BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 20 2014, pages 2891–2898
doi:10.1093/bioinformatics/btu404

Structural bioinformatics

Advance Access publication June 27, 2014

Efficient initial volume determination from electron microscopy images of single particles

Javier Vargas[1,*], Ana-Lucia Álvarez-Cabrera[1], Roberto Marabini[2], Jose M. Carazo[1] and C. O. S. Sorzano[1]

[1]Biocomputing Unit, Centro Nacional de Biotecnología-CSIC, C/Darwin 3 and [2]Escuela Politécnica Superior, Universidad Autónoma de Madrid, C/Francisco Tomás y Valiente, 28049, Cantoblanco (Madrid), Spain

Associate Editor: Robert Murphy

ABSTRACT

Motivation: Structural information of macromolecular complexes provides key insights into the way they carry out their biological functions. The reconstruction process leading to the final 3D map requires an approximate initial model. Generation of an initial model is still an open and challenging problem in single-particle analysis.

Results: We present a fast and efficient approach to obtain a reliable, low-resolution estimation of the 3D structure of a macromolecule, without any a priori knowledge, addressing the well-known issue of initial volume estimation in the field of single-particle analysis. The input of the algorithm is a set of class average images obtained from individual projections of a biological object at random and unknown orientations by transmission electron microscopy micrographs. The proposed method is based on an initial non-lineal dimensionality reduction approach, which allows to automatically selecting representative small sets of class average images capturing the most of the structural information of the particle under study. These reduced sets are then used to generate volumes from random orientation assignments. The best volume is determined from these guesses using a random sample consensus (RANSAC) approach. We have tested our proposed algorithm, which we will term 3D-RANSAC, with simulated and experimental data, obtaining satisfactory results under the low signal-to-noise conditions typical of cryo-electron microscopy.

Availability: The algorithm is freely available as part of the Xmipp 3.1 package [http://xmipp.cnb.csic.es].

Contact: jvargas@cnb.csic.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 15, 2014; revised and accepted on June 23, 2014

1 INTRODUCTION

Single-Particle Analysis (SPA) techniques can obtain 3D maps of biological complexes at near-atomic resolution by combining tens of thousands of projection images obtained with a transmission electron microscopy (TEM) (Frank, 1996; Zhang and Zhou, 2011). In general, the reconstruction process leading to the final 3D map requires the use of an approximate initial model. The fully automatic and efficient determination of this initial volume, either for symmetric or asymmetric structures, is still an open and challenging problem in SPA, as indicated by the existence of an ample literature on this matter.

Many previous attempts to the 'initial model problem' have been reported. On the one hand, in the Random Conical Tilt and orthogonal tilt methods (Leschziner and Nogales, 2006; Radermacher *et al.*, 1987), the alignment problem is simplified by acquiring micrographs as tilt pairs, or by using multiple different tilts with known tilt angles. On the other hand, we observe a large variety of methods based on common lines (Castón *et al.*, 1999; Crowther *et al.*, 1970; Elmlund and Elmlund, 2012; Elmlund *et al.*, 2008, 2010; Liu *et al.*, 2007; Ogura and Sato, 2006; Penczek and Zhu, 1996; van Heel, 1987; Thuman-Commike and Chiu, 1997) that, in principle, they allow for an initial model estimation without tilting of the specimen. Computer-generated shapes (Baker and Cheng, 1996; Bilbao-Castro *et al.*, 2004; Ludtke *et al.*, 2004), or reconstructions from one image of a particle assuming a certain symmetry (Cantele *et al.*, 2003; Castón *et al.*, 1999), have also been used to define initial volumes. Another line of research is based on the introduction of a 'random model' strategy based on first assigning random orientations to class averages (Harauz and van Heel, 1985; van Heel, 1984). In this latter case, an initial 3D reconstruction is obtained from these random angular assignments, which is finally refined by a projection matching strategy. Following this approach, Sanz-Garcia *et al.*, 2010 presented a random-model method that allows *ab initio* generation of starting models from raw experimental images. Several initial models were generated, assigning initially a random orientation to each imaged particle. Recently, Elmlund *et al.*, 2013, have presented a method based on a probabilistic initial 3D model generation procedure, which uses projection images instead of class averages. Furthermore, in Lyumkis *et al.* (2013), it presented OptiMod, a method that incorporates multiple automated algorithms for determining orientations using common-lines methodologies and, at the same time, provides criteria for scoring their results. This approach generates multiple maps using algorithm-specific randomization.

The 'initial model problem' is still, and in spite of the multiple algorithms so far proposed, widely accepted as a real issue in SPA (Taylor and Glaeser, 2008, Voss *et al.*, 2010), with methods demanding no trivial choices of input parameters and being computationally expensive. Indeed, in the way of defining high-throughput approaches in 3D-EM, we really need fast, simple and accurate methods, which are, indeed, the motivation of this

*To whom correspondence should be addressed.

work. In this way, we will describe in detail 3D-RANSAC, presenting its good performance on a wide range of specimens, using the same parameters for all cases (making the case of 'simplicity'), and obtaining final results in a matter of minutes on a typical laptop computer. Our proposed approach consists in a novel random modelling strategy based on an initial dimensionality-reduction method together with the RANSAC algorithm, which makes this approach efficient from a computational point of view. The method can be used to produce low-resolution initial volumes of symmetric or asymmetric biological complexes.

2 METHODS

2.1 Class average images

First, a set of class average images are obtained from the particle dataset using any image classification algorithm [in this article, we will use CL2D (Sorzano *et al*., 2010), included in the Xmipp 3.1 package]. Observe that the classification process is required in any usual SPA processing workflow and is not a special requirement of the proposed approach. Electron microscopy datasets commonly contain more than one different structures, as projections of different conformations of a given molecule, or projections of different molecules in the specimen preparation. These class images are then the input of the new proposed method and, typically, ~20–50 class images are sufficient. There is nothing in the algorithm that precludes using experimental projections instead of classes. However, we find that using a sufficient number of class average images has several practical advantages, such as (i) increasing the signal-to-noise ratio of the input images, (ii) introducing a desired smoothing (a '*de facto*' regularization) in the landscape of solutions and (iii) reducing the total processing time. Of course, we note that we are only interested in a low-resolution initial model.

2.2 Random model generation strategy

Our random model generation strategy is based on the following eight steps:

(1) The class averages are low-pass filtered, and their size is reduced according to user parameters (basically, the desired resolution in the initial model).

(2) The local tangent space alignment (LTSA) non-linear dimensionality reduction approach (Zhang and Hongyuan, 2005) is applied to automatically select a random, smaller and appropriate (in the sense that contains the most of the structural information of the biological complex under study) subset of class images. This approach non-linearly projects the class averages onto a lower-dimensional space [in our case, two-dimensional (2D)], where the projections of the structure at similar orientations appear close. LTSA method is essentially a local principal component analysis (PCA) approach that can efficiently 'learn' about non-linear manifolds by taking into account that non-linear manifolds can be considered to be locally linear in small neighborhoods. Observe that PCA cannot deal with non-linear manifolds, as it is a linear method, and the 2D projection of a 3D structure is, intrinsically, a non-linear process (Giannakis *et al*, 2012). Additionally, we note that the computer time required for LTSA and PCA are comparable, making LTSA clearly the method of choice for this task. As an example, in Figure 1, we show a projection of a set of image class averages into a 2D space. As can be seen from Figure 1, and as intuitively expected, similar projections of the biological object appear close in the 2D space, whereas different ones appear far apart. We can now use this 2D space to select a smaller image dataset containing the most of the structural information, and then assign random orientations to these images. To automatically select a representative image set, the 2D map is partitioned by a 2D regular grid of dimensions (with typically $n = 9$), and only one image class average is randomly selected each time from any of the grid squares. In Figure 1, we show an example of a smaller image dataset with a red square, and the 2D regular grid appears in gray color. Therefore, n class images are randomly selected in this way. Obviously, there is a trade-off with respect to the number of images composing this reduced dataset. As the number of images gets smaller, the processing time of subsequent steps is reduced and, additionally, the probability of assigning correct angles by random assignment is higher. However, if this number is too small, we can lose important structural information. In our algorithm, this number (n) is an input parameter and normally ranges between 4 and 9, with 9 being the 'default' value (all results presented in this article have been obtained with $n = 9$).

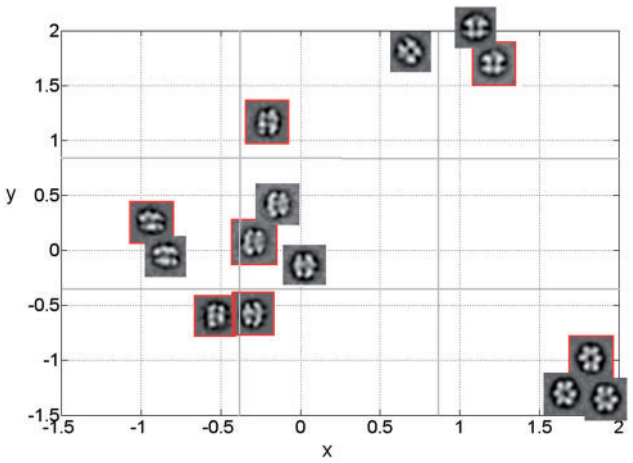

Fig. 1. Projection of a set of unsorted image class averages of a certain MCM467 complex into a 2D space using the LTSA dimensionality reduction approach. In this figure, x and y are the axis of the feature space learned from the input

(3) A 3D reconstruction is performed from the smaller image subset ($n = 9$) by random angular assignment. The reconstruction algorithm takes advantage of symmetry information when available. Reconstruction is made by interpolation in Fourier space. For each experimental image, the Fourier Transform is computed and placed in the corresponding plane in the 3D space as well as in all planes related by symmetry. This random 3D model is then projected at regular angular intervals determined by the angular step-size, which is an input parameter with a default value of 7°. Each initial class average is now compared with the projections of the 3D random model, and the best assignment is determined looking for the largest correlation coefficient.

(4) Steps 2–3 are repeated N times producing N different 3D random models. Therefore, N is also an input parameter of the method, but in the Supplementary Material, we present a statistical derivation for an informed selection of N. In this way, we set the default value of N to 380, as in this way, we know that the probability of obtaining a still better 3D model by increasing N is <0.01. This value of $N = 380$ has been used in all the examples presented in this article.

(5) For each generated random model, we define its inliers as the initial class average images that have a large enough correlation coefficient with respect to the reprojections of the random 3D map.

We will refer to these 'good' initial classes as 'inliers', and we say that they support this 3D model. The rest of the initial classes are referred to as 'outliers'. The practical way in which this selection is done is by establishing a threshold in the correlation coefficient obtained as the percentile p of all obtained correlations, with p typically between 75 and 80%. We define the score of each random 3D map as the sum of the correlations of its inliers. In steps 2–5, we are using Random Sample Consensus (RANSAC) approach (Fischler and Bolles, 1981) to capture 'correct' models. RANSAC is an iterative method to estimate a model (in this case a 3D map) from a set of observed data that contains a large amount of outliers. RANSAC approach consists in the following steps: (i) Randomly selecting a subset of the dataset. (ii) Fitting a model to the selected subset. (iii) Determining a score to each model, typically the number of outliers/inliers. (iv) Repeating steps 1–3 for a prescribed number of iterations. A basic assumption is that the data consist of inliers, data that can be explained by the model, though they may be subject to noise, and outliers, that are data that do not fit the model. The outliers may come, e.g. from extreme values of the noise, from erroneous measurements or incorrect hypotheses about the interpretation of data, for instance, wrongly assigned Euler angles. RANSAC also assumes that, given a (usually small) set of inliers, there exists a procedure that can estimate the model that optimally explains or fits this data. Additionally, this algorithm is probabilistic and non-deterministic, in the sense, that it produces always good results with a certain probability if the number of inliers is larger than the number of outliers, with this probability increasing as more iterations are allowed (Fischler and Bolles, 1981). Our RANSAC approach is performed using the following combination of steps: (i) Automatic selection of a small number of class averages following a dimensionality reduction approach. (ii) Random assignment of angles to each of the selected class averages and computation of a 3D model using them. (iii) Calculation of a score for each model as the sum of inliers correlation. (iv) Repeat steps 1–3 for a prescribed number of iterations. In the Supplementary Material, we show that with the number of RANSAC iterations >380, the probability of finding a better model is <0.01.

(6) The k random 3D models ($k \sim 5$–10) with largest number of inliers (largest score) are selected and new 3D reconstructions are obtained using as input classes only the inliers.

(7) The previous k 3D random models are now refined against all initial classes through a model refining strategy. In this article, we have used a projection matching approach (de la Rosa-Trevín *et al.*, 2013), using typically 10 iterations for refinement. However, other approaches can be used as well, such as the recent method of Elmlund *et al.*, 2013. Observe that we refine the best K models independently through projection matching to add more robustness to our approach. As the previous angular assignment is random, refining K models improves the probability of getting at least some good structures at the end.

(8) The resulting k volumes are scored taking into account the sum of the inliers correlation coefficients. Finally, the model with highest score is automatically selected.

In Figure 2, we show a diagram of the different processing steps.

3 RESULTS

In this section, we provide results obtained with simulated and experimental data that show the effectiveness of the proposed method.

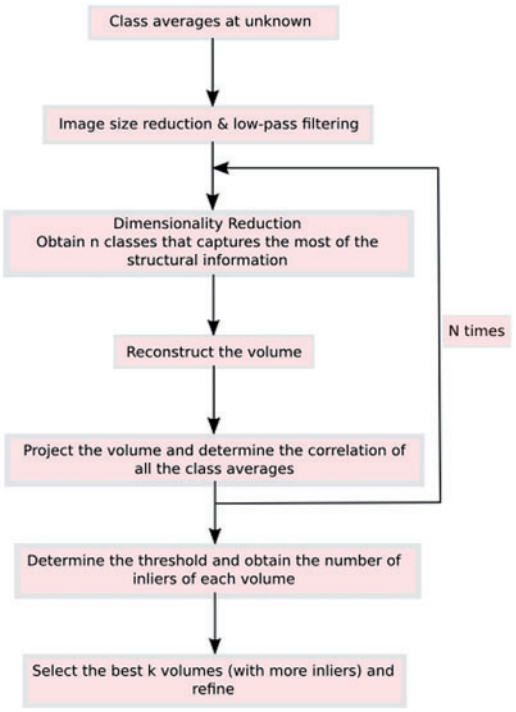

Fig. 2. Diagram of the different processing steps the input

3.1 Simulations

In the first simulation, we used the structure of the Bacteriorhodopsin as a phantom (PDB entry: 1BRD, Henderson *et al.*, 1990) that is projected at 200 unknown and random orientations. The projections are affected by a Gaussian noise with a signal-to-noise ratio (SNR) of 0.1. The size of the images is 100×100 px and the sampling rate is 1 Å/px. Note that, in this case, we did not obtain class averages and we use our proposed method directly with the noisy projections to show its robustness.

In Figure 3, we present 7 of the 200 noisy projections used, together with the phantom (left) and best obtained initial (right) 3D maps at three different orientations. The initial volume was obtained with the following parameters, $N = 380$, $n = 9$, $P = 0.77$ and $k = 10$. We used an angular sampling rate for retroprojection of 7°. The projections were low-pass filtered to a resolution of 5 Å. The required processing time for obtaining the $k = 10$ volumes is of 35 min with a 2.5 GHz laptop and using two processors.

To quantize the resolution of the obtained initial volumes, we have obtained the Fourier Shell Correlation (FSC) curves using the PDB volume as reference. The resolution at FSC = 0.5 and FSC = 0.143 are 4.6 and 4.5 Å, respectively (Fig. 4), that is consistent with the previously performed low-pass filtering, and shows that using perfectly aligned projections in presence of moderate noise, the proposed method can retrieve high-resolution models. This situation is not usual in experimental cases. However, with this simulation, we want to show that there is no theoretical restriction that limits the proposed method in high-resolution analysis.

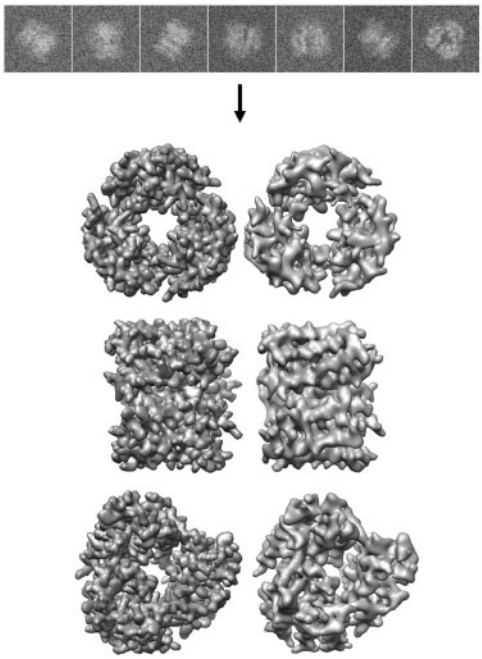

Fig. 3. Seven projections of Bacteriorhodopsin (PDB entry: 1BRD, Henderson *et al.*, 1990) phantom map with SNR of 0.1 are shown at top of the figure. The phantom 3D map at three different orientations is presented on the left-hand side, whereas the best obtained volume, using the proposed approach, is on the right-hand side

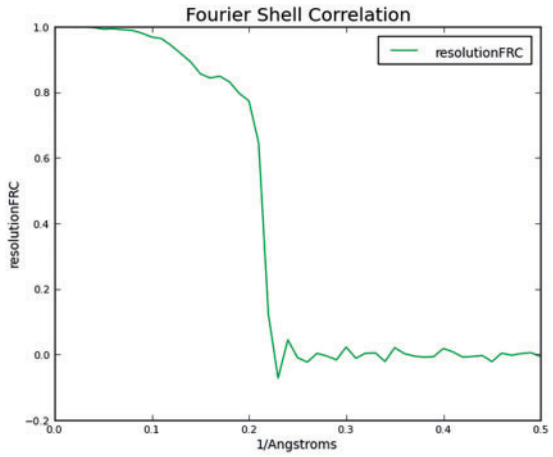

Fig. 4. FSC curve between the Bacteriorhodopsin phantom of the best initial 3D map obtained (highest score) using the proposed approach

3.2 Experimental results

3.2.1 Case 1: Bovine Papillomavirus The first case consists of images of Bovine Papillomavirus (BPV) (Wolf *et al.*, 2010) kindly made accessible by Drs Wolf and Grigorieff. The dataset consists of 49 micrographs of an approximate size of 10 000 × 10 000 pixels. The sampling rate is 1.237 Å/pixel, the microscope voltage 300 kV and the magnification × 56 588. A total of 5317 particles of size 120 × 120 px were identified using the method presented in (Abrishami *et al.*, 2013), from which 32 classes were determined

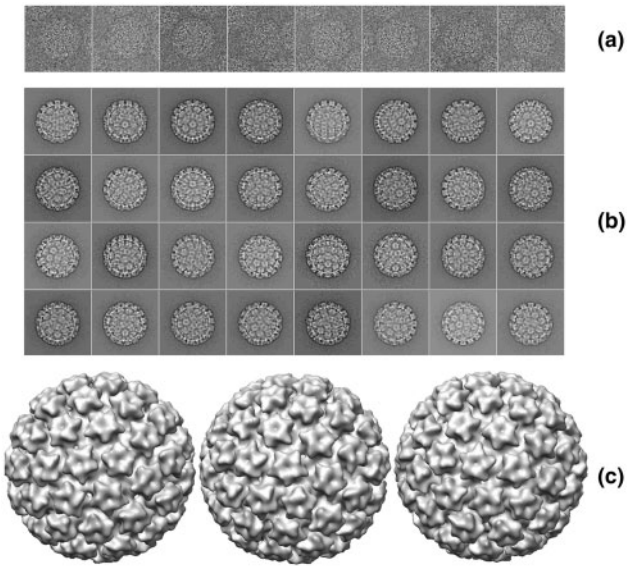

Fig. 5. Eight experimental projections of the BPV (**a**), 32 classes obtained using CL2D (**b**) and the best initial volume obtained by the proposed method at three different orientations (**c**)

using CL2D (Sorzano *et al.*, 2010). These classes were low-pass filtered to a resolution of 5 Å. In Figure 5, we show eight experimental projections (a) and the 32 obtained classes (b). The parameters used to obtain the initial volume by the proposed method are the same ones as in the case presented before. In Figure 5c, we show the best 3D-RANSAC map at three different orientations. The processing time for obtaining the k = 10 volumes is of ∼20 min, with the same computer as before using two processors. Observe that the processing time of CL2D is of 330 min, and then the total time CL2D + 3D-RANSAC is of 350 min.

3.2.2 Case 2: Eukaryotic ribosome Moreover, we have also performed an asymmetric reconstruction using ∼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 have a size of 130 × 130 px. We initially obtained 16 class average images using CL2D, which were then low-pass filtered to a resolution of 25 Å. We obtained the 3D-RANSAC map with the same parameters as in the cases before. In Figure 6, we show eight initial experimental projections of the ribosome (a) and the low-pass filtered class averages (b). Finally, in Figure 6c, we also show the 3D reconstruction using PRIME (Elmlund *et al.*, 2013) and one obtained using 3D-RANSAC at three different orientations. The FSC curves between the PRIME and the 3D-RANSAC maps at FSC = 0.5 and FSC = 0.143 are 19 and 12 Å, respectively, which means that both structures are similar, as visually suggested already in Figure 6c, certainly enough for any of them to be used as a low-resolution initial 3D map. However, the processing time for obtaining the 3D-RANSAC map with the same parameters and the same laptop computer than in previous cases is only 50 min. The processing time of CL2D is of 435 min, and then, the total time

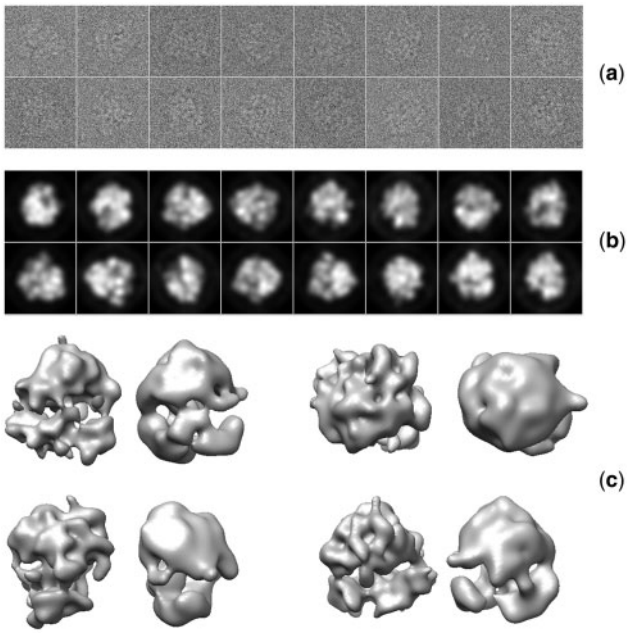

Fig. 6. Eight experimental projections of the eukaryotic ribosome (**a**), 16 class average images low-pass filtered (**b**) and 3D map obtained by PRIME approach (left) together with our best obtained initial volume (right) (with highest score) at four different orientations (**c**)

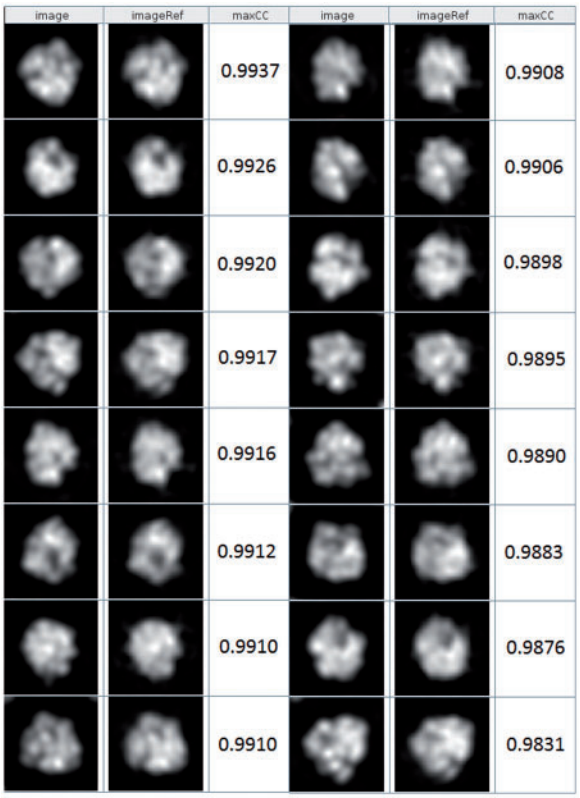

Fig. 7. Experimental class averages (image) of the asymmetric eukaryotic ribosome particles, corresponding initial volume projection at the same orientations (imageRef) and normalized cross correlation between both images (maxCC)

CL2D + 3D-RANSAC is of 485 min, whereas PRIME (or virtually any other method so far proposed) requires more than 4 days of computation (5760 min) in a laptop setting. To present the good agreement between the obtained initial volume and the used class averages, we show in Figure 7, the experimental class averages (labeled as 'image'), the corresponding initial 3D map projection at the same orientation ('imageRef') and the normalized cross correlation between both images ('maxCC'). As can be seen from Figure 7, there is a good agreement between the class averages and the corresponding projections.

3.2.3 Case 3: GroEL Additionally, we have used a GroEL dataset, kindly made available by Dr Ludtke (Ludtke *et al.*, 2004), composed by 26 micrographs of size 4082×6278 pixels. The sampling rate is 2.10 Å/pixel and the microscope voltage is of 200 kV. From this dataset, we have detected 4123 particles of size 128×128, using the method presented in (Abrishami *et al.*, 2013), and 16 classes were determined using CL2D (Sorzano *et al.*, 2010). In Figure 8a, we show the obtained 16 classes and the best volume (b) recovered by 3D-RANSAC map using default parameters, as before. In this case, the required processing time is ~5 min with the same computer as before. The processing time of CL2D is of 318 min, and then, the total time CL2D + 3D-RANSAC is of 324 min. The FSC curves between the phantom EMDB (EMDB code 1081, Ludtke *et al.*, 2004) and the 3D-RANSAC map at FSC = 0.5 and FSC = 0.143 are 10.2 and 9.7 Å, respectively, which means that both structures are similar. To further show the good agreement between the obtained initial volume and the used class averages, we present in Figure 9, the experimental class averages in the first row labeled as 'image', the corresponding initial volume projections at the same orientation in the second row ('imageRef') and the normalized cross correlation between both images in the third row ('maxCC'). As can be seen from Figure 9, there is a good agreement between the class averages and the corresponding projections.

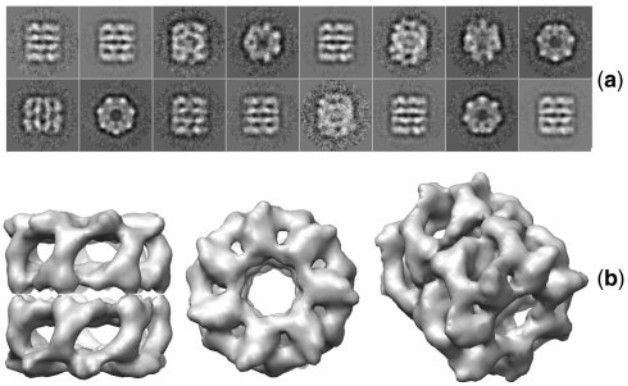

Fig. 8. Sixteen experimental class average images obtained using CL2D of GroEL experimental projections (**a**) and the best volume recovered by the proposed method at three different orientations (**b**)

3.2.4 Case 4: MCM467 Finally, for the sake of completeness, we show the behavior when dealing with small complexes in the

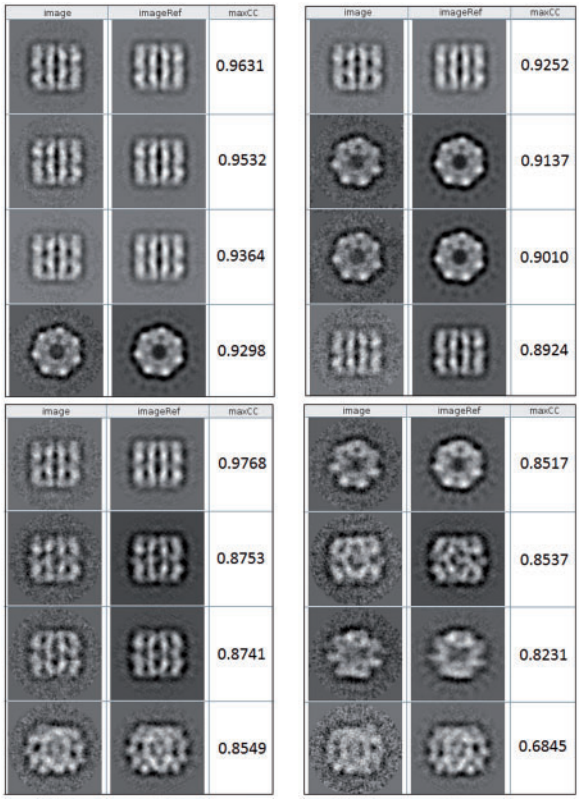

Fig. 9. Experimental class average images of the GroEL particles (image), corresponding initial volume projection at the same orientations (imageRef) and normalized cross correlation between both images (maxCC)

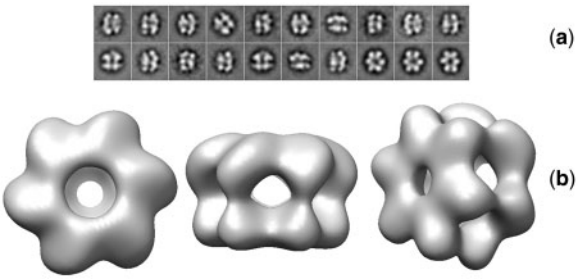

Fig. 10. Twenty obtained classes from experimental projections of MCM467 complex using CL2D approach (**a**) and three different orientations of the best initial volume obtained by the proposed method (**b**)

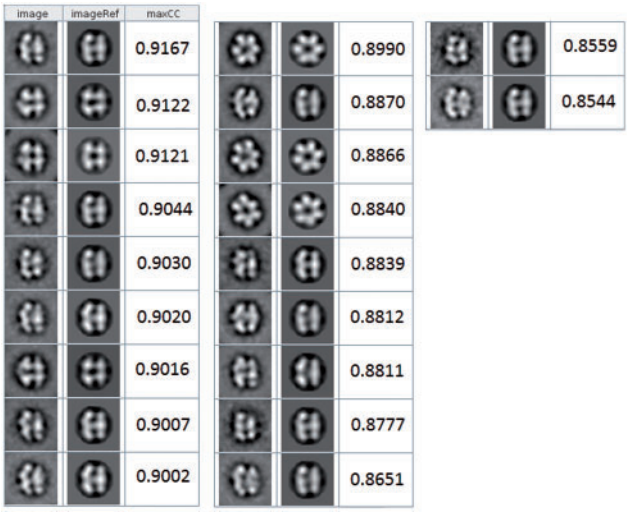

Fig. 11. Experimental class averages of the MCM467 complex (image), corresponding initial volume projection at the same orientations (imageRef) and normalized cross correlation between both images (maxCC)

order of a half million daltons using negative staining. This is an internally acquired dataset, corresponding to a certain MCM467 complex. The microscope is a JEOL JEM-1230, and the accelerating voltage is 100 kV. The nominal magnification is ×40 000 and the sampling rate 2.28 Å/pixel. We have obtained ∼6000 particles from 200 micrographs, using the method presented in (Abrishami *et al.*, 2013). From the picked particles, we obtained 128 class averages using CL2D (Sorzano *et al.*, 2010), which are manually curated to a smaller homogeneous dataset of 20 class averages.

In Figure 10a, we show the obtained class averages, after the curation process. As in the cases shown before, we have used the same input parameters for the determination of the 3D-RANSAC map. In this case, the required processing time is of 6 min with the same computer as before. The processing time of CL2D is of 720 min, and then, the total time CL2D + 3D-RANSAC is of 726 min. In Figure 10b we show the 3D-RANSAC map at different orientations.

As in the case before, to show the good agreement between the obtained initial volume and the used class averages, we present in Figure 11, the class averages, the corresponding projection of the initial volume at the same orientation and the normalized cross correlation between both images. As can be seen from Figure 11, there is a good agreement between the class averages and the corresponding projections.

4 DISCUSION AND CONCLUSION

Obtaining a reliable low-resolution initial map is a well-known current challenging problem in single particle electron microscopy. This statement can easily be supported considering the large number of recent publications about this topic (Elmlund and Elmlund, 2012; Elmlund *et al.*, 2013; Lyumkis *et al.*, 2013; Voss *et al.*, 2010). Existing current approaches are mainly based on common lines (Castón *et al.*, 1999; Crowther *et al.*, 1970; Elmlund and Elmlund, 2012; Elmlund *et al.*, 2010; Thuman-Commike and Chiu, 1997) or random model generation (Harauz and van Heel, 1985; Sanz-Garcia *et al.*, 2010; van Heel, 1984). However, in general, these methods are not easy to use, in term of the selection of input parameters, and are computing intensive. Naturally, the combination of a non-easy choice of input parameters together with long execution times complicates their practical use, requiring expert processing. Aware of these problems, we have set our aim at developing a simple-to-use method for which default parameters work well in most cases, at the same time as when the required computing time is minimized to minutes on a standard laptop

computer. Naturally, most methods can be trapped into local minima, and 3D-RANSAC is not an exception, but in this case, the landscape of solutions is particularly smooth, the use of RANSAC algorithm and, additionally, its short execution time reduce considerably this risk and also open the venue to future developments involving close-to-global optimization techniques for particularly challenging problems. Finally, we have developed 3D-RANSAC for the case of homogenous image populations, and at this stage this may be considered a limitation of the method. Observe that the input of the algorithm are homogeneous class average images, and therefore, the problem of separating structurally heterogeneous image sets into homogeneous classes has to be already solved in the previous classification approach. The extension of 3D-RANSAC to the non-homogenous case will be considered in future works, requiring a far from trivial extension of many concepts behind RANSAC.

In this article, we have presented a fast and efficient approach to obtain a reliable low-resolution initial volume from sets of macromolecular projection images without a priori information. The proposed method, instead of trying to explore the entire space of projection orientations, a task that is computationally intractable even for a few hundred images, uses a novel random modeling strategy based on an initial non-linear dimensionality reduction and RANSAC algorithm, which makes this approach efficient from a computational point of view. Observe that the probability of assigning the correct orientation to one projection corresponds to a standard uniform distribution. As the process of assigning the orientation to projections is statistically independent, the probability of giving correctly the angles to n particles corresponds to p^n (with P the probability to assign correctly the orientation to one projection). Therefore, if the number of images is high, this probability is low. Therefore, to increase this probability, we have used smaller sets of images (of typically nine projections) but at the same time capturing most of the structural information of the volume (note that this process is accomplished by the dimensionality reduction approach). 3D-RANSAC is a two-step survival algorithm. In the first step, a large number of models are generated (~380) but only the best k survive, which are the ones with the highest number of inliers (largest score). After this first selection process, these k models are refined again using all the initial classes by a projection matching approach, and at the end of this second selection step they are ranked again so that there is only one winner. We have tested our proposed method with synthetic (Bacteriorhodopsin) and experimental data (*BPV*, Eukaryotic Ribosome, GroEL and MCM467). In all cases, we have obtained fast and satisfactory results. The algorithm is freely available as a part of the Xmipp 3.1 package (de la Rosa-Trevín *et al.*, 2013).

Funding: The authors would like to acknowledge economical support from the Spanish Ministry of Economy and Competitiveness through grants AIC–A–2011–0638 and BIO2010-16566, the Comunidad de Madrid through grant CAM(S2010/BMD-2305) and the NSF through grant 1114901, 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.

Conflict of interest: none declared.

REFERENCES

Abrishami,V. *et al.* (2013) A pattern matching approach to selection of particles from low-contrast electron micrographs. *Bioinformatics*, **29**, 2460–2468.

Bilbao-Castro,J.R. *et al.* (2004) Phan3D: design of biological phantoms in 3D electron microscopy. *Bioinformatics*, **20**, 3286–3288.

Baker,T.S. and Cheng,R.H. (1996) A model-based approach for determining orientations of biological macromolecules imaged by cryo-electron microscopy. *J. Struct. Biol.*, **116**, 120–130.

Cantele,F. *et al.* (2003) The variance of icosahedral virus models is a key indicator in the structure determination: a model-free reconstruction of viruses, suitable for refractory particles. *J. Struct. Biol.*, **141**, 84–92.

Castón,J.R. *et al.* (1999) A strategy for determining the orientations of refractory particles for reconstruction from cryo-electron micrographs with particular reference to round, smooth-surfaced, icosahedral viruses. *J. Struct. Biol.*, **125**, 209–215.

Crowther,R.A. *et al.* (1970) Three dimensional reconstructions of spherical viruses by fourier synthesis from electron micrographs. *Nature*, **226**, 421–425.

de la Rosa-Trevín,J.M. *et al.* (2013) Xmipp 3.0: An improved software suite for image processing in electron microscopy. *J. Struct. Biol.*, **13**, 256–256.

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

Elmlund,H. *et al.* (2008) A new cryo-EM single-particle ab initio reconstruction method visualizes secondary structure elements in an ATP-fuelled AAA+ motor. *J. Mol. Biol.*, **375**, 934–947.

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

Elmlund,H. *et al.* (2013) PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy. *Structure*, **21**, 1299–1306.

Fischler,M.A. and Bolles,R.C. (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, **24**, 381–395.

Frank,J. (1996) *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic, New York, NY.

Giannakis,D. *et al.* (2012) The symmetries of image formation by scattering. I. Theoretical framework. *Opt. Express*, **20**, 12799–12826.

Harauz,G. and van Heel,M. (1985) Direct 3D reconstruction from projections with initially unknown angles. In: Gelsema,E.S. and Kanal,L.N. (eds) *Pattern Recognition in Practice II*. Elsevier, North-Holland Publishing, Amsterdam, pp. 279–288.

Henderson,R. *et al.* (1990) Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.*, **213**, 899–929.

Leschziner,A.E. and Nogales,E. (2006) The orthogonal tilt reconstruction method: an approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. *J. Struct. Biol.*, **153**, 284–299.

Ludtke,S.J. *et al.* (2004) Seeing GroEL at 6 Å resolution by single particle electron cryomicroscopy. *Structure*, **12**, 1–20.

Liu,X. *et al.* (2007) Averaging tens to hundreds of icosahedral particle images to resolve protein secondary structure elements using a multi-path simulated annealing optimization algorithm. *J. Struct. Biol.*, **160**, 11–27.

Lyumkis,D. *et al.* (2013) Optimod–an automated approach for constructing and optimizing initial models for single-particle electron microscopy. *J. Struct. Biol.*, **184**, 417–426.

Ogura,T. and Sato,C. (2006) A fully automatic 3D reconstruction method using simulated annealing enables accurate posterioric angular assignment of protein projections. *J. Struct. Biol.*, **156**, 371–386.

Penczek,P.A. *et al.* (1996) A common-lines based method for determining orientations for N > 3 particle projections simultaneously. *Ultramicroscopy*, **63**, 205–218.

Radermacher,M. *et al.* (1987) Three-dimensional reconstruction from a single-exposure random conical tilt series applied to the 50S ribosomal subunit of *Escherichia coli*. *J. Microsc.*, **146**, 113–136.

Sanz-Garcia,E. *et al.* (2010) The random-model method enables ab initio 3D reconstruction of asymmetric particles and determination of particle symmetry. *J. Struct. Biol.*, **171**, 216–222.

Scheres,S.H.W. *et al.* (2007) Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nat. Methods*, **4**, 27–29.

Sorzano,C.O.S. *et al.* (2010) A clustering approach to multireference alignment of single-particle projections in electron microscopy. *J. Struct. Biol.*, **171**, 197–206.

Thuman-Commike,P.A. and Chiu,W. (1997) Improved common-line-based icosahedral particle image orientation estimation algorithms *Ultramicroscopy*, **68**, 231–255.

Taylor,K.A. and Glaeser,R.M. (2008) Retrospective on the early development of cryo-electron microscopy of macromolecules and a prospective on opportunities for the future. *J. Struct. Biol.*, **163**, 214–223.

Voss,N.R. *et al.* (2010) A toolbox for ab initio 3-D reconstructions in single-particle electron microscopy. *J. Struct. Biol.*, **169**, 389–398.

Wolf,M. *et al.* (2010) Subunit interactions in bovine papillomavirus. *Proc. Natl Acad. Sci. USA*, **107**, 6298–6303.

van Heel,M. (1987) Angular reconstitutiona posteriori assignment of projection directions for 3-D reconstruction. *Ultramicroscopy*, **21**, 111–123.

van Heel,M. (1984) Three-dimensional reconstruction from projections with unknown angular relationships. In: Csanády,Á. *et al.* (ed.) *Eighth European Congress on Electron Microscopy*. Vol. 2. Programme Committee of the Eighth European Congress on Electron Microscopy, Budapest, Hungary, pp. 1347–1348.

Zhang,Z. and Hongyuan,Z. (2005) Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM J. Sci. Comput.*, **26**, 313–338.

Zhang,X. and Zhou,H.Z. (2011) Limiting factors in atomic resolution cryo electron microscopy: no simple tricks. *J. Struct. Biol.*, **175**, 253–263.