Rapid, routine structure determination of macromolecular assemblies using electron microscopy: current progress and further challenges

Bridget Carragher, Denis Fellmann, Francisco Guerra, Ronald A. Milligan, Fabrice Mouche, James Pulokas, Brian Sheehan, Joel Quispe, Christian Suloway, Yuanxin Zhu, and Clinton S. Potter

Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA, 92037, USA

Although the methodology of molecular microscopy has enormous potential, it is time consuming and labor intensive. The techniques required to produce a three dimensional (3D) electron density map of a macromolecular structure normally require manual operation of an electron microscope by a skilled operator and manual supervision of the sometimes complex software needed for analysis and calculation of 3D maps. We are developing systems to automate the process of data acquisition from an electron microscope and integrating these systems with specimen handling operations and post acquisition data processing. We report here on the current performance of our existing systems and the future challenges involved in substantially improving both the sustained throughput and the yield of automated data collection and analysis.

Keywords: cryoEM; automation.

1. Introduction

The development of large-scale laboratory automation and high throughput technologies pioneered in the semi-conductor industry is now being driven at an extremely fast pace by the needs of the pharmaceutical industry. In particular there has been a dramatic scaling up of automation for protein crystallography. The laborious operations that used to be carried out by specialists are now being handled by integrated systems that control every aspect of the process, from purification of the proteins to the data acquisition and analysis. A complementary technique to x-ray crystallography is molecular microscopy in which a transmission electron microscope is used to examine large protein complexes, usually preserved in vitreous ice. The technique has great promise for routinely and efficiently providing structural information at a resolution sufficient to resolve the secondary structure in proteins. It could thus be used in conjunction with the high resolution x-ray structures of individual proteins to interpret very large complexes to near atomic resolution. The techniques of molecular microscopy are however both time consuming and labor intensive. This includes almost every aspect of the process; the preparation of suitable specimens, the acquisition of the required very large numbers of electron micrographs, and the supervision of the sometimes-complex software needed for analysis and reconstruction of the three dimensional electron density maps.

The challenge then is to transform EM structure determination into a high throughput methodology. Success in this endeavor will not only facilitate the process of molecular microscopy but has the potential to expand the scope of accessible problems and make possible investigations that are presently deemed too high risk because of the inordinate effort involved.

To this end we are focused on the development of technologies to address automation for specimen handling, image acquisition, data processing and data information integration. Several years ago we began to develop a system, called Leginon (Carragher, et al, 2000), which automatically collects electron micrographs of macromolecular structures under low dose conditions. This system has been integrated with automated particle selection algorithms (Zhu et al. 1999, Zhu et al., 2003) and analysis and processing packages. We demonstrate here the results of applying this system to a variety of specimens, representing typical specimen classes encountered in molecular microscopy. Our current challenge is to substantially improve both the sustained throughput and the yield of automated data collection and analysis.

2. Automated molecular microscopy

Automating the process required to acquire low dose micrographs of macromolecules preserved in vitreous ice is a multiscale imaging problem (see figure 1). In a typical setup the process starts with the preparation of the frozen specimen suspended over a perforated carbon film that is supported by a copper mesh grid. The specimen grid is inserted into the microscope using a cryo-stage that maintains the specimen at liquid nitrogen temperatures. The system that we have been developing, called Leginon, is then responsible for automatically analyzing the grid at scales increasing over three orders of magnitude and for making decisions designed to emulate the actions of a trained microscopist. For the specimen shown in figure 1, these include identifying individual grid squares; finding holes with ice of suitable thickness and quality; selecting one or more targets within each hole; adjusting the focus settings of the microscope; and acquiring a final high magnification image. A variety of targeting options are available, including one that automatically identifies filaments in the hole and selects the longest and straightest one as the best target. The details of these procedures and the system performance are described in Carragher et al., 2000.

Once the high magnification images have been acquired, individual macromolecular structures must be automatically identified and segmented out of these images (Zhu et al, 1999, Zhu et al., 2003) and then passed along to further analysis routines.

3. Results

We have demonstrated the feasibility of automated acquisition and analysis using three test specimens that represent a range of specimen classes of interest in macromolecular microscopy, viz. helical filaments, single particles and icosahedral viruses.

3.1. Helical filaments: TMV

Tobacco Mosaic Virus (TMV), a helical filament about 180Å in diameter and 3000 Å in length, is a well-known standard in electron microscopy. It makes an excellent specimen for prototyping the automated methods in that it's structure has been previously determined to high resolution by x-ray fiber diffraction (Namba and Stubbs, 1986). We have shown that we can automatically calculate an electron density map of TMV to a resolution of ~8Å within 24 hours of inserting a grid into the microscope. An example of one such map is shown in figure 2a; it consists of an average of approximately 100,000 individual copies of the TMV molecule. Every step in the process of creating this map was automated and required no human intervention; this included image acquisition, identification and segmentation of the filaments from the high magnification images, and helical reconstruction using the Phoelix analysis package (Carragher et al., 1997). The estimated resolution of the map is based on the presence of significant signal on the 7.6 A layer line in the final average of the helical layer lines as well as the ability to resolve individual alpha helices in the electron density

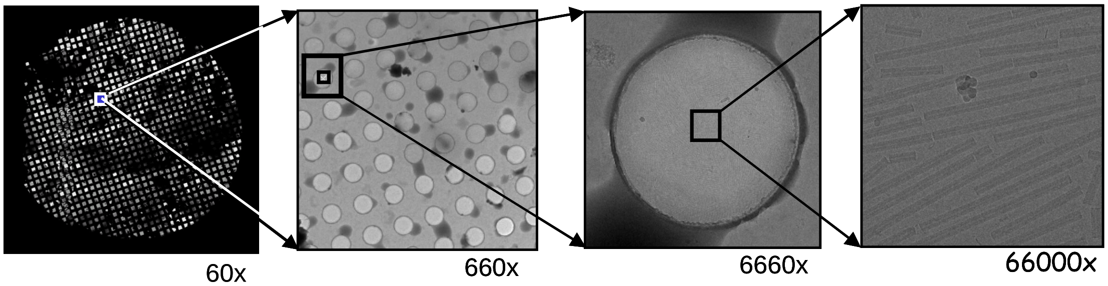

Figure 1 Multiscale acquisition and targeting algorithms are used in the Leginon system to emulate the actions of a trained microscopist in acquiring low dose images from specimens preserved in vitreous ice.

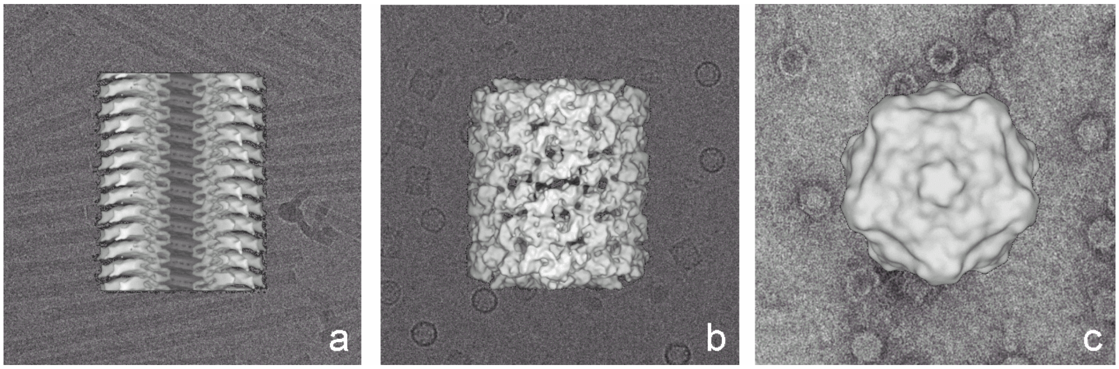

Figure 2 Reconstructed electron density maps of (a) tobacco mosaic virus, (b) keyhole limpet hemocyanin and (c) cow pea mosaic virus. Images in the background represent typical high magnification images acquired using Leginon.

map. These alpha helices are in the location expected from the interpretation of the high resolution map (Namba and Stubbs, 1986).

3.2. Single particles: hemocyanin

We have used hemocyanin from the keyhole limpet (KLH) as a test specimen for data collection and analysis from single particles (Orlova, et al.,1997). The predominant KLH species is a didecamer forming a hollow cylinder with an external diameter of 370 Å, a height of 400 Å and a molecular weight of 8 MDa. Data were collected from frozen hydrated specimens, using the Leginon system, and particles were automatically identified and selected from these images. This process was completed within 24 hours of inserting the specimen grid. The SPIDER image-processing package (Frank et al., 1996) was then used to calculate the final reconstructed map as shown in Figure 2b, a process that required approximately 20 days of CPU processing time (dual 2.8MHz P4 system). Approximately 50,000 individual copies of the KLH macromolecule contributed to the map. The estimated resolution of the map is ~15 Å as determined from the 0.5 Fourier Shell Correlation (FSC) criterion or ~11.5 Å by the 3σ criterion.

3.3. Icosahedral viruses: CPMV

Cow Pea Mosaic Virus (CPMV), a single stranded RNA icosahedral comovirus approximately 300Å in diameter, is currently being developed as an addressable nanoparticle (Wang et al., 2002). Assessing the effectiveness of various strategies to attach a wide range of molecules to the surface of the virus is achieved by reconstructing 3D maps of the modified virus. Typically, fairly low resolution maps obtained from negatively stained images are adequate for making this assessment. We have thus used negatively stained specimens of CPMV as our test bed for icosahedral virus reconstructions. Data were collected using the Leginon system at a rate of approximately 1000 particles per hour. Particles, automatically identified and selected from the images, were passed to the EMAN reconstruction package (Ludtke et al., 1999). A total of approximately 500 particles contributed to the final map shown in Figure 2c, which required about 2 hours of CPU time (2.4MHz P4 system). The final resolution of the map is estimated to be about 40Å (0.5 FSC criterion).

4. Improving the performance and throughput of the system

We have demonstrated the feasibility of automating the process of macromolecular microscopy using a variety of test specimens. Our overall goal is to transform EM structure determination into a high throughput methodology that will be accessible to the general scientific community. To achieve the latter goal we are currently expanding the Leginon system to provide a large library of targeting and acquisition schemes so as to provide for applications that can be customized for particular data collection protocols. Improving the overall throughput of the system is also essential if we are to acquire sufficient data during a single session at the microscope to reconstruct density maps in which secondary structure is interpretable. While an exact correlation between the resolution of the final map and the number of molecules contributing to the average is not well determined, based on available evidence it is reasonable to assume that interpretation of secondary structure will require on the order of 100,000 copies of the individual macromolecule (Henderson, 1995). In the case of helical or icosahedral structures the symmetry of the structures greatly reduces the number of individual images that must be acquired to achieve this copy number but for single particles lacking symmetry this is still a challenging task.

We propose that the throughput of a data collection system be defined by (i) the *rate* at which data can be collected (e.g. particles/hour), (ii) *yield* - the fraction of the data that is useful and (iii) *sustainability* - a measure of the duration over which data can be continuously acquired. As an example, for one of the typical TMV experiments, data were acquired at a rate of approximately 50,000 particles/hour, with an overall yield of ~15%, sustained through 20 hours of continuous operation. The total throughput was 150,000 particles for the entire experiment. In contrast, for the KLH reconstructions, we collected approximately 25,000 individual images of the didecamer during the duration of a typical experiment lasting approximately 20 hours. Of these a total of ~40% were used in the final average density map, representing a total throughput for the experiment of about 50,000 individual copies of the macromolecule. One of the most obvious ways of improving the throughput of data collection is to increase the size of the CCD cameras used to collect the data. We have used a 2Kx2K CCD system up to now but we are currently investigating whether the new 4Kx4K systems will have equivalent performance. If so, we will immediately quadruple the overall throughput of each experiment. We will seek additional improvements in throughput by optimizing both the design and the efficiency of the data collection protocols as well as in engineering efforts to stabilize the specimen stage. Instability in the specimen stage is currently one of the principal bottlenecks in the data collection protocols. We are confident that under the right conditions, we will be able to achieve our goal of acquiring 100,000 particles in a single session at the microscope.

This work was supported by the National Institutes of Health (NCRR17573 and GM61939) and the National Science Foundation (DBI-0296063).

References

Carragher, B., Whittaker, M., Milligan, R. (1996). J. Struct.Biol. **116,** 107-112.

Carragher, B., N. Kisseberth, D. Kriegman, R.A. Milligan, C.S. Potter, J. Pulokas, and A. Reilein. (2000). J. Struct. Biol., **132**, 33-45.

Frank, J., et al. (1996). J Struct. Biol., 1996. **116** 190-9.

Henderson, R. (1995). Quart. Rev. Biophysics, **28** 171-193.

Namba, N. and Stubbs, G. (1986). Science **231**, 1401-1406.

Orlova, E.V., Dube, P., Harris, R.J., Beckman, E., Zemlin, F., Markl, J., va Hel, M.. (1997) J. Mol. Biol., **271**, 417-437.

Wang, Q., Kaltgrad, E., Lin, T., Johnson, J.E., Finn, M.G. (2002) Chem. Biol. **9**, 805.

Ludtke, S.J., Baldwin, P.R., and Chiu, W. (1999) J. Struct. Biol. **128**, 82-97.

Zhu, Y., B. Carragher, D. J. Kriegman, R. A. Milligan, and C. S. Potter, (2001). J. Struct. Biol. **135**, 302-312.

Zhu, Y., B. Carragher, F. Mouche, and C. S. Potter. (2003). IEEE Trans. Med. Imaging, In press.

Received 30 April 2003 • Accepted 15 October 2003