MIT Libraries | DSpace@MIT

MIT Open Access Articles

Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

As Published: 10.1038/S41592-019-0575-8

Publisher: Springer Science and Business Media LLC

Persistent URL: https://hdl.handle.net/1721.1/136321

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike

Massachusetts Institute of Technology

HHS Public Access
Author manuscript
Res Comput Mol Biol. Author manuscript; available in PMC 2018 April 26.

Published in final edited form as:
Res Comput Mol Biol. 2018 April ; 10812: 245–247.

Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs

Tristan Bepler[1,2], **Andrew Morin**[2,6], **Alex J. Noble**[3], **Julia Brasch**[4], **Lawrence Shapiro**[4,5], and **Bonnie Berger**[1,2,6,*]

[1]Computational and Systems Biology, MIT, Cambridge, MA, USA

[2]Computer Science and AI Laboratory, MIT, Cambridge, MA, USA

[3]National Resource for Automated Molecular Microscopy, Simons Electron Microscopy Center, New York Structural Biology Center, New York, NY, USA

[4]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA

[5]Mortimer B. Zuckerman Mind Brain Behavior Institute, New York, NY, USA

[6]Department of Mathematics, MIT, Cambridge, MA, USA

Background

Structure determination with cryoEM involves reconstructing a 3D molecule from 2D projections. This process often requires tens to hundreds of thousands of experimental projections, or particles. Locating these particles in cryoEM micrographs, referred to as particle picking, is a major bottleneck in the current protein structure determination pipeline. This pipeline generally consists of sample and EM grid preparation, imaging, particle picking, and eventually structure determination. Labeling a sufficient number of particles to determine a high resolution structure can require months of effort – even with the use of existing methods designed to automate the process. Limitations of these tools include high false positive rates, requiring many hand-labeled training examples, and poor performance on non-globular proteins.

In order to better automate particle picking, and thus accelerate structure determination, we newly frame the particle picking problem as an instance of positive-unlabeled classification. In our framework, for a set of micrographs containing particles of interest with a small number labeled for training, we learn a convolutional neural network (CNN) to classify particles from background using a novel generalized-expectation criteria [1] to regularize the model's posterior over the unlabeled micrograph regions. This advance allows us to achieve state-of-the-art particle detection results with minimal hand-labeling required.

*Correspondence: bab@mit.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Methods

We develop Topaz, the first particle picking pipeline to use CNNs trained using only positive and unlabeled examples and GE-binomial, a general objective function for learning classifier parameters from positive and unlabeled data. The GE-binomial objective penalizes the negative log-likelihood of the labeled data points while regularizing the classifier's posterior over the unlabeled data to match a binomial distribution prior on the number of unlabeled positives. Denoting the set of labeled positive data points by P, the probabilistic classifier as g, the classifier's posterior over the number of unlabeled positives as q, and the binomial prior as p, the GE-binomial objective function is: $-\mathop{\mathbb{E}}_{x \in P}[\log g(x)] + KL(q \| p)$, where KL is the Kullback-Leibler divergence.

In the Topaz pipeline, CNN classifiers are fit to labeled particles and the remaining unlabeled micrograph regions using minibatched stochastic gradient descent to minimize the GE-binomial objective. Predicted particle coordinates are next extracted by scoring each micrograph region with the trained classifier and then using the non-maximum suppression algorithm to greedily select candidate particle coordinates.

Results

We show that the Topaz pipeline is able to accurately detect particles when trained with very few labeled example particles. On the EMPIAR-10096 cryoEM data set [2], Topaz achieves 46% precision at 90% recall with only 1000 labeled particles. In contrast, at the same recall level, EMAN2's byRef method [3] only reaches 33% precision with the same set of labeled particles – corresponding to 71% more false positives than Topaz. Remarkably, Topaz still achieves better precision than EMAN2 at 90% recall with 1/10th and even 1/100th the number of labeled particles. At all numbers of labeled particles tested, we improve substantially over EMAN2's byRef method in area under the precision-recall curve. The relative improvement in particle detection provided by Topaz is even greater on a second, unpublished dataset provided by the Shapiro lab, containing stick-like particles with low signal-to-noise ratio. Furthermore, we show that combining a convolutional decoder with the convolutional feature extractor and classifier learned with GE-binomial to form a hybrid classifier+autoencoder can further improve generalization when very few labeled data points are available. Finally, we demonstrate that our GE-binomial objective function outperforms other positive-unlabeled learning methods never before applied to particle picking. Topaz runs efficiently, training in hours and predicting in seconds with a single consumer grade GPU. We expect Topaz to become an essential component of single particle cryoEM analysis and our GE-binomial objective function to be widely applicable to positive-unlabeled classification problems.

Acknowledgments

This work was partially supported by grants: NIH R01-GM081871, NIH R01-MH1148175, Simons Foundation (349247), NYSTAR, NIH NIGMS (GM103310), the Agouron Institute (F00316) and NIH S10 OD019994-01.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

References

1. Mann GS, McCallum A. Generalized Expectation Criteria for Semi-Supervised Learning with Weakly Labeled Data. J Mach Learn Res. 2010; 11:955–984.
2. Tan YZ, Baldwin PR, Davis JH, Williamson JR, Potter CS, Carragher B, Lyumkis D. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. Nat Methods. 2017; 14:793–796. DOI: 10.1038/nmeth.4347 [PubMed: 28671674]
3. Tang G, Peng L, Baldwin PR, Mann DS, Jiang W, Rees I, Ludtke SJ. EMAN2: An extensible image processing suite for electron microscopy. J Struct Biol. 2007; 166:205–213. DOI: 10.1016/j.jsb.2006.05.009