Towards on-the-fly Cryo-Electron Microscopy Data Processing by High Performance Data Analysis

Evgeny Pichkur\textsuperscript{1}, Timur Baimukhametov\textsuperscript{1}, Anton Teslyuk\textsuperscript{1,2}, Anton Orekhov\textsuperscript{1}, Roman Kamyshtinsky\textsuperscript{1,2}, Yury Chesnokov\textsuperscript{1}, Viacheslav Ilyin\textsuperscript{1,2}, Alexander Vasiliev\textsuperscript{1} and Vasily Velikhov\textsuperscript{1}

\textsuperscript{1} NRC Kurchatov Institute, Pl. Akademika Kurchatova, 1, Moskva, 123098, Russia
\textsuperscript{2} Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia

E-mail: anthony.teslyuk@grid.kiae.ru

Abstract. In this paper we present an approach to the on-the-fly Cryo-EM data acquisition and analysis using High Performance Data Analysis concept. An analysis of bottlenecks in the currently used data processing is presented, and the new iterative scheme for near real time implementation of the Cryo-EM data processing is discussed. The on-the-fly approach is proposed that will considerably reduce the time required for the protein (or other biological objects, e.g. viruses) 3D structure determination and allow eliminating manual interventions in the data processing.

1. Introduction

Last years, new trend, High Performance Data Analysis (HPDA), is developing fast for the compute-intensive and data-intensive applications in different fields of science (modeling and simulations, mega science projects like Large Hadron Collider, humanitarian disciplines such as linguistics and archeology etc.), medicine, financial and commercial analytics, manufacturing, and many others. In short HPDA maintains applications where Big Data problems need High Performance Computing. Detailed discussion of the HPDA conception is presented in the book [1]. One of the notable HPDA example is structural biology where problem of 3D structure reconstruction of biomolecules and viruses is one of the most important methods for the design of new medicines. In this context reconstruction of the 3D structure of these biological objects is the serious computational problem that demands the high performance computing in the combination with big data treatment within the complicated workflows.

The Transmission Cryo Electron Microscopy (Cryo-EM) technique, in its modern development, becomes one of the most effective tools in the field of structural biology with the spatial resolution close to the traditional X-ray analysis [2, 3]. The progress in this method has been confirmed recently by the 2017 Nobel Prize award to J. Dubochet, R. Frank and R. Henderson. In contrary to the X-ray crystallography, Cryo-EM has good prospects to offer a unique option to investigate proteins and viruses in the native state with a possibility to reveal different conformations, which assume determination of the structure at different steps of biochemical reactions. Today the number of installations of modern Cryo-EM hardware is growing rapidly. Nowadays about 100 Hi-end cryogenic transmission electron microscopes are
working in research centers around the world. The only installation in Russia is located in NRC Kurchatov Institute - Titan Krios TEM (FEI, USA) equipped with Falcon II electron detector and Cs-corrector.

Recent developments in the Cryo-EM data processing software allow to perform 3D reconstruction using the single particle analysis (SPA) method [4]. This method requires the acquisition of 105-106 projections of the individual researched objects (we will use the term particles for such projections). The images are acquired using highly sensitive and quick direct electron detecting (DED) cameras [5]. The typical size of such cameras is 4096 x 4096 pixels and the size of the single image is up to 32 MB for 16 bits pixel depth. The specialized software for the single particle data acquisition, e.g. EPU (FEI, USA), generates a stack of 7 images recorded during the exposure. Thus, the total data volume for one image stack is about 240 MB including metadata. Typical time between exposures may reach 1.5 min used for the setup of the next exposure, while duration of the typical microscope session is about a few days. The spatial resolution of the reconstructed model depends on the number of single particle images involved in the analysis. Therefore, for improving of the resolution and extracting fine details of the 3D model we need raw data up to few TB for each microscope session. Typical Cryo-EM lab performs up to several dozens of microscope sessions per year. Thus, the data volume per year could be of the PB scale for a Cryo-EM facility. One can note that manual work is required at many stages of the 3D reconstruction pipeline. As a result, the total time needed to complete the 3D reconstruction of the data generated during one microscope session can be about few months.

The paper is organized as follows.

In the next section, we describe the basic characteristics of input data coming from Cryo-EM.

In 3-rd section, typical scheme of the 3D reconstruction pipeline used by the Cryo-EM research groups at present time is discussed on example of the Cryo-EM lab and HPC facilities in NRC Kurchatov Institute.

In 4-th section, we formulate an approach to the on-the-fly implementation of the Cryo-EM data processing in fully automatic way following the HPDA concept. This approach will allow to eliminate bottlenecks in the state-of-the-art Cryo-EM data-processing, caused by extensive manual operations at various stages. Furthermore, we believe that this approach will also considerably reduce the time spent on the 3D reconstruction procedure.

In Conclusion, we discuss key advantages of the on-the-fly approach which might be valuable to Cryo-EM researchers.

2. Input data for 3D reconstruction

Below we discuss key details of the data acquisition process used in the 3D characterization of biological samples on Titan Krios TEM (FEI, USA), equipped with Falcon II electron detector, which is installed in NRC Kurchatov Institute. Sample preparation follows the standard procedures and details of our technique were described in [6, 7, 8]. The current data collecting rate is about 1000 image stacks per day [9]. In our current setup each image stack is transmitted to the NFS storage server over 1Gb Ethernet, which takes about 2 seconds. Next image stack acquisition starts 1.5 min later. During this time the acquisition of an image with lower magnification, centring and finding the foil hole, autofocusing procedure and timeouts for stage relaxation are taking place in the automatic mode. A DED Falcon II generates stacks of 4096x4096 images each. Up to 7 physical images per second are stored on the disk. Output image stacks are saved either as a stack of the 16 bits/pixel image or as a 32 bits/pixel stack after gain correction. A typical raw stack is about 240MB or up to 450MB after converting to 32bit and the motion correction.

Raw data volumes need to be stored. Dedicated disk storage is used of the total volume of 40 TB. The storage is accessible from computing nodes via NFS. For data processing two different
types of computing facilities are applied. For interactive operations as well as for debugging we use dedicated servers equipped with Tesla M2070 GPU cards. Each server has 2 Xeon X5660 CPUs and 24 GB RAM. For the intensive computing the parallel supercomputer installed in Kurchatov institute is utilized. It is equipped with 23 nodes each having Xeon Haswell (v3) CPU’s, 128 GB RAM and three NVIDIA K80 GPU cards on board. 56Gb/s Infiniband fabric interconnects the computing nodes. The peak computing performance of this CPU/GPU cluster is 130 TFLOPs.

3. Current implementation of the image processing
Here we describe a current implementation of the data processing which includes a lot of manual operations. A representing scheme of this implementation is shown in Fig. 1.

Full dataset (about 4000 image stacks per microscope session) is collected and stored to the Raw Image Repository. Then full dataset is transferred to the Processing Pipeline session. It utilizes several image processing packages:

- MotionCor2 to correct sample motion, align image stacks and generate averaged output [10]
- Gctf to estimate the contrast transfer function and refine local defocus for each particle [11, 12]
- Gautautomatch or EMAN2 for the particle picking [13, 14]
- RELION and CryoSPARC for the reference-free 2D classification, 3D classification and high-resolution 3D refinement [15, 16, 17]

Most of 2D classification strategies are based on the 2D multi-reference alignment [18], which is a straightforward implementation of iterative K-means clustering algorithm [19, 20]. There are different approaches to obtain high-quality 3D reconstruction: some of them rely on the 3D initial model (RELION), while others suggest stochastic gradient descent optimization scheme for ab initio structure determination and 3D classification without any prior models (CryoSPARC).

A typical 3D structure determination utilizes 4 nodes with peak performance of approximately 22 TFlops (double precision). This step may require up to 30 hours, depending on the size of the object of interest and amount of particles. The 2D/3D classification and 3D refinement are compute-intensive steps. A lot of manual operations are required to be performed by researchers. The particle selection is also a time consuming step and requires manual operations.

4. On-the-fly Cryo-EM 3D structure determination
Here we formulate an approach to the on-the-fly Cryo-EM 3D structure determination, corresponding flowchart is shown in Fig. 2. Key processing stages are as following.

4.1. Preprocessing
Image stacks generated by Cryo-EM are stored in Raw Image Repository and transmitted to the dedicated GPU-accelerated workstation or node of the cluster for preprocessing, including motion correction and CTF estimation. The implementation of this step may be based on the existing system [9] or in-house software, which has to be developed. The cross-correlation of the simulated CTF with the observed experimental spectra (per image) is used for the automatic resolution cut-off after the setting up the required threshold. Also, the machine learning techniques could be applied for the identification of bad images in terms of presence of ice contaminations, large carbon film areas or, for example, very small particle concentration. All operations listed above have to be computed at near real time regime. The averaged images are stored in the fast Averaged Images Local Storage, which is connected to the computing nodes.
4.2. **Processing Cycle**

This cycle is a core of the on-the-fly approach. Main steps of the **Processing Cycle** iterations are explained below. The processing cycle starts as soon as a chunk of images have been preprocessed and stored, yet the exact number depends on the sample and may vary. The chunk size is a parameter to be optimized as a trade-off between computing time and quality of analysis.

4.2.1. **Step 1** The particle picking is performed in a fully automatic manner using machine learning segmentation methods. Simulated re-projections of existing structures taken from EMDataBank [22] and Protein Data Bank [23] as well as public datasets with manually picked particles are used as a training dataset for the algorithm (software is to be developed). With 200 particles per image and the total amount of 2000 particles is a good estimate for the dataset size at the first iteration of the cycle. Unsupervised machine learning techniques like clustering methods or autoencoder neural networks [24] could be used to detect informative images. This is a subject for further research.

Nowadays this step requires manual operations, that is why automation of the picking is essential for the on-the-fly processing. Re-picking of the particles may be performed using embedded loop between **Step 1** and **Step 2** in the case of poor results. Machine learning tools may be used to check whether or not the picking routine is successful - if yes the re-picking loop is stopped. All available information after previous iterations, including the 2D class averages, the stacks of actual particles and irrelevant objects, re-projections of the 3D model as well as all averaged images stored in **Averaged Images Local Storage**, is used for re-picking. A predefined criterion should be designed to stop the re-picking loop as soon as the population and number of 2D classes are not changed significantly.

The re-picking loop is eliminated from the **Processing Cycle** by predefined criteria when coordinates of the actual particles don’t change considerably between iterations.

4.2.2. **Step 2** After the fully automatic particle picking the particle stacks are stored in the **Iteration Cache** and the reference free 2D classification is performed. The purpose of the **Step 2** is to split the dataset of particles into a number of classes. Each class is formed by particles in close orientations. Particles are averaged inside each class to produce a class average image with significantly higher signal to noise ratio (SNR). The classification is used to select classes with actual particles and filter out the rest. The possibility of machine learning based classification of particle images in random orientations in Fourier space was recently studied [25].

4.2.3. **Step 3** After the 2D classification and elimination of the irrelevant images, containing contamination and noise, an ab-initio 3D reconstruction and classification [26] is performed. The 3D classification associates projections with 3D conformational classes using the expectation maximization (EM) algorithm. As soon as the 3D model is obtained, half-maps are used to calculate Fourier shell correlation (FSC) and determine the resolution of the model [27]. The decision to stop or continue the **Processing Cycle** (and, as a result, to finalize the microscope session) is based on the obtained resolution of the 3D model estimated in reciprocal or in real space [21].

All output data after **Steps 1, 2 and 3** is stored in the **Iteration Cache**, including particle stacks, coordinates of the particles, 2D class averages, 3D models, refined angles and offsets of the particles. One of the key aspects of the proposed approach is the usage of the 2D offsets and 3D rotations, determined on all previous iterations, for reconstruction of the 3D model in current iteration of the **Processing Cycle**. It will result in significant reducing of the computational time. All intermediate results obtained during **Processing Cycle** should be available for experts. This will allow a fast response to unexpected events during the experiment and reconstruction.
Next iteration (Steps 1, 2, 3) is started as soon as a new chunk of averaged images and corresponding metadata arrive after the Preprocessing.

It is worth to note that all decision-making criteria mentioned above will combine data-driven heuristics and expert knowledge. Further development of the software is required to bring the proposed concept to the Cryo-EM practice.

5. Conclusion
We propose the on-the-fly approach for cryo-EM structure determination, which relies on high level of automation and eliminates bottlenecks of the methods currently used. The use of machine learning techniques promises a significant improvement in particle picking and allows performing fast sorting of the good 2D and 3D class averages without human intervention. Once brought into practice, the proposed approach will incorporate all intermediate 3D rotations and 2D translation values, determined on previous iterations of the Processing Cycle significantly reducing computational time. The final model is expected to be ready in near real time when the data collection ends, thus reducing the whole duration of the experiment.

Acknowledgments
This work has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-Science Facilities at NRC Kurchatov Institute, http://ckp.nrcki.ru/.

The work was supported by the Russian Science Foundation grant No. 18-41-06001.

References
[1] A. Osseyran and M. Giles, Industrial Applications of High Performance Computing, CRC Press, Taylor & Francis Group, ISBN 13:978-1-4665-9681-8, 2015.
[2] I.S. Jacobs and C.P. Bean, Fine particles, thin films and exchange anisotropy, in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
[3] Callaway, Ewen, The revolution will not be crystallized: A new method sweeps through structural biology, Nature. 525 (7568): 1724 (2015)
[4] A. Doerr, Single-particle cryo-electron microscopy, Nature Methods, 13, 23 (2016)
[5] G. McMullan, et al, Comparison of Optimal Performance at 300 keV of Three Direct Electron Detectors for Use in Low Dose Electron Microscopy, Ultramicroscopy 147 (2014): 156163. PMC. Web. 20 Oct. 2017
[6] A. L. Kwiatkowski, et al, Impact of Salt Co- and Counterions on Rheological Properties and Structure of Wormlike Micellar Solutions, J. Phys. Chem. B 2016, 120, 1254712556
[7] A. V. Shibaev, et al, Viscoelastic Synergy and Microstructure Formation in Aqueous Mixtures of Nonionic Hydrophilic Polymer and Charged Wormlike Surfactant Micelles, Macromolecules 2017, 50, 339348
[8] T. Shtam, R. Samsonov, R. Kamshinsky, R. Pantina, N. Verlov, A. Vasiliev, A. L. Konevega, and A. V. Malek, Exosomes: Some approaches to cancer diagnosis and therapy, AIP Conference Proceedings 1882, 020006 (2017)
[9] X. Li, S. Zheng, D. A. Agard and Y. Cheng, Asynchronous data acquisition and on-the-fly analysis of dose fractionated cryoEM images by UCSFImage, J Struct Biol. 2015 November , 192(2): 174178
[10] S. Q. Zheng, E. Palovcak, J.-P. Armache, K. A. Verba, Y. Cheng and D. A Agard, MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy, Nature Methods 14, 331332 (2017)
[11] Zhang K., Gctf: real-time CTF determination and correction, Journal of structural, Biology, 2016; 193 (1), 1-12
[12] Wade R.H, A brief look at imaging and contrast transfer, Ultramicroscopy, 1992, 46, pp. 145156
[13] http://www.mrc-lmb.cam.ac.uk/kzhang/
[14] G. Tang, L. Peng, P.R. Baldwin, D.S. Mann, W. Jiang, I. Rees & S.J. Ludtke, EMAN2: an extensible image processing suite for electron microscopy, J Struct Biol. 157, pp. 38-46, (2017)
[15] Sjors H.W. Scheres, RELION: Implementation of a Bayesian approach to cryo-EM structure determination, J Struct Biol. 2012 Dec, 180 (3): 519530
[16] D. Kimanius, Bjorn O Forsberg, Sjors HW Scheres and Erik Lindahl, Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2, eLife 2016, 5: e18722
Figure 1. Current implementation of the image processing
Figure 2. On-the-fly structure determination scheme