimensional structure of herpes simplex virus from cryo-electron tomography. Science *302*, 1396–1398.

Healy, D.M., Rockmore, D.N., Kostelec, P.J., and Moore, S. (2003). FFTs for the 2-sphere—improvements and Variations. J. Fourier Anal. Appl. *9*, 341–385.

Heumann, J.M., Hoenger, A., and Mastronarde, D.N. (2011). Clustering and variance maps for cryo-electron tomography using wedge-masked differences. J. Struct. Biol. *175*, 288–299.

Hrabe, T., Chen, Y., Pfeffer, S., Cuellar, L.K., Mangold, A.-V., and Förster, F. (2012). PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. J. Struct. Biol. *178*, 177–188.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., and Wu, A.Y. (2004). A local search approximation algorithm for k-means clustering. Comput. Geom. *28*, 89–112.

Kostelec, P.J. (2008). FFTs on the rotation group. J. Fourier Anal. Appl. *14*, 145–179.

Kuybeda, O., Frank, G.A., Bartesaghi, A., Borgnia, M., Subramaniam, S., and Sapiro, G. (2013). A collaborative framework for 3D alignment and classification of heterogeneous subvolumes in cryo-electron tomography. J. Struct. Biol. *181*, 116–127.

Lučić, V., Förster, F., and Baumeister, W. (2005). Structural studies by electron tomography: from cells to molecules. Annu. Rev. Biochem. *74*, 833–865.

Mastronarde, D.N. (2005). Automated electron microscope tomography using robust prediction of specimen movements. J. Struct. Biol. *152*, 36–51.

Melnikov, S., Ben-Shem, A., Garreau de Loubresse, N., Jenner, L., Yusupova, G., and Yusupov, M. (2012). One core, two shells: bacterial and eukaryotic ribosomes. Nat. Struct. Mol. Biol. *19*, 560–567.

Pfeffer, S., Brandt, F., Hrabe, T., Lang, S., Eibauer, M., Zimmermann, R., and Förster, F. (2012). Structure and 3D arrangement of endoplasmic reticulum membrane-associated ribosomes. Structure *20*, 1508–1518.

Pfeffer, S., Dudek, J., Gogala, M., Schorr, S., Linxweiler, J., Lang, S., Becker, T., Beckmann, R., Zimmermann, R., and Förster, F. (2014). Structure of the mammalian oligosaccharyl-transferase complex in the native ER protein translocon. Nat. Commun. *5*, 3072.

Roseman, A.M. (2003). Particle finding in electron micrographs using a fast local correlation algorithm. Ultramicroscopy *94*, 225–236.

Scheres, S.H.W., Melero, R., Valle, M., and Carazo, J.-M. (2009). Averaging of electron subtomograms and random conical tilt reconstructions through likelihood optimization. Structure *17*, 1563–1572.

Schur, F.K.M., Hagen, W.J.H., de Marco, A., and Briggs, J.A.G. (2013). Determination of protein structure at 8.5Å resolution using cryo-electron tomography and sub-tomogram averaging. J. Struct. Biol. *184*, 394–400.

Stölken, M., Beck, F., Haller, T., Hegerl, R., Gutsche, I., Carazo, J.-M., Baumeister, W., Scheres, S.H.W., and Nickell, S. (2011). Maximum likelihood based classification of electron tomographic data. J. Struct. Biol. *173*, 77–85.

Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., and Ludtke, S.J. (2007). EMAN2: an extensible image processing suite for electron microscopy. J. Struct. Biol. *157*, 38–46.

Wilson, D.N., and Doudna Cate, J.H. (2012). The structure and function of the eukaryotic ribosome. Cold Spring Harb. Perspect. Biol. *4*, 4.

Winkler, H. (2007). 3D reconstruction and processing of volumetric data in cryo-electron tomography. J. Struct. Biol. *157*, 126–137.

Xu, M., Beck, M., and Alber, F. (2012). High-throughput subtomogram alignment and classification by Fourier space constrained fast volumetric matching. J. Struct. Biol. *178*, 152–164.

Yu, L., Snapp, R.R., Ruiz, T., and Radermacher, M. (2010). Probabilistic principal component analysis with expectation maximization (PPCA-EM) facilitates volume classification and estimates the missing data. J. Struct. Biol. *171*, 18–30.