Journal of Structural Biology 187 (2014) 10–19

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

M-free: Scoring the reference bias in sub-tomogram averaging and template matching

Zhou Yu, Achilleas S. Frangakis *

Goethe University Frankfurt, Buchmann Institute for Molecular Life Sciences and Institute for Biophysics, Max-von-Laue Str.15, 60438 Frankfurt am Main, Germany

ARTICLE INFO

Article history:
Received 19 March 2014
Received in revised form 23 April 2014
Accepted 15 May 2014
Available online 22 May 2014

Keywords:
Template matching
Sub-tomogram averaging
Assessment of reference bias
Reference bias
Template bias

ABSTRACT

Cryo-electron tomography provides a snapshot of the cellular proteome. With template matching, the spatial positions of various macromolecular complexes within their native cellular context can be detected. However, the growing awareness of the reference bias introduced by the cross-correlation based approaches, and more importantly the lack of a reliable confidence measurement in the selection of these macromolecular complexes, has restricted the use of these applications. Here we propose a heuristic, in which the reference bias is measured in real space in an analogous way to the R-free value in X-ray crystallography. We measure the reference bias within the mask used to outline the area of the template, and do not modify the template itself. The heuristic works by splitting the mask into a *working* and a *testing area* in a volume ratio of 9:1. While the *working area* is used during the calculation of the cross-correlation function, the information from both areas is explored to calculate the M-free score. We show using artificial data, that the M-free score gives a reliable measure for the reference bias. The heuristic can be applied in template matching and in sub-tomogram averaging. We further test the applicability of the heuristic in tomograms of purified macromolecules, and tomograms of whole Mycoplasma cells.

1. Introduction

Cryo-electron tomography is currently the only imaging technique that can visualize large macromolecular assemblies in an unperturbed cellular environment. The tomograms are three-dimensional (3D) images with a resolution in the range of a few nanometers. In contrast to fluorescence-based microscopy that displays only a few fluorescently labeled macromolecules together, electron tomograms visualize the entire proteome of a cell and contain the information of spatial relationships of the involved macromolecules (Nickell et al., 2006). The most promising way of extracting this information from the tomograms is by utilizing template-based approaches, which use known macromolecular structures and scan the tomogram for their spatial positions (Baumeister, 2005; Frangakis et al., 2002; Lucic et al., 2013). In particular, with the current availability of direct-detector cameras and exponentially growing computational resources, template-based approaches can become more applicable. The primary prerequisite for applying template-based approaches is a figure of merit that assesses the quality of the search, limits the reference bias, and reliably estimates the amount of false positives.

There is a growing awareness that cross-correlation based approaches can introduce a significant amount of reference bias that leads to uncertain or wrong interpretations. This is because high cross-correlation values may be assigned to features that are not the sought macromolecular structure or that the search has delivered the correct macromolecule but in the incorrect orientation. Thus, confidence that the highest cross-correlation values truly represent the properly aligned macromolecular complex is low. When the signal is sufficiently strong, correlation-based sub-tomogram averaging of individual molecules has provided a multitude of high-resolution structures, sometimes with a resolution of ~1 nm (Schur et al., 2013). While sub-tomogram averaging is less prone to reference bias than template matching, the bias still plays a significant role. Approaches such as the gold-standard method conceptually address the bias introduced by overfitting in high frequencies (Penczek, 2002; Scheres and Chen, 2012). However, in the case of low-resolution structures such as those in electron tomography, if the bias is already introduced at low frequencies, the outcome is already compromised.

Previous studies have shown that reference bias is also present in single particle approaches, and they have addressed ways and means to minimize it (Scheres and Chen, 2012; Shaikh et al., 2003). Fortunately, in single particle analysis there are numerous benefits that may act favorably towards a reduced reference bias

* Corresponding author.
E-mail address: achilleas.frangakis@biophysik.org (A.S. Frangakis).

http://dx.doi.org/10.1016/j.jsb.2014.05.007
1047-8477/

in comparison to tomography. Firstly, the higher SNR and the higher resolution achieved by this method reduce the possibility of errors. Secondly, the reduced number of degrees of freedom (DOF) during the alignment search is advantageous. Thirdly, since the particles analyzed are generally purified, the heterogeneity is less compared to sub-tomogram averaging and template matching, which are usually applied to entire cells. Thus, in combination with well-developed classification techniques such as RELION (Scheres, 2012a,b) and FREALIGN (Lyumkis et al., 2013), the number of false positives and wrongly oriented particles is much less or can be reduced during the alignment process.

Reference bias was previously addressed in single particle analysis by utilizing the Fourier Shell Correlation (FSC) (Shaikh et al., 2003). However, the FSC and subsequent approaches such as the gold standard method (Penczek, 2002; Scheres and Chen, 2012), measure the self-consistency of two halves of the data set and do not assess the amount of reference bias. While data can be removed or substituted by noise in order to assess the amount of bias during the refinement process and during flexible fitting (Chen et al., 2013; Falkner and Schroder, 2013), these methods operate in Fourier space and are difficult to transfer to sub-tomogram averaging (Briggs, 2013), and even more so to template matching. For template matching and sub-tomogram averaging, approaches that operate in real space have been suggested that investigate the signal recovery in an area excluded from the search and the classification process (Yu and Frangakis, 2011). These approaches were shown to increase the specificity of the template matching when the signal is strong (Hrabe et al., 2012).

The reason for reference bias is complex. Firstly, template matching will, by definition, indicate positions marking putative objects of interest such as macromolecular complexes. Whether this recognition truly corresponds to a macromolecular complex with the correct translation and rotation depends primarily on the signal-to-noise ratio (SNR), which cannot be measured. Furthermore, it depends on normalization of the cross-correlation function, which is affected by the choice of the mask and the molecular crowding. Finally, the missing wedge additionally plays an important role, as characteristic features of a macromolecule can be missing, making the identification in one particular direction especially complicated. Thus, in summary, reference bias can be introduced by (i) noise alone, (ii) misaligned particles of the macromolecule of interest, (iii) falsely identified features in the tomograms.

Here we develop a heuristic in real space, named the "M-free score", which allows an estimation of the amount of reference bias, without the knowledge of the genuine underlying signal. The name "M-free" originates from "Mask-free", indicating that a particular region of the mask is excluded during the template matching or sub-tomogram averaging process, and is only used to estimate the amount of reference bias. In contrast to previous approaches, which are based on measuring the recovery of frequencies in Fourier space, we investigate how the signal behaves and recovers in a defined area of the mask in real space, which contains information from all the frequencies in Fourier space. In this regard, the mask is split into a *working area* and a *testing area*. Comparable to the R-free value in X-ray crystallography, the *testing area* is not "seen" during the search or alignment process correspondingly (Brunger, 1992). In our heuristic, the behavior of the signal in the *testing area* is compared to the signal in *working area* after the alignment. The main argument for using a real space heuristic is that the medium and high frequencies, which are typically used for the Fourier based approaches (Chen et al., 2013; Falkner and Schroder, 2013; Shaikh et al., 2003), are almost absent in the raw data used for template matching and sub-tomogram averaging. Further, the contrast of the signal in the *testing* and *working areas* provides a good visual impression that helps users to validate the results.

We will first derive the mathematics on which the M-free heuristic is based. Next we will apply the heuristic on artificial data sets, and compare the results with the cross-correlation coefficient. We will show that the M-free score estimates the amount of reference bias, independently of the underlying signal, the SNR and the search range. To provide more realistic recording conditions, we use a tomographic data set of purified GroEL particles. We will show how the heuristic behaves when the tomogram is searched with an incorrect template. Lastly, we will show using the example of the ribosome in cellular data sets, that when the signal is improved, the M-free score is smaller allowing an estimation of the reference bias.

2. Materials and methods

The template matching and sub-tomogram averaging algorithms are implemented in C/C++ for massive parallel processing. The script for M-free score calculation implemented in Matlab (The MathWorks, Inc.) is available upon request. Visualization was performed with the Amira package (Pruggnaller et al., 2008).

2.1. Simulated data set

For the simulated data we used the same procedures and data sets first used in Foerster et al. (2008), which have subsequently also been used in several other studies. The simulated images include all artefacts present in electron tomography, and provide tomograms with a very realistic impression.

2.2. Biological data sets

We used previously published publicly available data sets (cited at each particular experiment), recorded on various microscopes and under various conditions. We use in particular an *in vitro* GroEL data set, which has already been used in many other publications. We used the results of sub-tomogram averaging from the SIV-spike data sets in (Zanetti et al., 2006) for validating the M-free score. Further, we used the Mycoplasma data set from (Seybert et al., 2006) to compare with cellular data sets recorded on modern direct detector cameras.

Mycoplasma cells were grown as described in (Hayflick, 1965). Prior to vitrification by plunge freezing in liquid ethane, 10 nm colloidal gold particles were applied.

Single-axis tilt-series were collected covering an angular range from −66° to +66° with 1.5° angular increment on a Titan Krios microscope operated at 300 kV (FEI, Eindhoven, The Netherlands) cooled to liquid nitrogen temperature and equipped with a Quantum energy filter (Gatan, Pleasanton, CA, USA). Data acquisition was carried out under low-dose conditions using the Digital Micrograph Software (Gatan, Pleasanton, CA, USA). Images were recorded on a 4×4 k pixel K2 Summit direct detector (Gatan, Pleasanton, CA, USA) at a nominal defocus of −5 μm. The pixel size at the specimen level is 0.34 nm.

3. Theory

In the core of our heuristic is the separation of the mask applied on the reference into a *working area* (W) and a *testing area* (T), in a similar fashion to (Brunger, 1992). W and T are disjoint and their conjunction is the full mask area, that is:

$$W \cap T = \emptyset \text{ and } W \cup T = \text{ALL}. \quad (1)$$

We consider the average of the sub-tomograms $p = s + n$, with $n = n_{\text{bias}} + n_{\text{ran}}$.

We dissect the signal into three types: the macromolecular signal s, the noise n which is considered to be composed of random noise n_{ran}, and biased noised n_{bias}. Since the signal in the *testing area* never "sees" the reference, the amount of biased noise in the *testing area* $n_{\text{bias},T} = 0$, while $n_{\text{bias},W}$ in the *working area* is what we aim to measure. The reference is considered to be uncorrelated with the random noise, and correlated to the biased noise, in a similar fashion to Chen et al. (2013).

The normalized cross-correlation coefficient between the reference r and the average of the sub-tomograms p within the *working area* can be expressed as:

$$\text{CCC}_W(r,p) = \frac{\sum_W (rp)}{\sqrt{\sum_W r^2 \sum_W p^2}} = \frac{\sum_W (rs + rn_{\text{bias},W})}{\sqrt{\sum_W r^2 \sum_W p^2}}, \tag{2}$$

with r and p being mean-value free within the area of the mask.

Within the *testing area* the information has not been used for the alignment thus $n_{\text{bias},T} = 0$, and the normalized cross-correlation coefficient can be expressed as:

$$CCC_T(r,p) = \frac{\sum_T rs}{\sqrt{\sum_T r^2 \sum_T p^2}} \tag{3}$$

The ratio of Eqs. (2) and (3) can be written as:

$$\frac{CCC_W(r,p)}{CCC_T(r,p)} = \frac{\sum_W (rs + rn_{\text{bias},W})}{\sqrt{\sum_W r^2 \sum_W p^2}} \cdot \frac{\sqrt{\sum_T r^2 \sum_T p^2}}{\sum_T rs} \tag{4}$$

We choose W and T such that the variance of the reference within each region is approximately the same. We also assume that the macromolecular signal s has approximately the same variance in both areas.The variance can be written in the following way:

$\sum_W p^2 = k \sum_T p^2$ and $\sum_W r^2 = k \sum_T r^2$, where k is a constant describing the quotient of the size of the *working area* to the *testing area*.

Thus Eq. (4) can be simplified as: $\frac{CCC_W(r,p)}{CCC_T(r,p)} = \frac{\sum_W (rs + rn_{\text{bias},W})}{k \sum_T rs}$, and we can write:

$$\frac{CCC_W(r,p) - CCC_T(r,p)}{CCC_T(r,p)} = \frac{\sum_W rn_{\text{bias},W}}{\sum_W rs} \tag{5}$$

with $\sum_W rs = k \sum_T rs$ assuming that a signal s results in the same cross-correlation coefficient between s and r in the *testing* and the *working area*.

Using the vector form of the cross-correlation function between the signals in the *working area*, Eq. (5) could be written as the ratio of the scalar products:

$$\frac{CCC_W(r,p) - CCC_T(r,p)}{CCC_T(r,p)} = \frac{\mathbf{R} \cdot \mathbf{N}_{\text{bias},W}}{\mathbf{R} \cdot \mathbf{S}} = \frac{||\mathbf{R}|| ||\mathbf{N}_{\text{bias},W}|| \cos \theta_{\mathbf{R},\mathbf{N}_{\text{bias},W}}}{||\mathbf{R}|| ||\mathbf{S}|| \cos \theta_{\mathbf{R},\mathbf{S}}} \tag{6}$$

If the reference **R** is similar to the signal **S**, which is valid for most of the real cases, the magnitude of the angle between the vector **R** and **S** is similar to the magnitude of the angle between **R** and $\mathbf{N}_{\text{bias},W}$, which by definition has to be similar. Thus in our effort to express the M-free score independently from the reference we assume the cosine of the angles $\theta_{\mathbf{R},\mathbf{N}_{bias,W}}$ and $\theta_{\mathbf{R},\mathbf{S}}$ to be equal. The right hand side of Eq. (6) can then be written as:

$$\frac{CCC_W(r,p) - CCC_T(r,p)}{CCC_T(r,p)} = \frac{||\mathbf{N}_{bias,W}||}{||\mathbf{S}||}. \tag{7}$$

Thus the term $\frac{CCC_W(r,p) - CCC_T(r,p)}{CCC_T(r,p)}$ that we name **M-free** gives an estimate of the reference bias $\mathbf{N}_{\text{bias},W}$ on a signal **S**. The CCC_T in the denominator of Eq. (7) can be negative, showing an anti-correlation between the reference and the average in this area, in which case the magnitude is irrelevant since it only reveals that there is no density in the testing area. Thus we set all negative values to be equal to zero resulting in an M-free score range of $(0\ +\infty)$, with 0 indicating no template bias, and $+\infty$ indicating non-existence of a macromolecular signal.Three assumptions were necessary in order to allow a description of the reference bias independent of the reference: (i) the signal s and the reference r in the *testing* and the *working area* relate in the following way: $\sum_W rs = k \cdot \sum_T rs$; (ii) the variance of the noise in the *working* and in the *testing area* relate in this way: $\sum_W n^2 = k \cdot \sum_T n^2$, assuming that after the alignment the total noise variance remains the same in the working and testing areas; and (iii) the magnitude of the angle between the reference and the biased noise is similar to the magnitude of the angle between the reference and the signal $\theta_{\mathbf{R},\mathbf{N}_{\text{bias},W}} = \theta_{\mathbf{R},\mathbf{S}}$. These assumptions only affect the correctness of the M-free score when the reference bias is very high. In the following section, we will show experimentally that the M-free score gives sensible results, and can be used for estimating the amount of reference bias present in the experiment.

4. Results

In sub-tomogram averaging and template matching, the area around the reference is masked. This is to delineate the boundaries of the template from the surrounding, as well as to define the area in which the variance for the denominator of the cross-correlation function is calculated. The masks can have arbitrary shapes, and are rotated together with the template. Here, we reserve ~10% of the mask as a *testing area* that does not contribute to the cross-correlation function (Fig. 1a). The *testing area* is selected such that the signal variance within the *testing area* is identical to the signal variance in the *working area* and can have an arbitrary shape.

4.1. Artificial data

We first measured the M-free score in sub-tomogram averaging of simulated GroELs with different SNR, random orientations and missing wedge. After sub-tomogram averaging using the original GroEL structure as a reference, the particles with a SNR above 0.003 produce an average that is fully symmetric with a well-defined density in the *testing area* (Fig. 1b–e). The average structures from the particles with a SNR below 0.003 do not have a density in the *testing area* (Fig. 1f, g). When the M-free score is measured for an SNR above 0.003, the M-free score is close to zero, while for lower SNR the M-free value rises quickly (Fig. 1h). This is a fundamentally different behavior to the cross-correlation coefficient between the reference and the sub-tomogram average, which slowly decreases in a fashion that is dependent on the SNR, and the number of sub-tomograms (with more sub-tomograms the decrease would be less). Examining the reason for the increased in the M-free score shows that a number of particles were misaligned in the final average (Fig. 1i).

We next explored the validity of the M-free score against contamination. For this we chose the particles with a SNR of 0.003 which is at the lower limit of SNR necessary for success of the alignment. We contaminated the simulated GroELs with Thermosomes having the same missing wedge and SNR as the GroELs. The M-free score rises linearly depending on the amount of contamination present in the data set, which corresponds to the amount of misaligned sub-tomograms (Fig. 2a).

When template matching is simulated on an artificial data set containing GroEL particles, and an exhaustive search with six DOF is performed, the M-free score is even more informative. While the cross-correlation value between the average and the GroEL reference stays approximately the same, the M-free score shows a completely different characteristic: When a GroEL reference is used to search for GroELs the M-free score is 0.16, but when

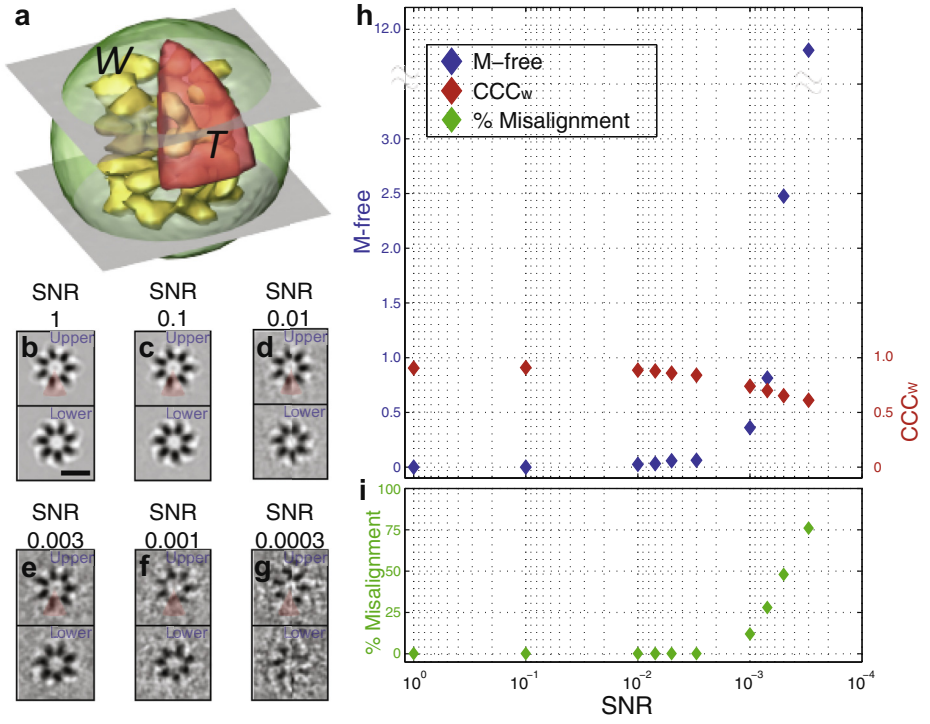

Fig.1. Results of the sub-tomogram averaging experiments with artificial GroEL sub-tomograms. (a) An illustration of the *working area W* in green and the *testing area T* in red encircling the isosurface of the GroEL in yellow. The two computational sections through the apical domains intersecting with the *testing* and *working area* respectively are shown in (b)–(g). (b)–(g) The slices indicated in (a) of the averages from data sets with SNR (b) 1, (c) 0.1, (d) 0.01, (e) 0.003, (f) 0.001, and (g) 0.0003. In the upper slices, the *testing area* is indicated with transparent red. The M-free scores measured are 0, 0, 0, 0.05, 0.4, and 12 respectively. The scale bar is 10 nm. (h) M-free scores measured for data sets with different SNRs aligned using GroEL as a starting reference. M-free scores remain at ~0 until the SNR reaches 0.003. When the SNR decreases further, the M-free score rises exponentially. In comparison, the cross-correlation values decrease slowly. More importantly, the cross-correlation value depends on the number of particles, while the M-free score is independent of this criterion. (i) The percentile of misaligned particles as a function of the SNR is shown. The M-free score appropriately reflects the amount of particles for which the orientation and/or translation was not found properly due to the SNR.

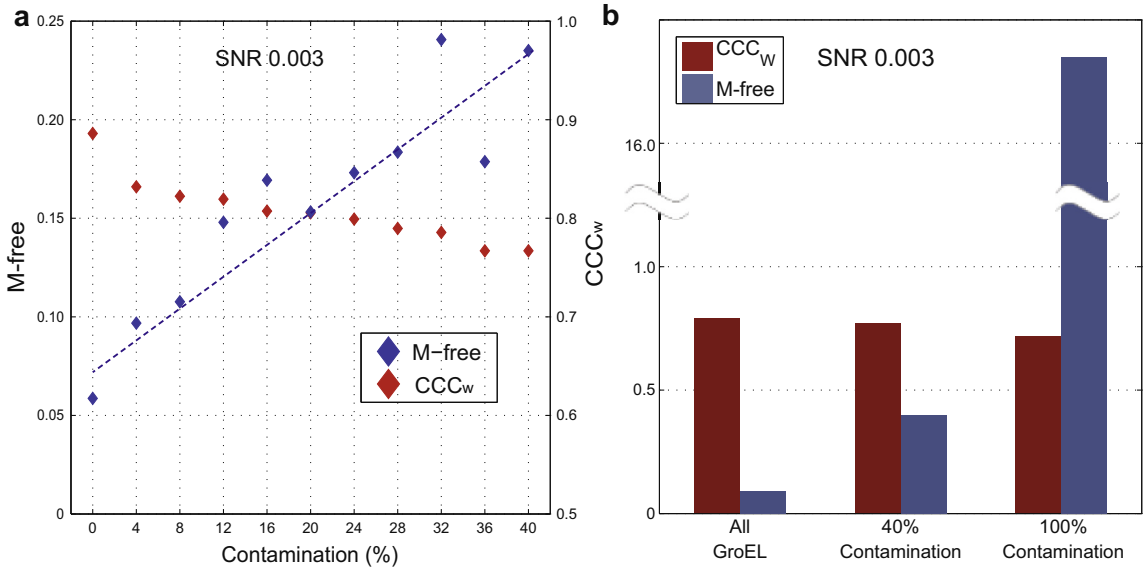

Fig.2. Results of sub-tomogram averaging and template matching experiments with contaminated artificial GroEL data sets. (a) The M-free scores and cross-correlation values measured from sub-tomogram averaged data sets with an SNR equal to 0.003 and a contamination of 0–40% in 4% steps. (b) Template matching on an artificial data set containing GroELs contaminated with Thermosomes. The bar chart shows M-free scores and cross-correlation values. While the cross-correlation value stays approximately constant the M-free score rises significantly.

the contamination is at ~40% the M-free score is close to 0.7, and when only Thermosomes are present the M-free score is 17. The corresponding cross-correlation coefficients in the above experiments are 0.81, 0.81 and 0.72 (Fig. 2b). Concluding, in defined computational conditions the M-free score gives meaningful results that reflect the quality of the alignment and the amount of contamination.

4.2. In vitro data sets

To measure the information content of the M-free score at realistic experimental conditions we performed template matching on a tomogram of purified GroELs (Fig. 3a). We visually selected ~100 sub-tomograms depicting putative GroEL particles and performed sub-tomogram averaging. In a similar manner to the artificial data

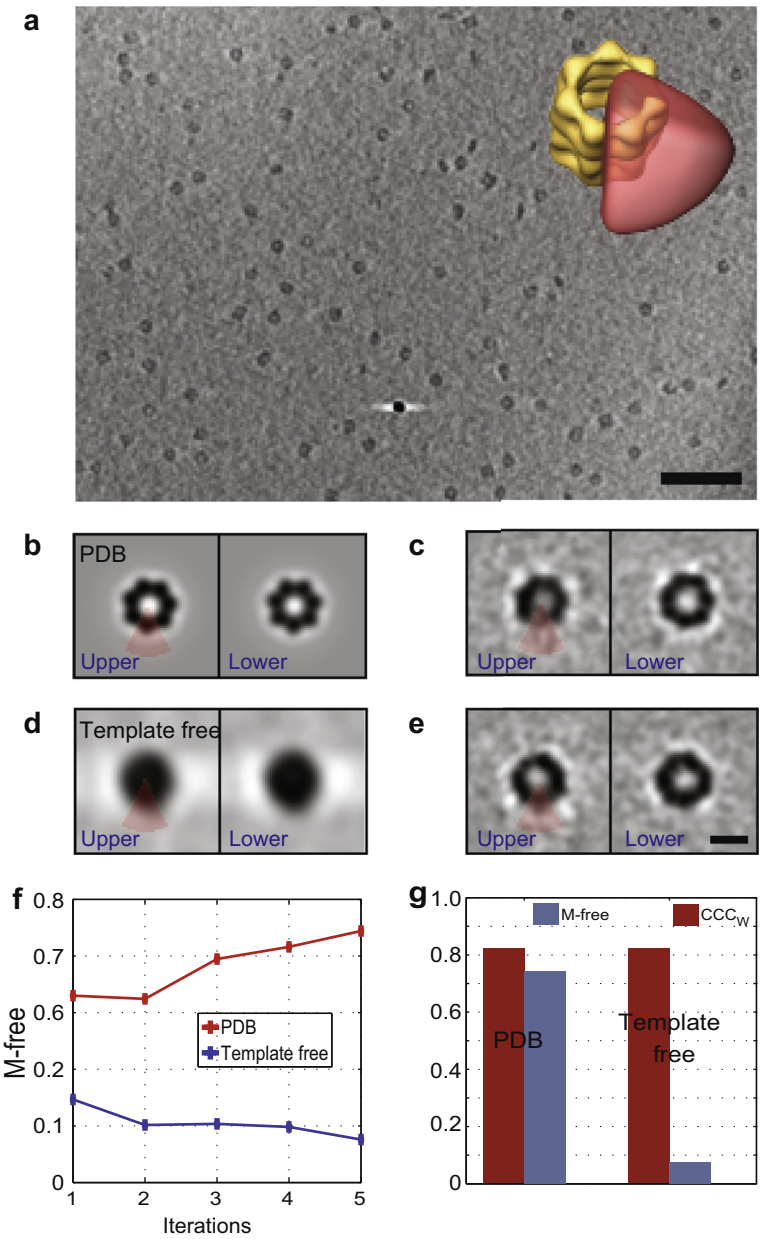

Fig.3. Results of sub-tomogram averaging with a tomogram depicting purified GroELs. (a) An 18 nm thick projection slice of the tomogram. The scale bar is 100 nm. In the inset, an illustration of the *testing area* is shown in red in relation to the isosurface of the GroEL derived from the crystal structure (PDB: 2C7E) in yellow. (b) Two slices through the upper and lower apical domains of the density map derived from the GroEL crystal structure. The *testing area* is marked with transparent red throughout all images. (c) The same slices through the sub-tomogram average after 5 iterations using the GroEL crystal structure as a starting reference. The M-free score is 0.74. (d) The slices of the Gaussian blob used for "template-free" sub-tomogram averaging. (e) The two slices of the result of "template-free" sub-tomogram averaging after 5 iterations. The M-free score is 0.08. Scale bar is 10 nm. (f) The development of the M-free score depending on the number of iterations. The red curve shows the M-free scores with GroEL crystal structure as a starting reference, while the blue curve shows the "template-free" approach. (g) A bar chart where M-free scores are shown in blue and the cross-correlation coefficients calculated in the *working area* (CCC_W) are shown in red, measured from the averages in (c) and (e). The CCC_W do not differ from each other in the different experiments, however the M-free scores are 10-fold higher in experiments using a crystal-structure-derived template rather than a Gaussian sphere.

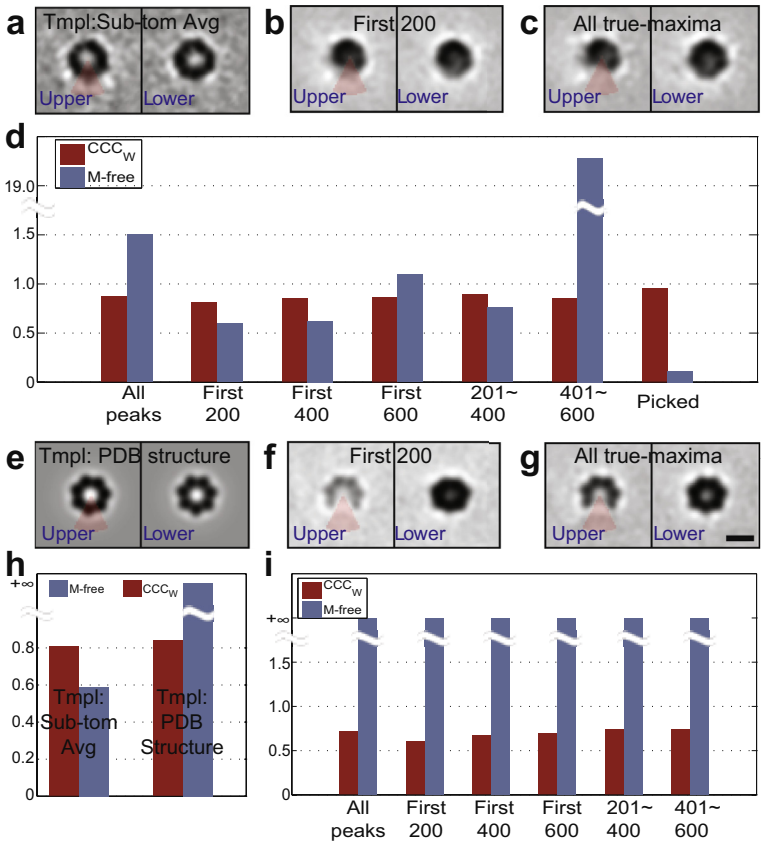

Fig.4. Result of template matching with a tomogram of purified GroELs. (a) Two slices through the upper and lower apical domains of the sub-tomogram average, which was used as the reference for template matching. (b) The outcome after template matching and averaging the positions corresponding to the highest 200 true maxima. The M-free score was calculated to be 0.59. (c) The same two slices using all true maxima and an M-free score of 1.5. Scale bar is 10 nm. (d) A bar diagram of the M-free scores in blue and the cross-correlation coefficients in red, measured from averages generated from batches of sub-tomograms selected according to the true maxima detected from the template matching with a GroEL template. The M-free scores are similar for the averages generated from the 200 highest true maxima, the 201st–400th true maxima and all the 400 highest true maxima together. The average created from the 401st–600th true maxima has a much higher M-free score. An average generated from the manually selected positions has the lowest M-free score of 0.08, which indicates that when the amount of contamination is low the M-free score is very low. In contrast, the cross-correlation coefficients in W are similarly high for all averages. (e) Two slices through the upper and lower apical domains of filtered PDB structure, that was used as the reference for the second template matching experiment. (f) The density after sub-tomogram averaging of the sub-tomograms corresponding to the highest 200 true maxima. The M-free value is infinite. (g) The density after sub-tomogram averaging all putative particles. The M-free value is infinite again. Scale bar is 10 nm in all slices. In both cases the density in the *testing area* is much more poorly recovered than in (c). (h) The bar chart for M-free and CCC_W for template matching shows that the cross-correlation values are similar to sub-tomogram averaging, while the M-free score efficiently discriminates between the different amounts reference bias. (i) The GroEL tomogram was subjected to template matching using a Thermosome in its open conformation as a template, which does not occur in this tomogram. The cross-correlation values again show similar values for all the averages, and are only slightly lower compared to the experiment where GroEL is used as a search template. The M-free scores however, are all infinitely high indicating that no energy from the macromolecular signature is present at all, thereby unambiguously rejecting the Thermosome as a potential macromolecule in this tomogram.

sets, we used a spherical mask from which a wedge representing 10% of the total volume was removed to provide the *testing area* (Fig. 3a inset). The position of the *testing area* was chosen in such a way so that the variance within the *testing* and *working area* was approximately the same. The sub-tomogram averaging was performed for 5 iterations. Firstly, the GroEL crystal structure from the PDB was used as a reference (Fig. 3b) (Ranson et al., 2001), where the sub-tomogram averages converged to a density showing clear sevenfold symmetry, although with less density in the *testing area* than elsewhere (Fig. 3c). Using a Gaussian blob (Fig. 3d) as the starting reference, the sub-tomogram average converges to a slightly different structure, with less prominent sevenfold symmetry, but with a clear density in the *testing area* (Fig. 3e). For both cases the M-free score was calculated after every iteration (Fig. 3f). In the last iteration, the M-free score for the PDB reference was 0.74 while the M-free score for the Gaussian blob reference was 0.08. The cross-correlation value is 0.82 for both approaches and shows no difference (Fig. 3g). A visual comparison between Fig. 3c and e shows that in the case of the PDB reference the density does not recover in the *testing area*. Furthermore, the M-free score reflects a common sense observation that the Gaussian blob reference does not introduce a bias, while the GroEL crystal structure does.

While for the sub-tomogram averaging only a selected number of particles are used, in template matching the complete tomogram is scanned for putative particles. For this experiment the complete tomogram was scanned exhaustively with two templates. We first searched with the density obtained from the sub-tomogram averaging using the Gaussian blob as a reference (Fig. 4a). Sorting the values of the local maxima of the cross-correlation function shows a steadily decreasing function for which no hard threshold for the putative positions of the sought macromolecule can be set. Additionally, the number of putative particles is not known. Therefore, we selected all 716 true maxima present in the cross-correlation map and calculated the average of the highest 200 first and of all 716 next (Fig. 4b, c), for which the M-free score was 0.59 and 1.5 respectively (Fig. 4d). When the 201st–400th true maxima are averaged, the M-free score increases slightly to 0.60 and for the average of the 401st–600th true maxima the score rises to 20 (Fig. 4d). The cross-correlation value between the template and

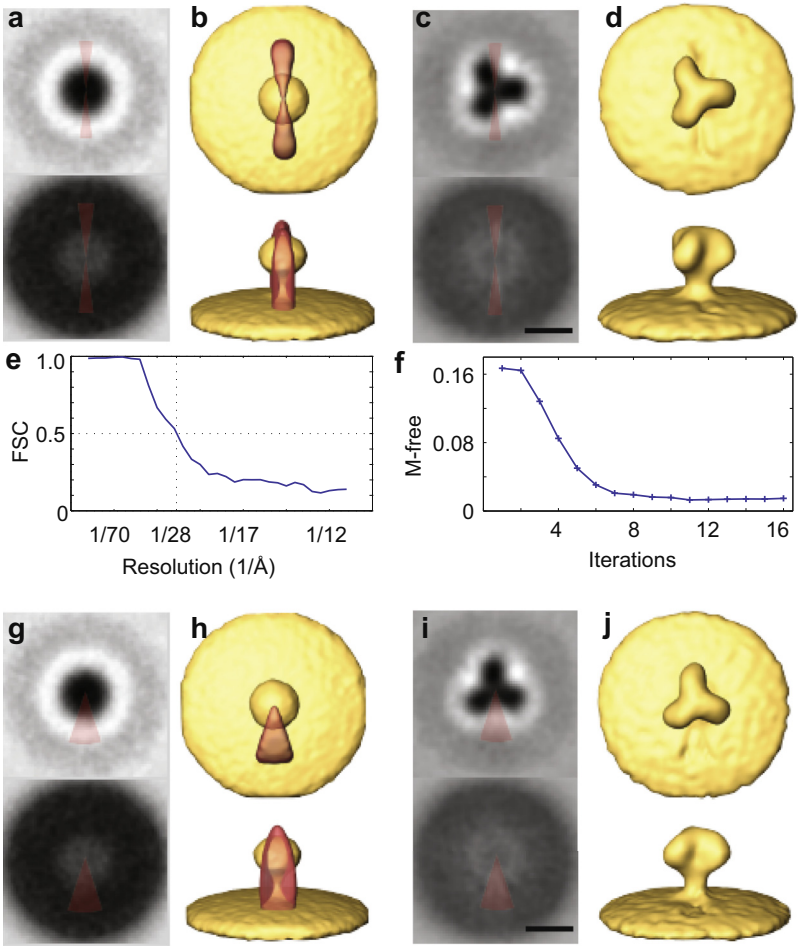

Fig.5. Results of the sub-tomogram averaging with SIV-spike data sets (Zanetti et al., 2006). (a) 2 xy-slices of the starting template generated using the information from the pre-alignment. The upper slice cuts through the spike and the lower slice cuts through the membrane. The *testing area* is marked in red. (b) The top view and side view of the isosurface from (a). (c) The same xy-slices as in (a) of the average after 16 iterations. The scale bar is 10 nm. (d) The top view and side view of the isosurface from the result in (c). The average clearly has a threefold symmetry and corresponds to the result as shown in Fig. 2 in Zanetti et al. (2006). (e) The Fourier Shell Correlation (FSC) shows that the average in (c) and (d) has a resolution of 28 Å judged using the 0.5 criterion, which matches the result in Zanetti et al. (2006). (f) The development of the M-free score depending on the number of iterations. After the first iteration the M-free score is relatively high with a value of 0.17, but it decreases linearly with the quality improvement of the average. The M-free score converges to 0.01 after 16 iterations. (g)–(j), the same as (a)–(d), with a different *testing area* that is shown in (g)–(i). The result is extremely similar to the one above. The scale bar in (i) is 10 nm.

the average increases slowly as more sub-tomograms are selected (Fig. 4d). Interestingly, the Fourier Shell Correlation also shows increasing resolution as more sub-tomograms are added (not shown), despite the fact that more misaligned or non-true particles are included in the average. Calculating the M-free score of the fraction of the true maxima corresponding to the manually selected sub-tomograms results in a score of 0.09 (Fig. 4d), which is similar to the score in the sub-tomogram averaging experiment indicating a similar amount of bias, showing that for classified data the alignment is successful.

Repeating the above experiment using the crystal structure as a template (Fig. 4e) the M-free score is infinite for the first 200 highest true maxima (Fig. 4f) and stays infinite when all positions are selected (Fig. 4g). The cross-correlation again shows almost no difference between the performances of the two templates (Fig. 4h).

To perform the negative experiment, we used a Thermosome template (pixel size and CTF adapted to the recording conditions of the tomogram with the same mask used for GroEL) to scan the tomogram containing only GroELs. Even though the distribution of the cross-correlation values and the selected positions are very similar compared to the ones of the GroEL template, the M-free score is infinite (Fig. 4i). Thus, while the cross-correlation value hardly discerns between GroEL and Thermosomes, the M-free score shows significant differences.

4.3. Sub-tomogram averaging of SIV spikes (Zanetti et al., 2006)

We next tested the information content on the sub-tomogram averaging of SIV spikes from Zanetti et al. (2006), especially with regard to misalignments, resolution short fallings and artifacts introduced by the shape of the *working area*. We used an elliptical mask from which a wedge of 10% the volume was carved out as the *testing area* (Fig. 5a, b). We used several masks and probed the outcome of the sub-tomogram averaging. Apart from the mask alterations, the processing was carried out as described in the original paper: the sub-tomograms were rotationally pre-aligned based on their position on the surface of the virus, which results in a globular structure on top of a membrane patch (Fig. 5b). With progressing iterations a structure resembling the one shown in the paper of Zanetti et al. (2006) emerges (Fig. 5c, d), with a resolution of 2.8 nm, as judged by the 0.5 criterion of the FSC curve (Fig. 5e). The shape of the *testing area* is not visible in the final structure at all. The M-free score decreases during the process and reaches a value of around 0.01 indicating no template bias (Fig. 5f).

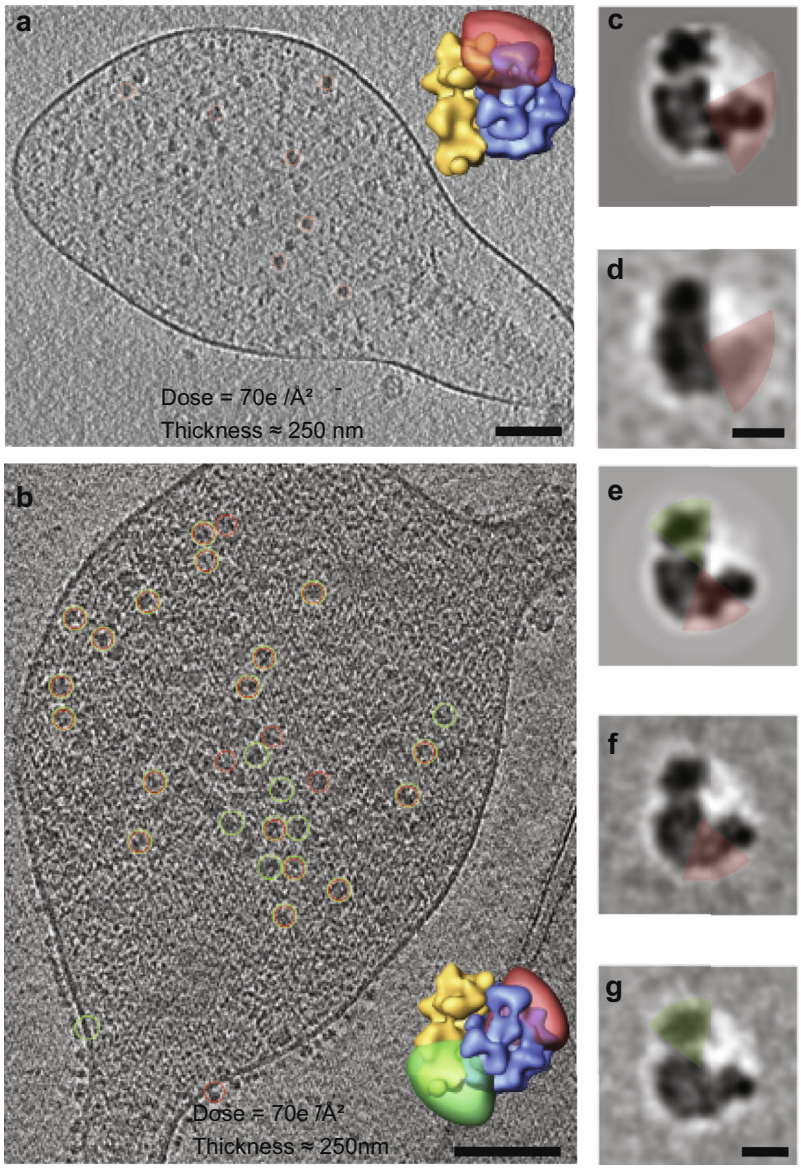

Fig.6. Results of the template matching with whole cell tomograms of Mycoplasma cells. (a) A 7 nm xy-slice of the tomogram of a ~250 nm thick Mycoplasma cell recorded on the Megascan camera. Inset in the upper right corner shows an isosurface of the ribosome template derived from the crystal structure (2AW7 in yellow and 2AWB in blue) with the *testing area* in transparent red. The top 150 detected positions that contain most of the ribosomes on this slice are shown in red circles. The scale bar is 100 nm. (b) A 7 nm xy-slice of a ~250 nm thick Mycoplasma cell recorded on the K2 direct detector camera. The X-ray derived template is shown in the upper right corner. Two different *testing areas* are marked in red and green. The top 150 detected positions of putative ribosomes on this slice are shown in red or green circles, corresponding to the red or green *testing areas*. The scale bar is 100 nm. (c) The central slice from the template derived from the crystal structure with the *testing area* shown in red. (d) The central slice from the average of the top 150 detected positions with the *testing area* shown in red. M-free score is 0.41. The scale bar is 10 nm. (e) The central slice from the template derived from X-ray structure and the *testing areas* in red or green. (f) The center slice from the average of the detected positions with the red *testing area*. M-free score is 0.09. (g) Same as (f), with the green *testing area*. The average shows minimal difference in comparison to (f). The amount of template bias is similar. The M-free score is 0.10. The scale bar is 10 nm.

We repeated the same process with a variety of masks using different regions of the starting reference of the SIV spikes (Fig. 5g, h). As long as the variance within the testing and *working area* was approximately the same, the resulting sub-tomogram average exhibited a similar resolution and a similar M-free score (Fig. 5i, j).

4.4. Template matching – whole Mycoplasma cells

We then tested the information quality of the M-free score in template matching of cellular tomograms. We thereby scanned two cryo-electron tomograms of Mycoplasma cells with a ribosome template derived from the crystal structure (Schuwirth et al., 2005). We used a spherical mask, from which a wedge representing 10% of the total volume was used as the *testing area*. In order to compare the M-free scores we performed the experiment multiple times with the *testing area* placed in random orientations. The cells were embedded in ice of similar thickness, and recorded at the same conditions with the same cumulative dose (70 $e^-/Å^2$). The only difference was that the first tomogram was recorded on a Gatan 2002 energy filter with a 2 × 2 k pixel Gatan Megascan 795 CCD camera (Short Megascan) (Fig. 6a), while the second was recorded on a Quantum energy filter with a 4 × 4 k pixel K2 Summit direct detector (Short K2) (Fig. 6b). The ribosome is by far the biggest complex within these cells, and occurs in high numbers. Indeed, template matching indicates numerous putative particles that are localized throughout the tomogram – sometimes

on the cell-boundaries and in some rare cases even outside the cell (Fig. 6a, b). The positions corresponding to the highest 150 true maxima after running the template matching with the filtered ribosome X-ray structure were selected for the Megascan (Fig. 6c). The M-free score was measured at 0.41 (Fig. 6d). In the slices of the sub-tomogram average the ribosome region under the testing area remains suspiciously electron lucent compared to the rest. For the tomogram recorded on the K2, M-free scores around 0.1 were measured for different *testing areas* (Fig. 6e). When visually examining the averages of the putative ribosome positions, the average created from data recorded on the K2 shows the similar electron density in the *testing area* as in the *working area* – which is not the case for the average created from data recorded on the Megascan (Fig. 6f, g). Comparison of these values with the values from the artificial data sets shows that the quality of ribosome detection on the tomogram recorded with the Megascan is very good, but the result is not as reliable for the tomogram recorded with the K2.

5. Discussion

With our heuristic, we suggest excluding a *testing area* from the mask used in template matching or sub-tomogram averaging in order to use the information in the area to estimate the amount of bias imposed by the reference. We show that sacrificing 10% of the mask area is sufficient to provide a reliable score for assessing the amount of template bias. Sacrificing signal from the sub-tomogram might worsen the alignment. Thus, the alignment could first be performed with a complete mask, and the result can be assessed later (for instance by a difference map) according to the same mask but excluding a *testing area*. Importantly, we show that for sub-tomogram averages that have a strong signal, the influence of the *testing area* is very small.

In our experiments we used arbitrarily oriented wedge shaped masks, which were automatically oriented such that the variance within the *testing* and *working area* is the same. For these cases we did not experience any major influence of the mask orientation or position in the magnitude of the M-free score. In the case of sub-tomogram averaging, the choice of the *testing area* was more complicated, because in some cases the average "moved-away" from the *testing area* and converged in such way, that the *testing area* was essentially empty, rendering it meaningless. For these cases the *testing area* had to be chosen in a way such that the variance equality in the *testing* and *working area* was fulfilled throughout the sub-tomogram averaging.

The M-free score focuses on real space rather than on Fourier space. In single particle approaches, the strong signal in medium–high frequencies allows analysis of the signal recovery in shells in Fourier space (Shaikh et al., 2003). In contrast, the strong signal in electron tomography for sub-tomogram averaging or template matching is only present in a few inner Fourier shells. At higher frequencies the signal can only be recovered after sub-tomogram averaging. Thus ignoring a few shells in the low frequency domains for testing purposes means rejecting a good portion of the information that is most necessary for the alignment. Excising an area in real space still keeps most of the information from the whole frequency domain that is essential for averaging purposes.

The smaller the M-free score, the smaller the amount of bias in the alignment process. In perfectly aligned artificial data sets without contamination, the M-free score is zero independent of the SNR; a property fundamentally different to the cross-correlation coefficient. When either the contamination increases or the SNR decreases below a certain threshold, the M-free score rises, thereby giving an estimate of the reference bias. Our experiments for real data sets show that an average generated either from template matching or sub-tomogram averaging with an M-free score below 0.1 indicates an acceptably small reference bias. If the M-free score is in the range of 0.1–0.5, the process is still partially reliable, while any value higher than 0.5 makes the outcome questionable. There is always some reference bias in the alignment due to the existence of noise, and a score of zero could never be attained for real data, even for the best possible selected data sets. However, the scores were very close to zero.

Template matching and sub-tomogram averaging are essential approaches to explore the spatial information of macromolecular complexes within their cellular context. The M-free score that we introduce here provides a reliability measurement for those approaches, which makes them more applicable to structural biology.

Acknowledgments

We thank A. Seybert for the Mycoplasma tomogram on the Megascan CCD camera, L. Gonzales for the Mycoplasma tomogram on the K2 direct detector, and J. Briggs for the sub-tomograms of the SIV-spikes. We also thank M. Kunz, D. Castano-Diez, M. Scheffer and C. Wigge for critical reading of the manuscript. We would like to thank Reiner Hegerl for detailed advice on the theoretical section of the manuscript. This work was motivated by CRC 902 and was supported by an ERC starting grant to ASF.

References

Baumeister, W., 2005. From proteomic inventory to architecture. FEBS Lett. 579, 933–937.
Briggs, J.A., 2013. Structural biology in situ – the potential of subtomogram averaging. Curr. Opin. Struct. Biol. 23, 261–267.
Brunger, A.T., 1992. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 355, 472–475.
Chen, S., McMullan, G., Faruqi, A.R., Murshudov, G.N., Short, J.M., Scheres, S.H., Henderson, R., 2013. High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. Ultramicroscopy 135, 24–35.
Falkner, B., Schroder, G.F., 2013. Cross-validation in cryo-EM-based structural modeling. Proc. Natl. Acad. Sci. U.S.A. 110, 8930–8935.
Foerster, F., Pruggnaller, S., Seybert, A., Frangakis, A.S., 2008. Classification of cryo-electron sub-tomograms using constrained correlation. J. Struct. Biol. 161, 276–286.
Frangakis, A.S., Bohm, J., Forster, F., Nickell, S., Nicastro, D., Typke, D., Hegerl, R., Baumeister, W., 2002. Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. Proc. Natl. Acad. Sci. U.S.A. 99, 14153–14158.
Hayflick, L., 1965. Tissue cultures and mycoplasmas. Tex. Rep. Biol. Med. 23 (Suppl. 1), 285+.
Hrabe, T., Chen, Y., Pfeffer, S., Cuellar, L.K., Mangold, A.V., Forster, F., 2012. PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. J. Struct. Biol. 178, 177–188.
Lucic, V., Rigort, A., Baumeister, W., 2013. Cryo-electron tomography: the challenge of doing structural biology in situ. J. Cell Biol. 202, 407–419.
Lyumkis, D., Brilot, A.F., Theobald, D.L., Grigorieff, N., 2013. Likelihood-based classification of cryo-EM images using FREALIGN. J. Struct. Biol. 183, 377–388.
Nickell, S., Kofler, C., Leis, A.P., Baumeister, W., 2006. A visual approach to proteomics. Nat. Rev. Mol. Cell Biol. 7, 225–230.
Penczek, P.A., 2002. Three-dimensional spectral signal-to-noise ratio for a class of reconstruction algorithms. J. Struct. Biol. 138, 34–46.
Pruggnaller, S., Mayr, M., Frangakis, A.S., 2008. A visualization and segmentation toolbox for electron microscopy. J. Struct. Biol. 164, 161–165.
Ranson, N.A., Farr, G.W., Roseman, A.M., Gowen, B., Fenton, W.A., Horwich, A.L., Saibil, H.R., 2001. ATP-bound states of GroEL captured by cryo-electron microscopy. Cell 107, 869–879.
Scheres, S.H., 2012a. RELION: implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol. 180, 519–530.
Scheres, S.H., 2012b. A Bayesian view on cryo-EM structure determination. J. Mol. Biol. 415, 406–418.
Scheres, S.H., Chen, S., 2012. Prevention of overfitting in cryo-EM structure determination. Nat. Methods 9, 853–854.
Schur, F.K., Hagen, W.J., de Marco, A., Briggs, J.A., 2013. Determination of protein structure at 8.5 A resolution using cryo-electron tomography and sub-tomogram averaging. J. Struct. Biol. 184, 394–400.

Schuwirth, B.S., Borovinskaya, M.A., Hau, C.W., Zhang, W., Vila-Sanjurjo, A., Holton, J.M., Cate, J.H.D., 2005. Structures of the bacterial ribosome at 3.5 angstrom resolution. Science 310, 827–834.
Seybert, A., Herrmann, R., Frangakis, A.S., 2006. Structural analysis of Mycoplasma pneumoniae by cryo-electron tomography. J. Struct. Biol. 156, 342–354.
Shaikh, T.R., Hegerl, R., Frank, J., 2003. An approach to examining model dependence in EM reconstructions using cross-validation. J. Struct. Biol. 142, 301–310.
Yu, Z., Frangakis, A.S., 2011. Classification of electron sub-tomograms with neural networks and its application to template-matching. J. Struct. Biol. 174, 494–504.
Zanetti, G., Briggs, J.A., Grunewald, K., Sattentau, Q.J., Fuller, S.D., 2006. Cryo-electron tomographic structure of an immunodeficiency virus envelope complex in situ. PLoS Pathog. 2, e83.