Journal of Structural Biology 187 (2014) 112–118

Contents lists available at ScienceDirect

Journal of Structural Biology

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

Robust evaluation of 3D electron cryomicroscopy data using tilt-pairs

Christopher J. Russo, Lori A. Passmore

MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK

ARTICLE INFO

Article history:
Received 7 May 2014
Received in revised form 17 June 2014
Accepted 20 June 2014
Available online 9 July 2014

Keywords:
Electron cryo-microscopy
Single-particle reconstruction
Structure validation
Cryo-EM
Protein structure

ABSTRACT

Determining the structure of a protein complex using electron microscopy requires the calculation of a 3D density map from 2D images of single particles. Since the individual images are taken at low electron dose to avoid radiation damage, they are noisy and difficult to align with each other. This can result in incorrect maps, making validation essential. Pairs of electron micrographs taken at known angles to each other (tilt-pairs) can be used to measure the accuracy of assigned projection orientations and verify the soundness of calculated maps. Here we establish a statistical framework for evaluating images and density maps using tilt-pairs. The directional distribution of such angular data is modelled using a Fisher distribution on the unit sphere. This provides a simple, quantitative and easily comparable metric, the concentration parameter κ, for evaluating the quality of datasets and density maps that is independent of the data collection and analysis methods. A large κ is indicative of good agreement between the particle images and the 3D density map. For structure validation, we recommend $\kappa > 10$ and a p-value <0.01. The statistical framework herein allows one to objectively answer the question: Is a reconstructed density map correct within a particular confidence interval?

1. Introduction

Single particle electron microscopy (EM) can be used for three-dimensional (3D) structure determination of biological macromolecules. With the advent of direct electron detectors, more stable stages and reliable microscopes with field emission guns, near atomic resolution structures are now possible in the best cases (Kuhlbrandt, 2014). Still, important biological information can be obtained from medium resolution (10–50 Å) density maps where the secondary structure of the molecules is not resolved.

In single particle EM, two dimensional (2D) projection images of biological specimens are recorded in an electron microscope, their relative orientations are determined using one of a number of alignment algorithms, and finally one or more 3D reconstructions are calculated (Frank et al., 1996; Van Heel et al., 1996; Marabini et al., 1996; Grigorieff, 2007; Tang et al., 2007; Scheres, 2012). With favourable datasets (high signal-to-noise, even particle distributions, homogeneous conformation, etc.), iterative refinement of the orientations assigned to each particle image will converge to the true 3D density map. But because biological specimens are radiation sensitive, imaging takes place under low-dose conditions resulting in low signal-to-noise images. Moreover, complex heterogeneity, blurring of particle images due to radiation-induced motion and unfavourable protein interactions with surfaces degrade image quality. Obtaining an initial model that is suitable for accurate refinement of orientation parameters also remains a major challenge, especially for molecules lacking distinct low-resolution structural features (Henderson et al., 2011; Henderson and McMullan, 2013; Elmlund et al., 2013). Thus, in unfavourable cases, the refinement procedure can converge to a local minimum with an incorrect 3D map (Stewart and Grigorieff, 2004; Scheres and Chen, 2012; Murray et al., 2013; Henderson, 2013). It is therefore important to independently validate whether the resultant 3D density map is correct.

Analysis of pairs of particle images recorded at different tilt angles (tilt-pairs) provides an objective measure of the accuracy of particle alignment and the validity of reconstructed maps that is not subject to the problems associated with over-fitting of noisy data (Rosenthal and Henderson, 2003; Henderson et al., 2011). Tilt-pair data are easily collected with any single particle dataset, and are evaluated by determining whether the independently assigned orientation parameters from each tilt-pair match the known tilt angle and direction (Wasilewski and Rosenthal, 2014). Ideally, the calculated tilt angle and tilt direction would be located close to the true tilt angle and direction of the goniometer for all particles. Although this is true for large complexes that align well (e.g. rotavirus with molecular weight ~50 MDa), many smaller specimens show a large scatter of directions (Henderson et al.,

E-mail addresses: crusso@mrc-lmb.cam.ac.uk (C.J. Russo), passmore@mrc-lmb.cam.ac.uk (L.A. Passmore)

http://dx.doi.org/10.1016/j.jsb.2014.06.006
1047-8477/

2011). In such cases, it can be difficult to decide whether the clustering of points is adequate to validate a given 3D map.

A robust statistical analysis of the tilt-pair data could provide a rapid assessment of image and map quality that could be used to improve data collection and processing, and could be reported along with the structure, much as the free R parameter is used to asses the quality of crystal structures (Brunger, 1992). The discrete angular data generated by tilt-pair analysis comprise a distribution of directions on the unit sphere, thus making them well suited for analysis using the calculus of directional statistics. The statistics of directions is well established in several fields, and can provide rigorous and quantitative answers to important questions about experimental data quality and validity (Fisher et al., 1987; Mardia and Jupp, 2000; Tauxe, 2010). With this in mind, given one or more tilt-pair datasets, we provide methods to answer the following practical questions using statistical tests:

1. Is a particular set of tilt-pair measurements randomly distributed (and therefore should the corresponding dataset or map be discarded due to poor quality)?
2. Given a set (or sets) of tilt-pairs, is dataset A better than dataset B? or is map A better than map B?
3. Does a given dataset and map show evidence of systematic bias not assumed during the generation of the map or angular assignments?
4. Is a reconstructed density map correct to within a specified level of confidence?

2. Methods

2.1. Statistical model

To analyse a particular set of tilt-pair measurements we model the distribution of directions as a Fisher distribution on the unit sphere (Fisher, 1953). The Fisher distribution is one in which the probability of an observed direction has a density

$$f(\omega) \propto e^{\kappa \cos \omega} \tag{1}$$

where ω is the angle between the observed and the true direction. The precision parameter κ is the concentration of the distribution and is analogous to the inverse of the width of the Gaussian distribution. A κ of 0 indicates a uniform probability in all directions; as $\kappa \to \infty$ the distribution becomes more sharply peaked around the mean direction. Four pseudo-random samples of 100 points, taken from Fisher distributions with $\kappa = \{1, 10, 100, 1000\}$, are shown in Fig. 1(a).

To find the mean direction given a set of N tilt-pair angles $(\theta_1, \phi_1) \ldots (\theta_N, \phi_N)$, where (θ_i, ϕ_i) is the azimuth and inclination of a particular tilt-pair, first we convert each of the angles from spherical polar coordinates $(\theta_i,\ \phi_i)$ to vectors in Cartesian coordinates on the unit sphere:

$$(x_i, y_i, z_i) = (\sin\theta_i \cos\phi_i, \sin\theta_i \sin\phi_i, \cos\theta_i) \tag{2}$$

Next, we calculate the magnitude of the sum of each of the vector components over all tilt-pair angles

$$R = \sqrt{\left(\sum_i x_i\right)^2 + \left(\sum_i y_i\right)^2 + \left(\sum_i z_i\right)^2} \tag{3}$$

The mean direction of the Cartesian component vectors is then

$$(\bar{x}, \bar{y}, \bar{z}) = \left(\frac{1}{R}\sum_i x_i, \frac{1}{R}\sum_i y_i, \frac{1}{R}\sum_i z_i\right) \tag{4}$$

We convert these back to an inclination and azimuth to find the mean tilt direction:

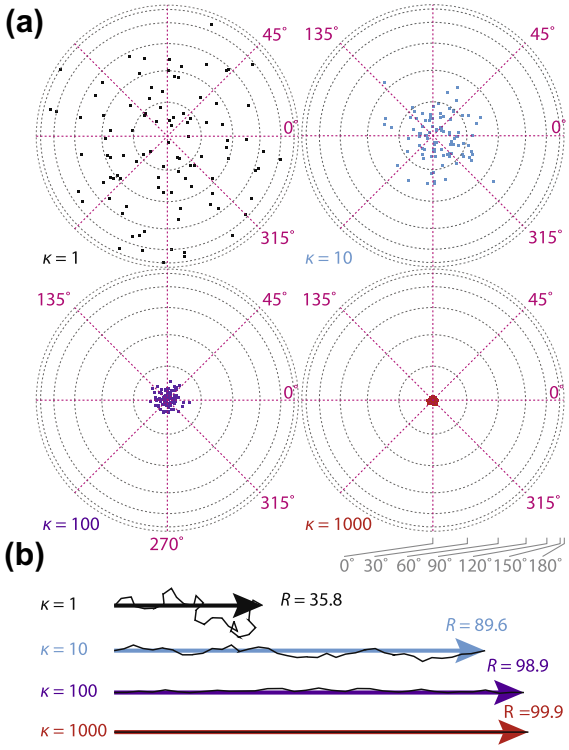

Fig.1. Fisher distributions using 100 simulated data points. Panel (a) shows four Fisher distributions on the unit sphere plotted using Lambert equal area projections for various concentration parameters, κ. For illustration, the mean direction is the pole of the sphere, which points out of the page. In the plots, the radius indicates the angle θ from 0° at the centre to 180° at the edge, and the azimuth indicates the direction of the tilt. Panel (b) shows a graphical construction of the R parameters for the same κ values in (a). Black segments are cartoons meant to illustrate how the individual direction vectors sum to a longer R as their directions become more correlated with each other. Lengths of R are proportional to the actual values for the distributions in (a), with the values indicated.

$$(\bar{\theta}, \bar{\phi}) = \left(\arccos \bar{z}, \arctan \frac{\bar{y}}{\bar{x}}\right) \tag{5}$$

The mean direction obtained from Eq. (5) represents an estimate of the true tilt direction based on the available data. Other estimates of the true direction are possible and we consider more below. The uncertainty in the mean direction as an estimate of the true direction can be represented by a confidence interval about the mean. Given that the data are taken from a Fisher distribution, we calculate the confidence interval for a given p-value, which is represented by a cone of solid angle around the mean direction that intersects the sphere in a circle with radius

$$\alpha_c = \arccos\left\{1 - \frac{N-R}{R}\left[\left(\frac{1}{p}\right)^{1/N-1} - 1\right]\right\}. \tag{6}$$

Next we calculate the concentration (precision) parameter of the distribution, κ, using the approximation (Fisher, 1953)

$$\kappa \simeq k = \frac{N-1}{N-R} \tag{7}$$

which we have tested using simulations (Section 2.3) and verified for $10 \leqslant N \leqslant 10^6$ and $1 \leqslant \kappa \leqslant 10^6$.

Finally, we calculate the median direction on the sphere (Fisher, 1985). Analogous to the linear median, the geometric median direction is defined as the location on the sphere where the sum of distances to all the points in the distribution is minimised. Various distance functions on the unit sphere can be used for this calculation; we chose the magnitude of the vector distance between