Journal of Structural Biology 183 (2013) 342–353

Contents lists available at ScienceDirect

Journal of Structural Biology

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

Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques

J. Vargas [a,*], V. Abrishami [a], R. Marabini [b], J.M. de la Rosa-Trevín [a], A. Zaldivar [a], J.M. Carazo [a], C.O.S. Sorzano [a]

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

ARTICLE INFO

Article history:
Received 7 May 2013
Received in revised form 10 July 2013
Accepted 31 July 2013
Available online 6 August 2013

Keywords:
Electron microscopy
Particle picking
Machine learning
Single particle analysis

ABSTRACT

Three-dimensional reconstruction of biological specimens using electron microscopy by single particle methodologies requires the identification and extraction of the imaged particles from the acquired micrographs. Automatic and semiautomatic particle selection approaches can localize these particles, minimizing the user interaction, but at the cost of selecting a non-negligible number of incorrect particles, which can corrupt the final three-dimensional reconstruction. In this work, we present a novel particle quality assessment and sorting method that can separate most erroneously picked particles from correct ones. The proposed method is based on multivariate statistical analysis of a particle set that has been picked previously using any automatic or manual approach. The new method uses different sets of particle descriptors, which are morphology-based, histogram-based and signal to noise analysis based. We have tested our proposed algorithm with experimental data obtaining very satisfactory results. The algorithm is freely available as a part of the Xmipp 3.0 package [http://xmipp.cnb.csic.es].

1. Introduction

Single particle analysis (SPA) techniques based on transmission electron microscopy (TEM) can obtain three-dimensional (3D) reconstructions of biological complexes near atomic resolution (Zhang and Zhou, 2011). However, this high resolution studies require acquiring tens of thousands of projection images. The typical particle selection or picking approach consists in locating the two-dimensional (2D) projections of the biological structure under study within the captured electron micrographs. This process may be cumbersome, laborious and time consuming, especially if done manually, and represents a major bottleneck for SPA of large datasets. However, the success of the reconstruction crucially depends on the number and the quality of the 2D picked particles. In order to develop high-throughput methods that minimize the user iteration in the different processing steps, a number of different automatic and semiautomatic particle picking approaches have been proposed. Automatic particle picking techniques (Chen and Grigorieff, 2007; Adiga et al., 2004; Huang and Penczek, 2004; Kumar et al., 2004; Ogura and Sato, 2005; Plaisier et al., 2004; Rath and Frank, 2004; Roseman, 2004; Singh et al., 2004; Wong et al., 2004) consist in image processing algorithms capable to detect and boxing out the particle projections without the need of any user interaction. These methods are usually fast and provide a large number of particles; however, they may have accuracy and robustness problems, providing a relatively large set of incorrect and erroneously picked particles (false positives). These false positives typically range, depending on the picking algorithm, from fractions of 10% to more than 25% (Zhu et al., 2004). Therefore, after the picking process, it is always required to perform a subsequent manual curation (screening) approach to reject false positives. In turn, semiautomatic picking approaches require the user to provide an initial set of particles, which are manually picked. From this training set, the algorithms learn the kind of objects to be detected and boxed from the micrographs (Arbeláez et al., 2011; Hall and Patwardhan, 2004; Mallick et al., 2004; Ogura and Sato, 2004; Plaisier et al., 2004; Short, 2004; Sorzano et al., 2009; Volkmann, 2004). These methods are halfway between manual and automatic methodologies but they also require a posterior manual particle screening process.

Incorrectly picked particles typically appear because the sample presents some degree of heterogeneity and/or the existence of overlapping or degraded particles in the dataset. Additionally, some of the picked particles usually are strongly affected by noise or may even contain only noise. Finally, the presence of image artifacts such as ice, dust and contaminations, can corrupt the detected particles. In all of these cases, these picked particles represent false positives and must be discarded. Fig. 1 shows an example of typical cases of correctly and incorrectly picked particles. In Fig. 1(a), we see a correctly picked particle (true positive) of *Bovine papillomavirus* (Wolf et al., 2010). In turn, Fig. 1(b), (c) and (d) show examples

* Corresponding author. Fax: +34 585 4506.
E-mail address: jvargas@cnb.csic.es (J. Vargas).

1047-8477/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jsb.2013.07.015

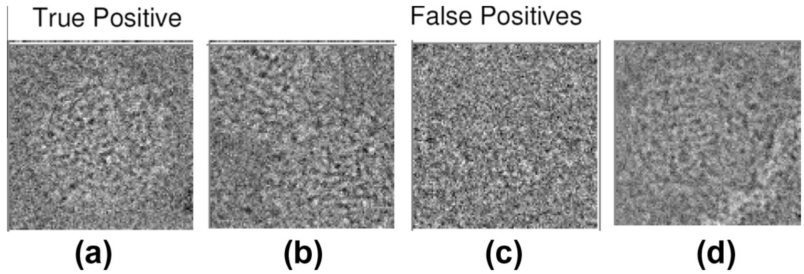

Fig.1. Example of three kinds of typical automatic or semiautomatic particle picking problems; (a) correctly picked particle (true positive) of *Bovine papillomavirus*, two overlapping particles (b), only noise image (c) and particle affected by an artifact (d).

of two overlapping particles, an only noise image and a particle affected by an artifact, respectively.

Post-picking methodologies, based on processing the output of automatic and/or semiautomatic particle picking methods, have been previously proposed to separate particle images from non-particle ones (Norousi et al., 2013; Sorzano et al., 2004). Since these two strategies are conceptually close to our newly proposed approach, we will describe them in some depth in the following. In (Norousi et al., 2013), a supervised discriminative post-picking approach based on characteristic features calculated from the boxed out images is presented. This method requires the user to provide a training dataset from the previously picked particles. The classification is based on the extraction of distinct features, such as: radially weighted average intensity, phase symmetry and dark dot dispersion. First, the radially weighted average intensity is calculated as a weighted sum of the pixel intensities. The weighting is inversely proportional to the pixel's Euclidean distance from the image center. Note that this descriptor does not obtain reliable values in cases where the particle centroid is not placed exactly at the image center. In these cases, the particles will be incorrectly considered as false positives. Additionally, the phase symmetry is computed from a set of 2D wavelets that extracts local frequency information (Kovesi, 1997). In (Norousi et al., 2013), it is claimed that non-particle images will have larger locally symmetric areas, but this consideration may not be general. Finally, the dark dot dispersion descriptor consists in convolving the image with a 2D symmetric Gaussian kernel and binarizing the resultant image using the 0.95 intensity quartile. The authors establish that the average distance between the dark detected regions is larger in particle images than in non-particle images. Observe that the good performance of this descriptor depends strongly on the signal-to-noise ratio (SNR) of the images. Additionally, in (Sorzano et al., 2004) it is presented a post-picking particle sorting method to determine the quality of the input boxed images and identifying erroneous particles. This method uses as descriptors both the radial average intensity and the image histogram. Note that as mentioned before, the radial average descriptor presents problems in cases where the particle centroid is not placed in the image center.

Building on the large body of experience in the field, which has been briefly presented in previous paragraphs, we present in this work a new approach to particle quality sorting that outperforms previous approaches while being computationally efficient. The main objective of this work is to detect outliers (incorrectly picked particles) from a previously picked dataset and not to be very accurate in the fine assessment of correctly picked particles. The input of the algorithm is a previously picked particle dataset that can be affected by outliers and may be coming from any manual, semiautomatic or automatic particle picking method. The unique requirement of our new approach is that there is a majority of correctly picked particles in the dataset. For each of the provided particles, several different types of descriptors are obtained, that are morphology, histogram and noise-based. Morphology descriptors encode information about the shape of the particles. Histogram descriptors give statistical intensity information of the particle images. Finally, noise-related descriptors allow for the separation of noise-only images from those containing signal and noise. For each particle and type of descriptor we compose a vector and, consequently, we will have three vectors per particle. Stacking the vectors of the same descriptor class and of all particles, we compose three descriptor matrices. Using a principal component analysis (PCA) dimensionality reduction approach (Roweis, 1998), we obtain an error score (z-score) for each particle taking into account the descriptor matrices. Furthermore, we study the statistical distribution of our proposed z-score under a set of simple hypothesis, reaching the conclusion that z-score values around 3 should be appropriated, specially when performing the sorting in a fully automatic, parameter-less, high throughput way. Low values of this z-score mean high reliability of the particle under study. Note that a reliable z-score measure is of high importance as it can be used to discriminate between true positive and false positive particles. Additionally, this reliable particle measure can also be utilized as a weighting parameter of the different projections in further processing steps, as for example the three-dimensional density reconstruction, although we have not exploited this issue in the present work.

The paper is organized as follows. In next section, we present the particle quality assessment and sorting method. In Section 3, we show some experimental data results and, finally, the discussion and conclusions are given.

2. Methods

In this section, we present the proposed approach to obtain the morphology, noise-based and histogram descriptors, as well as, the method to compute the particle score from them.

2.1. Morphology descriptors

The morphology descriptors obtain image features that enable discrimination of incorrect particlesbased on their general morphology/shape while eliminating overlapping particles. A good example of the kind of incorrectly picked particles that these morphology descriptors will remove can be seen in Fig. 1. Note that the shape of the particles shown in Fig. 1(b) and (d) is very different from the particle shape shown in Fig. 1(a). Therefore, in these cases, the morphology descriptors will provide valuable information to differentiate between these correctly and incorrectly selected particles. In this work, we use two different sets of morphology descriptors. The first one is based on an image normalization process using the spiral phase transform (SPT) method (Larkin et al., 2001) and Fourier filtering. The second one is derived from the particle autocorrelation map, which is sensitive to the particle shape, at the same time that it is shift invariant.