EMHP: An accurate automated hole masking algorithm for single-particle cryo-EM image processing.

Zachary Berndsen[1], Charles Bowman[1*], Haerin Jang[1], and Andrew B. Ward[1]

[1]*Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA USA*

[*]Email: bowman@scripps.edu

1 Introduction

The recent surge in popularity of single-particle cryo-EM as a tool for molecular structure determination alongside advances in software that have reduced the computational infrastructure needed to process single-particle datasets (Kimanius et al, 2016) have created the need for a more streamlined suite of tools to help locally facilitate initial data treatment and make processing more attainable at the workstation level.

Current technical limitations inherent to the process of structure determination via single-particle cryo-EM require collecting very large data sets – often several thousands of images. This task is facilitated by automated imaging software, however downstream preprocessing steps such as quality assessment and masking of individual images are still performed manually by the researcher and can become quite cumbersome. EMHP focuses on streamlining this preprocessing stage – specifically image assessment, masking, and pick filtering in preparation for single-particle analysis.

The need for hole masking in images stems from the fundamentals of cryo-EM sample preparation and data processing. Samples are traditionally prepared for imaging by flash freezing a few microliters of solution on copper mesh grids coated in holey carbon. The hole patterns suspend particles in a thin meniscus of vitreous ice, ideal for high resolution imaging, and are often used by automated collection software to assist in exposure targeting. (Suloway et al, 2005) While images are ideally taken entirely over these holes of thin ice, it is still often necessary to collect around the edges of the holes due to low particle densities, preferential particle distribution, or cost transfer function (CTF) estimation limitations. For these reasons, many images in a cryo-EM dataset contain sections of thick carbon support present in one or more quadrants of the image.

While automated particle selection algorithms allow for rapid selection of single particle projections from within larger images, these algorithms have difficulty or are not designed to discriminate between particles on carbon and particles in ice, resulting in the inclusion of many false positives into the initial particle stack which can dramatically increase computation times and have adverse effects on downstream classification and refinement steps.

Here we present a fast and accurate algorithm for detecting carbon supports in cryo-EM images, along with a small suite of tools for image assessment and pick filtering that allow users to preprocess their data rapidly and with minimal overhead while piping the results into common formats that are readable by the most popular single-particle cryo-EM processing packages. Our algorithm shows improved performance in terms of time and accuracy over existing hole-finding algorithms and can be easily executed in a minimal or non-HPC environment making it more accessible to users.

2 Software Package Overview

The software included in the EMHP package is coded using Python 2.7 and relies on packages that are freely available. EMHP uses the .mrc file parser methods implemented in pyami via Appion, (Lander et al, 2009) which are included in the code repository for convenience. The package includes a Tkinter-based GUI

image assessor, an implementation of the automatic hole masker and particle filter, a Tkinter-based GUI for manual hole masking, and a script that applies already computed masks to images.

2.1 Algorithm Description

We based our hole-finding algorithm on two main assumptions: that the edges of the carbon hole are manufactured to certain consistent specifications and that the carbon holes themselves have more textural features than the vitreous ice and particles filling the holes. First, the image is loaded and a Gaussian filter is applied. After normalization via contrast stretching (Fig. 1A), a standard Sobel edge filter is applied (Fig. 1B). This edge filter accentuates the texture along the edge of the carbon hole. Two subsequent rounds of edge enhancement are performed using radial summing followed by intensity thresholding resulting in a signal representing only the carbon edge (Fig. 1C-E). The image is then binned 100 times and a circle with dimensions set by the pixel size of the image and hole diameter specified by the manufacturer is fit to the binary thresholded image. The coordinates of the circle are then extrapolated back to the unbinned image and the mask is shifted toward or away from the edge based on user input. This final binary mask is then used to filter out the particle picks that lie on top of the carbon support. (Fig. 1F)

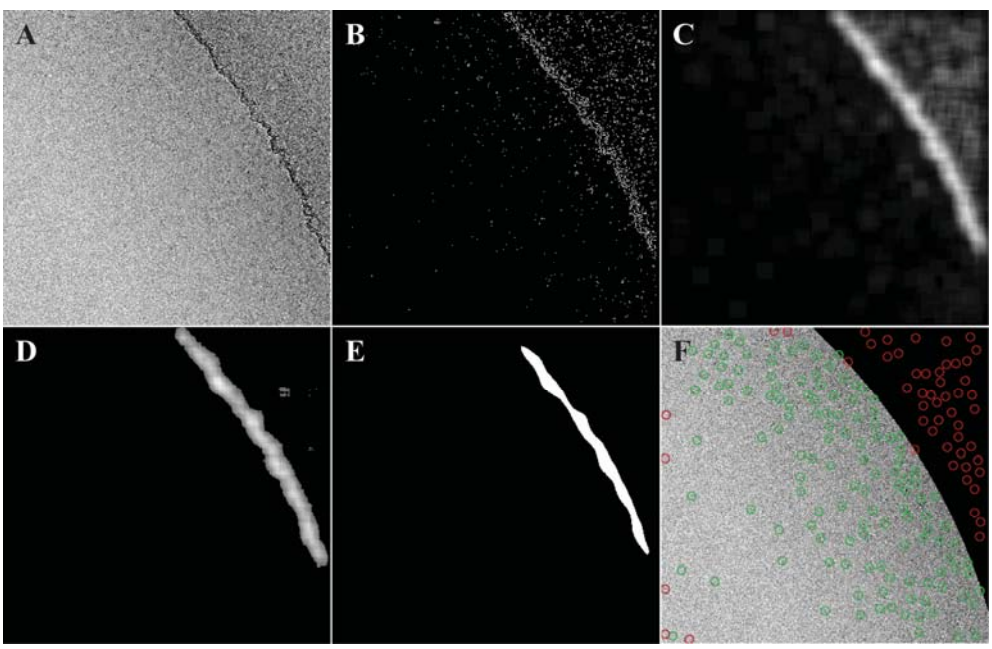

Fig. 1. EMHP automasking example. The image after (A) Gaussian filter and contrast stretching, (B) after applying a Sobel filter, (C) after one round of edge amplification by radial summing, (D) after applying a user-defined threshold and (E) after another round of edge amplification and binary thresholding. (F) The final image after circle fitting, masking, and pick filtering. Blue single particle picks are included, while red picks are excluded due to the mask, and green are excluded due to edge proximity.

2.2 Test Dataset and Methods

For a concrete evaluation of the algorithm's effectiveness, we conducted two performance tests on a set of 100 images chosen at random from a previously published dataset. (Lee et al, 2016) We compare performance of the most commonly used automasker, em_hole_finder, currently incorporated into the Appion (Lander et al, 2009) web based software suite with the EMHP automasking algorithm. To do this, masks were carefully constructed manually and a set of particle picks generated with DoG Picker (Voss et al, 2009) were assigned to either ice or carbon. Next, both algorithms were used to classify the same set of picks and the sensitivity (true positive rate) and specificity (true negative rate) were calculated for each (Table 1). These 100 test images have been included in the code repository for testing and benchmarking.

3 Conclusions

The EMHP automasking algorithm shows improved sensitivity compared to em_hole_finder. This highlights the main shortcoming of em_hole_finder, the tendency to over-mask thereby excluding a substantial number of true positives from the initial particle stack. The two algorithms performed equivalently in terms of specificity. We would like to note that over the course of testing and extensive in-house use, we found that the EMHP algorithm has trouble with edge detection in two specific situations: when carbon edges are poorly manufactured (overly jagged or un-circular), or when particle density at the very edge of the carbon holes is excessively high.

Rapid advances in single-particle cryo-EM data processing software and hardware capabilities are bringing state-of-the-art structure determination capabilities to the desktop. We saw the need for an alternative to current data pre-processing suites that addresses the unique image processing needs of the field. EMHP provides this in a simple python package that is accessible to users of all experience levels with minimal computational overhead.

Table 1. Benchmark comparisons using em_hole_finder and EMHP

Algorithm	Sensitivity	Specificity
EMHP	0.983	0.965
em_hole_finder	0.785	0.990

Funding

This work has been supported by the Bill and Melinda Gates Foundation CAVD (OPP1115782).

The authors have no conflicts of interest to declare.

References

Kimanius D, Forsberg BO, Scheres SH, & Lindahl E. (2016). Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. eLife, 5, e18722.

Lee JH, Ozorowski G, & Ward AB. (2016). Cryo-EM structure of a native, fully glycosylated, cleaved HIV-1 envelope trimer. Science, 351(6277), 1043-1048.

Lander GC, Stagg SM, Voss NR, Cheng A, Fellmann D, Pulokas J, ... & Lyumkis D. (2009). Appion: an integrated, database-driven pipeline to fa-cilitate EM image processing. J Struct Biol, 166(1), 95-102.

Suloway C, Pulokas J, Fellmann D, Cheng A, Guerra F, Quispe J, ... & Car-raghe, B. (2005). Automated molecular microscopy: the new Leginon sys-tem. J Struct Biol, 151(1), 41-60.

Voss NR, Yoshioka CK, Radermacher M, Potter CS, & Carragher B. (2009). DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. J Struct Biol, 166(2), 205-213.