doi:10.1017/S1431927619007566

Microsc. Microanal. 25 (Suppl 2), 2019

Building a Data-driven Workflow to Streamline Cryo-EM Data Processing

Yilai Li[1] and Michael A. Cianfrocco[1*]

[1.] Life Sciences Institute & Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, United States.
* Corresponding author: mcianfro@umich.edu

Cryo-electron microscopy (cryo-EM) is fasting-becoming a primary technique for structural biologists worldwide. Compared to X-ray crystallography and Nuclear Magnetic Resonance (NMR), the ability of cryo-EM to determine the atomic structures of large biological molecules - including protein-protein complexes and membrane proteins - has quickly drawn the attention of many in both academia and industry. In the PDB, the number of deposited structures determined with cryo-EM has been increasing at rates unmatched by the other two techniques since 2012. This rapid rise of cryo-EM has led to both universities and government-funding agencies investing in the required infrastructure, which will lead to more scientists leveraging the power of cryo-EM to aid in macromolecule structure determination.

The fast pace of cryo-EM development for sample preparation, data collection, and software development has led to a bottleneck in the knowledgebase required to determine cryo-EM structures. Even with the same dataset, a beginning user may end up with a structure with a much lower resolution compared to the same structure solved by an expert. This is due to differences in user-based decision-making between the beginning and expert user, where subjective decisions can affect the final result. For example, after raw movies are motion corrected, only the micrographs deemed 'good' will be allowed to go through the next steps. After this, the micrographs will be subjected to CTF (contrast transfer function) estimation and particle picking. Next, particles will be extracted and aligned in order to grouped into similar orientation classes by a 2D classification, allowing coherent averaging of noisy particles into high signal-to-noise class averages. At this point, another subjective step is encountered: only the particles in the "good" class averages which are determined by the user are kept and used in the following 3D reconstruction steps. Within each individual step mentioned above [1, 2], the monitoring and evaluation of the user are extremely important, and usually multiple rounds of trial and error are being performed in order to discover the best set of parameters. Such subjective user decisions might be obvious to the experts but are obstacles to the new users, and undermine the reproducibility of the research as well.

To make cryo-EM data processing more user-friendly and more reproducible, we developed a data-driven workflow which takes the motion corrected micrographs as the input, and outputs a particle stack that contains high-resolution particles that will be used in the following 3D reconstruction steps, a process that does not require any subjective user decisions. Specifically, our workflow can automatically detect bad micrographs, determine the best parameters for particle picking and 2D classification, and identify the good class averages that can be used in 3D reconstruction. In the workflow. the subjective user decisions are replaced with statistical models based on the features extracted with image processing methods and convolutional neural networks, along with the expert knowledge. Notably, we have built a database with thousands of "bad" micrographs, which can be used to learn the specific features of useless data with a deep learning model. Moreover, the workflow can also generate evaluation reports that quantify the quality of the data and the performance after each step, which will be a great benefit for the new users to have a deeper understanding of their cryo-EM data. Finally, using the good particle stack from the output

https://doi.org/10.1017/S1431927619007566 Published online by Cambridge University Press

of the workflow, we have developed novel classifiers capable of classifying micrographs and individual particles, further reducing time for processing datasets.

In order to make this workflow reproducible and user-friendly, we implemented the Common Workflow Language (CWL) [3], and the Rabix Composer (a CWL editor with a graphical user interface) [4]. Given that CWL was developed to support the reproducibility, portability and scalability of the genomics data processing, we believe CWL will help to make cryo-EM data processing more straightforward. The combined use of CWL with the Rabix Composer makes it possible to include all the processing steps and along with the parameters in a small text file, which significantly helps the reproducibility of the cryo-EM data processing. Additionally, using the simple drag-and-drop interface, custom workflows are very easy to build, making it possible for algorithm developers to test new algorithms in a more reproducible way.

References:

[1] FJ Sigworth, Microscopy **65.1** (2016), p. 57.
[2] R Fernandez-Leiro and SHW Scheres, Acta Crystallographica Section D: Structural Biology **73.6** (2017), p. 496.
[3] Common Workflow Language, https://www.commonwl.org, (accessed Feb 22, 2019).
[4] Rabix – Reproducible Analysis for Bioinformatics, http://rabix.io, (accessed Feb 22, 2019).

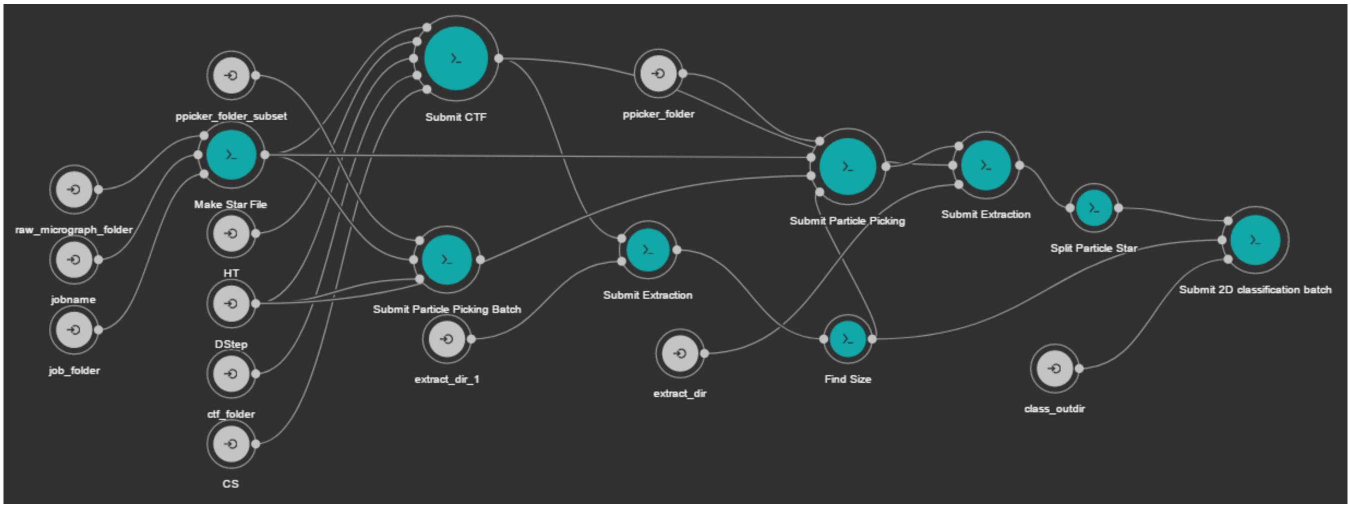

Figure 1. Proposed data-driven workflow in the Rabix Composer graphical interface. Green circles indicate different processing steps, grey circles indicate input and output directory paths for each step.

https://doi.org/10.1017/S1431927619007566 Published online by Cambridge University Press