Supplementary materials for *Bayesian inference of initial models in cryo-electron microscopy using pseudo-atoms*

Paul Joubert [1*] and Michael Habeck [1,2*]

[1]Felix-Bernstein Institute for Mathematical Statistics, Georg-August-Universität Göttingen
[2]Max Planck Institute for Biophysical Chemistry, Göttingen

1 FSC curves

Figure 1 shows the FSC curves between the references and the reconstructions.

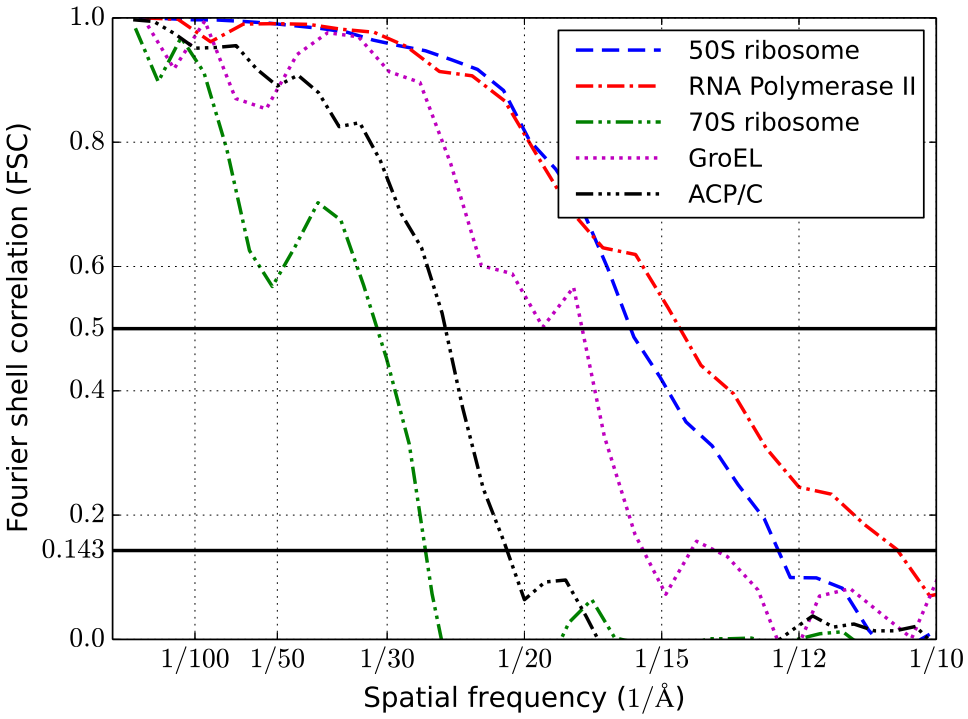

Figure 1: FSC curves comparing the references with the reconstructions. The references are the same as shown in the main text, i.e. for the 50S ribosome, RNA Polymerase II and GroEL they are created from the atomic structures at 25 Å, 20 Å and 20 Å respectively, for the 70S ribosome the reference was obtained by the PRIME algorithm and for the last structure (APC/C) the reference is from another publication (EMD-2354). The normalized cross-correlations for the same five pairs of structures are 0.990, 0.966, 0.900, 0.927 and 0.902 respectively.

*to whom correspondence should be addressed

2 Movies

The first movie (movie1.avi) shows multiple pseudo-atomic models of the same structure (RNA Polymerase II) with increasing numbers of pseudo atoms. It demonstrates that a small number of pseudo-atoms are sufficient for representing low-resolution structures, and that the number of model parameters required are orders of magnitude fewer than with the standard grid-based representation.

The second movie (movie2.avi) shows a reconstruction of the 50S ribosome, from the initial random model to the final model. The trajectories of the individual pseudo-atoms as well as those of the individual rotations can be seen. For the rotations, only the projection direction is shown (the first two Euler angles), not the in-plane rotation component (the third Euler angle).

3 Computing the posterior distribution

Here we give a more formal description of the different components of the Bayesian framework, starting with the data $\mathcal{D}$.

The i'th deconvolved class average is a non-negative grayscale image $(I_{ij})_{1\leq j\leq n}$ with 2D pixel coordinates[1] $(x^o_{ij})_{1\leq j\leq n}$, where the pixels are indexed by j, and n is the number of pixels per image. The I_{ij}'s for a given i are then multiplied by a constant scaling factor $\alpha \geq 0$, and rounded to the nearest integer to obtain $y_{ij} := \text{round}(\alpha I_{ij})$. The constant α is chosen such that $\sum_j y_{ij} \approx C$ for a previously fixed constant C. The I_{ij}'s are discarded, and we continue with the y_{ij}'s.

The observed data is now considered as 2D points, with y_{ij} points at each pixel centered at x^o_{ij}. Every pixel gives rise to a y_{ij}-dimensional data vector

$$d_{ij} = [x^o_{ij1}, \ldots, x^o_{ijl}, \ldots, x^o_{ijy_{ij}}]$$

with identical entries $x^o_{ijl} = x^o_{ij}$, where l runs from 1 to y_{ij}. All the data vectors d_{ij} together form the observed data $\mathcal{D} = x^o = \{x^o_{ijl}\}$.

Before describing the forward model, we introduce the latent variables. These are the assignments z as typically used for Gaussian mixture models, and the missing z-components x^m.

The assignments $z = \{z_{ijl}\}$ consist of one assignment z_{ijl} for each point x_{ijl}, indicating the mixture component responsible for generating the point. We use 1-of-K notation, whereby z_{ijl} is a length K vector $(z_{ijlk})_k$ with $z_{ijlk} \in \{0, 1\}$ and $\sum_k z_{ijlk} = 1$. I.e. the k for which $z_{ijlk} = 1$ indicates the component that generated x_{ijl}.

The missing components $x^m = \{x^m_{ijl}\}$ consist of the z-component x^m_{ijl} for each sampled 3D point $x_{ijl} = [x^o_{ijl}\ x^m_{ijl}]$. Since we only observe the first two coordinates, i.e. x^o_{ijl}, the z-component is referred to as missing.

We write $\mathcal{Z} = \{z, x^m\}$ for all latent variables together.

Given a model parameterized as described above, the observed data can be generated as follows: for a given direction i compute the 2D density $I_i(x)$. Sample C points from this density, and create a 2D histogram with bins centered at the grid points x^o_{ij}. Then y_{ij} is defined as the number of points in the j'th bin. If the grid is sufficiently fine, we can make the following assumption to simplify the forward model: all the points in the j'th bin are replaced by the center of the bin, x^o_{ij}. In other words we assume that we sampled y_{ij} copies of x^o_{ij}, for each j.

This forward model can also be described in a slightly different way which will be used below in formulating the sampling algorithm. Instead of first projecting the 3D density to 2D, and then sampling C points, we could equivalently first sample C 3D points, and then project them to 2D. The 3D points sampled from the rotated density are denoted by $x_{ijl} = [x^o_{ijl}\ x^m_{ijl}]$, where $x^o_{ijl} \in \mathbb{R}^2$, and $x^m_{ijl} \in \mathbb{R}$. Projection along the z-axis means discarding the z-component x^m_{ijl}, i.e. x_{ijl} is projected to x^o_{ijl}.

[1] The superscripts o and m stand for *observed* and *missing* respectively, as will be explained shortly.

The above forward model describes the extended likelihood for the data and latent variables:

$$p(\mathcal{D}, \mathcal{Z}|\theta) = p(x^o, x^m, z|\mu, s, w, R, t) \tag{1}$$

$$= \prod_{ijlk} w_k^{z_{ijlk}} f(x_{ijl}|R_i\mu_k + t_i|\frac{1}{s}I)^{z_{ijlk}}. \tag{2}$$

Marginalizing out the latent variables gives the data likelihood:

$$p(\mathcal{D}|\theta) = p(x^o|\mu, s, w, R, t) \tag{3}$$

$$= \iint p(x^o, x^m, z|\mu, s, w, R, t)\mathrm{d}x^m\mathrm{d}z \tag{4}$$

$$= \int p(\mathcal{D}, \mathcal{Z}|\theta)\mathrm{d}\mathcal{Z}. \tag{5}$$

The prior is assumed to factorize over the model parameters:

$$p(\theta) = p(\mu, s, w, R, t) = p(\mu)p(s)p(w)p(R)p(t), \tag{6}$$

where the means $p(\mu) = \prod_k f(\mu_k|0, \frac{1}{r}I)$ follow normal distributions, the precision $p(s) \propto s^{a-1}e^{-bs}$ a gamma distribution, the weights $p(w) \propto \prod_k w_k^{\lambda-1}$ a Dirichlet distribution, the rotations are distributed uniformly, and the translations $p(t) = \prod_i f(t_i|0, \frac{1}{r}I)$ follow normal distributions. The hyperparameters r, a, b and λ are kept fixed.

4 Gibbs sampling

Here we give the equations for performing Gibbs sampling. These are uniquely determined given the forward model and prior defined above. We use Gibbs sampling to sample from the extended posterior $p(\mathcal{Z}, \theta|\mathcal{D})$, and then discard the latent variables $\mathcal{Z}$ to obtain samples from the posterior. The extended posterior is proportional to the product of the extended likelihood (Equation 2) and the prior (Equation 6). To sample from this extended posterior using Gibbs sampling we compute the conditional distribution for each of the parameters, conditioned on all the other parameters. They are all standard distributions (Gaussian, multinomial, Dirichlet and gamma) except for the rotations, which are of the form $\exp[\mathrm{tr}(A^T R)]$.

The conditional for each assignment is a multinomial distribution:

$$p(z_{ijl}|x^o_{ijl}, w, \mu, s, R, t) = \prod_k w_k^{z_{ijlk}} f(x^o_{ijl}|P(R_i\mu_k + t_i), \frac{1}{s}I)^{z_{ijlk}}.$$

where

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

The conditional for the missing z-component for a single point is a 1D normal distribution:

$$p(x^m_{ijl}|x^o_{ijl}, z_{ijl}, \mu, s, R, t) = \prod_k f(x^m_{ijl}|P_z(R_i\mu_k + t_i), \frac{1}{s})^{z_{ijlk}},$$

where

$$P_z = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}.$$

The conditional for the weights is a Dirichlet distribution:

$$p(w|x^m, z, \mu, s, R, t) \propto \prod_k w_k^{n_k+\lambda-1}$$

where

$$n_k = \sum_{ijl} z_{ijlk}.$$

The conditional for each mean is a normal distribution:

$$p(\mu_k | x^m, z, s, w, R, t, x^o) = f(\mu_k | \mu, \Sigma)$$

where

$$\begin{aligned} \mu &= \frac{s}{sn_k + r} \sum_{ijl} z_{ijlk} R_i^T (x_{ijl} - t_i) \\ \Sigma &= \frac{1}{sn_k + r} I. \end{aligned}$$

The conditional for the precision is a gamma distribution:

$$p(s | x^m, z, \mu, R, t) \propto s^{a'-1} e^{-b's},$$

where

$$\begin{aligned} a' &= a + \frac{3}{2}N \\ b' &= b + \frac{1}{2} \sum_{ijkl} z_{ijkl} \left[\|x^o_{ijl} - P(R_i \mu_k + t_i)\|^2 + (x^m_{ijl} - P_z(R_i + t_i)\mu_k)^2 \right]. \end{aligned}$$

The conditional for each rotation is

$$p(R_i | t_i, x^m, \mu, s, x^o) \propto \exp\left[\mathrm{tr}(A_i^T R_i) \right],$$

where

$$A_i = s \sum_{jlk} z_{ijlk} (x_{ijl} - t_i) \mu_k^T.$$

The conditional for each translation is a normal distribution:

$$p(t_i | x, \mu, s, R) = f(t_i | \mu, \Sigma)$$

where

$$\begin{aligned} \mu &= \frac{\sum_{jlk} z_{ijlk} (x_{ijl} - R_i \mu_k)}{\sum_{jlk} z_{ijlk}} \\ \Sigma &= \frac{1}{s \sum_{jlk} z_{ijlk}} I. \end{aligned}$$

5 Prior hyperparameters

The prior distribution on the pseudo-atom size σ is given by a gamma distribution over the precision $s = 1/\sigma^2$:

$$p(s) = \frac{\beta^\alpha}{\Gamma(\alpha)} s^{\alpha-1} e^{-\beta s}.$$

The mean α/β and variance α/β^2 of this distribution encodes our prior knowledge about the size of the pseudo-atoms. In Figure 2 on the left are a few examples for different values of α and β, with β chosen such that the mean is $1/10^2$ (i.e. $\beta = 10^2\alpha$). In the same figure on the right are the effects of the different choices of the hyperparameters on a reconstruction of the 50S ribosome from simulated data. It shows that

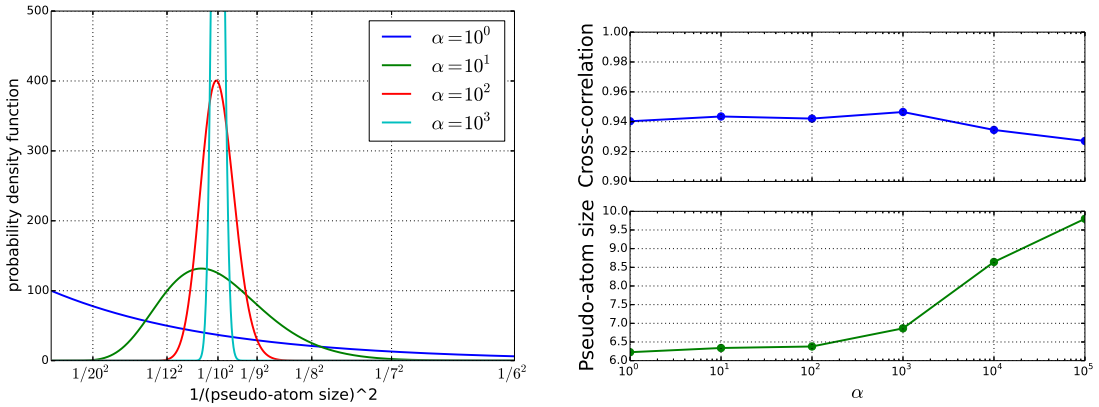

Figure 2: Varying the prior on the pseudo-atom size. On the left are different prior distributions over the pseudo-atom precision s, which is related to the pseudo-atom size σ by $s = 1/\sigma^2$. They correspond to a wide range of different choices of the hyperparameter α. The other hyperparameter β determining the prior distribution is chosen to ensure that the mean of s is $1/10^2$. Low values of α place effectively no restriction on the pseudo-atom size, while very high values of α restrict the pseudo-atom size to be very close to 10 Å. This can be seen on the right, where a 50S ribosome model inferred from simulated data is compared to a reference model. When a broad prior on the pseudo-atom size is used (i.e. α is low), then then final size is around 6 Å. As the prior becomes narrower (i.e. α is high), the final size tends to the mean prior value of 10 Å. The figure shows that for α in the range 1 to about 1000, the quality of the result as measured by the cross-correlation does not depend on the specific choice of α.

good results are obtained for all values of the hyperparameters, although the results deteriorate for very high values of α (above 1000). We conclude that the specific value of alpha is not very important for our algorithm, and recommend it to be chosen in the range 1 to 1000. We used $\alpha = 10$ for our experiments. The value of β can be chosen as was done here to ensure that the mean is 10, although a similar experiment shows that the exact value of β is also not very important. For our experiments we used $\beta = 10^2\alpha = 1000$.

The prior distribution on the pseudo-atom weights $w = \{w_k\}$ is a Dirichlet distribution

$$p(w) \propto \sum_{k=1}^{K} w_k^{\lambda-1}.$$

This distribution is parametrized by a single hyperparameter λ, which determines the allowable variation among the weights for the different pseudo-atoms. Higher values of λ lead to less variation.

In Figure 3 we show the effect of varying λ on the quality of a reconstruction using the same data as before. The figure shows that all values of λ in a wide range lead to similar results, although very small values of λ (0.1 and 1.0) give slightly worse results. We therefore recommend choosing λ in the range from 10 to 10^5. We used the value $\lambda = 1000$ for our experiments.

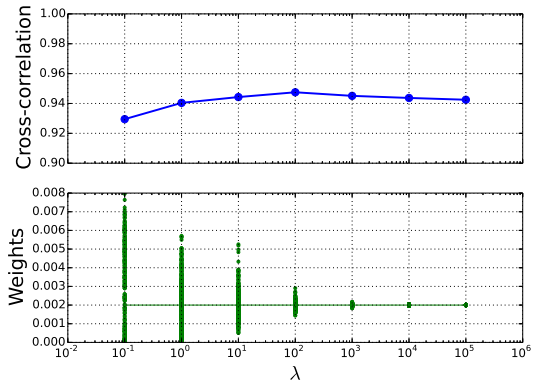

Figure 3: Varying the prior on the pseudo-atom weights. Multiple reconstructions using the same 50S ribosome data as before were performed using a range of values of λ, the hyperparameter for the weights. At the top are the resulting cross-correlations with the reference, measuring the quality of the inferred models. The cross-correlations are slightly lower for very low values of λ, but stay relatively constant for λ above 10. At the bottom are the individual pseudo-atom weights, showing that the variation in the weights decreases with increasing λ.