High-Performance Image Processing for Cryo-Electron Microscopy

Z. Hong Zhou, Ph.D. (Hong.Zhou@ucla.edu)
Director, Electron Imaging Center for Nanomachines, CNSI and Professor, Dept of MIMG

Recent advances have made electron imaging an indispensable tool for determining the three-dimensional (3D) structures and molecular interactions of macromolecular complexes or biological nano-machineries. The newly established Electron Imaging Center for Nanomachines (EICN) at UCLA (http://EICN.CNSI.UCLA.edu) aims to provide this emerging technology in its finest forms to both nano biology and nano-materials science researchers. Two slightly different modalities of electron imaging – single particle cryo-electron microscopy (**cryoEM**) and cryo-electron tomography (cryoET) – are commonly employed to visualize or "see" nano-biological machineries or particles of different structural property (Fig. 1). These structural methods provide exciting opportunities to both biomedical investigators and materials scientists to determine the 3D structures of sub-cellular assemblies that are either too large or too heterogeneous to be investigated by conventional X-ray crystallographic or NMR methods.

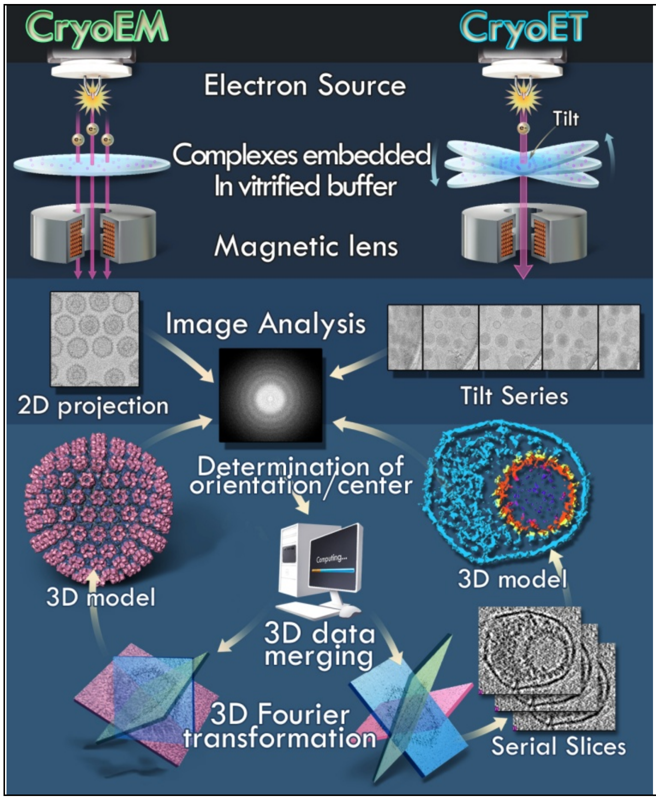

Fig. 1 Two modalities of electron imaging in biomedical research - cryo-electron microscopy (cryoEM) and tomography (cryoET). In cryoEM, a single or a pair of images are recorded with no-tilt and "single"/individual particle orientations are determined (e.g., Liang *et al.* 2002). In cryoET, multiple-shot tilt series is recorded for the same specimen area and aligned. In both cases, 2D Fourier transforms of the images are combined and inverse-transformed in 3D to reconstruct a 3D structure and iteratively refined.

Computational Challenges in CryoEM and need for -- CryoEM has the potential of revealing structural details of large macromolecular machines at near atomic scale without requiring crystallization, and unlike other techniques, can examine structural changes occurring during operation of the molecule (Henderson 1995; Chiu *et al.* 1997; Stark *et al.* 2000). Our group has shown the feasibility of imaging a plant virus particle out to 3-4 Å resolution. However, due to the intrinsic high level of cryoEM images, data processing presents some of the greatest computational challenges in modern biology. For a reconstruction of a moderately large, icosahedrally symmetric virus particle (e.g. 700 Å in diameter) approaching 3-4 Å resolution, thousands of micrographs, each with unique parameters are required. This produces a collective data set of hundreds of thousands of particles {Zhou, 2003 #7554}, which constitute over 200 Gigabytes of image data. Moreover, when no internal symmetry is present, the data requirements can easily expand to millions of particle images.

To cope with some of these challenges, our group has developed the software package, *IMIRS (Image Management and Icosahedral Reconstruction System)*, which is integrated with a distributed relational database for managing complex datasets (Liang *et al.* 2002; Dai *et al.* 2003). The modular programs in IMIRS are parallelized to run on high-performance multiprocessor computer servers (Johnson *et al.* 1997; Zhou *et al.* 1998). The integration of data management with processing in IMIRS automates the tedious tasks of data management, enables data coherence, and facilitates information sharing. The applications of these procedures (Zhou *et al.* 2000) to analyze tens of thousands of individual virus particle images in our latest attempts have led to near atomic resolution 3-D structure. The visualization of 3D structure coupled with bioinformatics data has allowed us to generate an atomic model of the protein components, hypothesize the structural basis of virus assembly and identify the possible binding site for the host receptor (Zhou *et al.* 2001).

However, due to the exponential increase of image data to be processed, improving the structure to atomic resolution is currently limited by computational speed. To get to a higher resolution, an order of (N^2) of images is needed. The memory usage also scale $O(N^2)$ to $O(N^3)$ due to increased image size. The computation time scales almost $O(N^4)$ of the resolution due to the product of the number and size of images. For example, a single run of 3D refinement and reconstruction takes over 2 weeks on the latest quadra-core Dell server with 8Gbyte RAM. Therefore, <u>high performance computing using massively parallel and graphics processors is highly desirable</u>.

Proposed high-performance computation research -- During the earlier years, the focus of the development of computer was on making higher-clock-rate processors. However, looking at the roadmaps of Intel and AMD, it has been recognized that future processors won't get performance gains by the gigahertz as it was during the last decades. Instead, the CPUs adopted the idea of being multi core. Consequently, multi core processors require new approaches to take advantage of the accumulated computation power in a single box. This could be considered as one of the major tasks of the next few years for computational scientists as well as scientists who depend on processing power. Implementing the basic image processing / 3D reconstruction packages under the most current technologies would be the aspiration for cryoEM based structural biologists. <u>We propose to enhance our IMIRS package to take advantage of the distributed and GPU computing at IDRE.</u>

<u>CryoEM image processing is a computation task that is well fit in the category of stream processing</u>. The basic idea is to align images with similar features into classes and determine their relative aspect angles in 3D. The current method to align images involves various manipulation of the Fourier transformed images, most of which are well <u>vectorizable</u>. Therefore, instruction level vectorization of the current code will vastly speedup its execution. At the same time, given the size of images, prefetching and caching of data is always a good idea that hide the data I/O time in the stream.

Of particulate note is on the overall distribution and supervision of individual computational processes. As an average project involves tens of thousands of images, it is no doubt desirable to distribute the workload onto multiple computer nodes. Cycle scavenging is also possible due to the small size of indivisible job units, provided that a local LAN allows for data transfer at several MB per second. Upon all the distribution ideas, a centralized control process that is highly error-tolerant is critical so that no part of the data is left behind.

We believe that the applications of cryoEM image processing well suit the project topics given by IDRE, under the category of porting well known algorithms to multi-core/cell/graphics processors and perhaps scaling up a program to a big number (if we can cycle scavenge all computers in a building during after-hours). The nature of images processing that belongs to the cryoEM applications give them the potential to be ported to those parallel vector machines that were initially designed for processing such data. The porting of these applications will take the image processing of cryoEM data to the next-generation computer platforms.

Significances -- Because CNSI is a share resource for the entire UCLA campus, the funding and success execution of the proposed research would benefit many groups and research projects across many institutes and departments of UCLA. It would also provide a showcase how IDRE impacts biomedical research through cross-disciplinary collaboration.

References:

Chiu, W., A. J. Avila-Sakar and M. F. Schmid (1997) Electron crystallography of macromolecular periodic arrays on phospholipid monolayers, *Adv. Biophys.* **34,** 161-172.

Dai, W., Y. Liang and Z. H. Zhou (2003) Web portal to an image database for high-resolution three-dimensional reconstruction, *J Struct Biol* **144,** 238-245.

Henderson, R. (1995) The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules, *Quarterly Reviews of Biophysics* **28,** 171-193.

Johnson, O., V. Govindan, Y. Park and Z. H. Zhou (1997). *Custom virtual memory policy for an image reconstruction application*, *in* Proceedings of the 4th International Conference on High Performance Computing, IEEE Computer Society Press, Los Alamitos, CA.**,** 517-521.

Liang, Y., E. Y. Ke and Z. H. Zhou (2002) IMIRS: a high-resolution 3D reconstruction package integrated with a relational image database, *J Struct Biol* **137,** 292-304.

Stark, H., M. V. Rodnina, H. J. Wieden, M. van Heel and W. Wintermeyer (2000) Large-scale movement of elongation factor G and extensive conformational change of the ribosome during translocation, *Cell* **100,** 301-309.

Zhou, Z. H., M. L. Baker, W. Jiang, M. Dougherty, J. Jakana, G. Dong, G. Lu and W. Chiu (2001) Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus, *Nat Struct Biol* **8,** 868-873.

Zhou, Z. H., W. Chiu, K. Haskell, H. Spears, Jr., J. Jakana, F. J. Rixon and L. R. Scott (1998) Refinement of herpesvirus B-capsid structure on parallel supercomputers, *Biophys J* **74,** 576-588.

Zhou, Z. H., M. Dougherty, J. Jakana, J. He, F. J. Rixon and W. Chiu (2000) Seeing the herpesvirus capsid at 8.5 Å, *Science* **288,** 877-880.