doi:10.1017/S1431927619007700
Microsc. Microanal. 25 (Suppl 2), 2019

Strategies for Data Flow and Storage for High Throughput, High Resolution Cryo-EM Data Collection

William J. Rice[1,2*], Anchi Cheng[1,2], Sargis Dallakyan[1], Swapnil Bhatkar[1], Shaker Krit[1], Edward T. Eng[1,2], Bridget Carragher[1,2,3] and Clinton S. Potter[1,2,3]

[1.] National Resource for Automated Molecular Microscopy, Simons Electron Microscopy Center, New York Structural Biology Center, New York, NY, USA.
[2.] National Center for Cryo-EM Access and Training, New York Structural Biology Center, New York, NY, USA.
[3.] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA.
* Corresponding author: rice@nysbc.org

The introduction of direct detectors to the field of electron microscopy has revolutionized the field of structural biology [1]. Structures of proteins resolved to 3-4 Å resolution are now commonly determined, resolution ranges between 2 Å and 3 Å are becoming more common, and resolutions beyond 2 Å are now possible in a few cases (Fig. 1). All of these single particle techniques require a large number of images to be taken.

The Simons Electron Microscopy Center (SEMC) currently has three Titan Krios microscopes on site and operating 24 hours per day, 7 days per week, 52 weeks per year. The use of Leginon software on all microscopes allows for unattended operation once collection queues are set up. Each microscope is currently equipped with both a Gatan K2 camera (Gatan Inc, Pleasanton CA) and a Falcon 3 camera (Thermo Fisher Scientific). The K2 cameras typically collect movies of length 6-12 s, with a frame time generally between 150 ms and 250 ms. The raw camera size is 3838x3710 pixels, and our most common movie size is 50 frames, which corresponds to 1400 MB if stored in uncompressed form. In addition, the aligned sum takes 55 MB of disk space and does not compress well under lossless schemes. Since installation of the first Titan Krios, movie collection has grown exponentially, and we are now approaching 5,000 movies per day (Fig. 2).

We pre-process our data “on the fly” using the Appion workflow [2]. The advantage of this workflow is that it provides an easy interface for staff scientists to align frames, determine CTF parameters, and start picking particles by entering only a few parameters, with most required parameters for these programs either set by default or pulled from the leginon database. The operator is therefore freed from much tedious decision making and can concentrate on the quality of the data itself as it comes down the pipeline.

Our standard data pipeline includes saving the raw frames as LZW compressed tiff stacks, rather than MRC format stacks. The LZW tiff compression is lossless and is performed in memory before saving to disk. This saves time in saving, since less data needs to be written, as well as in transferring to our buffer computer. In addition, image processing software such as Relion [3] can read the tiff-LZW stack natively. These points provide clear advantages over our previous method, which was to save as MRC and then compress with bzip. Bzip compressed stacks are slightly smaller, compressing to 15.8% of the original size versus 20.1% for tiff, but the processing advantages greatly outweigh the small space saving.

Movies are saved onto the K2 computer and soon afterwards automatically moved to a buffer computer over a dedicated fiber line. The buffer computer is equipped with 50 TB of disk space and 2 Nvidia 1080

GPU's for frame alignment. Operators run two instances of Motioncor2 [4] on these computers as data is collected, with the aligned and dose weighted summed images being stored on our high performance GPFS filesystem for further processing. We have found that this keeps up with collection of both counted and "super-resolution" images. A second script automatically copies the compressed raw frames to this same filesystem, and the frames are automatically removed from the buffer after 2 weeks. Users of the center are required to store their own frames at their local institutions, and raw movie frames are deleted after one month. In addition, all images are automatically now compressed to JPEG format after one year, which reduces the file size to approximately 7% of the original raw MRC [5]. We see this as the "lesser of two evils", with the alternative being deleting old files outright [6].

References:

[1] Y Cheng, Science **361(6405)** (2018), p. 876.
[2] GC Lander et al., J Struct Biol **166(1)** (2009), p. 95.
[3] SH Scheres, J Struct Biol **180(3)** (2012), p. 519.
[4] SQ Zheng et al., Nat Methods **14(4)** (2017), p. 331.
[5] ET Eng et al., bioRxiv (2018).
[6] This work was supported by grants from the Simons Foundation (349247), NYSTAR, and the NIH National Institute of General Medical Sciences (GM103310) with additional support from the Agouron Institute [Grant Number: F00316] and NIH S10 OD019994-01.

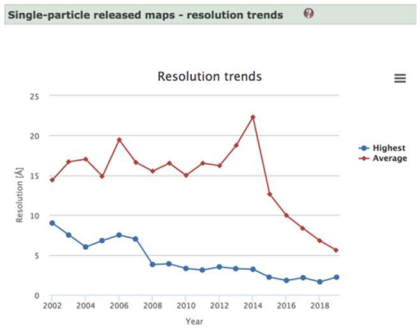

Figure 1. Resolution of released maps over time. Data downloaded from EMBD (http://www.ebi.ac.uk/pdbe/emdb/statistics_main.html/)

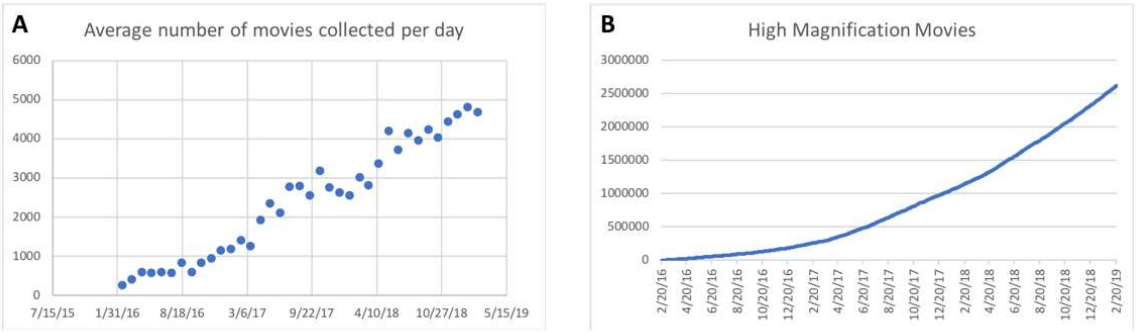

Figure 2. (A) Average number of high magnification movies collected per day at NYSBC across all 300 keV microscopes. (B) Total number of high magnification movies collected at NYSBC over time.

https://doi.org/10.1017/S1431927619007700 Published online by Cambridge University Press