Journal of Structural Biology 180 (2012) 254–258

Contents lists available at SciVerse ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Technical Note

Creating an infrastructure for high-throughput high-resolution cryogenic electron microscopy

Donald C. Shrum [a,c], Brent W. Woodruff [a,c], Scott M. Stagg [b,c,*]

[a] Department of Scientific Computing, 400 Dirac Science Library, Florida State University, Tallahassee, FL 32306, USA
[b] Institute of Molecular Biophysics, Chemistry and Biochemistry, 91 Chieftan Way, Florida State University, Tallahassee, FL 32306, USA
[c] Florida State University, Tallahassee, FL 32306, USA

ARTICLE INFO

Article history:
Received 8 March 2012
Received in revised form 29 June 2012
Accepted 16 July 2012
Available online 25 July 2012

Keywords:
Cryogenic electron microscopy
CryoEM
Database
Three-dimensional electron microscopy
High-throughput

ABSTRACT

New instrumentation for three-dimensional electron microscopy is facilitating an increase in the throughput of data collection and reconstruction. The increase in throughput creates bottlenecks in the workflow for storing and processing the image data. Here we describe the creation and quantify the throughput of a high-throughput infrastructure supporting collection of three-dimensional data collection.

Single particle three-dimensional electron microscopy (3DEM) is a powerful technique for determining the structures of biologically relevant macromolecules with several structures now approaching atomic resolution. One of the primary factors that limit resolution of single particle reconstructions is the number of particles that contribute to the reconstruction. So far, the structures that approach atomic resolution have masses of around 1 MDa or greater, and the number of asymmetric units contributing to the structures are from several hundred thousand to several millions of subunits (Cong et al., 2010; Ludtke et al., 2008; Zhang et al., 2008; Zhou, 2011), which agrees well with calculations of the dependence of resolution on numbers of particles (Henderson, 1995; LeBarron et al., 2008; Rosenthal and Henderson, 2003; Stagg et al., 2008). At the same time that the field is approaching atomic resolution for single particle reconstructions, techniques for dealing with heterogeneous data are being developed for tomographic data (Stölken et al., 2011; Winkler, 2007). In the tomographic case, tomographic subvolumes are aligned and classified to sort out the heterogeneity in three-dimensions. Like with single particle data, the quality of subvolume averaging depends on the total number of subvolumes that can be collected. Thus, both single particle and tomographic data collection are driving for an increasing amount of raw data to be able to derive the best possible 3D interpretations. The pressure for more and more data creates bottlenecks in the structure determination pipeline; disk space is required to store all the raw data, increased processing power is required to process the data in a reasonable amount of time, and the disk storage must be able to accommodate reads and writes from many different requests at the same time.

In addition to the techniques requiring more data, new detection devices are coming online such as cameras with large arrays of pixels (Ellisman et al., 2011), hybrid pixel detectors (Faruqi and Henderson, 2007), and monolithic active pixel sensor (MAPS) direct electron detectors (DDD) (Bammes et al., 2012; Milazzo et al., 2011). These developments have the potential to dramatically increase the demands for processing and storage. The commercial MAPS DDDs such as the Direct Electron DE-12, FEI Falcon, and Gatan K2 have fast readout rates with the latter device having a rate of up to 400 frames per second. In the simplest case, many DDD frames are integrated to produce a single EM exposure, and the individual frames contributing to the final exposure are discarded. However there are many potential reasons for storing the contributing frames including dose fractionation and monitoring specimen movement due to beam induced motion (Brilot et al., 2012). Thus, DDDs have the potential to both increase throughput and increase the amount of storage required for the raw data. At the same time that DDDs are being developed, the cameras are getting larger in pixel area (Ellisman et al., 2011). Doubling the linear dimensions of a camera quadruples the storage requirements for an individual image. These technological developments combine to dramatically

[*] Corresponding author at: Florida State University, Tallahassee, FL 32306, USA.
E-mail addresses: sstagg@fsu.edu, sstagg@mac.com (S.M. Stagg).

1047-8477/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jsb.2012.07.009

increase the demands on the processing pipeline and increase the pressure on the previously mentioned bottlenecks.

Dealing with the volume of data coming from EM platforms utilizing new technologies and high-throughput automated data collection requires a nonstandard approach to data storage and processing. Moreover, high-end instruments support many users each with unique data acquisition and storage requirements. The storage and processing facility must be flexible enough to accommodate the different needs of multiple users. This requirement increases the dependence on information technology and computational architecture expertise to acquire the appropriate hardware, software, and support multiple users. Utilizing expertise already in place at a high performance computing (HPC) center facilitates supporting a high-throughput kind of device. However, because high-throughput depends on the robust performance of both the microscope and the processing machines, the considerations described here will be the same even for labs with in-house clusters or that run other automated data collection applications (Mastronarde, 2005; Nickell et al., 2005; Zheng et al., 2007).

Here we describe the throughput and methods for integration of an high performance computing infrastructure with a Titan Krios (FEI Company) equipped with a 4 × 4 k pixel CCD camera with automated data collection and processing with Leginon (Suloway et al., 2005) and Appion (Lander et al., 2009). Though we describe our setup using these specific tools, the considerations described are generalizable to any resource running 24 h-a-day data collection. We describe the considerations for hardware and the tools and methodologies used to ensure seamless integration and ensure dependencies on the processing machines do not adversely impact the availability of the microscope. The setup is scalable and is described with enough detail that our setup can be replicated at other locations by individuals with modest system administration expertise.

1. Quantitation of throughput

Data collection statistics were acquired for several single particle and tomographic data collection sessions on the Titan Krios equipped with a Gatan 4 × 4 k Ultrascan CCD with four port readout using automated data collection with Leginon. With single particle data collection, the throughput depends on several factors such as the readout rate of the camera, the stability of the goniometer (drift-rate after a move), and the number of images that can be acquired per target area. We measured the throughput for two data collection sessions with typical Leginon data collection parameters. Dataset 1 was a COPII complex preserved in vitreous ice over a holey carbon film and was collected at 59,000× magnification (1.5 Å/pix) for final exposures. We were able to collect three images per hole for this session. The overall exposure rate determined as the total number of high magnification exposures over the total session time for this dataset was 95 exposures/hour. Dataset 2 was an adeno-associated virus (AAV) dataset at 120,000× (0.65 Å/pix), and we collected 93 exposures/hour for this session. For both sessions, the setup time before fully automated data collection was ~3 h. The diameter of the holes in the support film for both datasets was 2 μm and the diameter of the e^- beam was 1.4 μm. This resulted in some beam overlap in the center of the holes, but the area that was overlapping was not imaged by the camera at those magnifications. Given that the beam diameter is required to be greater than 1.3 μm in order to maintain parallel illumination with our imaging conditions, three exposures per hole is the maximum we can attain. The structures associated with these data are being published elsewhere, but the AAV data reconstructed to 4.5 Å resolution, which shows that the data collection conditions are sufficient for high-resolution (Lerch et al., 2012). The images are 4 × 4 k pixels and are stored as 16 bit signed floats in MRC format that results in an image that takes up 64 MB of disk space. Collecting single particle data in this way for 24 h requires ~144 GB of disk space.

The situation is similar for tomographic data. In a tomographic data collection session with typical Leginon data collection parameters, we collected 119 images per tilt series and could collect 1.85 series per hour. In 24 h, we can collect 44 tilt series, which in turn takes up 340 GB of disk space. The Titan Krios can be operated 24 h-a-day for 6 days-a-week. If we assume 3 days of single particle collection and 3 days of tomographic collection, we would require ~1.5 TB of disk storage per week. These data are summarized in Table 1.

Given the throughput afforded by automation and a high-end microscope, it is unfeasible to store the raw data locally on the computer that is driving the data collection. Moreover, processing this much data takes some time, and in a high-throughput scenario, the data is processed immediately after it is acquired. This means that data collection and processing are occurring simultaneously on the same disk volume. Depending on the number of processing jobs, this can be quite taxing on the disk and network that is serving the data. Some of these problems are solved by hosting the data on a distributed file system, but then the limit on the rate of data acquisition becomes dependent on the bandwidth and traffic load of the network. These considerations led us to create a setup where data are staged locally on the computer that runs Leginon and then moved in real time to an off-site secondary data storage system that is connected to the processing computers.

Hosting the data in two physical locations presented a problem for the high-throughput multi-user scenario. The standard setup for data acquisition and processing with the Leginon/Appion software requires a LINUX computer that runs Leginon and is connected to the microscope computer. The Leginon computer also requires network connectivity to a MySQL database to host the metadata and a disk volume to host the image data (Fig 1A). Both the computer driving data collection and the computers doing processing require access to the same image data and database. If we hosted a single database off-site and the network went down, the data collection would go down with it. This problem and the problem of how to store the large volume of image data were solved through the use of a data replication scheme. The computer that is directly connected to the microscope computer hosted a copy of the MySQL database and a small volume of the most recently acquired microscope images, and the image data and metadata database were replicated to a high-performance computing facility with high capacity for disk space and network traffic (Fig. 1B). The duties performed by the HPC computers are split into two functions: (1) pre-processing which includes tasks like particle picking, CTF estimation, and preliminary image classification, and

Table 1
Data collection and throughput statistics.

Collection type	Images/target	Time between exposures (s)	Time between holes (s)	Overall exposure rate	Images/day	Disk consumption/day
Single particle data collection	3	16.5	76	94 images/hr	2256	144 GB
Tomographic data collection	119	19.5	162	1.83 series/hr	5226	340 GB