



HHS Public Access

Author manuscript

J Struct Biol. Author manuscript; available in PMC 2016 November 01.

Published in final edited form as:

J Struct Biol. 2015 November ; 192(2): 279–286. doi:10.1016/j.jsb.2015.06.016.

The Caltech Tomography Database and Automatic Processing Pipeline

H. Jane Ding^a, Catherine M. Oikonomou^a, and Grant J. Jensen^{a,b,*}

^aDivision of Biology, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA, 91125

^bHoward Hughes Medical Institute

Abstract

Here we describe the Caltech Tomography Database and automatic image processing pipeline, designed to process, store, display, and distribute electron tomographic data including tilt-series, sample information, data collection parameters, 3D reconstructions, correlated light microscope images, snapshots, segmentations, movies, and other associated files. Tilt-series are typically uploaded automatically during collection to a user's "Inbox" and processed automatically, but can also be entered and processed in batches via scripts or file-by-file through an internet interface. As with the video website YouTube, each tilt-series is represented on the browsing page with a link to the full record, a thumbnail image and a video icon that delivers a movie of the tomogram in a pop-out window. Annotation tools allow users to add notes and snapshots. The database is fully searchable, and sets of tilt-series can be selected and re-processed, edited, or downloaded to a personal workstation. The results of further processing and snapshots of key results can be recorded in the database, automatically linked to the appropriate tilt-series. While the database is password-protected for local browsing and searching, datasets can be made public and individual files can be shared with collaborators over the Internet. Together these tools facilitate high-throughput tomography work by both individuals and groups.

Keywords

electron tomography; structural biology database; image database; automatic processing; Caltech Tomography Database

INTRODUCTION

In electron tomography (ET), samples are repeatedly imaged in an electron microscope as they are tilted incrementally around an axis, producing a "tilt-series" of projection images.

*correspondence: tel. (626) 395-8827, jensen@caltech.edu.

SOFTWARE AVAILABILITY

The software is freely available upon request for academic users.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Three-dimensional reconstructions are then calculated from the tilt-series. ET is currently being applied to both biological and “materials science” samples, at both room- and cryo-temperatures (Gan and Jensen, 2012; Van Tendeloo et al., 2012). In the case of biological samples, cryotomography can produce 3-D views of cellular ultrastructure to ~4–6nm resolution in a near-native, “frozen-hydrated” state. When multiple copies of structures of interest are present in the tomograms, sub-tomogram averaging can overcome dose limitations and reveal details reliably at even higher resolution (Briegel et al., 2012; Briggs, 2013).

Over the past decade our lab has collected more than forty thousand electron tomograms of cells, viruses, and purified macromolecular complexes. While most of the tomograms were collected with the aim of characterizing a particular cellular ultrastructure, most also contain a plethora of interesting structural information about other known and unknown structures. As the number of different purifications, species, strains, and growth conditions we have imaged has increased, unanticipated discoveries have come from cross-comparisons. As examples, after we obtained the first 3-D images of complete flagellar motors within intact cells in *Treponema primitia* (Murphy et al., 2006), we later compared that structure to the structures of flagellar motors from 10 other species (Chen et al., 2011). In similar fashion, after we first discovered that bacterial chemoreceptor arrays were arranged in 12-nm hexagonal arrays inside *Caulobacter crescentus* cells (Briegel et al., 2008), we later found that this architecture is universally conserved across bacteria (Briegel et al., 2009) and archaea (Briegel et al., 2015). Other structures went unidentified, sometimes for years, until new clues emerged. This was the case, for instance, for CTP synthase filaments (Ingerson-Mahar et al., 2010) and the bacterial type VI secretion system (Basler et al., 2012). Many of the biological questions we are now tackling require recognizing patterns across hundreds of tomograms. This requires long-term access and facile comparison of tomograms taken over periods of many years by many different users.

Before we built the database, each user organized and stored their own tilt-series and reconstructions on disks attached to their personal workstations. In addition to the risk of such individual drives failing, this made it time-consuming for subsequent users to mine previous tomograms for new purposes. In order to address these challenges, we decided to establish a centralized database for all lab tomography data.

Many databases for structural data have been described (e.g. the BioImage database (Carazo et al., 1999; Lindek et al., 1999), EMEN2 (Ludtke et al., 2003; Rees et al., 2013), the Carragher Lab database (Fellmann et al., 2002), IMIRS (Dai et al., 2003), E-MSD (Boutselakis et al., 2003), PDB (Berman et al., 2003; Bernstein et al., 1977), and the EMDB (Henrick et al., 2003; Lawson et al., 2011; Tagari et al., 2002)), and we first explored whether an existing program might meet our needs. We installed and tested the Cell Centered Database developed at UC San Diego, a sophisticated and complex tool designed to house microscopy data of many different types in a distributed fashion across disk drives spread across the world (Martone et al., 2002; Martone et al., 2008). We found, however, that we preferred something simpler, local, and more customizable, so we designed and developed our own database, which we call the Caltech Tomography Database. Currently,

this database holds more than 46,000 3D images including more than 22,000 tomographic tilt-series of more than 200 different species/specimens (~70 terabytes of data).

An important goal in developing the Caltech Tomography Database was to facilitate high-throughput image processing. Thanks to advances in data acquisition, users who might have collected a handful of tilt-series per imaging session a decade ago can now routinely collect a tilt-series every 10–20 minutes automatically and continuously over multiple days. To facilitate the analysis of large volumes of data, we established a pipeline to manage imaging data as it comes off the microscope, including uploading a large number of tilt-series to the database, automatically processing them, and generating 2D images and 3D movies for rapid initial analysis.

Here we describe the goals of the database, its basic design, and how it has facilitated discovery in our lab over the last several years.

MATERIALS AND METHODS

Database construction

The Caltech Tomography Database is designed for tomographic data, and is therefore organized around tilt-series. Sample details, microscope collection parameters, and raw tilt-series files are included. Unlimited associated processing files, including 3D reconstructions, subvolumes, segmentations, relevant 2D images, correlated light microscopy images, movies, and other analysis files can be deposited and associated with a given tilt-series. Searchable text notes may be added. Structural features observed, such as filaments, flagella, carboxysomes, etc., can be recorded. Table 1 illustrates the information and types of files that the database may hold for each tilt-series.

The Caltech Tomography Database was constructed in MySQL, the second-most widely used open-source relational database management system after the mobile-packaged SQLite. MySQL was chosen for its fast performance, ease of use, and cost efficiency. The relational database contains tables corresponding to various features. This architecture is designed for ease of expansion, since additional tables can be added to handle additional features. Each dataset, consisting of a single- or dual-axis tilt-series and associated files, is identified by a unique code specifying the user (by initials) and experiment date, along with a sequence number. This is the “tilt-series ID.” There is no limit to the number or type of files that can be associated with a single tilt-series ID. To reduce the size of the database, 3D images and other uploaded files are not stored directly in the MySQL database. Since these datasets can be quite large (a single tilt-series is typically between 1.5 and 5 GB and can reach 10 GB), only the file type and a database ID number, from which the paths to the files can be generated, are recorded in the database. This allows flexibility in file storage, which can be carried out using different disk media. At Caltech, the database is stored on individual redundant arrays of independent disks (RAID) and periodically backed up in archived copies stored in different locations.

Database Access

The Caltech Tomography Database can be accessed through an Internet browser, providing flexibility in viewing and downloading data. The database web interface code was written in PHP, a widely-used server-side scripting language well suited to web development with the MySQL extension. Web interface functions include uploading, downloading, editing, browsing, and searching. 2D images are displayed if they are in a common web-friendly format such as .jpg, .gif, or .png. Files that cannot be embedded in the webpage are displayed as links to facilitate downloading. The interface is compatible with browsers on Linux, Windows, and Mac operating systems. On the administration side, there are a number of well-written front-end applications for MySQL database management such as phpMyAdmin, an open-source web-based tool. Additional scripts for managing and accessing the database were written in Python, including a script for uploading batches of reprocessed files to the database.

Automatic Processing Pipeline

The Automatic Processing Pipeline is run on a dedicated multi-core processing computer(s). We generated scripts in Python that allow automatic processing in one of two modes either – “real-time” (during data collection) or “off-line” (after data collection). To start the pipeline, users log into a processing computer and modify a starting script (a plain text file that is a standard Unix/Linux shell script) to input data collection information such as tilt-angles and diameter of fiducial markers and specify reconstruction parameters such as reconstruction method, binning factor, and dimensions of output file. When the starting script is executed, for each single-axis tilt-series in .mrc format, a tomographic reconstruction is automatically generated using command-line 3D reconstruction software. Currently the options in the Pipeline are RAPTOR (Amat et al., 2008) and Batchruntomo and automatic fiducial seeding and tracking in IMOD (Kremer et al., 1996; Mastronarde, 2013), but it can easily be expanded to include any command-line reconstruction software and/or filters. Key images are also automatically generated, consisting of an average of the five central slices of the tomographic reconstruction, or the zero tilt image of the tilt-series if no reconstruction is available. Finally, for each reconstruction, a key movie is generated. Embedded in the database browser display as a Flash video, this is a video clip of the contrast-adjusted reconstruction showing a slideshow of slices through the volume. Raw tilt-series, reconstructions, key images, and movies are all automatically deposited in the database. Intermediate processing files are kept on the local disk of the processing computer during processing and deleted after completion by default. However, when a user starts the pipeline, they are given the option of keeping alignment files on a shared disk array. The log files from the reconstruction software are kept with the datasets in the database.

Our processing computer is equipped with a queuing system. This is particularly useful when a large number of tilt-series have been collected off-line, since the queuing system can control how many tilt-series can be processed in parallel while keeping the rest in the queue. We are currently using the TORQUE Resource Manager (a free Portable Batch System (PBS)) with the Maui Cluster Scheduler (a free job scheduler). Multiple processing pipelines can be run simultaneously on one or multiple computers to increase the efficiency of processing. Currently, a small portion of the pipeline Python code integrates the Maui job

scheduler, but it can be modified easily to execute processes sequentially, in which case a queuing system/job scheduler is not needed.

The Automatic Processing Pipeline assumes that a tilt-series is represented as a single .mrc stack. If the acquisition software produces tilt-series in a different format, an extra conversion step can be added. The current “real-time” mode runs with the UCSF Tomography acquisition software (Zheng et al., 2007). The part of the code that detects newly completed tilt-series may need minor modifications if other acquisition software is used. Additional image processing procedures can be added to the pipeline as long as there are command-line tools for the processes.

The time needed to process one tilt-series depends on the size of the tilt-series, the size of the output reconstruction, the reconstruction software, the speed of the processing computer and, if the data needs to be transferred through a network, the speed of the network. Typically, the collection time for a tilt-series on our microscope is around 20 minutes. We have found that a multi-core Dell Precision computer with 64 GB RAM is easily able to keep up with “real-time” processing.

If multiple command-line reconstruction software packages are installed, pipeline users can prioritize their preferred software, and instruct the pipeline to try different software automatically in case the initial software fails to produce a reconstruction. If all software fails on a particular tilt-series, a data record marked “failed” appears in the user’s Inbox with no reconstruction. The user can then examine the raw tilt-series and log file in the database and decide how to proceed. In our experience, such failures are rare so we do not automatically track success rates; however, it would be straightforward to implement a tool to do this. In the future, we may also add a feature to record alignment error, which is not currently part of the pipeline.

The flow of data in the Database and Processing Pipeline is summarized in Figure 1.

Installation Requirements

The Caltech Tomography Database is designed to be lightweight and easy to install and use. The only required equipment for hosting the database itself is a computer running MySQL and a web server, with available disk space to host data files. The lightweight design of the database means that the server computer does not have to be particularly powerful. We recommend a separate computer to run the processing pipeline, although it can theoretically be run on the server machine. Requirements for the processing computer depend on the size of the tilt-series and reconstruction files. The computer should be powerful enough to run the reconstruction software. Software packages required for tomographic reconstruction and image processing are Python/MySQLdb; RAPTOR (Amat et al., 2008); IMOD (Kremer et al., 1996); Bsoft (Heymann, 2001; Heymann and Belnap, 2007; Heymann et al., 2008); and movie-making software components (mencoder, ffmpeg, convert) from Mplayer and ImageMagick. If running multiple jobs simultaneously, a PBS/Maui job scheduler is needed. All of these are available for free online. Server-side codes were tested on 64-bit Redhat Enterprise Linux 5 and 6 systems with various versions of MySQL (up to 5.1.73) and PHP (up to 5.3.3).

RESULTS AND DISCUSSION

The Caltech Tomography Database (Figure 2) was designed to manage the increasingly large volume of imaging data produced by an ET lab. Developed over seven years, the database evolved in response to testing and suggestions by lab members. Ultimately, we envisioned a central repository of all tomographic data generated in the lab that would facilitate high-throughput tomography, be easily searchable by any internal user and facilitate distribution to colleagues. We discuss the main features in detail below.

Upload

First, users must be able to upload files to the database. Designed for tomography, the database is organized around tilt-series. Each dataset consists of a single tilt-series and all associated files. Figure 3 shows a typical tilt-series record with associated files and key movie in a pop-out Flash window. There is no limit to the number or type of files that can be associated with a single tilt-series. There are three ways to upload data.

1. Automatic Processing Pipeline (Inbox)

At Caltech, the vast majority of data is uploaded through the automated pipeline. In this case, tilt-series and automatically-generated 3D-reconstruction files are imported automatically. Once uploaded, datasets appear in a user's "Inbox," where users can perform an initial screening of their data. They may delete failed datasets or accept successful datasets, adding additional information as desired.

2. Script-based uploading

A collection of Python scripts allows users to upload data from their workstations, singly or in batches, without using the automated pipeline. This is useful when a user has collected tilt-series on a remote instrument, reconstructs a tilt-series "manually" using interactive software, or wants to upload additional analysis files such as sub-tomogram averages or segmentations.

3. Browser-based uploading

Alternatively, users can upload individual datasets through a web interface that creates a tilt-series ID and prompts the user to choose files from their local workstation to be uploaded and associated with that tilt-series ID. A script for transferring the files is generated, which the user then executes. This interface also allows users to upload additional files to an existing dataset in the database.

Workbox

Workboxes (Figure 2) are designed to hold all the datasets associated with a specific project for ease of viewing, processing, editing, or downloading. Workboxes are essentially just lists of tilt-series IDs, so datasets can be part of multiple workboxes. Since one of the main functions of the database is to facilitate mining of past and current data, users are not restricted to datasets they collected. Users can create new workboxes or delete existing ones, include or exclude datasets from a workbox, input associated notes, and rename workboxes. The Workbox provides tools for executing certain functions in batch: editing, downloading,

re-processing, and generating lists of file URLs for sharing. All files associated with a Workbox can be downloaded to a personal workstation in a way that maintains their directory relationships. This makes it possible for users to re-process or otherwise further analyze data and then upload key final results back to the existing records in the database easily *en masse* via scripts. These functionalities are also available in the user's Inbox.

Annotation

A major goal of the Caltech Tomography Database is to exploit fully all the structural detail of the tomograms we collect. To that end, it is important to be able to flag interesting details for future study. We implemented several functions to do this, including "Text Notes," "Features Observed," and "Snapshots."

1. Text Notes

Text notes can be recorded for each tilt-series dataset. The Edit page allows users to enter text corresponding to "Notes, Keywords, and Tags." New notes are appended to, but do not replace, existing notes, and all notes are searchable, as described below.

2. Features Observed

"Features Observed" consists of a list of common structural details found in cells, one or more of which can be checked for a given dataset. Users can subsequently search for a given feature to pull out all tomograms with that tag. Current cellular features being tracked include carboxysomes, chemoreceptors, filaments, flagella, granules, internal membranes, pili, surface layers and nucleoids, but this list frequently expands.

3. Snapshots – IMOD Plugin

"Snapshots" are 2D images associated with a dataset that highlight interesting features in a tomogram. We worked with David Mastronarde at the University of Colorado, Boulder to implement a "Grab with Note" plugin for the visualization program 3dmod in IMOD, of which he is an author (Kremer et al., 1996). When viewing tomograms with 3dmod, users can now mark cellular features of interest with a brief text description and arrows and record the result as a "snapshot." Additionally, each snapshot records the coordinates, zoom, and rotation parameters to allow subsequent users to instantly access the relevant view within an often visually complicated tomogram. For users without a database, these snapshots are saved by 3dmod to their local desktop. For Caltech Tomography Database users, snapshots are uploaded to the database automatically, linked to the appropriate tilt-series, and can be browsed from the database webpage as a thumbnail grid or slideshow (Figure 4). An icon on the database record links to a script that will reopen the exact view in 3dmod on the user's local computer. Additional searchable text notes can be added to a snapshot in the database by any user.

Browse and Search

Once data files and annotations are in the database, it is important to allow users, both current and future, to mine the database for information relevant to their interests. To that end we implemented browse and search functions that allow navigation of the database. “Browse” allows viewers to choose which datasets are displayed and organize them by user and/or species/specimen, ordered by “Time Modified,” “Date Taken,” or “Most Visited.” Each dataset is represented by a thumbnail image, which links to the full-size image in a new window, along with additional information including the tilt-series ID and a short descriptive name of the sample chosen by the user. In addition, a Flash movie displaying the reconstruction can be accessed with a video camera icon. The Browse page also provides several simple filters that allow viewers to specify the date range and/or words or phrases in associated text fields. Selecting the tilt-series ID opens the full record of the dataset, containing all metadata and associated files (Figure 3).

The “Search” function allows more sophisticated navigation of the database. Tilt-series can be filtered by species/specimen, user, features observed, acquisition details (microscope, software, single- or dual-axis tilt), and/or date of acquisition. Viewers can also perform text-based searches for terms in one or more metadata fields, or can search for datasets associated with specific publications. Search results are displayed as for the “browse” function described above. A separate search function allows viewers to query the text notes associated with snapshots taken with the “Grab with Note” IMOD plug-in. As mentioned above, all snapshots from a given set of tomograms or date range or other selection criteria can be browsed either in a montage or slideshow (Figure 4). This allows users to quickly browse through all the interesting structures noted in the tomograms so far, or for group leaders to stay abreast of all new findings emerging week by week, for instance.

Security and Distribution

While password protection can be used to restrict access when needed, the database also allows users to share data easily with collaborators. Users can select one or more tilt-series and then designate which files to share: binned movies, raw tilt-series, and/or reconstructions. Lists of URLs are then generated that allow anyone with the link to download the relevant file from a separate server with protected directories. A “public” flag has also been implemented that can be used to make selected tilt-series in the database accessible to the general public.

CONCLUSIONS

As the number and diversity of complex 3D datasets increases, databases such as the Caltech Tomography Database will become increasingly important. We hope our description of the goals, design, and use of this database will inform and accelerate similar development efforts in the future. We are also pleased that several other ET labs have already found the present implementation useful both as a working tool and as a basis for further expansion.

Acknowledgments

We thank David Mastronarde for assisting with the “Grab with Note” plugin and integrating it into the IMOD software. This work was supported in part by NIH grant 2P50GM082545 to GJJ and the Beckman Institute at Caltech. This project/publication was made possible through the support of a grant from the John Templeton Foundation as part of the Boundaries of Life project. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

References

- Amat F, Moussavi F, Comolli LR, Elidan G, Downing KH, Horowitz M. Markov random field based automatic image alignment for electron tomography. *Journal of structural biology*. 2008; 161:260–275. [PubMed: 17855124]
- Basler M, Pilhofer M, Henderson GP, Jensen GJ, Mekalanos JJ. Type VI secretion requires a dynamic contractile phage tail-like structure. *Nature*. 2012; 483:182–186. [PubMed: 22367545]
- Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nature structural biology*. 2003; 10:980. [PubMed: 14634627]
- Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank. A computer-based archival file for macromolecular structures. *European journal of biochemistry / FEBS*. 1977; 80:319–324. [PubMed: 923582]
- Boutselakis H, Dimitropoulos D, Fillon J, Golovin A, Henrick K, Hussain A, Ionides J, John M, Keller PA, Krissinel E, McNeil P, Naim A, Newman R, Oldfield T, Pineda J, Rachedi A, Copeland J, Sitnov A, Sobhany S, Suarez-Uruena A, Swaminathan J, Tagari M, Tate J, Tromm S, Velankar S, Vranken W. E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic acids research*. 2003; 31:458–462. [PubMed: 12520052]
- Briegel A, Li X, Bilwes AM, Hughes KT, Jensen GJ, Crane BR. Bacterial chemoreceptor arrays are hexagonally packed trimers of receptor dimers networked by rings of kinase and coupling proteins. *Proceedings of the National Academy of Sciences of the United States of America*. 2012; 109:3766–3771. [PubMed: 22355139]
- Briegel A, Ortega DR, Huang A, Oikonomou CM, Gunsalus RP, Jensen GJ. Structural conservation of chemotaxis machinery across Archaea and Bacteria. *Environmental microbiology reports*. 2015
- Briegel A, Ding HJ, Li Z, Werner J, Gitai Z, Dias DP, Jensen RB, Jensen GJ. Location and architecture of the *Caulobacter crescentus* chemoreceptor array. *Molecular microbiology*. 2008; 69:30–41. [PubMed: 18363791]
- Briegel A, Ortega DR, Tocheva EI, Wuichet K, Li Z, Chen S, Muller A, Iancu CV, Murphy GE, Dobro MJ, Zhulin IB, Jensen GJ. Universal architecture of bacterial chemoreceptor arrays. *Proceedings of the National Academy of Sciences of the United States of America*. 2009; 106:17181–17186. [PubMed: 19805102]
- Briggs JA. Structural biology in situ--the potential of subtomogram averaging. *Current opinion in structural biology*. 2013; 23:261–267. [PubMed: 23466038]
- Carazo JM, Stelzer EH, Engel A, Fita I, Henn C, Machtynger J, McNeil P, Shotton DM, Chagoyen M, de Alarcon PA, Fritsch R, Heymann JB, Kalko S, Pittet JJ, Rodriguez-Tome P, Boudier T. Organising multi-dimensional biological image information: the BioImage Database. *Nucleic acids research*. 1999; 27:280–283. [PubMed: 9847201]
- Chen S, Beeby M, Murphy GE, Leadbetter JR, Hendrixson DR, Briegel A, Li Z, Shi J, Tocheva EI, Muller A, Dobro MJ, Jensen GJ. Structural diversity of bacterial flagellar motors. *The EMBO journal*. 2011; 30:2972–2981. [PubMed: 21673657]
- Dai W, Liang Y, Zhou ZH. Web portal to an image database for high-resolution three-dimensional reconstruction. *Journal of structural biology*. 2003; 144:238–245. [PubMed: 14643226]
- Fellmann D, Pulokas J, Milligan RA, Carragher B, Potter CS. A relational database for cryoEM: experience at one year and 50 000 images. *Journal of structural biology*. 2002; 137:273–282. [PubMed: 12096895]
- Gan L, Jensen GJ. Electron tomography of cells. *Quarterly reviews of biophysics*. 2012; 45:27–56. [PubMed: 22082691]

- Henrick K, Newman R, Tagari M, Chagoyen M. EMDep: a web-based system for the deposition and validation of high-resolution electron microscopy macromolecular structural information. *Journal of structural biology*. 2003; 144:228–237. [PubMed: 14643225]
- Heymann JB. Bsoft: image and molecular processing in electron microscopy. *Journal of structural biology*. 2001; 133:156–169. [PubMed: 11472087]
- Heymann JB, Belnap DM. Bsoft: image processing and molecular modeling for electron microscopy. *Journal of structural biology*. 2007; 157:3–18. [PubMed: 17011211]
- Heymann JB, Cardone G, Winkler DC, Steven AC. Computational resources for cryo-electron tomography in Bsoft. *Journal of structural biology*. 2008; 161:232–242. [PubMed: 17869539]
- Ingerson-Mahar M, Briegel A, Werner JN, Jensen GJ, Gitai Z. The metabolic enzyme CTP synthase forms cytoskeletal filaments. *Nature cell biology*. 2010; 12:739–746. [PubMed: 20639870]
- Kremer JR, Mastronarde DN, McIntosh JR. Computer visualization of three-dimensional image data using IMOD. *Journal of structural biology*. 1996; 116:71–76. [PubMed: 8742726]
- Lawson CL, Baker ML, Best C, Bi C, Dougherty M, Feng P, van Ginkel G, Devkota B, Lagerstedt I, Ludtke SJ, Newman RH, Oldfield TJ, Rees I, Sahni G, Sala R, Velankar S, Warren J, Westbrook JD, Henrick K, Kleywegt GJ, Berman HM, Chiu W. EMDataBank.org: unified data resource for CryoEM. *Nucleic acids research*. 2011; 39:D456–464. [PubMed: 20935055]
- Lindek S, Fritsch R, Machtynger J, de Alarcon PA, Chagoyen M. Design and realization of an on-line database for multidimensional microscopic images of biological specimens. *Journal of structural biology*. 1999; 125:103–111. [PubMed: 1022267]
- Ludtke SJ, Nason L, Tu H, Peng L, Chiu W. Object oriented database and electronic notebook for transmission electron microscopy. *Microscopy and microanalysis : the official journal of Microscopy Society of America, Microbeam Analysis Society, Microscopical Society of Canada*. 2003; 9:556–565.
- Martone ME, Gupta A, Wong M, Qian X, Sosinsky G, Ludascher B, Ellisman MH. A cell-centered database for electron tomographic data. *Journal of structural biology*. 2002; 138:145–155. [PubMed: 12160711]
- Martone ME, Tran J, Wong WW, Sargis J, Fong L, Larson S, Lamont SP, Gupta A, Ellisman MH. The cell centered database project: an update on building community resources for managing and sharing 3D imaging data. *Journal of structural biology*. 2008; 161:220–231. [PubMed: 18054501]
- Mastronarde, DN. Automated Tomographic Reconstruction in the IMOD Software Package. *Microscopy & Microanalysis Conference; Indianapolis, IN*. 2013.
- Murphy GE, Leadbetter JR, Jensen GJ. In situ structure of the complete *Treponema primitia* flagellar motor. *Nature*. 2006; 442:1062–1064. [PubMed: 16885937]
- Rees I, Langley E, Chiu W, Ludtke SJ. EMEN2: an object oriented database and electronic lab notebook. *Microscopy and microanalysis : the official journal of Microscopy Society of America, Microbeam Analysis Society, Microscopical Society of Canada*. 2013; 19:1–10.
- Tagari M, Newman R, Chagoyen M, Carazo JM, Henrick K. New electron microscopy database and deposition system. *Trends in biochemical sciences*. 2002; 27:589. [PubMed: 12417136]
- Van Tendeloo G, Bals S, Van Aert S, Verbeeck J, Van Dyck D. Advanced electron microscopy for advanced materials. *Advanced materials*. 2012; 24:5655–5675. [PubMed: 22907862]
- Zheng SQ, Keszthelyi B, Branlund E, Lyle JM, Braunfeld MB, Sedat JW, Agard DA. UCSF tomography: an integrated software suite for real-time electron microscopic tomographic data collection, alignment, and reconstruction. *Journal of structural biology*. 2007; 157:138–147. [PubMed: 16904341]



Figure 1. Workflow of the Caltech Tomography Database and Automatic Processing Pipeline
This schematic shows the flow of data (arrows) through hardware elements (boxes) resulting from the indicated processes (red). Briefly, the Automatic Processing Pipeline collects tilt-series from a data source, either the collecting microscope (“real-time” mode) or a hard drive (“off-line” mode). Tomographic reconstructions, as well as key images and movies, are then generated by the processing computer, a basic workstation equipped with reconstruction and movie-making software. These automatically-generated files are then uploaded to the database, which consists of servers (web and MySQL) and a disk array. Individual users can browse the database, add annotations, download data to their personal computers, and/or upload additional data files. See Materials and Methods for detailed specifications.

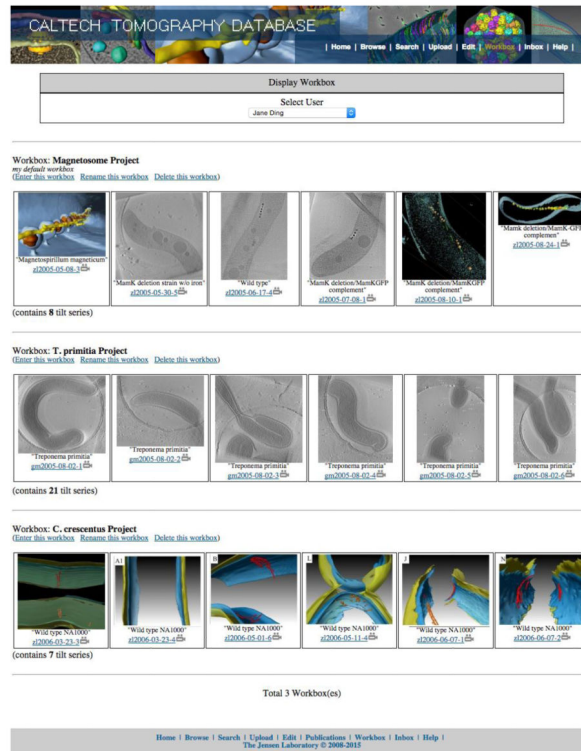


Figure 2. The Caltech Tomography Database Interface

A screenshot of the database interface, which allows users to search and/or browse tilt-series datasets, collecting them in Workboxes to facilitate processing for individual projects. Datasets are represented by key images and tilt-series IDs link to the full tilt-series record.

Tilt Series date: 08/11/2005
Data Taken By: Gavin Murphy
Descriptive Title: Treponema primitia
Description: Treponema primitia The rest of the files are in ~/images/050811
 (Original path: /mnt/fe/Meta/gavmm/Backup/Up/images/gm2005-08-11-04)
Species / Specimen: Treponema primitia ZAS-2 545694
Collaborators and Roles: Cells provided by Eric Matson, Jared Leadbetter; Data collected and processed by Gavin Murphy
Purification, Growth Conditions & Treatment: Cells grown in standard media around exponential growth. Removed from anaerobic tube with syringe. Directly applied to grids, no centrifugation
Sample Preparation: quantifoil grid, 10nm gold, 100% humidity...
Tilt Series Setting: single tilt, constant angular increment, step: 1, tilt range: (-60, 60) degrees, dosage: 30/A2, defocus: -12um.
Acquisition Software: UCSF Tomo
Processing Software Used: imod, bsoft
Publications:
 - Gavin E. Murphy, Jared R. Leadbetter, Grant J. Jensen, "In situ structure of the complete Treponema primitia cell wall," *Nature*, 2006, 442, 1062-1064
3D Image (#10): [T4a_S497G2_22k18d110e.mrc](#) Raw Data (pixel size: 0.98nm)
3D Image (#11): [T4B2s.mrc](#) 3d Reconstruction
3D Image (#12): [T4B2FadBotSiz.mrc](#) Subvolume
3D Image (#13): [T4B2sBR_28_partTrn.tif](#) Subvolume
File (#49): [T4B2sBR_28_partTrn.tif](#) Subvolume
 Notes: Files under Other have not been uploaded

Images: (click on image to see full version)

central slice file associated to 3D Image #12 file associated to 3D Image #12 file associated to 3D Image #12

file associated to 3D Image #12 file associated to 3D Image #12 file associated to 3D Image #12 file associated to 3D Image #12

file associated to 3D Image #12 file associated to 3D Image #12 file associated to 3D Image #12

file associated to 3D Image #12 file associated to 3D Image #12




Figure 3. Tilt-Series Record

A screenshot of a tilt-series record shows information about sample preparation and data acquisition, as well as links to the raw data and 3D reconstructions. Thumbnails link to associated processing files, as well as a key movie, which is shown at right in a pop-out Flash window.

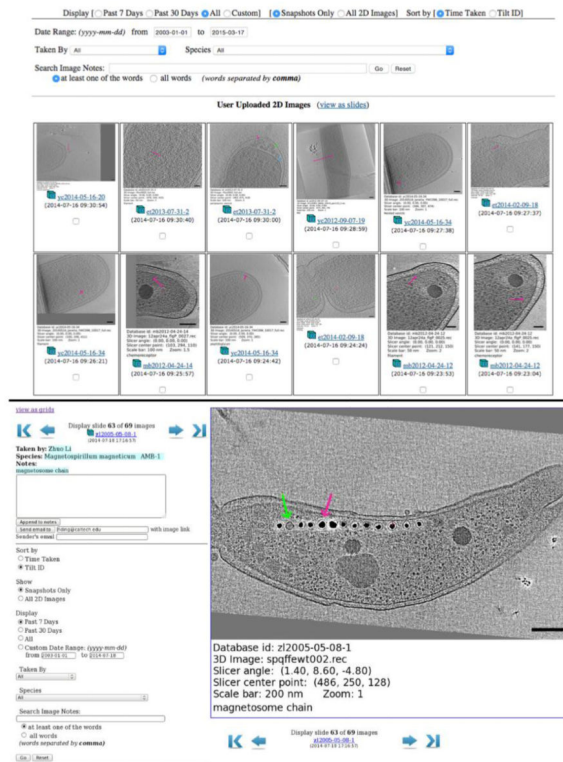


Figure 4. Snapshot Function

Database users can use the “Grab with Note” plugin in the IMOD viewing software to capture a snapshot of a relevant feature in a tomogram. Snapshots are automatically uploaded to the database, associated with the correct tilt-series, and can be searched or browsed in a thumbnail grid (top) or slideshow (bottom) view. Coordinates and rotation details are stored so that a link reopens the tomogram in IMOD to the same view.

Table 1

Types of information and files that can be associated with a tilt-series in the database.

Tilt Series					
Sample Descriptions	Data Collection Values	3D Images *	2D Images *	Associated Files *	Automatic Generated *
species & specimen sample preparation notes, keywords, and tags descriptive title features observed principle experimenter & collaborators experiment date publications	single or dual-axis magnification defocus tilt scheme tilt range and step dosage microscope acquisition software	raw tilt series reconstruction sub-volumes other 3D images	(displayed) snapshots figures	segmentations movies any other type of processing or analysis files	thumbnail image featured movie pipeline log

* database records the paths to the actual files