Contents lists available at ScienceDirect

Journal of Structural Biology

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

Automated particle picking for low-contrast macromolecules in cryo-electron microscopy

Robert Langlois [a], Jesper Pallesen [a,b], Jordan T. Ash [a,c,1], Danny Nam Ho [d], John L. Rubinstein [e,f], Joachim Frank [a,b,d,*]

[a] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, United States
[b] Howard Hughes Medical Institute, Columbia University, New York, NY 10032, United States
[c] Department of Biomedical Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, United States
[d] Department of Biological Sciences, Columbia University, New York, NY 10027, United States
[e] The Hospital for Sick Children Research Institute, Toronto M5G 0A4, Canada
[f] Departments of Biochemistry and Medical Biophysics, University of Toronto, Toronto M5S 1A8, Canada

ARTICLE INFO

Article history:
Received 2 January 2014
Received in revised form 11 February 2014
Accepted 2 March 2014
Available online 6 March 2014

Keywords:
Particle selection
Machine-learning
Automation
High-resolution
Cryo-EM

ABSTRACT

Cryo-electron microscopy is an increasingly popular tool for studying the structure and dynamics of biological macromolecules at high resolution. A crucial step in automating single-particle reconstruction of a biological sample is the selection of particle images from a micrograph. We present a novel algorithm for selecting particle images in low-contrast conditions; it proves more effective than the human eye on close-to-focus micrographs, yielding improved or comparable resolution in reconstructions of two macromolecular complexes.

1. Introduction

The term single-particle reconstruction refers to the reconstruction of a macromolecule from multiple projections, each presenting a single, freestanding copy of the macromolecule. These projections are obtained by cryo-electron microscopy (cryo-EM). The plunge-freeze procedure traps the molecules in a thin layer of vitreous ice. A low-dose electron beam captures a low-contrast, two-dimensional projection image (referred to as a micrograph) containing a collection of the molecules trapped in random orientations. The images of the molecules are then subjected to a computational workflow commonly referred to as single-particle analysis, which results in a 3D density map of the macromolecule.

A single high-resolution reconstruction of a 3D macromolecular complex requires the collection of thousands of micrographs, which typically yield hundreds of thousands of particle images. In cases where contrast is extremely low (e.g. with low electron exposures and low defocus settings), a researcher currently spends a substantial amount of time picking particle images from the micrographs. From an image-processing standpoint, the particle-picking problem can be broken down into two steps. First, candidate particle images must be selected from the micrograph; this step historically has been referred to as *particle selection*. Second, the “true” particles (i.e. those representing biological molecules) must be identified among those candidates that may contain falsely discovered non-particles such as contaminants or noise; this step is commonly referred to as *particle verification*. This effort is often compounded by specimen heterogeneity, i.e. multiple conformational states coexisting within the same sample. This problem makes it necessary to collect a larger dataset to ensure there is sufficient relevant data left, after classification, to build a high-resolution map of the structure of interest. Hence, particle picking, especially the second step of particle verification, represents a significant barrier to a completely automated, reproducible single-particle analysis workflow.

Considerable effort has been made to develop algorithms that aid the human eye in selecting good particle images in these extremely low-contrast micrographs (Glaeser, 2004; Langlois et al., 2011; Zhu et al., 2004). A strategy often used is to employ a

[*] Corresponding author at: Howard Hughes Medical Institute, Columbia University, Department of Biochemistry and Molecular Biophysics, 650 West 168th Street, Black Building 2-221, New York, NY 10032, United States. Fax: +1 212 305 9500.
E-mail address: jf2192@columbia.edu (J. Frank).
[1] present address: Department of Computer Science, Princeton University, Princeton, NJ 08540, United States.

http://dx.doi.org/10.1016/j.jsb.2014.03.001
1047-8477/© 2014 Elsevier Inc. All rights reserved.

cross-correlation search over the micrograph in identifying data windows containing candidate particles and then manually verify each window (Rath and Frank, 2004; Roseman, 2003). Another approach to limit the false discovery rate (Langlois and Frank, 2011) is to use hand-tuned thresholds, which can be applied on a micrograph-by-micrograph basis (Chen and Grigorieff, 2007; Tang et al., 2007) or over the entire set (Voss et al., 2009). The elements of subjectivity can be reduced using a machine-learning algorithm referred to as a *classifier*, a supervised learning tool which requires the user to define an initial selection comprising several hundred examples of "good" and "bad" windows (Arbeláez et al., 2011; Langlois et al., 2011; Zhao et al., 2013). Alternatively, candidate particle images identified can be aligned in 2D and then clustered into classes based on intrinsic information; this enables the user to look at the average of each class and either verify or reject the entire class, or further inspect individual particles within that class (Arbeláez et al., 2011; Shaikh et al., 2008). Nevertheless, current methods still require significant effort by the user to verify particles.

We envision a new type of tool that uses unsupervised learning to select particles from the micrograph with minimal user intervention. The user is only required to provide the approximate size of the macromolecule. Unsupervised learning leverages the observation that images of physical objects have limited complexity, and thus, can be described by a compact representation. We seek to further reduce this compact representation by exploiting the fact that the views of the macromolecules are linked by rigid-body transformations: azimuthal rotation and translation.

In the present study, we introduce a two-step automated particle-picking procedure. The first step is a modified template-matching procedure, termed AutoPicker, which identifies a set of candidate particle images from a collection of micrographs and rejects high-contrast contamination and noise using an unsupervised learning procedure. The second step employs an unsupervised one-class classifier, termed View Classifier or ViCer, which exploits the similarity among *aligned* true particles to reject outliers. To assess the quality of the final particle selection, we have applied the algorithm to identify and verify particles from two independent datasets recorded under low-contrast conditions: one of micrographs containing 70S ribosomes from *Escherichia coli* and the second containing molecules of the V/A-ATPase from *Thermus thermophilus*. The density maps obtained using the automatically selected particle images were compared to maps derived from manually selected particle images, which led to high-quality structures. We demonstrate that the particle images selected from of the AutoPicker/ViCer workflow lead to density maps with comparable, if not better, resolved features, and find that this outcome is in part a consequence of AutoPicker/ViCer's ability to identify additional true particles in close-to-focus micrographs.

2. Methods

2.1. Proposed particle-picking algorithm

The proposed automated particle-picking algorithm naturally reduces to two steps: (1) identification as well as an initial verification of potential particles with AutoPicker and (2) further verification using outlier rejection with ViCer.

2.1.1. AutoPicker

The AutoPicker algorithm, as outlined in Supplemental Fig. 1a, uses template matching to identify windows that contain candidate particle images in a micrograph and classification by unsupervised learning to reject both high contrast contaminants and noise windows. Template matching alone provides an excellent ranking of low-contrast, noisy particle (SNR ~0.06) windows over noise, yet provides no means for selecting the optimal threshold to distinguish these two groups. In addition, a micrograph may contain high-contrast contaminants such as ice crystals and bubbles in the ice after radiation damage of the specimen; depending on their size, windows containing contaminants are ranked, according to the cross-correlation score between each window and a template, higher than, or at the same level as, those containing particles. The unsupervised learning algorithm introduced by AutoPicker handles both of these limitations.

First, AutoPicker employs principal component analysis (PCA) over the power spectra of the extracted image windows, reducing each image to a single principal component. Then, assuming a Gaussian distribution, it rejects windows that fall in the tail, i.e. more than 4 standard deviations from the mean. While this cutoff might seem extreme, in practice only the noise windows follow a Gaussian distribution, whereas contaminants tend to follow a more skewed distribution on the tail. This cutoff targets only a specific type contaminant that proves deleterious to the next step. AutoPicker then repeats this procedure over the background surrounding the particle as defined by a ring around the particle; the size of the ring is defined as the particle radius multiplied by the exclusion multiplier and the width is the exclusion distance. Large contaminants and aggregation violate this ring of exclusion, and consequently, become outliers. This step eliminates the most obvious high-contrast contaminants.

Second, AutoPicker applies Otsu's algorithm (Otsu, 1979) on the cross-correlation scores of the remaining windows with the template in order to determine the optimal threshold that separates candidate particles from noise. Note that the order of these two steps is important because high-contrast contaminants tend to skew the cross-correlation histogram, causing Otsu's method to find a suboptimal threshold. In this work, the template was chosen as a disk with a radius corresponding to the particle size and its edges softened by application of a kernel with a Gaussian falloff.

2.1.2. ViCer 2.0

For relatively clean micrographs lacking ice crystals and other artifacts, the AutoPicker algorithm is sufficient to ensure good particle selection. However, many contingencies can contrive to produce less than ideal micrographs and in such cases additional contaminant removal proves necessary. The View Classifier (ViCer) can then be used to further clean the candidate particles of contaminants.

The original ViCer outlier rejection algorithm (Langlois et al., 2012), as outlined in Supplemental Fig. 1b, works by maximizing the similarity between true particles and, as a byproduct, is able to recognize contaminants as outliers. ViCer requires that the particle images have been aligned and grouped into views; it then uses the translation-invariant bispectral transforms of the particle images to further increase the similarity among true particles. Next, PCA is used to represent the bispectral transforms in a two-dimensional feature space. Visual inspection of this space revealed that the true projections tend to form a single cluster, surrounded by outlier contaminants.

The new ViCer algorithm includes two substantial improvements over the original algorithm. First, the PCA is replaced with an outlier-robust version of PCA called DHR-PCA (Feng et al., 2012). This robust PCA prevents corruption of the covariance matrix by contaminants, and as a consequence, yields principal components that better separate contaminants from true particles. Second, the Mahalanobis distance score (a multivariate *z*-score) replaces the ad hoc multivariate extension of the median absolute deviation (MAD) score (Hoaglin et al., 1983) to define the decision boundary between true particles and outlier contaminants. The Mahalanobis distance is defined as follows: