Real-time and decision taking selection of single-particles during automated cryo-EM sessions based on neuro-fuzzy method

David Gil-Carton [a,*], Miguel Zamora [a], James D. Sutherland [c], Rosa Barrio [c], Izaskun Garrido [b], Mikel Valle [a], Aitor J. Garrido [b]

[a] Structural Biology Unit, Cooperative Center for Research in Biosciences CIC bioGUNE, Bizkaia Technology Park, 48160 Derio, Spain
[b] Automatic Control Group – ACG, Department of Automatic Control and Systems Engineering, University of the Basque Country (UPV-EHU), 48013 Bilbao, Bizkaia, Spain
[c] Functional Genomics Unit, Cooperative Center for Research in Biosciences CIC bioGUNE, Bizkaia Technology Park, 48160 Derio, Spain

ARTICLE INFO

ABSTRACT

Cryo-electron microscopy (cryo-EM) is a three-dimensional (3D) averaging technique that makes use of two-dimensional (2D) images of biological macromolecules preserved in a thin layer of vitreous ice. Recent advances in the field have facilitated the evolution of cryo-EM towards atomic resolution, and the technique provides 3D maps with detailed description of biological macromolecules. Data acquisition at the transmission electron microscope (TEM) is the first crucial step during the single-particle analysis workflow in cryo-EM. In order to exploit the potential of this structural technique for atomic or near-atomic resolution, the initial collection must allow recording of large datasets and, hence, requires operating the TEM in automated mode. The quality of the acquired dataset relies, however, on the expertise of researchers and unsupervised operations might result in low data quality. This work presents the first expert system integrated in a novel scheme to automate cryo-EM data acquisition in a TEM. This development takes advantage of fuzzy logic systems to integrate the working mode of an expert in a linguistic manner and to learn from acquired data through an adaptive network. A new method based on different image-processing algorithms and on adaptive neuro-fuzzy inference systems (ANFIS) identifies, in an unsupervised manner, the single-particles present in cryo-EM images during the automated acquisition on a TEM. This single-particle identification system is integrated in a new intelligent control scheme to automate cryo-EM data acquisition. A classic fuzzy inference system (FIS) was programmed to make appropriate decisions during the session. The designed system can be trained for a specific sample and allows for unsupervised but efficient data collection imitating the working mode of an experienced microscopist.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

In recent years cryo-EM single-particle analysis (Frank, 2006) has become a high-resolution technique (Kühlbrandt, 2014) due to the development of direct electron detectors and new procedures in image/video processing (Bai, McMullan, & Scheres, 2015; Bammes, Rochat, Jakana, Chen, & Chiu, 2012; Campbell et al., 2014; McMullan, Faruqi, Clare, & Henderson, 2014; Ruskin, Yu, & Grigorieff, 2013). Consequently, there is a need for high-throughput and automation of modern microscopes to generate large datasets

of images to exploit this potential for near atomic resolution. Current available methods for automated acquisition in cryo-EM, such as the commercial EPU (E Pluribus Unum) software, or other academic software packages (Carragher et al., 2000; Zhang et al., 2009), allow collecting images from pre-selected areas. However, none of these softwares makes decisions during an automated session according to the quality of the cryo-EM images being recorded. Until now, the automatic evaluation of the data being collected during cryo-EM data acquisition sessions has not been available and, in the case of an unsupervised TEM session, the final output might be of poor quality. The presence of contamination, an inadequate ice thickness, empty holes due to broken carbon film, aggregation, poor concentration of particles, or damaged specimens, might severely reduce the final size of useful dataset. In this context, the development of new intelligent control schemes that imitates the behavior of expert researchers during cryo-EM data

* Corresponding author. Tel.: +34 946572516; fax: +34 944061301.
E-mail addresses: dcarton@cicbiogune.es, davidgilcarton@yahoo.es (D. Gil-Carton), mzamora@cicbiogune.es (M. Zamora), jsutherland@cicbiogune.es (J.D. Sutherland), rbarrio@cicbiogune.es (R. Barrio), izaskun.garrido@ehu.es (I. Garrido), mvalle@cicbiogune.es (M. Valle), aitor.garrido@ehu.es (A.J. Garrido).

acquisition seems to be essential to improve acquisition outcomes. Over the last 15 years there has been a great deal of research attempting to automate single-particles selection process in cryo-EM images using different algorithms divided into two categories: feature-based or template-based approaches (Chen, & Grigorieff, 2007; Norousi et al., 2013; Roseman, 2003; Scheres, 2015; Voss, Yoshioka, Radermacher, Potter, & Carragher, 2009; Woolford et al., 2007; Zhao, Brubaker, & Rubinstein, 2013). These methods perform the single-particle detection process after obtaining the complete dataset of cryo-EM images. Feature-based approaches have the difficulty of measuring acceptable local features to distinguish the areas of low contrast single-particles from the background in noisy cryo-EM images. The weakness of template-based approaches is the high rate of false-positive selection (Zhu et al., 2004). Furthermore, often cryo-EM datasets present heterogeneous populations of particles, due to limitations in biochemical sample preparation, or may include several contaminants (small ice crystals). With this in mind, we have implemented a neuro-fuzzy single-particle identification tool that can be trained to evaluate the distribution of particles and the quality of the cryo-EM images during the acquisition and to change the collection strategy accordingly. The main goal is to obtain a hybrid methodology that combines computer vision pattern recognition algorithms (Parker, 1997) with an adaptive network-based fuzzy inference scheme (Azeem, Hanmandlu, & Ahmad, 2000; Jang, 1993) to imitate the behavior of expert researchers. In summary, while Mamdani-type FIS makes decisions during the automatic acquisition of images, ANFIS monitors the presence of good single-particles or ice crystals within the recorded files and provides an output that feeds the FIS system to define the following steps in automatic data collection.

2. Automatic single-particle identification system based on neuro-fuzzy methods

2.1. Theory and calculations

The proposed procedure allows the automation of valid single-particles selection in each recorded cryo-EM image, avoiding ice contamination, aggregates, or incomplete particles. The method is based on the sequence of the following steps, commonly used in material science research (Hernández, Sáez, & Mery, 2004; Johnsrud, Yang, Nayak, & Punyasena, 2013; Papa et al., 2013; Yang, & Ahuja, 2014; Zapata, Vilar, & Ruiz, 2011): image pre-processing, extraction/segmentation of areas of interest, conversion into binary form, extraction of discriminatory features, and an identification process. The nature and quality of different samples in cryo-EM images depends on the experiment. Therefore, the most important property of the single-particle selection method must be the adaptability to any type of sample. The automatic selection system proposed in this work is formed by different ANFIS models that can be programmed and trained to recognize and select separately any type of single-particle present in the cryo-EM images. Our single-particle selection system starts with digital image pre-processing techniques to improve the contrast, so that single-particles in the cryo-EM image become more apparent than the background. Later, different segmentation processes, including a classical threshold selection method from gray-level histograms, Canny edge detection algorithm and texture analysis, were combined to isolate the areas of single-particles from the background. Then, eight discriminatory features describing the size, morphology (spherical, cylindrical or icosahedral shape), gray scale contrast, and aggregation state (isolated versus aggregated) were extracted from these segmented areas. These discriminatory features form the input pattern for the ANFIS part of the single-particle selection system. Each ANFIS is trained for the identification of a unique type of single-particle and the output of

each ANFIS model is a crisp value between {−1; +1}, range that indicates the likelihood ratio of belonging to a particular type of single-particle for each ANFIS.

2.1.1. Image pre-processing to improve the signal-to-noise ratio of noisy cryo-EM images

Cryo-EM images usually are noisy and do not exhibit strong contrast between single-particle areas and the background due to the low dose necessary to preserve the original structure of biological objects (Baxter, Grassucci, Gao, & Frank, 2009). Therefore, after cryo-EM digital image acquisition, it is necessary to adopt a pre-processing stage to reduce the noise and improve the contrast before segmentation. A combination of different filters and histogram adjustment produced satisfactory results with biological macromolecules of different size (from groEL chaperones to large viruses). Two pre-processing steps were carried out. In the first step an adaptive Wiener filter (Lim, 1990; Parker, 1997) with a kernel of 7-by-7, a nonlinear median filtering (Huang, Yang, & Tang, 1979), and a 5-by-5 Gaussian low-pass filter were applied for smoothness, noise reduction and the elimination of small artifacts in the cryo-EM images. The adaptive Wiener filter eliminates white additive noise while it simultaneously preserves the edges and high spatial frequency details in the images. Nonlinear median filters replace each pixel value by the median value of the neighboring pixels. This type of filter reduces salt-and-pepper noise, preserving the edges of the objects in the image. Gaussian low-pass filters only affect to regions with abrupt changes, with the cutoff frequency being determined by the kernel size and its coefficients (Kalman, 1960). In a second step, a normalization process by histogram stretching to increase the dynamic range and contrast-limited adaptive histogram equalization (CLAHE) (Zuiderveld, 1994) method accomplishes the contrast enhancement. CLAHE is useful to enhance the details in an image by improving the local contrast.

2.1.2. Segmentation process of single-particle areas

The goal of the segmentation processes is to convert the gray scale image into a binary image separating pixels for each single-particle from the background pixels. The combination of three different segmentation processes results in a very efficient algorithm for any type of macromolecular complex. First, a classical threshold selection method (Otsu, 1979) chooses the optimal threshold k to binarize the image, so that one part has mean gray value $\mu_0 = \sum_{i=1}^{k} x_i p_i$ and probability $\omega_0 = \sum_{i=1}^{k} p_i$, while the other part has mean gray value $\mu_1 = \sum_{i=k+1}^{n} x_i p_i$ and probability $\omega_1 = \sum_{i=k+1}^{n} p_i$. The optimal threshold value k is chosen to minimize the intra-class variance, defined as a weighted sum of variances of the two classes, calculated as:

$$\sigma^2(k) = \omega_0(\mu_0 - \mu)^2 + \omega_1(\mu_1 - \mu)^2 = \frac{(\mu\omega(k) - \mu(k))^2}{\omega(k)(1 - \omega(k))}$$

where $\mu_0 = \frac{\mu(k)}{\omega(k)}$ and $\omega_0 = \omega(k)$ so that $\mu_1 = \frac{\mu - \mu(k)}{1 - \omega(k)}$ and $\omega_1 = 1 - \omega(k)$. Second, a Canny edge detection method is used to avoid the selection of incomplete particles corresponding to ice artifacts in the background (Canny, 1986). The Canny algorithm uses the calculus of variations where the optimal function is described by the summation of four exponential terms, but it can be approximated by the first derivative of a Gaussian. Canny algorithm searches for horizontal, vertical and diagonal edges. It uses two different thresholds to detect both strong and weak edges, including only the weak edges in the output if they are connected to strong edges. Third, a texture segmentation process that measures the entropy value of the neighborhood improves the efficiency of the total segmentation (Gonzalez, Woods, & Eddins, 2003). Morphological erosion was also applied to remove very small regions corresponding to noise (Soille, 1999). Then, a morphological dilation