

Lawrence Berkeley National Laboratory

Lawrence Berkeley National Laboratory

Title

Historical background: Why is it important to improve automated particle selection methods?

Permalink

<https://escholarship.org/uc/item/4rz5v8p5>

Author

Glaeser, Robert M.

Publication Date

2003-08-14

HISTORICAL BACKGROUND: WHY IS IT IMPORTANT TO IMPROVE AUTOMATED PARTICLE SELECTION METHODS?

Robert M. Glaeser, Department of Molecular and Cell Biology, University of California,
Berkeley, CA 94720-3206 and Physical Biosciences Division, Lawrence Berkeley National
Laboratory, University of California, Berkeley, CA 94720

ABSTRACT

A current trend in single-particle electron microscopy is to compute three-dimensional reconstructions with ever-increasing numbers of particles in the data sets. Since manual - or even semi-automated - selection of particles represents a major bottleneck when the data set exceeds several thousand particles, there is growing interest in developing automatic methods for selecting images of individual particles. Except in special cases, however, it has proven difficult to achieve the degree of efficiency and reliability that would make fully automated particle selection a useful tool. The simplest methods such as cross correlation (i.e., matched filtering) do not perform well enough to be used for fully automated particle selection. Geometric properties (area, perimeter-to-area ratio, etc.) and the integrated “mass” of candidate particles are additional factors that could improve automated particle selection if suitable methods of contouring particles could be developed. Another suggestion is that data be always collected as pairs of images, the first taken at low defocus (to capture information at the highest possible resolution) and the second at very high defocus (to improve the visibility of the particle). Finally, it is emphasized that well-annotated, open-access data sets need to be established in order to encourage the further development and validation of methods for automated particle selection.

INTRODUCTION

Production of three-dimensional (3-D) reconstructions of large macromolecular assemblies from images of single particles is a well established (Frank, 1996) and rapidly advancing tool in biochemistry and cell biology (Nogales, 2001; Sali, 2003). Only relatively small amounts of specimen material are needed in order to obtain an accurate 3-D reconstruction, the material does not need to be highly purified (as long as the desired particle can be distinguished from other components), and structural imaging can often be done in physiological buffer conditions rather than in a precipitant (i.e., crystallization) buffer. These unique features make single-particle cryo-EM an attractive alternative to x-ray crystallography for the study of multiple functional states of complex macromolecules. Even though single-particle cryo-EM is not yet able to produce 3-D reconstructions at high enough resolution to build chain-trace models of the structure of monomeric particles, rigid-body atomic models of component proteins - and even domains of individual proteins - can nevertheless be docked very accurately into macromolecular envelopes of multiprotein complexes at resolutions of ~ 2 nm (e.g., Volkman, 1999; Wriggers, 1999; Rossmann, 1999; Wriggers, 2001; Volkman, 2003). The ability to detect changes in tertiary structure, such as those seen in different functional states of the Ca-ATPase (Xu, 2002), is improved even further when the resolution extends below 1 nm, at which point separate helices become well defined.

It is already possible to collect in a single day, at least for favorable specimens, the image data needed to generate 3-D reconstructions at ~ 2 nm resolution. Automation of the data-collection step (Carragher, 2000; Zhang, 2001) can even reduce the time needed to collect data for a < 1 nm reconstruction to less than one day (Zhu, 2001). These advances now make it practical to envision collecting data sets of a million particles or more, the number that is

currently required to produce a 3-D reconstruction in the range 0.3-0.4 nm resolution (Henderson, 1995). Much smaller numbers of single particles would be needed for high-resolution reconstructions if image-formation of every particle would be free of beam-induced movement effects (Glaeser, 1999). Images of nearly perfect quality occur only at random, however, with low frequency, and even then only within small areas of the entire recorded field (Typke, In press). Thus, until matters improve significantly, one is forced to start with very large data sets (approximately 100,000 particles or more) if the final goal is to produce a 3-D reconstruction at a resolution better than ~ 1 nm. In favorable cases it then may be possible to select just the very best particles from an initially large data set, thus reducing the size of subsequent computations.

The need to use at least a semi-automated selection process becomes apparent when one starts to expand the size of a data set beyond a few thousand particles. When the size of the data set has to be increased beyond a few tens of thousands of particles, however, even a semi-automated process is not really good enough, and the use of a fully automated process of particle selection becomes increasingly important. The demands placed on the performance of a fully automated selection process will normally be quite high, however.

FULLY AUTOMATED PARTICLE-SELECTION ALGORITHMS HAVE ENCOUNTERED UNACCEPTABLE TRADE-OFFS BETWEEN EFFICIENCY AND RELIABILITY

The effort to develop fully automated, computational particle-selection tools has been recently reviewed (Nicholson, 2001). As is described in that review and in the more recent work that is reported in this special issue, a variety of approaches have been investigated, including sophisticated variations on matched filtration, the use of edge-detection algorithms, and the use

of trained neural networks. The current outcome for the work just mentioned is that efficient and reliable tools have been developed primarily for favorable cases, such as spherical particles or linear fibers. The advantage in such cases, of course, is that the predictable appearance of the particle can be incorporated as prior knowledge in the algorithm that is used to select particles.

It is tempting to think that the prior knowledge provided by an existing, low-resolution reconstruction of any, general particle could similarly be used as a constraint in the selection of further particles of the same type. While this is probably a sound idea in principle, approaches of that type have not been given great attention because they are seen as being computationally too intensive. As an example, one might use a complete set of projections of an existing 3-D reconstruction to generate a corresponding set of matched filters. This approach is not effective, however, unless the convolution (matched filtration) is done for all possible in-plane rotations of each reference-projection. For example, the letter “capital A” will only show a high cross correlation with other “capital A” letters when the reference and the particle are in the same angular orientation. Since convolution with a large number of references, each at a large number of rotational angles, becomes computationally expensive, approaches that are based on the idea of matched filtering have generally resorted to the use of a circularly averaged reference.

The use of a circularly averaged particle (or projection of a 3-D model) as a reference “particle” is itself problematic. Much of the information that might otherwise have been used to distinguish a particle is necessarily lost in the averaging step. Nevertheless, a circular average does give a good estimate of the size of the particle if the particle itself is globular in shape, and thus gives a good starting point for at least selecting particles of the correct size. In many cases, even a simple, two-dimensional Gaussian “blob” is found to serve just as well for the same purpose.

As may thus be expected, if the reference model that is used for cross correlation contains only little of the information that is in the image of an authentic particle, the convolution (i.e., matched filtration) process will generate a large number of peaks that are actually false positives in addition to the desired “true hits.” In general, one can set the threshold in a peak-search high enough to eliminate most of the false positives, but if that is done, then a high percentage of the true hits is also eliminated. Conversely, if the threshold is set low enough to capture most of the true hits, then a rather large number of false positives will also be included.

The most general method for selecting (boxing out) images of very large numbers of single particles has thus been found to be a two-stage, semi-automatic approach. The first, “automated” stage involves computing the two-dimensional (2-D) cross correlation function (CCF) between an image-field (containing many single particles) and a circularly symmetrical reference “particle.” This step can be regarded either as the use of a matched filter or as a noise-suppression (smoothing) operation. The reference “particle” can be something as simple as a two-dimension Gaussian function whose width is similar to that of the authentic particles, or it may be a circularly averaged version of the images of one or more authentic particles. Multiple (circularly averaged) reference particles can also be used, corresponding to different views of the particle. The positions of candidate particles are then identified by applying a peak-search algorithm to the 2-D cross correlation function. In the most sophisticated version, these candidate particles are further sorted into “particles”, “junk”, and “noise” based upon their statistical and textural qualities as well as an estimate of the particle size (Lata, 1995). The second, manual stage then involves human editing of the gallery of candidate particles that have made it through the automated phase. As many as a third to half of the candidate particles may be deleted during the manual editing phase of the semi-automated approach.

The use of semi-automated particle-selection approaches has been widely adopted in projects that aim to select more than just a few thousand particles. While particle selection is seen to be less of a challenge when the task is merely to edit, or “prune” a pre-selected gallery of candidate particles, even this stage of the work can be a major bottleneck in the process of obtaining a 3-D density map when the size of the data set expands to include many tens of thousands of particles. As a result, there remains an ever-growing interest in improving automated particle selection tools to a level where human editing of the boxed selections is no longer judged to be necessary.

FULLY AUTOMATED PARTICLE-SELECTION TOOLS NEED NOT BE PERFECT

In order to be useful, a fully automated selection process must be an efficient one, identifying most of the same particles that would have made it through the human-editing phase of the semi-automatic process, for example. Most users might agree that a fully automated selection process would be preferable to a semi-automated one if the fully automated tool were able to pick at least 75% of the particles that an expert human would select from the same micrograph. Although an efficiency of at least 75% would imply that one must start with a larger number of electron micrographs than would otherwise be the case, the effort needed to generate more micrographs would presumably be far less than the effort needed to edit the very large gallery of candidate particles that would be generated by a semi-automated process. It thus seems likely that an efficiency of 75% or more is a good goal to aim for.

In addition, of course, the fully automated process must be a highly reliable one. The number of false hits, or “false boxes” that are selected must be small enough that these non-particles do not distort the 3-D reconstruction. Although there have been no quantitative tests of

how many false positives (boxes that the human expert would never have included) can be tolerated, it seems reasonable to demand that these should be less than 10% of the total. If the number of false positives is small enough, of course, they will contribute so little weight (“noise”) to the reconstruction that their presence will go unnoticed. In fact, the situation is even better than that, since false positive “particles”, if present as only a minority, will presumably merge only poorly with the 3-D reconstruction that is determined by the majority of the data, and thus they will tend to be rejected as statistical outliers during the refinement of the 3-D reconstruction.

THE SCOPE OF POSSIBLE IMPROVEMENTS IS FAR FROM BEING EXHAUSTED

It is natural to ask what an experienced human being does during the manual-editing phase of the semi-automated approach, that is not done by a simple CCF peak-search that is followed by an assessment of various measures of texture in the area surrounding the peak. It is commonly assumed that the human editor is using the unique ability of the brain to recognize complex shapes and to do so in all orientations. A human can easily recognize a “capital A” in any orientation, for example, but it is possible that the answer is not as complicated as that. Personal experience will confirm that one of the first things that a human does is to reject objects that are either too large or too small (do not have the correct area). A second thing that we immediately do is to reject things that have too much or too little contrast relative to the background (the integrated intensity is not within “normal bounds.”) Finally, we readily reject things that do not have the correct perimeter-to-area ratio, or do not meet other geometry-related criteria that could be quite easily codified. A case can thus be made that a great deal of the

expertise of a human editor could be captured in a small number of computable tests, if it were first possible to generate a sufficiently accurate contour of each candidate particle.

The ability of a human editor to reliably and efficiently select particles is also greatly improved when images are taken at quite high defocus. One might anticipate that all aspects of automated particle selection, including contouring the particle, will also be more reliable and more efficient for highly defocused images rather than for images recorded close to focus. The advantage of using highly defocused images for automated particle selection need not conflict with the need to take images rather close to focus in order to capture high-resolution data, since it is always possible to take the high-resolution image first, and then follow that with a second image taken at high defocus. The fact that high-resolution features of the particle will be heavily damaged at the end of the second exposure to the electron beam will be unimportant, since the second image need be used for nothing more than to automatically select the particles. The throughput of data collection would be cut nearly in half, of course, by stopping to take two images rather than just one for every field of particles. Nevertheless, the time lost in doubling the amount of data that is collected will be more than compensated if reliable and efficient particle selection should become possible through the use of highly defocused images.

In order to encourage as many groups as possible to develop improvements in automated particle selection, it will be extremely helpful for experimentalists to create a number of open-source, well-annotated data sets. These data sets should consist of the raw, digitized images so that each group can develop their own preprocessing strategies. The location of every manually (or semi-automatically) selected particle should be provided as a gold-standard, against which the fully automated method can then be compared. The idea here is not so much that manual selection is perfect, even when done by someone who is quite expert as regards the particle in

question. Instead, the idea is that the first goal of automated particle selection should be to perform as closely as possible to how that human expert would perform, so that the human will feel confident in turning over the task (of particle boxing) to the computer program.

Very useful information would also be provided to those developing algorithms if the experimentalist would identify particles that were rejected, either during the initial selection or subsequently during refinement of the Euler angles, and annotate as fully as possible the reason why those particles were rejected. Additional “double blind” data sets should also be provided in which the particle coordinates are not available to the group who want to test a new algorithm. In this case, one should still have the ability to submit the coordinates of automatically selected particles to a server, and get in return the statistics of reliability and efficiency in the selected data set. The application to standard-reference data sets would provide an important benchmark to be cited in the publication of new algorithms and strategies.

The possibility exists that automated particle selection might be more efficient, more reliable, or both, than the efficiency of the human who manually (or semi-automatically) generated the “gold standard” data set. If the differences between the gold standard and the fully automated selections are small in number, it may be difficult to validate which has done the better job, and it will be irrelevant to do so. If the number of particles selected by a fully automated approach is much larger than the number of particles identified in the gold standard, however, there are a number of things that can be done to validate the excess particles. The excess particles could be used to generate an independent 3-D reconstruction, for example, starting with an independent 3-D reference (Grigorieff, 2000), and the Fourier Shell Correlation would then tell whether the particles obtained (in addition to the “gold-standard” set) were ones that should have been collected in the first place.

REFERENCES

- Carragher, B., N. Kisseberth, D. Kriegman, R.A. Milligan, C.S. Potter, J. Pulokas, and A. Reilein. 2000. Leginon: an automated system for acquisition of images from vitreous ice specimens. *J Struct Biol.* 132:33-45.
- Frank, J. 1996. Three-dimensional electron microscopy of macromolecular assemblies. *Academic Press, San Diego, CA.*
- Glaeser, R.M. 1999. Review: electron crystallography: present excitement, a nod to the past, anticipating the future. *J Struct Biol.* 128:3-14.
- Grigorieff, N. 2000. Resolution measurement in structures derived from single particles. *Acta Crystallogr D Biol Crystallogr.* 56 (Pt 10):1270-7.
- Henderson, R. 1995. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q Rev Biophys.* 28:171-93.
- Lata, K.R., P. Penczek, and J. Frank. 1995. Automatic particle picking from electron micrographs. *Ultramicroscopy.* 58:381-91.
- Nicholson, W.V., and R.M. Glaeser. 2001. Review: automatic particle detection in electron microscopy. *J Struct Biol.* 133:90-101.
- Nogales, E., and N. Grigorieff. 2001. Molecular Machines: putting the pieces together. *J Cell Biol.* 152:F1-10.
- Rossmann, M.G., and Y. Tao. 1999. Cryo-electron-microscopy reconstruction of partially symmetric objects. *J Struct Biol.* 125:196-208.
- Sali, A., R. Glaeser, T. Earnest, and W. Baumeister. 2003. From words to literature in structural proteomics. *Nature.* 422:216-25.
- Typke, D., K. Downing, and R. Glaeser. Electron microscopy of biological macromolecules: Bridging the gap between what physics allows and what we currently can get. *Microscopy & Microanalysis.* In Press.
- Volkman, N., and D. Hanein. 1999. Quantitative fitting of atomic models into observed densities derived by electron microscopy. *J Struct Biol.* 125:176-84.
- Volkman, N., G. Ouyang, K.M. Trybus, D.J. DeRosier, S. Lowey, and D. Hanein. 2003. Myosin isoforms show unique conformations in the actin-bound state. *Proc Natl Acad Sci U S A.* 100:3227-32.
- Wriggers, W., and S. Birmanns. 2001. Using situs for flexible and rigid-body fitting of multiresolution single-molecule data. *J Struct Biol.* 133:193-202.
- Wriggers, W., R.A. Milligan, and J.A. McCammon. 1999. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *J Struct Biol.* 125:185-95.
- Xu, C., W.J. Rice, W. He, and D.L. Stokes. 2002. A structural model for the catalytic cycle of Ca(2+)-ATPase. *J Mol Biol.* 316:201-11.
- Zhang, P., A. Beatty, J.L. Milne, and S. Subramaniam. 2001. Automated data collection with a Tecnai 12 electron microscope: applications for molecular imaging by cryomicroscopy. *J Struct Biol.* 135:251-61.
- Zhu, Y., B. Carragher, D.J. Kriegman, R.A. Milligan, and C.S. Potter. 2001. Automated identification of filaments in cryoelectron microscopy images. *J Struct Biol.* 135:302-12.