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/

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

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

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)

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

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:

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

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

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