Journal of Structural Biology 189 (2015) 213–219

Contents lists available at ScienceDirect

Journal of Structural Biology

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

A statistical approach to the initial volume problem in Single Particle Analysis by Electron Microscopy

C.O.S. Sorzano [a,b,*], J. Vargas [a], J.M. de la Rosa-Trevín [a], J. Otón [a], A.L. Álvarez-Cabrera [a], V. Abrishami [a], E. Sesmero [b], R. Marabini [c], J.M. Carazo [a]

[a] National Center of Biotechnology (CSIC), c/Darwin, 3, Campus Univ. Autonoma de Madrid, 28049 Cantoblanco, Madrid, Spain
[b] Bioengineering Lab., Univ. San Pablo CEU, Campus Urb. Monteprincipe s/n, 28668 Boadilla del Monte, Madrid, Spain
[c] Escuela Politecnica Superior, Univ. Autonoma de Madrid, Campus. Univ. Autonoma de Madrid, 28049 Cantoblanco, Madrid, Spain

ARTICLE INFO

Article history:
Received 22 October 2014
Received in revised form 30 December 2014
Accepted 17 January 2015
Available online 28 January 2015

Keywords:
3D reconstruction
Initial volume

ABSTRACT

Cryo Electron Microscopy is a powerful Structural Biology technique, allowing the elucidation of the three-dimensional structure of biological macromolecules. In particular, the structural study of purified macromolecules –often referred as Single Particle Analysis(SPA)– is normally performed through an iterative process that needs a first estimation of the three-dimensional structure that is progressively refined using experimental data. It is well-known the local optimisation nature of this refinement, so that the initial choice of this first structure may substantially change the final result. Computational algorithms aiming to providing this first structure already exist. However, the question is far from settled and more robust algorithms are still needed so that the refinement process can be performed with sufficient guarantees.

In this article we present a new algorithm that addresses the initial volume problem in SPA by setting it in a Weighted Least Squares framework and calculating the weights through a statistical approach based on the cumulative density function of different image similarity measures. We show that the new algorithm is significantly more robust than other state-of-the-art algorithms currently in use in the field.

The algorithm is available as part of the software suite Xmipp (http://xmipp.cnb.csic.es) and Scipion (http://scipion.cnb.csic.es) under the name "Significant".

1. Introduction

Single Particle Analysis using the Electron Microscope is a powerful experimental technique to elucidate the three-dimensional structure of macromolecular complexes (Frank, 2006; Sorzano et al., 2007). Thousands of two-dimensional projections of the structure under study are collected with the Electron Microscope, which are then used in most cases within iterative algorithms that have as initial input a first estimation of the three-dimensional structure. However, refinement algorithms are known to behave as local optimizers (Sorzano et al., 2006; Henderson et al., 2012), so that the dependence of the final result on the initial volume is a major concern in the field. This situation is known as the "initial volume problem". There exist several algorithms addressing the task of reconstructing a 3D volume compatible either with the 2D experimental images or with their image class averages (Penczek et al., 1996; Ogura and Sato, 2006; Singer et al., 2010; Coifman et al., 2010; Elmlund et al., 2010; Sanz-García et al., 2010; Singer and Shkolnisky, 2011; Elmlund and Elmlund, 2012; Elmlund et al., 2013; Vargas et al., 2014). However, the problem is far from settled due to several reasons: (1) It is an optimisation problem in a high-dimensional space; (2) There are many local minima and algorithms may get trapped into them. Except for Elmlund et al. (2013), most algorithms aim at trying to avoid local minima. Elmlund et al. (2013) takes a soft optimisation probabilistic approach, in which an image can take multiple 3D orientations with different weights calculated from some heuristically determined function within a subset of so-called feasible directions. This idea is somehow similar to the one in Maximum Likelihood and Bayesian reconstruction (Scheres et al., 2005, 2007; Scheres, 2012a), in which all projections can take all directions with different weights (in this case, calculated from the assumed *a priori* distribution of noise (ML) and signal coefficients (Bayesian)). In turn, Vargas et al. (2014) adopts a statistical approach with the goal of

* Corresponding author at: National Center of Biotechnology (CSIC), c/Darwin, 3, Campus Univ. Autonoma de Madrid, 28049 Cantoblanco, Madrid, Spain. Fax: +34 91 585 4506.

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

http://dx.doi.org/10.1016/j.jsb.2015.01.009
1047-8477/© 2015 Elsevier Inc. All rights reserved.

also avoiding the local minima by strongly reducing the search space using image subsets, randomly assigning Euler angles and checking which of the assignments was more successful. Unfortunately, current practice shows that, despite the availability of all these possibilities, more robust algorithms are still in need, since there are occasions in which the existing programs fail to produce a satisfactory result. Some recent approaches (like Optimod (Lyumkis et al., 2013) or MyFirstMap) take the pragmatic approach of generating many different volumes (preferably with different algorithms) and rank the volumes according to their fit to the experimental data.

The algorithm presented in this paper, which we will refer to as Significant, follows previous approaches in the field in which an image is allowed to have different projection directions with different weights. However, instead of setting the problem as a closed form optimisation of a given functional under a simplified set of assumptions, which may be violated in practical works, it considers more realistic models at the expense of mathematical tractability. We rely on the theory of Weighted Least Squares (WLS) optimisation rather than, for instance, on Maximum Likelihood (ML) optimisation. The rationale for this choice is that we are more free to choose a different weight scheme in which we incorporate more criteria evaluating the quality of the fitting between a given particle and its candidate projection direction. The fact that the functional is changed along iterations complicates its mathematical properties in the limit, so that the algorithm cannot be understood as an iterative algorithm to solve a Weighted Least Squares problem because the weights change from iteration to iteration. In principle, no weighting scheme is better than another, and the proof of its correctness can only be based on the results it produces.

Following the rational just introduced, Significant has been developed so that similarity measures are certainly addressed within statistically significant intervals; additionally, we have incorporated a number of new "desired properties" of a solution. In this way, we introduce the notion of "images being important for a projection" and of "projections being important for an image", the explicit consideration of the spatial neighbourhood of projection directions and, finally, the combined use of several image similarity measures (the correlation coefficient and the IMED (IMage Euclidean Distance) (Wang et al., 2005), an image metric that takes into account pixel neighbourhoods). In the Results section we compare our new algorithm with a number of common methods in the field.

2. Methods

Let us call I_i the ith image in a collection of N images (they can be experimental images or class averages, from the point of view of our algorithm the only difference is a larger execution time in the case of experimental images, since there are many more experimental images than class averages). In order to construct a first reference volume, we assign random angles to each one of the images and make a first reconstruction, that we will refer to as $V^{(0)}$. This first reconstruction normally looks as a smooth sphere whose radius coincides with the particle radius. If a better prior exists (the volume is approximately a cylinder, or even a previous 3D reconstruction of a related molecule), we may use it instead.

Let us now refine the first reconstruction using the following iterative method

$$V^{(k+1)} = \arg\min_V \sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}^{(k)} \|\widetilde{I}_{ij}^{(k)} - P_j V\|^2 \qquad (1)$$

where P_j denotes the projection operator along the direction j (assuming that we are exploring a discrete library of M projections), and $\widetilde{I}_{ij}^{(k)}$ is the image resulting of aligning, rotationally and translationally, the ith image to the jth projection of $V^{(k)}$. $w_{ij}^{(k)}$ is a weight (note that normally weights are between 0 and 1, and this is indeed the case in our method, although this is not strictly necessary) that controls whether the ith image should be considered to come from the jth direction at iteration k. Note that many of the 3D reconstruction formalisms can be set in this generic framework: Projection Matching (Scheres et al., 2008) has $w_{ij}^{(k)} = 1$ for only one of the M directions; in Maximum-Likelihood 3D (Scheres et al., 2007) all weights can, in principle, be different from 0 and they are calculated based on the *a priori* assumption of Gaussianly distributed noise; similarly, Relion calculates weights based on the previous assumption and the assumption that Fourier coefficients are Gaussianly distributed (Scheres, 2012a). This type of algorithms is referred as Weighted Least Squares (WLS).

In this article, we also adopt a probabilistic approach for the weight calculation, although in this case based on the concept of statistical significance. Let us consider the case of Projection Matching. It compares, after alignment, the ith image to all M projections generated from the volume at iteration k. This comparison is usually performed by calculating Pearson's correlation coefficient between the two images, $\rho_{ij}^{(k)}$, and the algorithm selects the direction with maximum correlation. However, since images are noisy, the correlation coefficient itself is a random variable. If both the experimental images and the reprojections were to follow a normal distribution, the one-sided confidence interval associated to their cross correlation could be easily computed through Fisher's transformation (Sheskin, 2004, Chap. 28)

$$\rho \in \left[\tanh\left(\tanh^{-1}\left(\max_j\{\rho_{ij}^{(k)}\}\right) - \frac{z_{1-\alpha^{(k)}}}{\sqrt{N-3}}\right), \max_j\{\rho_{ij}^{(k)}\}\right] \qquad (2)$$

where tanh is the hyperbolic tangent, α is the level of confidence, $z_{1-\alpha^{(k)}}$ is the $1-\alpha^{(k)}$ percentile of the Gaussian distribution, and N is the number of pixels on which the correlation has been calculated. The idea is that, because of the noise, all those directions whose correlation coefficient lay in this confidence interval are statistically indistinguishable from the maximum (with a confidence level $\alpha^{(k)}$), and consequently, they should all be kept as feasible solutions. However, the assumption of normality does not hold in practical cases (this issue will be further discussed along this work), which makes inaccurate the simple computation of Fisher's transformation. At this point Significant departs from other algorithms in the field in that it still uses Fisher's confidence interval as a first way to filter out direction candidates, but it subsequently explicitly considers the distribution of experimental correlation coefficients for the actual confidence assignment (note that this approach allows the use of other similarity measures besides cross correlation). This latter concept is what we will refer as "a direction being significant to an image" (with a confidence level $\alpha^{(k)}$). For doing so, we estimate the marginal probability density function of the $\rho_{i\cdot}^{(k)}$ variable (see Fig. 1), and we check whether $\rho_{ij}^{(k)}$ is larger than the $1-\alpha^{(k)}$ percentile:

$$\Pr\left\{\rho_{i\cdot}^{(k)} \leqslant \rho_{ij}^{(k)}\right\} \;\geqslant\; 1-\alpha^{(k)} \qquad (3)$$

Note that in this condition $\alpha^{(k)}$ plays a similar role to the Type I error (α) in Statistical Inference, and from that analogy we have chosen the name "Significant" for this method. Note that the role of this condition is to allow the contribution of an image to a number of "Significant" directions at the same time, while working with the experimental distribution of similarity measures, without being restricted to normality assumptions or the use of cross correlations.

We may also add the desired condition that the image is significant to the direction by testing whether

$$\Pr\left\{\rho_{\cdot j}^{(k)} \leqslant \rho_{ij}^{(k)}\right\} \;\geqslant\; 1-\alpha^{(k)} \qquad (4)$$

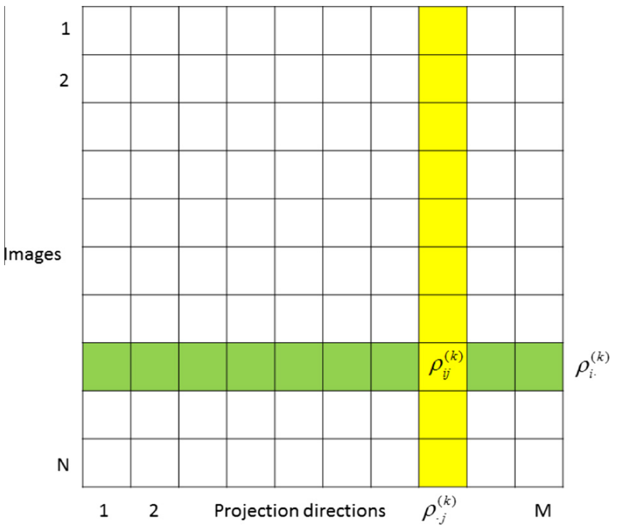

Fig.1. Graphical representation of the ρ_{ij} matrix and the marginal variables $\rho_{i\cdot}$ (the set of correlations for a given image) and $\rho_{\cdot j}$ (the set of correlations for a given direction).

(see Fig. 1). This condition may be used if we expect to have many outliers (empty images, images with no specimen,...), although its effects will naturally imply the selection of only a few images per projection direction and the rejection of the rest. We refer to this condition as the "strict" direction condition (the direction becomes "strict" about which images can contribute to it). In practical cases the "strict" conditions will be seldom used, with the exception of heavily contaminated data sets, for which "strict" can be very useful, as we will show in subsequent sections.

If the condition on direction significance is met, for which we will demand both the fulfilment of Fisher's condition and of the experimentally-determined significance interval, the weight is calculated as

$$w_{ij}^{(k)} = \frac{\rho_{ij}^{(k)}}{\max\limits_{\substack{i' \in 1,\ldots,N \\ j' \in \mathrm{Neigh}_\theta(j)}} \rho_{i'j'}^{(k)}} \Pr\left\{\rho_{i\cdot}^{(k)} \leqslant \rho_{ij}^{(k)}\right\} \Pr\left\{\rho_{\cdot j}^{(k)} \leqslant \rho_{ij}^{(k)}\right\} \tag{5}$$

where $\mathrm{Neigh}_\theta(j)$ is the set of projection directions that differ from the jth projection direction in less than θ degrees; otherwise, the weight is zero. Note that the use of $\mathrm{Neigh}_\theta(j)$ represents a new "criterion" of our solution: its correlation should be a good match also when compared to its surroundings. The rationale is that if an image is correctly assigned to a given direction, then its correlation should also be amongst the best in a neighbourhood of that direction. Note that Projection Matching is obtained in this scheme if $\alpha = \frac{1}{N}$ and we drop any reference to the distribution of correlations along a direction ($\rho_{\cdot j}$). In the Supplementary Material we show the effect of the conditions used in this method to select candidate directions.

Of course, if the condition of "strict" direction is enabled, we will also reinforce that images with similarity measures significantly lower than the maximum value will be assigned a weight equal to zero.

Cross-correlation among two different images is simply a way to measure their similarity. There have been other proposals to measure this similarity like IMED (Wang et al., 2005). IMED is, in fact, a generalisation of the Euclidean distance between two images X and Y that takes into account local neighbourhoods of each pixel:

$$\eta(X,Y) = \sum_{m=1}^{P}\sum_{n=1}^{P} \exp\left(\frac{\|\mathbf{r}_m - \mathbf{r}_n\|^2}{2}\right)(X(\mathbf{r}_m) - Y(\mathbf{r}_m))(X(\mathbf{r}_n) - Y(\mathbf{r}_n)) \tag{6}$$

where P is the number of pixels in images X and Y, $\mathbf{r}_n$ denotes the location of the nth pixel, and $X(\mathbf{r}_n)$ the pixel value at that location. We have observed that IMED has a better discriminatory power than cross-correlation. For instance, Fig. 2 shows a plot of the cross-correlation and IMED at the high-end of cross-correlation (images that are very similar to each other according to cross-correlation). The line plotted is a polynomial of degree 3 fitted to this data. Note the increase of slope of IMED with respect to cross-correlation at very high correlation values revealing the more discriminatory power of IMED (note also that IMED values decrease as cross-correlation values increase).

Finally, we calculate the weight as

$$w_{ij}^{(k)} = \left(\frac{\min\limits_{\substack{i' \in 1,\ldots,N \\ j' \in \mathrm{Neigh}_\theta(j)}} \eta_{i'j'}^{(k)}}{\eta_{ij}^{(k)}} \Pr\left\{\eta_{i\cdot}^{(k)} \geqslant \eta_{ij}^{(k)}\right\} \Pr\left\{\eta_{\cdot j}^{(k)} \geqslant \eta_{ij}^{(k)}\right\} \right) \left(\frac{\rho_{ij}^{(k)}}{\max\limits_{\substack{i' \in 1,\ldots,N \\ j' \in \mathrm{Neigh}_\theta(j)}} \rho_{i'j'}^{(k)}} \Pr\left\{\rho_{i\cdot}^{(k)} \leqslant \rho_{ij}^{(k)}\right\} \Pr\left\{\rho_{\cdot j}^{(k)} \leqslant \rho_{ij}^{(k)}\right\} \right) \tag{7}$$

Note that these weights are necessarily between 0 and 1, and that if an image is the best one for a given direction and that direction is the best for that image, then $w_{ij} = 1$ as in Projection Matching.

A new volume is reconstructed (Eq. (1)) for iteration $k+1$ using the just calculated weights, and the process is iterated for a fixed number of times. At each iteration we normally increase our level of confidence, $1-\alpha^{(k)}$, by using a monotonically decreasing sequence of $\alpha^{(k)}$ values. However, any other strategy could be used. Typically, our confidence levels range between 85% and 99.99%. At low confidence levels, Fisher's confidence interval is relatively small because we do not need to be very confident about it, while the number of candidate directions amongst the top $1-\alpha^{(k)}$ percentile is relatively large. As the confidence level increases, Fisher's confidence interval increases, to account for the larger confidence needed, while the number of candidate directions in the top list decreases (because the $1-\alpha^{(k)}$ percentile increases).

We may think of the new reconstruction algorithm as being half way between Projection Matching (only one direction has a weight different from zero) and Maximum Likelihood/Maximum *a posteriori*

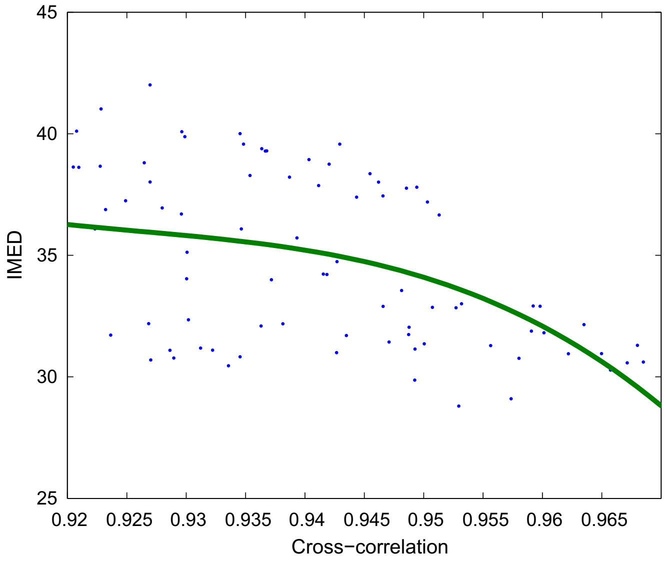

Fig.2. Scatter plot of cross-correlation and IMED values for the GroEL example. Only those images with cross-correlation higher than 0.92 are shown.

(all directions get a weight different from zero), but generalised to any type of similarity measures and to experimentally-determined similarity value distributions. In the new scheme, only a few directions get a non-zero weight (the number of projections range between 100 and 1, depending on the total number of projections, M, and the confidence level, $1-\alpha$), specifically those that are significantly more similar to each image. Interestingly, we have also introduced the notion of an image being significant for a direction. Typically, Electron Microscopy algorithms always assign a projection direction to any input image. However, in our scheme this may not be the case if the image at hand is not good enough (meeting our confidence conditions) for any of the projection directions. This prevents images with very low Signal-to-Noise Ratios, empty images and images corresponding to sufficiently different conformations from being used during the reconstruction.

3. Results

3.1. GroEL

GroEL (Ranson et al., 2001) is considered to be a difficult case for blind initial volume algorithms, because its top views and side views are approximately of the same size, and the algorithm does not always find their correct relative orientation. We used the GroEL dataset publicly available as the tutorial of EMAN2 (http://blake.bcm.edu/emanwiki/Ws2011/Eman2) (Tang et al., 2007). We automatically picked 8758 particles from 26 micrographs at a sampling rate of 4.2 Å/pixel using the algorithm described at Abrishami et al. (2013). We automatically evaluated their quality (Vargas et al., 2013) and kept 8589. Then, we performed a 2D classification (Sorzano et al., 2010) into 44 classes as a way to construct a "summary" of the collected data (see Fig. 3). It can be seen that some of the classes are of much better quality than others, with a quite different number of images assigned to them, as is normally the case in an experimental setting.

We compared the results of reconstruct_significant (run with 100 iterations and α linearly decreasing from 0.15 to 0.001; with a non-strict direction condition) to the results of EMAN2 (*e2initialmodel.py* run with 8 iterations), Simple 1.0 (Elmlund and Elmlund, 2012) (*origami* with low pass filtered to 15 Å, note that Simple 1.0 was originally introduced for raw images but that we are applying here to classes), RANSAC (Vargas et al., 2014) (run with 380 RANSAC iterations and an inlier threshold of 0.77), Relion (with autorefine) (Scheres, 2012b) and Projection Matching as implemented in Xmipp (Scheres et al., 2008). All these algorithms were run with their default parameters normally used in Xmipp. For those algorithms needing a starting volume, we constructed a "sphere" by assigning random angles to the 2D classes, performing a 3D reconstruction, and radially averaging the resulting volume. Relion cannot work with as few as 44 classes and we supplied it with the 8589 selected particles. Each algorithm produced 10 volumes (either by asking the algorithm to do so, or by repeating it 10 times) and found that the newly proposed algorithm constructed a correct model (understanding by correct a volume whose FSC = 0.5 frequency with respect to the EMDB volume is finer than 25 Å) 10 times, RANSAC 3 times, EMAN2 and Simple 1.0 2 times,

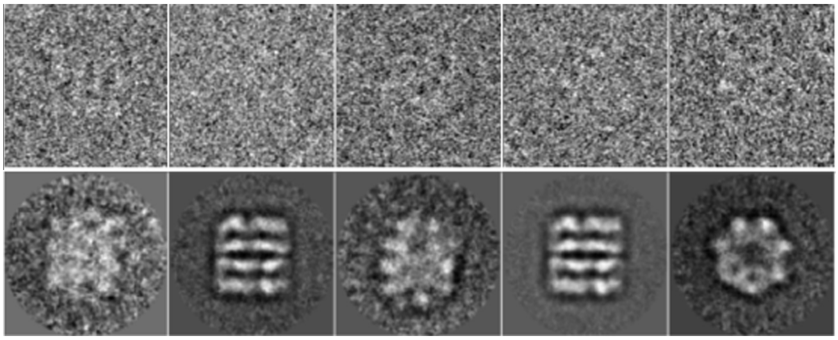

Fig.3. Sample images and classes of GroEL.

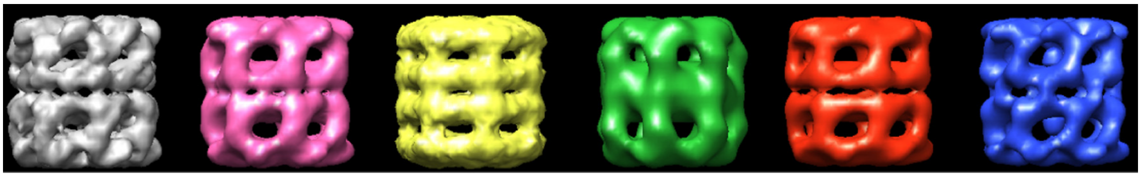

Fig.4. Examples of correct GroEL reconstructions. From left to right: GroEL structure deposited at EMDB (1042) at 10.3 Å; reconstruct_significant (cross-correlation, *cc*, to EMDB-1042: 0.841); Simple 1.0 (*cc* = 0.834); EMAN2 (*cc* = 0.825); Relion (*cc* = 0.809); RANSAC (*cc* = 0.786).

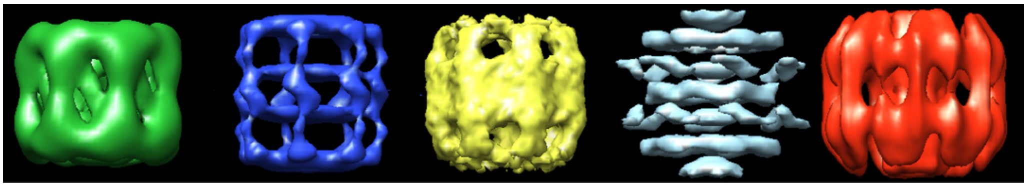

Fig.5. Examples of incorrect GroEL reconstructions. From left to right: EMAN2; RANSAC; Simple 1.0; Projection Matching; Relion.

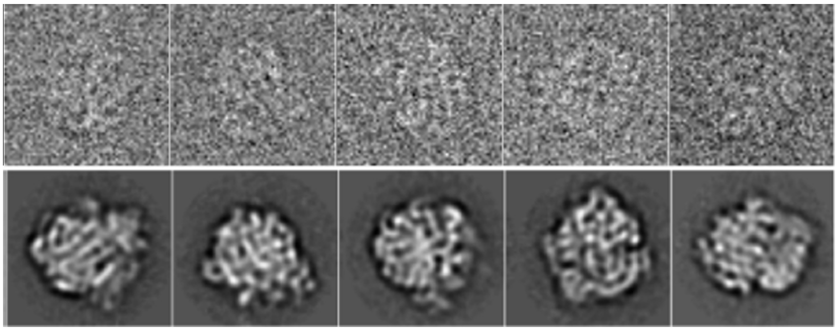

Fig.6. Sample images and classes of the eukaryotic ribosome.

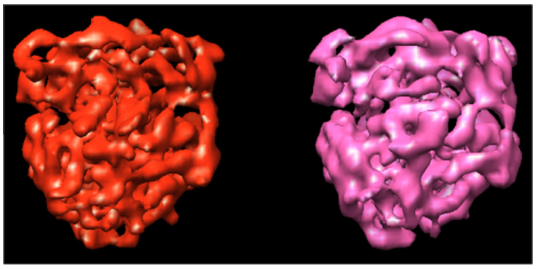

Fig.7. Examples of correct ribosome reconstructions. From left to right: Relion; reconstruct_significant (cross-correlation compared to Relion, $cc = 0.646$.).

Relion 1 time, Projection Matching 0 times. The execution time per volume using a single CPU was EMAN2 (2.7 min), RANSAC (2.8 min), Simple 1.0 (19 min), Projection Matching (25 min), reconstruct_significant (6 h), and Relion (33.6 days). Note that most of these algorithms are parallelized (including the newly proposed one) and the actual execution time must be divided by the number of processors available. Also, our algorithm produced a correct structure from iteration 4 ($\alpha = 0.856$) after only 24 min (in a single processor). Fig. 4 shows the correct GroEL structure as deposited at the Electron Microscopy Data Bank (entry 1042) and the best volumes reconstructed by each of the algorithms sorted by descending correlation coefficient while Fig. 5 shows some of the poorly reconstructed initial models.

In another experiment, we artificially added to the 44 class averages 396 images (=44 × 9) of pure noise with the same mean and standard deviation as the noise in the original images. Significant was capable of producing the correct structure if the "strict" direction condition was used (it was not without this condition). None of the other algorithms was able of producing a correct structure.

3.2. Eukaryotic ribosome

In our second experiment we have performed the blind *ab initio* reconstruction of 5000 cryo-EM projections of an eukaryotic ribosome, obtained from the EMDB test image data (http://www.ebi.ac.uk/pdbe/emdb/test_data.html) and originally used in the work of Scheres et al. (2007). The images had an original size of 130 × 130 pixels and we scaled them to a size of 64 × 64 pixels for speeding up the calculations. In our previous algorithm, RANSAC (Vargas et al., 2014), we needed to filter the 2D classes so that the algorithm was able to produce correct structures, the reason being probably that with high resolution information there were too many local minima in which the algorithm was getting trapped. In this experiment, we tested whether the new algorithm was able to produce good structures without any filtration and compared its results to the results of the rest of algorithms. Fig. 6 shows some of the images of the dataset and some of the classes calculated from them.

We estimated 32 2D classes using CL2D (Sorzano et al., 2010). We run the same set of programs as in the previous case, with the same parameters. This time, only Relion (only one run with the 5000 images) and Significant (in 100% of the 10 executions) were able to produce a correct structure (see Fig. 7). Fig. 8 shows the evolution along the iterations of the ribosome reconstructed by reconstruct_significant. The rest of the algorithms got trapped into local minima (see Fig. 9). The execution time per reconstructed model in a single CPU was: EMAN2 (3.4 min), Simple 1.0 (22 min), Projection Matching (36 min), RANSAC (10.5 h), Significant (16 h), Relion (37.3 days). Again, most of these algorithms are parallelized, so the actual wall clock time is much smaller.

To test whether the number of images played a role in this result, we provided EMAN2 and Simple 1.0 with the full set of raw images. None of the algorithms was capable of producing a good result after several days of execution.

4. Discussion

The determination of an initial volume that can be further refined in the context of iterative algorithms is a crucial step in the protocol of macromolecular structure determination from sets of Electron Micrographs. Practitioners in the field currently have a range of options, going from low-pass filtering a similar structure to *ab initio* 3D reconstruction passing by using random noise, geometrical models (Bilbao-Castro et al., 2004) and Random Conical

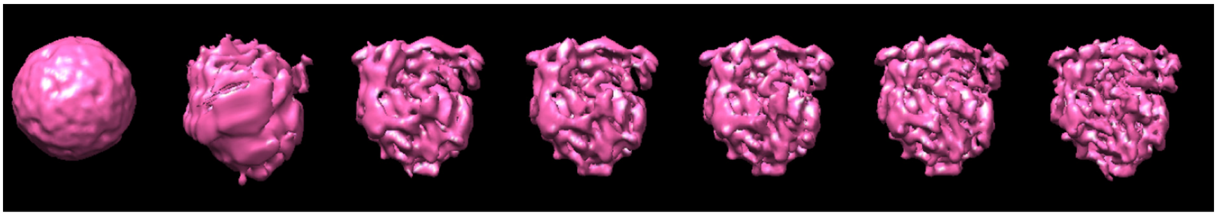

Fig.8. Evolution of the ribosome reconstruction along iterations (iteration 0, 15, 30, 45, 60, 75, and 90) using reconstruct_significant.

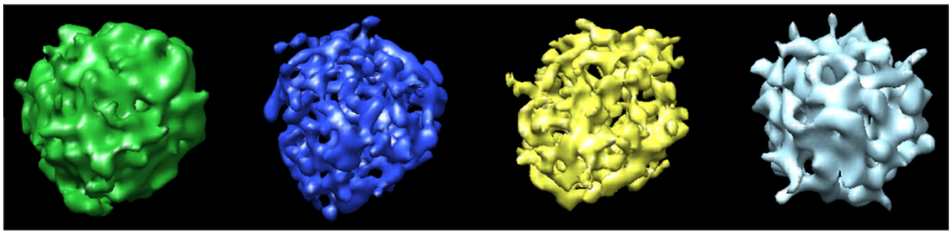

Fig.9. Examples of incorrect ribosome reconstructions. From left to right: EMAN2 ($cc = 0.297$); RANSAC ($cc = 0.217$); Simple 1.0 ($cc = 0.204$); Projection Matching ($cc = 0.196$).

Tilt reconstructions (Radermacher, 1988). *Ab initio* algorithms have extensively been explored by the EM community, as already shown in the introduction, with over 10 different methods. However, this large amount of possibilities may give the false impression that the problem is solved and that there is no need for yet another algorithm. However, reality is far from this. Indeed, there are many structures for which we can find an appropriate algorithm producing the correct result, however, there are also some other structures for which the existing algorithms still fall short. Even, as shown in our Results Section, there are structures for which a given algorithm may or may not produce a correct reconstruction. The algorithm proposed in this paper addresses these difficult situations. If the structure can be solved with fast methods like the one of EMAN2 or RANSAC, there is no need for an extra check with an algorithm such as the one proposed. However, if there are doubts about the plausibility of the initial model, running Significant may show worthy. Of course, we do not mean that the new algorithm is the definitive one, and there will always be room for improvement. However, in the Results Section we have shown that the algorithm is robust enough in situations with many (eukaryotic ribosome) or deep local minima (GroEL). The design of the algorithm borrows several ideas from many of the existing algorithms and combine them in a unique way with the aim of smoothing the landscape of solutions and finding the global minimum of the reconstruction problem. Specifically,

- It shares with Projection Matching and Maximum Likelihood (and partially with its Bayesian evolution in Relion) a Weighted Least Squares scheme. However, our weights are computed based on the cumulative distribution function of the correlation and IMED (a more robust distance measure between images) as well as the local structure of these two quantities around a given projection direction.
- It shares with Maximum Likelihood, Relion and Simple 1.0, the possibility that an image may contribute to multiple projection directions (with different weights). Again, there are differences with respect to the other algorithms in the way we choose those directions to which each image contributes to. We feel that our choice of comparing always each image to all possible projection directions (within the specified angular discretization limits) helps to not get trapped within local minima by taking decisions about the final projection direction too soon.
- It shares with Simple 1.0 and Simulated annealing the slow "cooling" scheme, in our case, the slow decrease of the Type I error (α). For each α we may think of the landscape of solutions of the modified Weighted Least Squares problem as one of a surrogate optimisation problem (we substitute the original landscape with many local minima by a smoother landscape with much fewer). As we get closer to the solution, we may go for fewer Type I errors (and consequently, more local minima). Obviously, reducing the number of iterations (in our examples, 100) for linearly going from α_0 to α_F would reduce the computing time, but it would also increase the risk of getting trapped into local minima. However, it is also true that the α sequence does not need to be linear and that faster sequences could be explored in the future as soon as we detect that we are in a sufficiently good minimum (which may occur rather early in the iterations, as was the case of GroEL).
- It introduces a new concept in EM that is the fact that the distribution of correlation and distance measures for a given projection direction also influences on the weight of the experimental image, and may even prevent it from participating in the 3D reconstruction ("strict" direction condition).

5. Conclusion

In this paper we have presented a new algorithm for the estimation of initial models that can be further refined by any of the already existing algorithms in EM. The algorithm is based on a Weighted Least Squares approach in which the weights are calculated using the cross correlation and IMED distance of the individual image to a specific projection direction as well as its relationship to neighbour directions and the comparison of these two quantities with respect to the rest of images available in the dataset (through the cumulative density functions defined in the Methods Section). All our design goes into the statistical direction of "being significant" (the projection direction must be significant for the image and vice versa, considering also the neighbourhood of that projection direction). We have experimentally shown that our algorithm succeeds in producing a correct initial guess in rather difficult cases. The conceptual bases of this new method can be expanded to other topics, such as 3D refinement at high resolution and 3D classification, although these extensions fall outside the scope of the present work.The algorithm is available through the open-source package Xmipp (http://xmipp.cnb.csic.es) (Sorzano et al., 2004; De la Rosa-Trevín et al., 2013) since version 3.2 (note that the official stable release is currently 3.1) under the name `xmipp_reconstruct_significant`.

Acknowledgments

Funding: The authors would like to acknowledge economical support from the Spanish Ministry of Economy and Competitiveness through Grants AIC-A-2011-0638, BFU2013-41249-P and BIO2013-44647-R, the Comunidad de Madrid through Grant CAM (S2010/BMD-2305), as well as postdoctoral Juan de la Cierva Grant with reference JCI-2011-10185. C.O.S. Sorzano is recipient of a Ramón y Cajal fellow.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jsb.2015.01.009.

References

Abrishami, V., Zaldívar-Peraza, A., de la Rosa-Trevín, J.M., Vargas, J., Otón, J., Marabini, R., Shkolnisky, Y., Carazo, J.M., Sorzano, C.O.S., 2013. A pattern matching approach to the automatic selection of particles from low-contrast electron micrographs. Bioinformatics 29 (19), 2460–2468.

Bilbao-Castro, J.R., Sorzano, C.O.S., García, I., Fernández, J.J., 2004. Phan3D: design of biological phantoms in 3D electron microscopy. Bioinformatics 20, 3286–3288.

Coifman, R.R., Shkolnisky, Y., Sigworth, F.J., Singer, A., 2010. Reference free structure determination through eigenvectors of center of mass operators. Appl. Comput. Harmon. Anal. 28 (3), 296–312.

De la Rosa-Trevín, J.M., Otón, J., Marabini, R., Zaldívar-Peraza, A., Vargas, J., Carazo, J.M., Sorzano, C.O.S., 2013. Xmipp 30: one step forward in scientific computing for electron microscopy. J. Struct. Biol. 184, 321–328.

Elmlund, D., Elmlund, H., 2012. Simple: software for ab initio reconstruction of heterogeneous single-particles. J. Struct. Biol. 180 (3), 420–427.

Elmlund, D., Davis, R., Elmlund, H., 2010. Ab initio structure determination from electron microscopic images of single molecules coexisting in different functional states. Structure 18 (7), 777–786.

Elmlund, H., Elmlund, D., Bengio, S., 2013. Prime: probabilistic initial 3D model generation for single-particle cryo-electron microscopy. Structure 21 (8), 1299–1306.

Frank, J., 2006. Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State. Oxford Univ. Press, New York, USA.

Henderson, R., Sali, A., Baker, M.L., Carragher, B., Devkota, B., Downing, K.H., Egelman, E.H., Feng, Z., Frank, J., Grigorieff, N., Jiang, W., Ludtke, S.J., Medalia, O., Penczek, P.A., Rosenthal, P.B., Rossmann, M.G., Schmid, M.F., SchrÃüder, G.F., Steven, A.C., Stokes, D.L., Westbrook, J.D., Wriggers, W., Yang, H., Young, J., Berman, H.M., Chiu, W., Kleywegt, G.J., Lawson, C.L., 2012. Outcome of the first electron microscopy validation task force meeting. Structure 20 (2), 205–214.

Lyumkis, D., Brilot, A.F., Theobald, D.L., Grigorieff, N., 2013. Likelihood-based classification of cryo-em images using freealign. J. Struct. Biol. 183 (3), 377–388.

Ogura, T., Sato, C., 2006. Posterior Euler angle assignment using simulated annealing. J. Struct. Biol. 156, 371–386.

Penczek, P.A., Zhu, J., Frank, J., 1996. A common-lines based method for determining orientations for N > 3 particle projections simultaneously. Ultramicroscopy 63, 205–218.

Radermacher, M., 1988. Three-dimensional reconstruction of single particles from random and nonrandom tilt series. J. Electron Microsc. Tech. 9, 359–394.

Ranson, N., Farr, G., Roseman, A., Gowen, B., Fenton, W., Horwich, A., Saibil, H., 2001. ATP-bound states of GroEL captured by cryo-electron microscopy. Cell 107, 869–879.

Sanz-García, E., Stewart, A.B., Belnap, D.M., 2010. The random-model method enables ab initio 3D reconstruction of asymmetric particles and determination of particle symmetry. J. Struct. Biol. 171 (2), 216–222.

Scheres, S.H.W., 2012a. A Bayesian view on cryo-EM structure determination. J. Mol. Biol. 415 (2), 406–418.

Scheres, S.H.W., 2012b. Relion: implementation of a bayesian approach to cryo-em structure determination. J. Struct. Biol. 180 (3), 519–530.

Scheres, S.H.W., Valle, M., Núñez, R., Sorzano, C.O.S., Marabini, R., Herman, G.T., Carazo, J.M., 2005. Maximum-likelihood multi-reference refinement for electron microscopy images. J. Mol. Biol. 348, 139–149.

Scheres, S.H.W., Gao, H., Valle, M., Herman, G.T., Eggermont, P.P.B., Frank, J., Carazo, J.M., 2007. Disentangling conformational states of macromolecules in 3d-em through likelihood optimisation. Nat. Methods 4 (1), 27–29.

Scheres, S.H.W., Núñez Ramírez, R., Sorzano, C.O.S., Carazo, J.M., Marabini, R., 2008. Image processing for electron microscopy single-particle analysis using xmipp. Nat. Protocols 3, 977–990.

Sheskin, D.J., 2004. Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC.

Singer, A., Shkolnisky, Y., 2011. Three-dimensional structure determination from common lines in cryo-em by eigenvectors and semidefinite programming(). SIAM J. Imaging Sci. 4 (2), 543–572.

Singer, A., Coifman, R.R., Sigworth, F.J., Chester, D.W., Shkolnisky, Y., 2010. Detecting consistent common lines in cryo-em by voting. J. Struct. Biol. 169, 312–322 (under review).

Sorzano, C.O.S., Marabini, R., Velázquez-Muriel, J., Bilbao-Castro, J.R., Scheres, S.H.W., Carazo, J.M., Pascual-Montano, A., 2004. XMIPP: a new generation of an open-source image processing package for electron microscopy. J. Struct. Biol. 148, 194–204.

Sorzano, C.O.S., Marabini, R., Pascual-Montano, A., Scheres, S.H.W., Carazo, J.M., 2006. Optimization problems in electron microscopy of single particles. Ann. Oper. Res. 148, 133–165.

Sorzano, C.O.S., Jonic, S., Cottevieille, M., Larquet, E., Boisset, N., Marco, S., 2007. 3D electron microscopy of biological nanomachines: principles and applications. Eur. Biophys. J. 36, 995–1013.

Sorzano, C.O.S., Bilbao-Castro, J.R., Shkolnisky, Y., Alcorlo, M., Melero, R., Caffarena-Fernández, G., Li, M., Xu, G., Marabini, R., Carazo, J.M., 2010. A clustering approach to multireference alignment of single-particle projections in electron microscopy. J. Struct. Biol. 171, 197–206.

Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., Ludtke, S.J., 2007. Eman2: an extensible image processing suite for electron microscopy. J. Struct. Biol. 157, 38–46.

Vargas, J., Abrishami, V., Marabini, R., de la Rosa-Trevín, J.M., Zaldivar, A., Carazo, J.M., Sorzano, C.O.S., 2013. Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. J. Struct. Biol. 183 (3), 342–353.

Vargas, J., Álvarez-Cabrera, A.L., Marabini, R., Carazo, J.M., Sorzano, C.O.S., 2014. Efficient initial volume determination from electron microscopy images of single particles. Bioinformatics 30, 2891–2898.

Wang, L., Zhang, Y., Feng, J., 2005. On the euclidean distance of images. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1334–1339.