Proceedings of the 6th International Conference on Statistics: Theory and Applications (ICSTA'24)
Barcelona, Spain – August 19 - 21, 2024
Paper No. 135
DOI: 10.11159/icsta24.135

A Novel Particle Picking Pipeline for Cryo-EM Using Semantic Segmentation and Conditional Random Field

Szu-Chi Chung[1], Po-Cheng Chou[1]
[1] Department of Applied Mathematics, National Sun Yat-sen University
No.70 Lien-hai Road,Kaohsiung,Taiwan
phonchi@math.nsysu.edu.tw

Extended Abstract

Since the resolution revolution in cryogenic electron microscopy (cryo-EM), ignited by advancements in equipment and improved image analysis algorithms [1], the method has transformed the landscape of structural biology. Among the computational pipeline, selecting high-quality particles from micrographs, known as particle picking, is a critical first step in achieving high-resolution protein structures. However, this process is laborious and challenging due to the low signal-to-noise ratio (SNR), the presence of contaminants, contrast variations from differing ice thickness, and the absence of well-separated particles.

Recent developments in automated particle picking based on deep learning object detection have shown promise in overcoming these challenges [2,3]. However, these methods often produce off-center particles, requiring a translational search for alignment in downstream analyses, which limits their application scenarios. Consequently, researchers have shifted their focus to image segmentation networks that can distinguish particles from the background at the pixel level [4,5]. Nonetheless, the low SNR in cryo-EM images complicates the establishment of an objective, automated process for generating accurate, pixel-level annotations necessary for training these supervised models. Such benchmarking is crucial for understanding the strengths and limitations of existing methods and for encouraging further development and validation.

To bridge this gap, this study introduces a procedural framework for generating segmentation maps from cryo-EM data, serving as ground truth labels for model training. We first present a novel workflow for generating synthetic micrographs and their corresponding segmentation maps to validate these computational approaches. Then, we develop a workflow to automatically generate segmentation maps from real cryo-EM datasets by leveraging EMPIAR [6] and CryoPPP [7], examining the methods' applicability to experimental datasets. Our results demonstrate that models trained on our labeled dataset, including Fully Convolutional Networks [8] and DeepLabv3 [9], achieve over 90% accuracy, recall, precision, Intersection over Union (IoU) metrics, and F1-score on the synthetic dataset. Notably, these models identify particles not labeled in the original CryoPPP experimental dataset, proving their effectiveness in distinguishing signals from background noise.

On the other hand, a significant challenge in semantic segmentation is the absence of spatial regularization in segmentation maps, leading to isolated regions or irregular boundaries — a challenge exacerbated by the high noise levels and outliers in cryo-EM. To address this, a Conditional Random Field (CRF) is integrated into our pipeline to refine weak and coarse pixel-level predictions, producing sharp boundaries and fine-grained segmentation [10]. Our experiments reveal that CRF integration improves the classification of particles labeled in the training set and enhances the recall score. It is noteworthy that CRF integration can be highly flexible, and we are currently experimenting with the use of higher layers of the neural network, which contains more class-discriminative information to replace the predefined intensity features. We anticipate our novel pipeline, integrating dataset generation and CRF-based particle picker, will yield more accurate results and address more challenging datasets. We have organized our framework into a modular package aimed at fostering further investigation, available at https://github.com/cyanazuki/CryoParticleSegment/.

References

[1] A. Singer and F. J. Sigworth, “Computational methods for single-particle electron cryomicroscopy,” *Annual Review of Biomedical Data Science*, vol. 3, pp. 163-190, 2020.
[2] T. Wagner, F. Merino, M. Stabrin, T. Moriya, C. Antoni, A. Apelbaum, P. Hagel, O. Sitsel, T. Raisch, D. Prumbaum, D. Quentin, D. Roderer, S. Tacke, B. Siebolds, E. Schubert, T. R. Shaikh, P. Lill, C. Gatsogiannis, and S. Raunser, “Sphire-cryolo is a fast and accurate fully automated particle picker for cryo-em,” *Communications biology*, vol. 2, no. 1, p. 218, 2019.
[3] T. Bepler, A. Morin, M. Rapp, J. Brasch, L. Shapiro, A. J. Noble, and B. Berger, “Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs,” *Nature Methods*, vol. 16, no. 11, pp. 1153–1160, 2019.
[4] J. Zhang, Z. Wang, Y. Chen, R. Han, Z. Liu, F. Sun, and F. Zhang, “Pixer: An automated particle-selection method based on segmentation using a deep neural network,” *BMC bioinformatics*, vol. 20, pp. 1–14, 2019.
[5] B. George, A. Assaiya, R. J. Roy, A. Kembhavi, R. Chauhan, G. Paul, J. Kumar and N. S. Philip, “Cassper is a semantic segmentation-based particle picking algorithm for single-particle cryo-electron microscopy,” *Communications biology*, vol. 4, no. 1, p. 200, 2021.
[6] A. Iudin, P. K. Korir, S. Somasundharam, S. Weyand, C. Cattavitello, N. Fonseca, O. Salih, G. J. Kleywegt, and A. Patwardhan, “Empiar: The electron microscopy public image archive,” *Nucleic Acids Research*, vol. 51, no. D1, pp. D1503–D1511, 2023.
[7] A. Dhakal, R. Gyawali, L. Wang, and J. Cheng, “A large expert-curated cryo-em image dataset for machine learning protein particle picking,” *Scientific Data*, vol. 10, no. 1, p. 392, 2023.
[8] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.
[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
[10] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” *Advances in neural information processing systems*, vol. 24, 2011