Published in final edited form as:
Procedia Comput Sci. 2014 ; 20: 2295–2305. doi:10.1016/j.procs.2014.05.214.

EPiK-a Workflow for Electron Tomography in Kepler*

Ruijuan Chen[#1], **Xiaohua Wan**[#1,2], **Ilkay Altintas**[3,‡], **Jianwu Wang**[3], **Daniel Crawl**[3], **Sébastien Phan**[1], **Albert Lawrence**[1], and **Mark Ellisman**[1]

[1]National Center for Microscopy and Imaging Research, University of California, San Diego, La Jolla, CA, USA

[2]Key Lab of Intelligent Information Processing and Advanced Computing Research Lab, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

[3]San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA, USA

[#] These authors contributed equally to this work.

Abstract

Scientific workflows integrate data and computing interfaces as configurable, semi-automatic graphs to solve a scientific problem. Kepler is such a software system for designing, executing, reusing, evolving, archiving and sharing scientific workflows. Electron tomography (ET) enables high-resolution views of complex cellular structures, such as cytoskeletons, organelles, viruses and chromosomes. Imaging investigations produce large datasets. For instance, in Electron Tomography, the size of a 16 fold image tilt series is about 65 Gigabytes with each projection image including 4096 by 4096 pixels. When we use serial sections or montage technique for large field ET, the dataset will be even larger. For higher resolution images with multiple tilt series, the data size may be in terabyte range. Demands of mass data processing and complex algorithms require the integration of diverse codes into flexible software structures. This paper describes a workflow for Electron Tomography Programs in Kepler (EPiK). This EPiK workflow embeds the tracking process of IMOD, and realizes the main algorithms including filtered backprojection (FBP) from TxBR and iterative reconstruction methods. We have tested the three dimensional (3D) reconstruction process using EPiK on ET data. EPiK can be a potential toolkit for biology researchers with the advantage of logical viewing, easy handling, convenient sharing and future extensibility.

Keywords

Electron Tomography; Scientific workflows; EPiK; TxBR; Kepler

*This work was supported by grants from an International Community Foundation, San Diego, the National Institutes of Health (GM103412, GM103426), the National Science Foundation (NSF DBI-1062565) and the National Science Foundation for China (61232001, 61202210, and 60921002)

Correspondence to: Ilkay Altintas.

‡Corresponding author
chenruijuan1@gmail.com, bjwxiaohua@gmail.com, altintas@sdsc.edu, jianwu@sdsc.edu, crawl@sdsc.edu, sph@ncmir.ucsd.edu, albert.rick.lawrence@gmail.com, mark@ncmir.ucsd.edu

1 Introduction

Many scientific applications are now data and information driven, and structured as pipelines or workflows with a large number of distinct computations. In general, workflow applications put together data sets from one or more data sources, transform the data into a format amenable for processing, analyze the data to produce useful results, and store the data and results in a repository which scientists can access [1]. Many of the steps in the processing and the data sets accessed are distributed across different execution sites, requiring data to be moved across a network for subsequent processing by the next step(s) in the workflow. Thus, scientific workflows are graphs of analytical steps that may involve, e.g., database access and querying steps, data analysis and mining steps, and many other steps including computationally intensive jobs on high performance cluster computers [2]. Kepler [2]-[4], as one of the main workflow management systems, provides a visual interface that can be used to define and build the processing required in a workflow, and generally raise the level of abstraction employed in a workflow application solution. Scientists working in many computation-intensive domains, ranging from astronomy to bioinformatics, have adopted workflows successfully by integrating large-scale, distributed and heterogeneous resources.

In the field of medicine, the head of National Institutes of Health (NIH) has pointed out that genomics has ushered in a new era of personalized medicine. In order to exploit the advances of genomics and molecular biology new research tools will be created, with a major input from the computational side. Relevant to this effort, there are historical similarities with the scientific discovery of antibiotics and their introduction into medicine. We are about half way through similar developments originating from the discovery of the structure of DNA-a stage comparable to the scientific development of antibiotics in the early 1900s. The next step is to explore advances in genomics and molecular biology. Startups such as Celera and Human Lorgevity, Inc herald the next phase of this activity.

One major challenge is to relate the molecular science to cellular structure. Much remains to be done. Siting or partitioning of the metabolic network, signaling, and transport all have many unsolved issues. For the future, in order to relate the molecular science to cell structure, we must operate across many spatial scales, employ several types of microscopy and work in diverse scientific disciplines. Digital codes for image processing, structure determination and modeling of biological processes will all play a role. This will require the integration of diverse codes into flexible software structures.

Key to much of this effort, Electron tomography (ET) is a powerful technology providing three-dimensional (3D) imaging of cellular ultrastructure. These structures are reconstructed from a set of micrographs taken at different sample orientations, and the final volume is the solution of a general inverse problem. Large-field high-resolution ET facilitates visualization and understanding of global structures, such as the cell nucleus, extended neural processes, or even whole cells on scales approaching molecular resolution. There are many electron microscope tomography packages, such as TxBR[5]-[10], IMOD[11], Xmipp[12][13], EMAN[14] and so on.

This paper describes an “Electron Tomography Programs in Kepler” (EPiK) workflow which is a response to the need to include TxBR and other codes in this stream of development. It combines several processes including tracking, alignment and reconstruction. We have tested the 3D reconstruction process using EPiK on an ET dataset. We demonstrate that EPiK workflow can provide a semi-automatic platform to obtain the 3D structure of components.

EPiK facilitates ET image processing because the algorithms are complex and often consist of multiple processing steps. Complicating this situation, there are many ET research groups from different areas in the world. Scientists at each site may develop and use their own resources and codes to perform their research. Contrary to published claims, it is difficult or impossible to make fair comparisons of different algorithms on common data sets. EPiK can integrate many packages and make it possible to compare and cooperate on methods that are proposed by different research groups.

The need for high-resolution tomography of complex biological specimens gives rise to the requirement for large reconstructed files. This requires extensive use of computational resources and considerable processing time [5]. In response to this requirement, TxBR which is in production use has been adapted for various parallel computers, computer clusters and processors with multiple graphical processor unit (GPU) boards. By using fast recursion algorithms and other parallel processing with GPU for algorithms such as backprojection, TxBR can achieve significant speed-ups on relatively inexpensive hardware that can be put together using commercial off-the-shelf components[6]. EPiK can also give users different choices of available resources according to the demand of different sizes of datasets.

2 Description of EPiK

In this section, the main structure and components of EPiK are introduced. The details will be explained respectively for each step including tracking, alignment and reconstruction.

2.1 Introduction of Components in EPiK

In ET, a 3D structure reconstruction of a sample is obtained from series of projection images recorded in a microscope by imaging a sample in various discrete orientations [15]. Generally the sample is tilted around one or more axes to produce the various orientations. In the process of 3D reconstruction, there are three main modules in EPiK: tracking, alignment and reconstruction.

In order to reconstruct the 3D structure of a sample, we need to know the relationship between an object and its projections. Ideally, all images should be aligned so that each represents a known projection of a 3D object at a known projection angle[16]. In the process of preparing a sample, the researchers may add a number of gold particles on both sides of the sample before the images are collected. Those particles provide fiducial markers that are used to derive the relationship between the sample and its projections.

The general interface of EPiK is shown in Figure 1 (a). First, tracking is carried out by IMOD [11][17]. This produces markers for alignment. Second, images alignment is completed using the bundle adjustment algorithm in TxBR [5]-[10]. TxBR was originally developed to compensate for curvilinear trajectories, sample warping and provide better alignment and reconstruction quality. The final step is to normalize the gray scales and reconstruct the specimen.

2.2 Tracking

As the first part of EPiK, fiducial tracking is carried out by IMOD package [11][17]. In order to make the workflow more simple and convenient, a composite actor of tracking that contains four main steps as shown in Figure 1 (b) substitutes the interface of IMOD.

1. Erase—The first preprocess is to erase X-ray artifacts arising from high-energy electron collision as well as defects and fiducial markers in microscope images from CCD cameras by looking for “peaks” or pixels whose intensity deviate from the surrounding pixels.

2. Cross-Correlation—As a preliminary alignment step, cross-correlation (CC) is used to find an initial translational pre-aligned stack between successive images of a tilt series. This is a marker-free coarse alignment step, which adjusts the images pairwise to align the series of images well enough for the automatic tracking of fiducial markers.

3. Select—After a coarse alignment map is obtained by cross-correlation, it is necessary to select a number of fiducial markers in order to get more precise projection. In EPiK, when the image stack is firstly shown by IMOD, it is required to select fiducial markers to generate a seed model. In order to obtain a precise alignment, markers must be located on both sides. Furthermore, the more markers that are selected, the more precisely the alignment model will be. After a sufficient number of seeds are selected, the seed model is saved.

4. Track—The last step is to track the selected fiducial gold beads on a series of images with different tilt views. In EPiK, after the seed model is saved, a new image stack will appear. This stack contains all the projected images and the positions of the selected fiducials that are calculated by the previous coarse alignment. Users need to fill the gaps and correct the position errors of the fiducials if necessary by using the bead helper which is in the pull-down menu of “Special”.

Finally, after all the selected markers are tracked, a “.fid” file will be generated in the directory defined by the “dataDir” workflow parameter. This fiducial file will be trans mitted into the alignment step.

2.3 Alignment

Alignment of the individual images of a tilt series is a critical step in obtaining high-quality electron microscope reconstructions. Electrons moving in a magnetic field will generally orbit the magnetic field lines, thus, they will move along helical trajectories. TxBR was developed to compensate for curved trajectories by using a high order polynomial projection map[5]-[10]. The alignment is completed on the platform of MATLAB, which is shown as the second actor in Figure 1 (a).

Alignment provides a 3D model of marker projections for which is consistent over all tilt projections and provides a separate projection model for each tilt image. TxBR employs nonlinear bundle adjustment simultaneously calculating marker positions and projection maps via conjugate gradient optimization[5].

Alignment is via conjugate gradient optimization using mean square reprojection error as the objective function, as shown in Eq. (1) for a general polynomial projection model.

$$E=\Sigma_{l,m}\left[\left(\Sigma_{i,j,k=0}^{i+j+k=n}C_{ijkm1}X_l^iY_l^iZ_l^k-x_{ml}\right)^2+\left(\Sigma_{i,j,k=0}^{i+j+k=n}C_{ijkm2}X_l^iY_l^jZ_l^k-y_{ml}\right)^2\right] \quad (1)$$

Where C is the projection map whose highest order is defined by users through the interface of EPiK. (x_{ml}, y_{ml}) is the projected coordinate of $X_lY_lZ_l$ on the mth tilt angle.

Various alternative models are possible for the alignment including polynomial functions, ratios of polynomials and mixed trigonometric-polynomial projection maps. We have implemented polynomial projection maps to 5th degree in our previous work[5]-[10].

Two output files will be generated automatically after the alignment process. One is a figure file that gives the coordinates of markers in 3D volume. The other file is a text file that will be used in the following reconstruction processing. It contains the coefficients of polynomial projection map from a 3D object to its 2D projections.

2.4 Reconstruction

2.4.1 Normalization—After alignment process, we can reconstruct the 3D structures. Because image statics should follow the cosine law, raw projection data should be normalized before being used to reconstruct, as shown in Figure 1 (a). In process of normalization, we adopt the following functions to readjust the gray value of each projection image.

1. Adjust grey scale to make all distributions equal in variance;
2. Adjust grey scale to make all distributions equal in mean;
3. Adjust distributions to follow cosine law;
4. Log transform image pixel values.

After that, we will get a “.st” file that has been normalized by the functions mentioned above.

2.4.2 Reconstruction Workflow—We have implemented two common reconstruction methods in EPiK. The first general method is filtered backprojection (FBP) which is relatively simple robust and fast. FBP is widely used in ET softwares such as IMOD [17] and TxBR. Iterative methods is the other reconstructed algorithm used in EPiK which is constituted by a class of alternatives to FBP in 3D reconstruction of ET. These methods both give good performance in handing incomplete, noisy data. In general, iterative methods are real-space reconstruction algorithms that formulate the 3D reconstruction problem as a large system of linear equations, as shown in Eq. (2).

$$v_i = \Sigma_{j=1}^{N} w_{ij} u_j, 1 \le i \le M \quad (2)$$

Iterative methods begin with an initial $u^{(0)}$ and repeat the iterative processes [18]. Here, we use an FBP solution as an initial value, which will generally improve the convergence rate, as simple backprojection is subject to a high degree of smoothing which obliterates the details, and in place of the starting point further from an actual solution. In each iteration cycle, the residuals, *i.e*. the differences between the actual projections ν and the computed projections ν' of the current approximation $u^{(k)}$ (*k* is the iterative number), are calculated and then $u^{(k)}$ is updated by the backprojection of these discrepancies. Thus, the algorithm produces a sequence of N-dimensional column vectors $u^{(k)}$. Eq. (3) below gives a typical iteration step.

$$u_j^{(k+1)} = u_j^{(k)} + \Sigma_{I=1}^{M} w_{ij} \left(v_i - \Sigma_{h=1}^{N} w_{ih} u_j^{(k)} \right) \quad (3)$$

In EPiK workflow, the generic iterative reconstructing process is described as follows:

1. Initialization: calculate initial value for $u_j^{(0)}$ by FBP;
2. Reprojection: estimate the projection data ν based on the current approximation $u_j^{(k)}$;
3. Backprojection: backproject the discrepancy $\Delta\nu$ between the experimental ν and calculated projections ν', and refine the current approximation *u* by incorporating the weighted backprojection Δ*u*.

2.4.3 Parallel Executio—Three-dimensional reconstruction in ET entails large computational costs and resources that are a function of the computational complexity of the reconstruction algorithms and the size of the projection images involved. This is especially true for wide-field tomography. Traditionally, high-performance computing has been used to address such computational requirements by means of parallel computing on supercomputers [19], large computer clusters [20], and multicore computers [21].

In EPiK workflow, we also use a parallel strategy to complete reconstructions on clusters. Our method permits the decomposition of the reconstruction problem into independent slabs along the Z - axis and makes the process well suited for parallelization. We have a natural choice for a parallel computation, in which the reconstruction along each Z-slice is calculated on a different processor. Thus, we can implement a parallel strategy where the sub-reconstruction along each Z-slice is calculated at the same time. This strategy makes use of message passing interface (MPI), standard in parallel programming. We can also apply a single program multiple data (SPMD) approach to perform the parallelization of the reconstruction on each Z-slice. The 3D volume is decomposed into several slabs with equal heights along the Z-axis. These slabs are assigned and reconstructed on an individual node on a cluster. The number of slabs equals to the number of nodes. Here, we adopt several actors including “SSH Session”, “SSH File Copier” and “GenericjobLauncher” to implement the parallel reconstruction strategy discussed above, as shown in Figure 1 (c).

3 Experiment

In this section, we report results of the reconstruction technique described in the previous sections.

In order to show the improvement of reconstruction by multiple tilt series, we collected 16 fold tilt series projection images and reconstructed the same sample by single tilt data and 16 tilt series data respectively. The sample is from the electric organ of an eel. The electric organ generates a large voltage pulse for defense of the eel. The reconstruction of its structure can greatly help biologists to understand its physiological functions, for example, the trans migration of ions across cell membranes in synchrony to san to large pulses.

The micrographs were taken in a 300kV FEI Titan TEM. The tilt series in this example is composed of 121 micrographs, each micrograph being 1024×1024. The size of reconstructions is 1035 × 997 × 66. The specimen is tilted from −60 to +60 degree in one-degree increment for each tilt. And the angle increment for two adjacent rotation tilts is 11.25 degree.

To get the reconstruction by EPiK, users just need to set the parameters shown in the interface and choose the reconstruction method. Subsequently, the processes will be automatically completed. EPiK has greatly facilitated the users to test and compare the results with different raw data and by different algorithms. In addition, it is convenient for developers to extend EPiK with new algorithms or functions.

Figure 2 shows one X-Y slice of the reconstruction along Z-axis. Figure 2 (a) is one X-Y section of the volume reconstructed by ASART after 5 iterations; and Figure 2 (b) is the section of the volume reconstructed by 16 fold tilt series. Chromatin coils are clearly visible in the second image.

The results show that by using EPiK, the 3D structure of the eel sample can be reconstructed successfully. From both figures, we can recognize different components of the sample, such as nucleus and filaments in the cytoplasm. Compared to Figure 2 (a), information is more widely distributed on the Orloff sphere[22], the reconstruction has higher quality in Figure 2 (b).

Note that multiple tilt series reconstruction requires more computational resources. For instance, the size of a 16 fold tilt series is about 65 Gigabyte with each projection image including 4096 by 4096 pixels. If we use higher resolution images collection or more tilt series data to reconstruct a sample, the data size may be in terabyte range.

Furthermore, serial sectioning is generally employed to reconstruct thicker samples. Montaging is commonly employed to increase the effective field of view. A montage consists of images of regions of the specimen that overlap in order to generate a larger map

[9][10]. With montaging, the dataset will be even larger. Right now, montages has already been implemented within TxBR.

4 Conclusion and Future Work

We implemented an electron tomography workflow EPiK that includes three basic steps: tracking, alignment and reconstruction. With Kepler, EPiK provides a visual language that can organize the processes of electron tomography in a workflow, and generally raises the level of abstraction in a complicated computation instance. EPiK can achieve the reconstruction of a sample from its microscope tilt series images effectively, and it makes things easier for the users. In addition, parallel computation is used in reconstruction in the current version of EPiK. Finally, both data sets and final reconstructions have increased by many orders of magnitude over the past two decades. EPiK provides means for handwork-independent implementation of improved parallelization. Therefore, EPiK will be a potentially useful tool in ET for biologists or any other researchers in different areas.

Future plans include: (1) Production of functionality to create a realistic phantoms with the following: realistic object, random markers, multiple tilt axes, off-center tilt axes, projections including helicity, warping and noise. This will yield a “ground truth” on which to evaluate and compare algorithms. (2) Incorporate alternative approaches to tracking, alignment and reconstruction. (3) Implementation of pre and post processing, for example, integrate other alignment such as marker-free alignment algorithm[23] to EPiK, and the nonlinear post processing for artifact reduction [24]. EPiK provides a framework for different kinds of algorithms comparison and graceful incorporation of improvements.

With EPiK, it is convenient for developers to extend the process with new algorithms. EPiK provides a simple means to compare and test the error using realistic phantoms and common examples of tilt series obtained in the lab. In line with development of TxBR, the montage function can also be added to EPiK for large field ET[9]. Finally, the process may be scaled up by parallel processing with supercomputers, large computer clusters and multicore computers.

References

[1]. Chase, Jared; Gorton, Ian; Sivaramakrishnan, Chandrika; Almquist, Justin; Wynne, Adam; Chin, George; Critchlow, Terence. Kepler + MeDICi. Proc. of the 2009 IEEE Congress on Services; Los Angeles, CA. Jul 6-10, 2009; 2009. p. 275-282.

[2]. Ludäscher, Bertram; Altintas, Ilkay; Berkley, Chad; Higgins, Dan; Jaeger-Frank, Efrat; Jones, Matthew; Lee, Edward A.; Tao, Jing; Zhao, Yang. Scientific workflow management and the Kepler system. Concurrency and Computation: Practice and Experience. 2006; 18:1039–1065.

[3]. Wang, Jianwu; Crawl, Daniel; Altintas, Ilkay. A framework for distributed data-parallel execution in the Kepler scientific workflow system. Procedia Computer Science. 2012; 9:1620–1629.

[4]. Wang, Jianwu; Altintas, Ilkay. Early cloud experiences with the Kepler scientific workflow system. Procedia Computer Science. 2012; 9:1630–1634.

[5]. Lawrence, Albert; Bouwer, James C.; Perkins, Guy, et al. Transform-based backprojection for volume reconstruction of large format electron microscope tilt series. Journal of Structural Biology. 2006; 154:144–167. 2006. [PubMed: 16542854]

[6]. Lawrence, Albert; Phan, Sébastien; Singh, Rajvikram. Parallel Processing and Large-Field Electron Microscope Tomograph. World Congress on Computer Science and Information Engineering. 2009:339–343.

[7]. Lawrence, Albert; Phan, Sebastien; Ellisman, Mark. Electron tomography and multiscale biology. Theory and Applications of Models of Computation Lecture Notes in Computer Science. 2012; 7287:109–130.

[8]. Lawrence, Albert; Phan, Sébastien. 3D reconstruction from Electron Microscope images with TxBR. Biomedical Imaging: From Nano to Macro, ISBI '09. IEEE International Symposium on; 2009. p. 339-343.

[9]. Phan, Sébastien; Lawrence, Albert; Molina, Tomas, et al. TxBR montage reconstruction for large field electron tomography. Journal of Structural Biology. 2012; 180:154–164. [PubMed: 22749959]

[10]. Phan, Sébastien; Terada, Masako; Lawrence, Albert. Serial reconstruction and montaging from large-field electron microscope tomograms. 31st Annual International Conference of the IEEE EMBS; 2009.

[11]. Kremer, James R.; Mastronarde, David N.; Mcintosh, J. Richard Computer visualization of three-dimensional image data using IMOD. Journal of Structural Biology. 1996; 116:71–76. [PubMed: 8742726]

[12]. Marbini R, Masegosa IM, San Martin MC, et al. Xmipp: An image processing package for Electron Microscopy. Journal of Structural Biology. 1996; 116:237–240. [PubMed: 8812978]

[13]. Sorzano COS, Marabini R, Velazquez-Muriel J, et al. XMIPP: a new generation of an open-source image processing package for electron microscopy. Journal of Structural Biology. 2004; 148:194–204. [PubMed: 15477099]

[14]. Tang, Guang; Peng, Liwei; Baldwin, Philip R., et al. EMAN2: An extensible image processing suite for electron microscopy. Journal of Structural Biology. 2007; 157:38–46. [PubMed: 16859925]

[15]. Fernandez JJ. High performance computing in structural determination by electron cryomicroscopy. Journal of Structural Biology. 2008; 164:1–6. 164. [PubMed: 18675361]

[16]. Frank, Joachim. Electron Tomography: Methods for Three-Dimensional Visualization of Structures in the Cell. Springer; 2006.

[17]. http://bio3d.colorado.edu/imod/

[18]. Wan, Xiaohua; Phan, Sébastien; Lawrence, Albert, et al. Iterative Methods in Large Field Electron Microscope Tomography. SIAM J. Sci. Comput. 2012; 35:S402–S419.

[19]. Fernandez JJ, Garazo JM, García I. Three-dimensional reconstruction of cellular structures by electron microscope tomography and parallel computing. Journal of Parallel and Distributed Computing. 2004; 64:285–300.

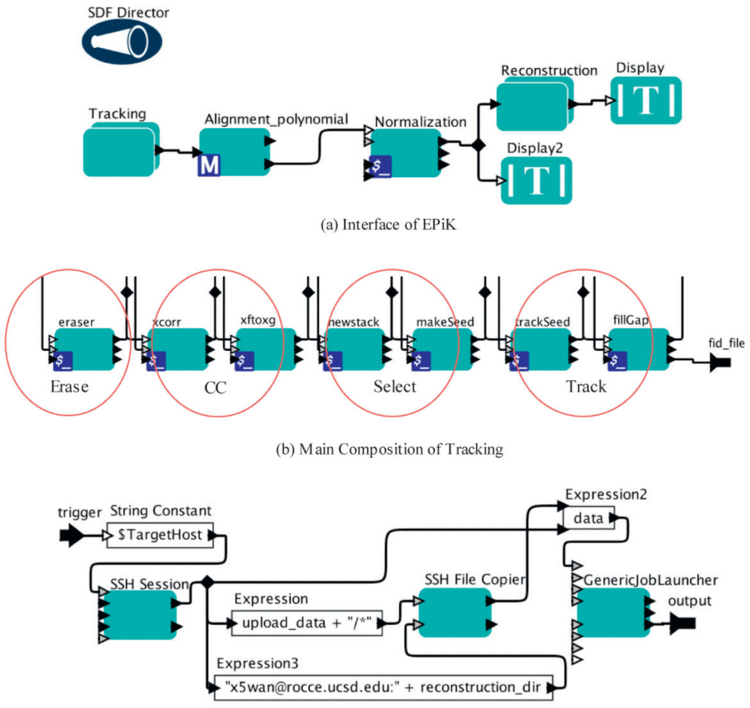
[20]. Wan, X.; Zhang, F.; Liu, Z. Modified simultaneous algebraic reconstruction technique and its parallelization in cryo-electron tomography. Proceeding of the International Conference on Parallel and Distributed Systems; 2009. p. 384-390.

[21]. Agulleiro JI, Fernandez JJ. Fast tomographic reconstruction on multicore computers. Bioinformatics. 2011; 27:582–583. [PubMed: 21172911]

[22]. Messaoudi, Cedric; de Loubresse, Nicole Garreau; Boudier, Thomas, et al. Multiple-axis tomography: applications to basal bodies from Paramecium tetraurelia. Biol. Cell. 2006; 98:415–425. [PubMed: 16499478]

[23]. Han, Renmin; Zhang, Fa; Wan, Xiaohua; Fernández, Jose-Jesus, et al. A marker-free automatic alignment method based on scale-invariant features. Journal of Structural Biology. 2014 in press.

[24]. Fernandez, J.; Agulleiro, J.; Bilao-Castro, J., et al. Image processing in electron tomography; Microscopy: science, technology, applications and education. 2010. p. 19-28.



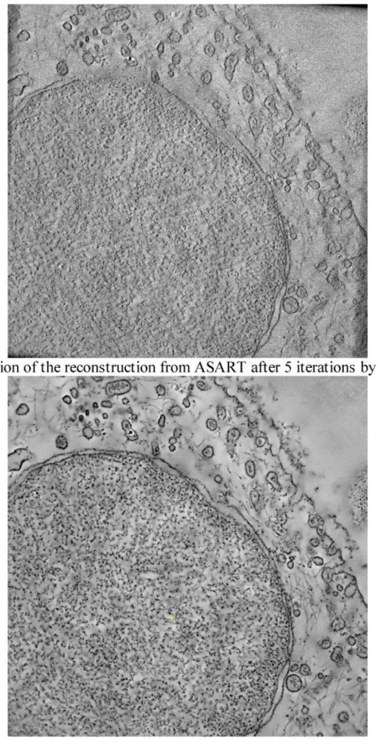
(a) Interface of EPiK

(b) Main Composition of Tracking

(c) Composition of Reconstruction

Figure 1.

EPiK Workflow. (a) The main interface of EPiK. It includes three main parts: tracking, alignment and reconstruction. Normalization is a pre-reconstruction step to insure that the grey-scale statistics are correct. (b) Main composition of tracking. IMOD is used for coarse tracking, and TxBR is used for fine tracking. All of the steps are integrated as a composite actor in EPiK. (c) Composition of reconstruction. There are parallel computings in this step. Multiple nodes in a cluster are used for large data sets.



(a) One X-Y section of the reconstruction from ASART after 5 iterations by single tilt images

(b) One X-Y section of the reconstruction from FBP by 16 fold tilt series images

Figure 2.

Reconstruction results of a cell nucleus from the electric organ of an eel sample. (a) The reconstruction by single tilt images; (b) The reconstruction by 16-fold tilt series images. With multiple tilt series images, the quality of volume reconstruction is greatly improved.