Journal of Structural Biology xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Structural Biology

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

A new algorithm for high-resolution reconstruction of single particles by electron microscopy

C.O.S. Sorzano[a,b,*], J. Vargas[c], J.M. de la Rosa-Trevín[d], A. Jiménez[a], D. Maluenda[a], R. Melero[a], M. Martínez[a], E. Ramírez-Aportela[a], P. Conesa[a], J.L. Vilas[a], R. Marabini[e], J.M. Carazo[a]

[a] Centro Nac. Biotecnología (CSIC), 28049 Cantoblanco, Madrid, Spain
[b] Univ. San Pablo – CEU, Campus Urb. Montepríncipe, 28668 Boadilla del Monte, Madrid, Spain
[c] Dept. Anatomy and Cell Biology, McGill Univ., Montreal, Canada
[d] SciLife Lab, Stockholm, Sweden
[e] Univ. Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain

ARTICLE INFO

Keywords:
Image processing
3D reconstruction
Angular assignment
Volume restoration

ABSTRACT

The Map Challenge organized by the Electron Microscopy Data Bank has prompted the development of an Xmipp high resolution reconstruction protocol (which we will refer to as *highres*) that is integrated in the software platform `Scipion`. In this work we describe the details of the image angular alignment and map reconstruction steps in our new method. This algorithm is similar to the standard projection matching approach with some important modifications, especially in the area of detecting significant features in the reconstructed volume. We show that the new method is able to produce higher resolution maps than the current *de facto* standard as measured by the Fourier Shell Correlation, the Monogenic Local Resolution and EMRinger.

1. Introduction

Single Particle Analysis of macromolecular structures by Electron Microscopy (EM) has become in the last few years one of the most successful techniques for Structural Biology (Nogales, 2016) due to its ability to achieve near-atomic resolution and to explore conformational flexibility, using low amounts of sample material. The automation of the image acquisition process at the microscope and the introduction of Direct Electron Detectors have allowed the recording of vast amounts of data whose analysis results in three-dimensional maps of the macromolecule under study from which structural models can be derived. However, the images acquired at the microscope are extremely noisy (Signal-to-Noise Ratios between 0.1 and 0.01). Such noisy measurements require robust data analysis methods.

The prerequisites to achieve near-atomic resolution include 1) a structurally homogeneous population of projection images obtained using the crucial 3D classification algorithms (Scheres, 2012), 2) a sufficiently good angular coverage to measure every region in Fourier space, and 3) a sufficiently good frequency coverage to measure every frequency and to preserve the microscope structural information at as high resolution as possible. To accomplish this last requirement, images are acquired at different defoci, especially at low defocus to preserve high frequency information. The 3D reconstruction process alternates between angular assignment and three-dimensional reconstruction to extract the maximum of structural information present at the micrographs. Overall, the whole problem can be seen as a regression in which the projection images are the data to be fitted, and the volume and the alignment parameters constitute the model.

All processing steps must be robust to high levels of noise and need to avoid overfitting (reconstruction artifacts that satisfy data constraints but that are either dominated by noise or that are far away from the best possible solution). Structural knowledge can also be incorporated in the analysis workflow, and most common ways are either using a Bayesian prior (as Relion did Scheres, 2012) or by regularization (for a review of regularization in 3D reconstruction, see Sorzano et al. (2017)).

Relion (Scheres, 2012), at the moment the most common method to refine maps in the field of Single Particle Analysis by EM, integrates the whole regression problem in a single functional that is optimized in a greedy fashion starting from an initial estimate of the volume to be reconstructed. This functional includes a Bayesian prior about the statistical distribution of the objects to be reconstructed (coefficients in Fourier space are independently distributed, with independent real and imaginary parts, Gaussianly distributed with zero-mean and a variance that is estimated from the data itself). Although in general, the prior is not accurate for macromolecular structures (Sorzano et al., 2015), it has

* Corresponding author.
E-mail address: coss@cnb.csic.es (C.O.S. Sorzano).

https://doi.org/10.1016/j.jsb.2018.08.002
Received 4 March 2018; Received in revised form 19 July 2018; Accepted 4 August 2018

the advantage that it is mathematically tractable and results in a low pass filter of the reconstructed volume similar to a Wiener filter (Scheres, 2012).

This Bayesian approach has currently dominated the cryoEM 3D reconstruction field, specially for its generalization to the very important task of multiple maps reconstruction as part of a 3D classification process. However, this approach is not the only valid strategy for map reconstruction. Actually, before the introduction in cryoEM of the Maximum Likelihood (Scheres et al., 2005) and Maximum *a posteriori* methods (Scheres, 2012), the standard approach to 3D angular alignment and reconstruction was the so-called "Projection Matching" (Penczek et al., 1992; Penczek et al., 1994). Recently, novel implementations of Maximum Likelihood based on GPU processing and stochastic gradient descent have significantly reduced the processing time (Punjani et al., 2017).

Let us refer to the parameters defining the angular alignment of the whole dataset as $\mathbf{\Theta}$ and to the reconstructed volume as $\mathbf{V}$. In an extremely simplified manner, we may think of the Maximum Likelihood method as an algorithm that minimizes

$$\mathbf{\Theta}^*, \mathbf{V}^* = \arg\min_{\mathbf{\Theta},\mathbf{V}} ||\widetilde{\mathbf{I}} - P_{\mathbf{\Theta}}\mathbf{V}||_W^2 \quad (P1)$$

where $\widetilde{\mathbf{I}}$ is the set of pixels from the acquired images (if the underlying algorithm allows an image to be at multiple angular orientations with different probabilities, then $\widetilde{\mathbf{I}}$ will contain multiple copies of the measured data and Θ will have several components devoted to the same image), $P_{\mathbf{\Theta}}$ is a projection operator that calculates the projections of the volume $\mathbf{V}$ along the directions and shifts specified by $\mathbf{\Theta}$ (depending on the specific implementation, this projection operator may include or not the aberrations caused by the electron microscope), and $||\cdot||_W$ is a weighted norm in which different pixels may be weighted differently according to some scheme adopted by the algorithm (the statistical distributions assumed for the noise and the alignment parameters automatically determine the form of this norm; this generic algorithmic framework may be adopted in real or Fourier space) (Sorzano et al., 2017). This problem is simply a data fidelity term (the reconstructed object has to be compatible with the acquired projections). The Bayesian approach adds *a priori* knowledge about the statistical distribution of the volumes being reconstructed that, in its turn, is translated into the minimization problem as an extra term that penalizes unlikely reconstructions

$$\mathbf{\Theta}^*, \mathbf{V}^* = \arg\min_{\mathbf{\Theta},\mathbf{V}} ||\widetilde{\mathbf{I}} - P_{\mathbf{\Theta}}\mathbf{V}||_W^2 + f(\mathbf{V}) \quad (P2)$$

The traditional approach in cryoEM, projection matching, decomposes the minimization in $P1$ in two subproblems that are minimized separately and iteratively (k denotes the iteration number)

$$\mathbf{\Theta}_k^* = \arg\min_{\mathbf{\Theta}} ||\widetilde{\mathbf{I}} - P_{\mathbf{\Theta}}\mathbf{V}_{k-1}^*||_{W_1}^2 \quad (P1a)$$

$$\mathbf{V}_k^* = \arg\min_{\mathbf{V}} ||\widetilde{\mathbf{I}} - P_{\mathbf{\Theta}_k^*}\mathbf{V}||_{W_2}^2 \quad (P1b)$$

In its most traditional approach, $\widetilde{\mathbf{I}}$ is restricted to have a single copy of the experimental measurements $\mathbf{I}$, that is, only one alignment parameter set is estimated per projection. The first subproblem ($P1a$) is called angular assignment (each experimental image is assigned a set of parameters that encodes its projection direction and in-plane alignment with respect to the current guess of the macromolecular structure). P1a strongly depends on the initial map used as reference. The second subproblem uses the assigned angles to update the 3D map of the macromolecule. In practice, the Maximum Likelihood (ML) and Maximum A Posteriori (MAP) problems are solved through a numerical technique called Expectation Maximization that boils down to an iterative scheme similar to angular assignment and 3D reconstruction iterations (Scheres et al., 2005; Scheres, 2012) (the prior in MAP affects the specific form of the iterative step in ($P1b$); the fact that a Gaussian distribution is the conjugate prior of the distribution chosen for the likelihood term helps to keep the mathematical complexity of the EM iteration tractable). In the approach with subproblems ($P1a$ and $P1b$), it is still possible to introduce *a priori* knowledge about the class of volumes being reconstructed through the so-called image restoration methods (Sorzano et al., 2017). An example of map restoration is to restrict maps to be members of certain subsets defining properties a good map should have (e.g., non-negative maps, maps with compact space support, etc.) (Sorzano et al., 2008). Projection onto Convex Sets (Carazo, 1992), which were early introduced in the EM field, was a form of incorporating this *a priori* information. However, statistical properties of the volumes (like a Bayesian prior Scheres, 2012) or any other known feature of the volume being reconstructed could also be used in a restoration scheme.

The ML approach introduced an important concept in the EM community: an experimental image may occupy more than one projection direction and in-plane alignment, but with different probability. This probability gives its weighting factor during the reconstruction process. The rationale behind this idea is that images are so noisy that the maximum correlation peak calculated during the alignment is prone to errors and allowing the image to sit at different angles gives it more opportunities to find its correct localization (ideally, the likelihood distribution for a single experimental image should converge to a delta, although this is not always the case in practice for all images). This idea of more than one location was further exploited for the blind construction of an initial model (Sorzano et al., 2015). On the other hand, the fact that all experimental images are, in principle, allowed to occupy all projection directions (with different probabilities) may cause that some projection directions with intrinsically more Signal-to-Noise Ratio are over-represented (Vargas et al., 2016; Vargas et al., 2017), what we refer to as the attraction problem (Sorzano et al., 2010).

In this work, instead of using the image likelihood as weight, we used its significance (which is the probability that a random image taken from the set has a cross-correlation smaller than the correlation of this experimental image), and we promote, as in projection matching, an angular assignment in which each image receives a single angular orientation. Despite the appeal of letting an image to occupy several angular orientations (due to the uncertainty introduced by noise), in reality an image is known to come from a single (although unknown) orientation. Angular assignment algorithms will very likely commit assignment errors. For the sake of argument, let us consider that the error rate of the angular assignment is 30% (that is, 30% of the particles are assigned an incorrect orientation). If we now allow for a second angular assignment, with different weights, we know that for the 70% of particles that were correctly assigned, this second assignment will be incorrect. And for the remaining 30%, only about 70% will be correctly assigned. From the whole set of assignments (that now is twice the size of the dataset because of the double angular assignment), only 45.5% ($=0.7{\cdot}0.5 + 0.3{\cdot}0.7{\cdot}0.5$) of it has a correct angular assignment. That is, by increasing the number of positions in which a particle might sit (as a measure to fight noise), we have decreased the accuracy of our angular assignment, possibly resulting in a low pass filtering of the reconstructed volume due to the incoherent averaging in the Fourier space. In principle, this situation should be alleviated by the fact that the different projection directions have different weights, and that this weight distribution would ideally be very spiky around the true angular assignment. However, this ideal situation is not always the case. In Fig. 1 we show the weight profile for an experimental projection of a ribosome when it is compared to the whole gallery of projections of the final reconstructed volume sampled every 5 degrees and with a maximum shift of 24 Å. More than 35 million combinations of orientation and positions were explored. Relion and *highres* both agreed on the angular assignment and shift of this particle, with a precision of less than 1° in the angles and 1 pixel in the shifts. Relion is based on the l_2 norm of the difference between the volume reprojection and the experimental projection with a frequency weight given by the noise variance, while *highres* is based on the correlation of the volume