Published in final edited form as:

J Struct Biol. 2009 May ; 166(2): 234–240. doi:10.1016/j.jsb.2009.02.007.

Maximum likelihood refinement of electron microscopy data with normalization errors

Sjors H.W. Scheres^{*a}, Mikel Valle^b, Patricia Grob^c, Eva Nogales^{c,d}, and José-María Carazo^a

^aCentro Nacional de Biotecnología - CSIC, Darwin 3, Cantoblanco, 28049, Madrid, Spain

^bCICBiogune, Parque Tecnológico de Bizkaia, 48160, Derio-Bizkaia, Spain

^cHoward Hughes Medical Institute, QB3/Molecular and Cell Biology Department, University of California at Berkeley, Berkeley, California 94720, USA

^dLawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA

Abstract

Commonly employed data models for maximum likelihood refinement of electron microscopy images behave poorly in the presence of normalization errors. Small variations in background mean or signal brightness are relatively common in cryo-electron microscopy data, and varying signal-to-noise ratios or artifacts in the images interfere with standard normalization procedures. In this paper, a statistical data model that accounts for normalization errors is presented, and a corresponding algorithm for maximum likelihood classification of structurally heterogeneous projection data is derived. The extended data model has general relevance, since similar algorithms may be derived for other maximum likelihood approaches in the field. The potentials of this approach are illustrated for two structurally heterogeneous data sets: 70S *E.coli* ribosomes and human RNA polymerase II complexes. In both cases, maximum likelihood classification based on the conventional data model failed, whereas the new approach was capable of revealing previously unobserved conformations.

Keywords

single particle analysis; structural heterogeneity; classification; expectation maximization

1. Introduction

Over the last decades, three-dimensional electron microscopy (3D-EM) has developed into a widely applicable technique for the structural characterization of biological complexes. On one hand, ever increasing resolutions are obtained for well-behaved (conformationally stable) macromolecular complexes, currently reaching up to 3.8 Å for icosahedral virus reconstructions (Zhang et al., 2008; Yu et al., 2008) and up to 5.4 Å for particles with low or no symmetry (e.g. see Stagg et al., 2008). On the other hand, 3D-EM techniques are being applied to ever more complicated samples. The structural characterization of highly flexible cellular machines is nowadays feasible through the single particle reconstruction approach of

*Corresponding author: Email address: E-mail: scheres@cnb.csic.es (Sjors H.W. Scheres).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

purified samples (Stark and Lüthmann, 2006; Grob et al., 2006; Nickell et al., 2007), while the characterization of the molecular atlas of whole cells is within reach of modern cryo-electron tomography (Nickell et al., 2006; Robinson et al., 2007). Together with numerous instrumental improvements, these advances have gone hand-in-hand with important developments in image processing techniques in the field.

With the image processing tasks becoming ever more complicated, there is a growing interest in the use of statistical methods in 3D-EM and in particular in maximum likelihood approaches. Perhaps the most important characteristic of the maximum likelihood approach is the natural way in which the noisy character of the experimental data may be modelled. This is especially relevant in the case of cryo-EM data, where a limited electron dose to prevent radiation damage results in extremely low signal to noise ratios. The maximum likelihood approach has now been applied to a range of different image processing tasks, such as 2D alignment (Sigworth, 1998), 2D classification (Pascual-Montano et al., 2001), 3D reconstruction of icosahedral viruses (Vogel and Provencher, 1988; Doerschuk and Johnson, 2000; Yin et al., 2003; Lee et al., 2007), alignment of 2D crystal images (Zeng et al., 2007), or 3D classification of heterogeneous projection data (Scheres et al., 2007a).

All these approaches employ the same statistical data model that assumes white gaussian noise in the data. Recently, we also introduced an alternative model for coloured noise (Scheres et al., 2007b), but the assumption of gaussian noise remains a common factor for all maximum likelihood approaches in the field. Although the explicit description of the experimental noise in the maximum likelihood approach offers general advantages over conventional approaches, it may also present important limitations in specific cases. The distance metric that underlies the gaussian model is based on the squared Euclidian distance between an experimental image and its template. In contrast to the conventional (normalized) cross-correlation coefficient, the Euclidian distance metric is highly sensitive to differences in image background and signal brightness. This means that any maximum likelihood approach based on this metric may suffer from variations in background mean or signal brightness among the data. In the case of image classification for example, the data may be separated in subsets with similar image backgrounds or signal brightness rather than in structurally homogeneous subsets.

In practice, one aims to minimize the variations in background mean and signal brightness by normalizing the data. Because the abundant noise in 3D-EM data makes it difficult to normalize the signal itself, it is common practice to normalize the noise instead. Typically, one subtracts a least-squares plane to obtain zero-mean backgrounds and subsequently divides by the standard deviation to obtain similar noise intensities among all images. To account for the fact that different orientations of an asymmetrical particle may yield projections with different signal powers, one often calculates this plane and standard deviation over an area of the image that presumably contains only noise (Sorzano et al., 2004a). However, the presence of neighbouring particles in this so-called background area, or variations in the signal-to-noise ratios (e.g. due to different ice thickness or defocus values) may lead to remaining variations in the signal intensity among the normalized images. Consequently, to some extent 3D-EM data sets always display non-zero background means and variations in signal brightness. Also the relatively common practice to high-pass filter the particles is not a remedy for this problem, as the additive and multiplicative variations in their underlying signal do not necessarily relate to the power of the entire image.

Therefore, we propose an extension of the commonly used data model of white gaussian noise that allows describing variations in background mean and signal brightness among the images. This model is generally applicable to any of the existing maximum likelihood approaches in the field, although here we will focus on the problem of 3D classification. This problem plays a crucial role in the single particle analysis of flexible macromolecular complexes that adopt

multiple conformations and may vary in subunit composition or ligand occupancy. The way these complexes work may often be inferred from their distinct structural states, but the difficulties in biochemically purifying these states make it cumbersome to study them. Cryo-EM allows recording projections of individual particles that are free to adopt any of their functional states. Thereby, one may obtain structural information about a whole range of conformations from a single cryo-EM experiment, provided that one can sort the data into subsets of projections from particles with identical 3D structures. However, this process of *in silico* purification currently still represents one of the major challenges in 3D-EM single-particle analysis (Leschziner and Nogales, 2007).

Based on the proposed statistical model for data with normalization errors, we have derived a maximum-likelihood algorithm for the 3D classification of structurally heterogeneous projection data. We call this algorithm MLn3D classification, to distinguish it from the previously introduced ML3D algorithm that is based on the commonly employed data model without normalization errors (Scheres et al., 2007a). Here we illustrate the usefulness of the new algorithm for two highly challenging cryo-electron microscopy data sets: a 70S *E.coli* ribosome data set and a data set on human RNA polymerase II in complex with human Alu RNA (Mariner et al., 2008). For both data sets, we show how the conventional maximum likelihood approach failed due to normalization errors in the data, whereas the MLn3D algorithm was capable of separating distinct, previously unobserved structural states.

2. Approach

2.1. The extended data model

We model 2D images $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$ as follows:

$$\mathbf{X}_i = s_i^o R_{\Phi_i} \mathbf{V}_{\kappa_i}^o + \mathbf{c}_i^o + \mathbf{N}_i, \quad (1)$$

where:

1. $\mathbf{X}_i \in \mathbb{R}^J$ are the recorded data.
2. κ_i is a random integer with possible values $1, 2, \dots, K$. Then, there are K unknown 3D structures, $\mathbf{V}_1^o, \mathbf{V}_2^o, \dots, \mathbf{V}_K^o$. These are the objects we wish to reconstruct from the data.
3. $R_{\Phi_i} \mathbf{V}_{\kappa_i}^o \in \mathbb{R}^J$ are the 2D projection data (uncontaminated by noise) of the unknown object $\mathbf{V}_{\kappa_i}^o$ in an unknown random orientation in space and position in the plane. The unknown orientation and position in the plane is parametrized by Φ_i , a 5D vector (three Eulerian angles, 2 plane coordinates). The parameter space is denoted by T .
4. s_i^o and $\mathbf{c}_i^o$ are constants related to suboptimal normalization of the individual experimental images: s_i^o is an overall scale factor for the signal brightness, and $\mathbf{c}_i^o$ describes non-zero background mean.
5. $\mathbf{N}_i \in \mathbb{R}^J$ is additive, independent and zero-mean Gaussian noise with standard deviation σ^o .

2.2. The optimization task

To make effective use of the data model in (1), we estimate $\mathbf{V}_1^o, \mathbf{V}_2^o, \dots, \mathbf{V}_K^o$ by way of maximum likelihood estimation. We view the estimation problem as a missing data problem, where the missing data associated with $\mathbf{X}_i$ are the position Φ_i and the random index κ_i . The associated $\mathbf{V}_{\kappa_i}^o$ is viewed as an unknown parameter, not as missing data. So, the complete data set is

$$(\mathbf{X}_i, \Phi_i, \kappa_i), \quad i=1, 2, \dots, N, \quad (2)$$

a random sample of $(\mathbf{X}, \Phi, \kappa)$.

Note that this random variable has one discrete component, to wit k , and two continuous components. The joint distribution may then be written as

$$P(\kappa=k) f(\mathbf{X}_i, \varphi|k) = \pi_k^o f(\varphi|k) f(\mathbf{X}_i|\varphi, k), \quad (3)$$

thus defining the probability vector π^o , which represents the unknown distribution of the data among the different classes. The distribution of the orientations and in-plane positions of the images is modelled by $f(\varphi|k)$. This distribution involves the assumption that particle picking has yielded roughly centred particles with residual offsets according to a two-dimensional Gaussian, centred at the origin. The corresponding formulae have been described in detail previously (Scheres et al., 2007a) and will not be repeated here. According to the noise model in (1), we calculate $f(\mathbf{X}_i|\varphi, k)$ as follows:

$$f(\mathbf{X}_i|\varphi, k) = (\sqrt{2\pi}\sigma)^{-J} \exp \left\{ \frac{\|\mathbf{X}_i - s_i R_\varphi \mathbf{V}_k - c_i\|^2}{-2\sigma^2} \right\}. \quad (4)$$

The marginal pdf of $\mathbf{X}_i$ is then a mixture,

$$f(\mathbf{X}_i) = \sum_{k=1}^K \pi_k^o \int_T f(\mathbf{X}_i|\varphi, k) f(\varphi|k) d\varphi, \quad (5)$$

and the maximum likelihood estimation problem is to find those parameters Θ^* that maximize the logarithm of the joint probability of observing the entire set of images $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$:

$$\Theta^* = \arg \max_{\Theta} \sum_{i=1}^N \log \sum_{k=1}^K \pi_k \int_T f(\mathbf{X}_i|\varphi, k) f(\varphi|k) d\varphi. \quad (6)$$

Note that, apart from the parameters describing $f(\varphi|k)$, the unknown parameter set Θ contains $\sigma^o, \pi^o, s^o, c^o$ and $\mathbf{V}^o$, with

$$\begin{aligned} \pi^o &\equiv (\pi_1^o, \pi_2^o, \dots, \pi_{K-1}^o), \\ s^o &\equiv (s_1^o, s_2^o, \dots, s_N^o), \\ c^o &\equiv (c_2^o, c_2^o, \dots, c_N^o), \\ \mathbf{V}^o &\equiv (\mathbf{V}_1^o, \mathbf{V}_2^o, \dots, \mathbf{V}_K^o). \end{aligned} \quad (7)$$

2.3. The MLn3D algorithm

The log-likelihood target function may be optimized using expectation maximization (Dempster et al., 1977). In the E-step of this iterative algorithm, a lower bound $Q(\Theta; \Theta^{\text{old}})$ to the log-likelihood is built based on the current model parameter set Θ^{old} :

$$Q(\Theta; \Theta^{\text{old}}) = \sum_{i=1}^N \sum_{k=1}^K \int_T \tau_{ik\varphi}^{\text{old}} \times \{ \log \pi_k + \log f(\varphi|k) + \log f(X_i|\varphi, k) \} d\varphi, \quad (8)$$

where, $\tau_{ik\varphi}^{\text{old}}$ is the probability distribution of the hidden variables conditioned on the observed measurements. This distribution may be calculated as:

$$\tau_{ik\varphi}^{\text{old}} = \frac{\pi_k^{\text{old}} f^{\text{old}}(\varphi|k) f(X_i|\varphi, k)}{\sum_{k'=1}^K \pi_{k'}^{\text{old}} \int_T f^{\text{old}}(\varphi'|k') f(X_i|\varphi', k') d\varphi'}. \quad (9)$$

In the subsequent M-step of the algorithm, we optimize the lower bound with respect to all model parameters.

The updates of the mixing proportions π^{new} may be calculated independently from the updates of the other model parameters:

$$\pi_k^{\text{new}} = \frac{1}{N} \sum_{i=1}^N \int_T \tau_{ik\varphi}^{\text{old}} d\varphi. \quad (10)$$

The update of the other model parameters is more complicated. Because the $\log f(\mathbf{X}_i|\varphi, k)$ term in (8) depends on σ^o , s^o , c^o and $\mathbf{V}^o$, a strict expectation-maximization algorithm would have to optimize the lower bound for all these parameters simultaneously. Instead of trying to solve the corresponding non-linear system of equations, we implemented the alternative that is outlined below. We note that a similar approach has previously been taken in other maximum likelihood approaches in the field, where the lower bound depends simultaneously on the model parameters for the signal and the standard deviation of the noise (e.g. see Sigworth, 1998; Zeng et al., 2007). For reasons of computational speed, one typically updates these two parameters separately, but in a strict sense this is not guaranteed to optimize the log-likelihood function. In practice however, these subtle differences do not seem to interfere significantly with the convergence behaviour of existing algorithms, and in our experience thus far the MLn3D algorithm is no exception.

First, we set the partial derivatives of (8) with respect to s_i and c_i to zero and solve for these variables respectively, yielding:

$$s_i^{\text{new}} = \frac{\sum_{k=1}^K \int_T \tau_{ik\varphi}^{\text{old}} (X_i - c_i^{\text{old}}) \cdot (R_\varphi \mathbf{V}_k^{\text{old}}) d\varphi}{\sum_{k=1}^K \int_T \tau_{ik\varphi}^{\text{old}} \|R_\varphi \mathbf{V}_k^{\text{old}}\|^2 d\varphi}, \quad (11)$$

and

$$c_i^{\text{new}} = \frac{\sum_{k=1}^K \int_T \tau_{ik\varphi}^{\text{old}} \sum_{j=1}^J (X_i - s_i^{\text{old}} R_\varphi V_k^{\text{old}})_j d\varphi}{J \sum_{k=1}^K \int_T \tau_{ik\varphi}^{\text{old}} d\varphi}, \quad (12)$$

where $(a) \cdot (b)$ denotes the dot product between a and b , and $(a)_j$ denotes the j^{th} pixel of a .

Secondly, we use the updated s_i^{new} and c_i^{new} in the updates of σ and $\mathbf{V}$. Again setting partial derivatives to zero and solving for σ yields:

$$(\sigma^{\text{new}})^2 = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K \int_T \tau_{ik\varphi}^{\text{old}} \times \|X_i - s_i^{\text{new}} R_\varphi V_k^{\text{old}} - c_i^{\text{new}}\|^2 d\varphi, \quad (13)$$

and the updated $\mathbf{V}$ may be obtained by solving the following K least-squares problems separately:

$$\min \sum_{i=1}^N \int_T \tau_{ik\varphi}^{\text{old}} \|X_i - s_i^{\text{new}} R_\varphi V_k^{\text{old}} - c_i^{\text{new}}\|^2 d\varphi, \quad (14)$$

to which purpose we use a modified ART algorithm as presented previously (Scheres et al., 2007a).

Finally since the overall brightness of $\mathbf{V}$ is directly correlated to the values of s_i^{new} , for each reference k we constrain the average image brightness to one (i.e.

$$\int_T \tau_{ik\varphi}^{\text{old}} s_i^{\text{new}} d\varphi / \int_T \tau_{ik\varphi}^{\text{old}} d\varphi = 1).$$

2.4. Implementation

The above exposed algorithm was implemented in the open-source package XMIPP (Sorzano et al., 2004b), and may be accessed conveniently as an expert option in the recently implemented standardized protocols (Scheres et al., 2008). Because the integration over T (which in practice is replaced by a Riemann sum over a discrete grid) is extremely computation-intensive, we also implemented a reduced-space approach as presented previously (Scheres et al., 2005). Furthermore, besides the proposed algorithm for 3D classification, we implemented a related 2D classification algorithm. In that case, instead of optimizing (8) with respect to 3D-structures $\mathbf{V}_1, \dots, \mathbf{V}_K$, one optimizes this function with respect to 2D images $\mathbf{A}_1, \dots, \mathbf{A}_K$. The algorithm remains basically the same, except for the fact that in this case R_φ represents an in-plane transformation (parametrized by a single rotation angle and two in-plane coordinates), and the least-squares problems in (14) are replaced by the following update formula:

$$A_k^{\text{new}} = \frac{\sum_{i=1}^N \int_T \tau_{ik\varphi}^{\text{old}} R_\varphi^{-1} s_i^{\text{new}} (X_i - c_i^{\text{new}}) d\varphi}{\sum_{i=1}^N \int_T \tau_{ik\varphi}^{\text{old}} (s_i^{\text{new}})^2 d\varphi}. \quad (15)$$

3. Results

3.1. 70S ribosome

To illustrate the usefulness of the MLn3D algorithm, we first show the results obtained with conventional ML3D classification (Scheres et al., 2007a) on a 70S ribosome complex from *E. coli* programmed with mRNA and containing deacylated tRNA^{fMet} in the P site and fMetLeu-tRNA^{Leu} in the A site. An initial data set of 69,262 individual particles was used to calculate a cryo-EM density map to 13 Å resolution. The overall configuration revealed an unratcheted ribosome with strong tRNA density in the P site, but scattered density in the A and E sites, not accounting for full tRNAs (not shown). The poor representation of the A and E-site tRNAs in the map suggested a low occupancy of these sites and/or the presence of a mixture of different positions. The latter possibility prompted us to perform an unsupervised ML3D classification of these data. However, no apparent conformational differences could be observed among the resulting maps when using four classes (Figure 1a). Instead, pairwise difference maps consisted of positive or negative density throughout the ribosome particle.

Starting from the same four seeds, the MLn3D algorithm yielded maps representing ribosomes in distinct structural states (Figure 1b). Three of the classes (together accounting for approximately 80% of the particles) showed the ribosome in an unratcheted state, while the fourth class (the remaining 20% of the particles) revealed a ratcheted ribosome. Separate refinements of the two classes to higher resolution revealed that the tRNAs in the unratcheted ribosome are positioned at the classical A and P sites, while the ratcheted ribosome showed a previously unobserved conformation with tRNAs in the hybrid A/P and P/E sites (see (Julián et al., 2008) for details).

The MLn3D-refined values for the background mean and signal brightness of every experimental particle were then used to analyse *a posteriori* why the ML3D run had failed. Histograms of these values for all images assigned to each of the four classes indicated that the conventional algorithm had indeed separated the images based on background mean as well as on image brightness, while, as expected, no such separation could be detected for the MLn3D algorithm (Figure 2a-d). A similar observation could also be made without the MLn3D-refined values of the normalization parameters. Radial average density profiles of all unaligned experimental images assigned to each of the four ML3D classes already hinted at a separation based on differences in signal brightness and/or background mean (Figure 2e-f). Finally, a visual inspection of images with relatively high or low refined values for the signal brightness or background mean suggested that neighbouring particles may be related to variations in background mean as well as image brightness, while differences in ice thickness or defocus values mainly affect image brightness (results not shown).

3.2. RNA polymerase II

The second test case concerns human RNA polymerase II in complex with the inhibitory human Alu RNA (Mariner et al., 2008). Application of the conventional ML3D algorithm with two classes yielded the maps that are depicted in Figure 3a. In this case, some putative conformational variability could be discerned between the resulting maps, but the absolute differences were relatively small. Much larger differences were obtained with the MLn3D algorithm (Figure 3b), which was again started from the same seeds as used for the conventional ML3D classification. In this case, the difference map showed specific regions of strong positive and negative density, which are indicative of a separation of the data according to conformational variability. The largest differences are located at the clamp of RNA polymerase II and around its DNA/RNA hybrid binding site. Smaller differences can be seen in the stalk domain and between the clamp and the stalk. These differences in conformation could be relevant to the different binding and inhibiting properties of the Alu RNA. Further

interpretation of the functional significance of the different hRNAPII/Alu RNA conformers will be presented elsewhere.

In this case, the posterior analysis of the refined normalization parameters showed that the conventional ML3D algorithm had, at least partially, separated the data based on differences in background mean alone rather than also on signal brightness (Figure 4). Again, no signs of separation based on normalization errors could be detected for the MLn3D algorithm. Furthermore, radial average density profiles of the two ML3D-classes showed a marked discontinuity at the radius used for the background circle in the normalization protocol (see Methods), directly linking the ML3D classification results with the normalization of the individual images. Also in this case, a visual inspection of the images with relatively high or low refined values for the background mean indicated that this variation may be related to the presence of neighbouring particles (not shown).

4. Discussion

The key to the advantage of maximum likelihood approaches over conventional refinement techniques lies in a more adequate statistical data model for 3D-EM images. In an intuitive manner, the explicit description of the abundant experimental noise allows to discern between situations where one is confident about the assignment of missing data items (e.g. the unknown orientation of a particle with respect to its template) and situations where based on the current model such confidence is not justified. Instead of taking “hard” decisions in the form of discrete assignments, in the maximum likelihood approach one calculates probabilities for all possible assignments, and the model parameters are obtained as a probability-weighted averages over all possibilities. However, if the statistical model does not describe the experimental data adequately, incorrect probability distributions will lead to suboptimal behaviour of the refinement approach. Therefore, a careful consideration of the underlying data model is of crucial importance for the potentials of the statistical approach.

As mentioned in the introduction, the squared distance metric that underlies all currently employed maximum likelihood approaches in the field may be seriously affected by variations in background mean or signal brightness among the data. Such variations may be relatively common in cryo-EM data, where abundant levels of noise complicate the process of image normalization. In particular, differences in ice thickness or defocus value yield different signal-to-noise ratios in the particles, which upon normalization of the noise results in variations in the signal brightness. In addition, the presence of neighbouring particles or other artefacts in those areas used to estimate the power of the noise may affect both the background mean and the image brightness. The presence of normalization errors presents a handicap for the maximum likelihood approach compared to refinement techniques based on cross-correlation coefficients. In the latter, the normalized cross-correlation coefficient is invariant to the background mean and signal brightness. Therefore, although these variations in theory still result in ill-posed 3D reconstructions, in practice their effects on conventional refinement may often be ignored. Unfortunately, this is not the case for maximum likelihood refinements, as is illustrated by the results presented in this paper. For two structurally heterogeneous cryo-EM data sets we showed that normalization errors may affect ML3D classification to such an extent that they prevent the separation of the data into structurally homogeneous subsets.

This was our main motivation to propose an extended data model that accounts for normalization errors and to derive a corresponding expectation-maximization (-like) algorithm for the maximum likelihood classification of structurally heterogeneous projection data. The successful classification of the two cases shown indicates that the extended data model and the proposed algorithm may be useful assets to the field. Given this example, it should be relatively easy to derive similar algorithms for other maximum likelihood approaches in the field, like

the 3D reconstruction of icosahedral viruses (Yin et al., 2003) or the alignment of 2D crystal images (Zeng et al., 2007). In addition, these principles could also be useful for maximum likelihood approaches that are yet to be proposed, for example for sub-tomogram averaging (Förster et al., 2008).

In conclusion, we foresee that the growing importance of statistical approaches in 3D-EM image processing will be accompanied by an increasing interest in their underlying data models. Experimental data may contain many more surprises that make our currently employed data models suboptimal. In that context, we hope that this paper may contribute to a continuing, community-wide discussion on better statistical models for 3D-EM image formation.

5. Materials and Methods

5.1. Ribosome preparation and electron microscopy

Ribosome samples were prepared as described in (Julián et al., 2008) and diluted to 32 nM final concentration. Cryo-EM grids were prepared following standard procedures and micrographs were taken in low-dose conditions on a JEM-2200FS electron microscope. Images were recorded on a 4k×4k CCD camera at a magnification of 67,368×, resulting in a 2.2 Å pixel size. Semi-automated particle picking from the SPIDER package (Frank et al., 1996) yielded 69,262 boxed particles of 160 × 160 pixels.

5.2. RNA polymerase II/ Alu RNA complex preparation and electron microscopy

Human RNA polymerase II (hRNAPII) was immunopurified from HeLa cell nuclei as previously described (Kostek et al., 2006). Alu RNA was provided by James Goodrich's laboratory (Mariner et al., 2008). hRNAPII was diluted to a final concentration of 60nM and incubated with 120nM Alu RNA. Cryo-EM grids were prepared according to standard procedures. EM data were collected on films (Kodak SO163) in a Tecnai 20F microscope (FEI) operated at 200kV and 50,000× magnification, under low dose conditions. Micrographs were digitized with a Nikon Super Coolscan 8000 with a 12.71 μm raster size resulting in a pixel size of 2.54 Å. The boxer software from EMAN (Ludtke et al., 1999) was used to pick semi-automatically 31,219 (120 × 120 pixel) particle images.

5.3. Image processing

All subsequent image processing operations were performed in the Xmipp package (Sorzano et al., 2004b). To reduce the computational costs of the maximum likelihood refinements, all data were downsampled using B-spline interpolation. The ribosome data were scaled to images of 64 × 64 pixels with a final pixel size of 5.5 Å/pixel; the RNA polymerase II data were scaled to 60 × 60 pixels with a final pixel size of 5.08 Å/pixel. All downsampled images were normalized using the following protocol for every image: (i) a background area was defined as those pixels outside a central, circular area of the image with a user-defined radius; (ii) a least-square plane was fitted through the pixels in the background area and subtracted from the entire image; and (iii) the resulting image was divided by the remaining standard deviation of the pixels in the background area. The radius of the background area circle was set to 30 pixels for the ribosome data and to 28 pixels for the hRNAPII data.

ML3D classifications were performed as described previously (Scheres et al., 2007a). For the seed generation the initial, average 3D reconstruction of all ribosome particles was low-pass filtered to 80 Å, the initial map for the hRNAPII data was filtered to 75 Å. The MLn3D runs were started from the same seeds as the conventional ML3D classifications, and all multi-reference refinements were stopped after twenty iterations.

Acknowledgments

We thank the Barcelona and the Galicia Supercomputing Centers (BSC-CNS and CESGA) for providing computer resources, James Goodrich for providing the human Alu RNA and Cameron L. Noland for his contribution to data collection in the hRNAPII study. Funding was provided by the Spanish Ministry of Science (CSD2006-00023, BIO2007-67150-C03-1/3) and Comunidad de Madrid (S-GEN-0166-2006), the European Union (FP6-502828), the US National Heart, Lung and Blood Institute and the National Institutes of Health (R01 HL070472, R01 GM63072). E.N. is a Howard Hughes Medical Institute investigator. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung and Blood Institute or the National Institutes of Health.

References

- Dempster A, Laird N, Rubin D. Maximum-likelihood from incomplete data via the em algorithm. *J Royal Statist Soc Ser B* 1977;39(1):1–38.
- Doerschuk PC, Johnson JE. Ab initio reconstruction and experimental design for cryo electron microscopy. *IEEE Transactions on Information Theory* 2000;46(5):1714–29.
- Förster F, Pruggnaller S, Seybert A, Frangakis AS. Classification of cryo-electron sub-tomograms using constrained correlation. *J Struct Biol Mar*;2008 161(3):276–286. [PubMed: 17720536]
- Frank J, Radermacher M, Penczek P, Zhu J, Li Y, Ladjadj M, Leith A. Spider and web: processing and visualization of images in 3d electron microscopy and related fields. *J Struct Biol Jan/Feb*;1996 116(1):190–199. [PubMed: 8742743]
- Grob P, Cruse MJ, Inouye C, Peris M, Penczek PA, Tjian R, Nogales E. Cryo-electron microscopy studies of human tftid: conformational breathing in the integration of gene regulatory cues. *Structure Mar*; 2006 14(3):511–520. [PubMed: 16531235]
- Julián P, Konevega AL, Scheres SHW, Lázaro M, Gil D, Wintermeyer W, Rodnina MV, Valle M. Structure of ratcheted ribosomes with trnas in hybrid states. *Proc Natl Acad Sci U S A Nov*;2008 105(44):16924–16927. [PubMed: 18971332]
- Kostek SA, Grob P, De Carlo S, Lipscomb JS, Garczarek F, Nogales E. Molecular architecture and conformational flexibility of human rna polymerase ii. *Structure Nov*;2006 14(11):1691–1700. [PubMed: 17098194]
- Lee J, Doerschuk PC, Johnson JE. Exact reduced-complexity maximum likelihood reconstruction of multiple 3-d objects from unlabeled unoriented 2-d projections and electron microscopy of viruses. *IEEE Trans Image Process* 2007;16(12):2865–78. [PubMed: 18092587]
- Leschziner AE, Nogales E. Visualizing flexibility at molecular resolution: Analysis of heterogeneity in single-particle electron microscopy reconstructions. *Annu Rev Biophys Biomol Struct* 2007;36:43–62. [PubMed: 17201674]
- Ludtke SJ, Baldwin PR, Chiu W. Eman: semiautomated software for high-resolution single-particle reconstructions. *J Struct Biol Dec*;1999 128(1):82–97. [PubMed: 10600563]
- Mariner PD, Walters RD, Espinoza CA, Drullinger LF, Wagner SD, Kugel JF, Goodrich JA. Human alu rna is a modular transacting repressor of mrna transcription during heat shock. *Mol Cell Feb*;2008 29(4):499–509. [PubMed: 18313387]
- Nickell S, Beck F, Korinek A, Mihalache O, Baum eister W, Plitzko JM. Automated cryoelectron microscopy of “single particles” applied to t he 26s proteasome. *FEBS Lett* 2007;581(15):2751–6. [PubMed: 17531228]
- Nickell S, Kofler C, Leis AP, Baumeister W. A visual approach to proteomics. *Nat Rev Mol Cell Biol Mar*;2006 7(3):225–230. [PubMed: 16482091]
- Pascual-Montano A, Donate LE, Valle M, Brcena M, Pascual-Marqu RD, Carazo JM. A novel neural network technique for analysis and classification of em single-particle images. *J Struct Biol* 2001;133(23):233–45. [PubMed: 11472094]
- Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. *Nature Dec*;2007 450(7172):973–982. [PubMed: 18075576]
- Scheres SHW, Gao H, Valle M, Herman GT, Eggermont PPB, Frank J, Carazo JM. Disentangling conformational states of macromolecules in 3d-em through likelihood optimization. *Nat Methods* 2007a;4(1):27–9. [PubMed: 17179934]

- Scheres SHW, Nunez-Ramirez R, Gomez-Llorente Y, San Martin C, Eggermont PPB, Carazo JM. Modeling experimental image formation for likelihood-based classification of electron microscopy data. *Structure* 2007b;15(10):1167–77. [PubMed: 17937907]
- Scheres SHW, Nunez-Ramirez R, Sorzano COS, Carazo JM, Marabini R. Image processing for electron microscopy single-particle analysis using xmipp. *Nat Protoc* 2008;3(6):977–90. [PubMed: 18536645]
- Scheres SHW, Valle M, Carazo JM. Fast maximum-likelihood refinement of electron microscopy images. *Bioinformatics* 2005;21:ii243–ii244. [PubMed: 16204112]
- Sigworth FJ. A maximum-likelihood approach to single-particle image refinement. *J Struct Biol* 1998;122(3):328–39. [PubMed: 9774537]
- Sorzano COS, de la Fraga LG, Clackdoyle R, Carazo JM. Normalizing projection images: a study of image normalizing procedures for single particle three-dimensional electron microscopy. *Ultramicroscopy* 2004a;101(24):129–38. [PubMed: 15450658]
- Sorzano COS, Marabini R, Velquez-Muriel J, Bilbao-Castro JR, Scheres SHW, Carazo JM, Pascual-Montano A. Xmipp: a new generation of an open-source image processing package for electron microscopy. *J Struct Biol* 2004b;148(2):194–204. [PubMed: 15477099]
- Stagg SM, Lander GC, Quispe J, Voss NR, Cheng A, Bradlow H, Bradlow S, Carragher B, Potter CS. A test-bed for optimizing high-resolution single particle reconstructions. *J Struct Biol* Jul;2008 163(1): 29–39. [PubMed: 18534866]
- Stark H, Lührmann R. Cryo-electron microscopy of spliceosomal components. *Annu Rev Biophys Biomol Struct* 2006;35:435–457. [PubMed: 16689644]
- Vogel RH, Provencher SW. Three-dimensional reconstruction from electron micrographs of disordered specimens. ii. implementation and results. *Ultramicroscopy* 1988;25(3):223–39. [PubMed: 3212838]
- Yin Z, Zheng Y, Doerschuk PC, Natarajan P, Johnson JE. A statistical approach to computer processing of cryo-electron microscope images: virion classification and 3-d reconstruction. *J Struct Biol* 2003;144(12):24–50. [PubMed: 14643207]
- Yu X, Jin L, Zhou ZH. 3.88 Å structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. *Nature* May;2008 453(7193):415–419. [PubMed: 18449192]
- Zeng X, Stahlberg H, Grigorieff N. A maximum likelihood approach to two-dimensional crystals. *J Struct Biol* Dec;2007 160(3):362–374. [PubMed: 17964808]
- Zhang X, Settembre E, Xu C, Dormitzer PR, Bellamy R, Harrison SC, Grigorieff N. Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction. *Proc Natl Acad Sci U S A* Feb;2008 105(6):1867–1872. [PubMed: 18238898]

Figure 1.

Classification of the ribosome particles. **a** Results obtained with the conventional ML3D classification; **b** results obtained with the MLn3D algorithm. The first to fourth columns from the left show the maps obtained for classes 1-4, respectively. To facilitate comparison between the different classes and runs, all density maps are displayed at the same threshold. The 50S ribosomal subunits are coloured blue, 30S subunits yellow. The fourth class of the MLn3D run yields a different structure compared to the other classes: the 30S subunit is in a ratcheted conformation (indicated with an arrow) and shows tRNA density at the ribosomal exit site (shown in orange and indicated with an ellipse). The fifth column shows the positive (green) and negative (red) difference maps at 4 σ between the corresponding maps from classes 3 and 4.

Figure 2.

Posterior analysis of the classes for the ribosome data. **a** Histograms of the refined background means for the four classes obtained with ML3D classification of the ribosome data (class 1 in solid black; class 2 in solid grey; class 3 in dashed black; class 4 in dashed grey). **b** As in a, but for the MLn3D classification. **c** Histograms of the refined signal brightness for the four classes obtained with ML3D classification. **d** As in c, but for the MLn3D classification. **e** Radial average density profiles for the four classes obtained with ML3D classification. **f** As in e, but for the MLn3D classification.

Figure 3.

Classification of the RNA polymerase II/Alu RNA complex. **a** Results obtained with the conventional ML3D classification; **b** results obtained with the MLn3D algorithm. The first and second columns from the left show the maps obtained for classes 1 and 2, respectively. To facilitate comparison between the different classes and runs, all density maps are displayed at the same threshold, and the major characteristics of the complex are indicated in b. The third column shows the positive (green) and negative (red) difference maps at 4σ between the maps from the two classes of both runs.

Figure 4.

Posterior analysis of the classes for the RNA polymerase II data. **a** Histograms of the refined background means for the two classes obtained with ML3D classification of the RNA polymerase II data (class 1 in black; class 2 in grey). **b** As in a, but for the MLn3D classification. **c** Histograms of the refined signal brightness for the classes obtained with ML3D classification. **d** As in c, but for the MLn3D classification. **e** Radial average density profiles for the two classes obtained with ML3D classification. **f** As in e, but for the MLn3D classification.