HHS Public Access

Author manuscript

Methods Enzymol. Author manuscript; available in PMC 2017 March 15.

Published in final edited form as:

Methods Enzymol. 2016 ; 579: 393–412. doi:10.1016/bs.mie.2016.04.015.

Databases and archiving for cryoEM

Ardan Patwardhan[a] and **Catherine L. Lawson**[b]

[a]Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton CB10 1SD, United Kingdom

[b]Department of Chemistry and Chemical Biology and Research Collaboratory for Structural Bioinformatics, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA

Abstract

Cryo-EM in structural biology is currently served by three public archives – EMDB for 3DEM reconstructions, PDB for models built from 3DEM reconstructions and EMPIAR for the raw 2D image data used to obtain the 3DEM reconstructions. These archives play a vital role for both the structural community and the wider biological community in making the data accessible so that results may be reused, reassessed and integrated with other structural and bioinformatics resources. The important role of the archives is underpinned by the fact that many journals mandate the deposition of data to PDB and EMDB on publication. The field is currently undergoing transformative changes where on the one hand high-resolution structures are becoming a routine occurrence while on the other hand electron tomography is enabling the study of macromolecules in the cellular context. Concomitantly the archives are evolving to best serve their stakeholder communities.

In this chapter we describe the current state of the archives, resources available for depositing, accessing, searching, visualising and validating data, on-going community-wide initiatives and opportunities and challenges for the future.

Keywords

Databases; EMDataBank; EMDB; Protein Data Bank; EMPIAR; Validation

1. Introduction

In recent years cryo-electron microscopy (cryoEM) and electron tomography (cryoET) have become indispensable tools for molecular and cellular structural biology. In the past they were commonly used to complement the more established techniques of X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy. Single-particle EM enables the study of large macromolecular assemblies and complexes in a close-to-native environment without the need for generating large amounts of purified material, forming crystals, or isotopic labelling. Single-particle cryoEM even forgoes the requirement for

Correspondence should be addressed to: Patwardhan: Phone: + 44 1223 492649; Fax: + 44 1223 494 468; ardan@ebi.ac.uk. Lawson: Phone: +1 848 445 5494; Fax: +1 732 445 4320; cathy.lawson@rutgers.edu.

extreme sample homogeneity, either compositional or conformational, since multiple states can be computationally separated into different 3D classes. CryoET is the method of choice when studying pleomorphic structures such as the HIV virus or structures in the cellular context.

Traditionally, EM based techniques have yielded 3D structures with limited resolution, preventing the direct unambiguous interpretation of the data in terms of biological entities. Electron diffraction and imaging have been used successfully on helical and 2D crystalline arrays to overcome this hurdle and obtain structures to atomic resolution, e.g., the αβ tubulin dimer (Nogales et al., 1998), but researchers using diffraction methods face the traditional challenge of obtaining well ordered crystals. These issues have prevented wider use but electron crystallography has found a niche in structure determination of membrane proteins e.g., aquaporin at 1.9Å resolution (Gonen et al., 2005). More typically, interpretation of lower resolution 3D maps has been aided by fitting of atomic coordinate models derived from other experiments, by comparing maps of related structures, or by segmenting the structure using other biochemical information or prior knowledge. For cryoET the problem of limited resolution has been even more severe owing to intrinsic limitations of the technique such as missing wedges, radiation damage from imaging the same specimen area multiple times, and specimen tilting. However, the resolution may be adequate for the purpose of the experiment, for example to examine the distribution or organization of a complex and molecular assembly in the cell (Brandt et al., 2010). In cases where there is ambiguity, other methods can be used to robustly identify targets, for example correlative light microscopy with fluorescent tagging (Kukulski et al., 2012). Sub-tomogram averaging (see chapter by Briggs) – a technique similar to single-particle methodology but involving the averaging and classification of 3D sub-volumes, can be used to improve resolution and overcome tomographic artefacts. Using classification techniques, sub-tomogram averaging enables visualization of structural variability in a cellular context.

In the past few years there have been major technological advances in the field, including the introduction of the direct electron detector, that have enabled the determination of single-particle structures to atomic resolution, and cryoET has also benefitted from the improved resolution. At the same time there has been an increased emphasis on combining different structural techniques to build up a holistic understanding of the biological problem at hand. Here electron tomography and correlative light microscopy has been vital in providing the cellular context to the macromolecular world (Zeev-Ben-Mordehai et al., 2014). Other notable developments include that for the first time there are phase plate technologies sufficiently robust for routine adoption (Danev and Baumeister, 2016) and that 3D electron diffraction has been successfully used to determine structures to 1.4 Å resolution (Rodriguez et al., 2015).

The structural biology community was one of the first to recognize the value of providing public open access to data from X-ray crystallography with the inception of the Protein Data Bank (PDB) as an archive for atomic coordinate models in 1971 (Berman et al., 2012). Open access to data provides a means to independent validation, re-use and integration of structural information. The PDB has served as a source of data for methods development and teaching, driving the field forward. It has also been a focal point for community wide efforts

on many issues including standardisation and validation that have benefitted the field. Today the PDB archive comprises over 110000 structures, including over 1000 structures determined using EM-based techniques (Figure 1, green bars). Deposition of experimentally derived atomic coordinate model structures to PDB is mandatory upon publication for most relevant journals. The PDB is managed by the members of the Worldwide Protein Databank (wwPDB; wwpdb.org; (Berman et al., 2003)): the Research Collaboratory for Structural Bioinformatics PDB (RCSB PDB), the Protein Data Bank in Europe (PDBe) at the European Bioinformatics Institute (EMBL-EBI), the Protein Data Bank Japan (PDBj) at the Institute for Protein Research in Osaka University, and the Biological Magnetic Resonance Bank (BMRB) at the University of Wisconsin-Madison.

In the same vein, in the late 1990's and early 2000's there was a growing realisation by EM researchers of the need for a similar resource for EM derived structures. At that time, most EM structures were not solved to a resolution where an atomic coordinate structure could be built from the 3D EM volume so it was critical for the volume itself to be stored. The Electron Microscopy Data Bank (EMDB) was set up in 2002 at EMBL-EBI as an archive for 3DEM reconstructions (Tagari et al., 2002). It now comprises over 3400 structures (Figure 1, purple bars) from a variety of EM techniques including single-particle, electron tomography, sub-tomogram averaging and 2D and 3D electron diffraction (Figure 2). Current trends in the field reflect directly on depositions to EMDB. Figure 3 shows how the number of structures deposited at better than 4 Å resolution has increased dramatically in the past few years and Figure 4 highlights the importance of the direct electron detector in advancing the field. Map volume deposition rates for published 3DEM structures have been gradually increasing as the potential of 3DEM methods is recognized. Many journals have implemented policies requiring experimental data to be deposited for EM-based studies. Nowadays many EM experiments involve coordinated depositions of the 3DEM volume to EMDB and fitted or built atomic coordinates to PDB. Another trend is towards hybrid experiments where constraints from several different methods are combined to obtain a structure. Notable examples include the nuclear pore complex (Alber et al., 2007), and amyloid fibrils (Fitzpatrick et al., 2013). The current data archives do not fully support the full range of possible hybrid experimental data; the challenges are discussed in more detail by Sali *et al.* (Sali et al., 2015).

The EMDB archives the final 3D reconstructions (map volumes) from EM experiments. There have been growing calls from the EM community for the public archiving of the raw EM image data, both to serve as benchmarks (Henderson et al., 2012) and to allow others to perform a full validation of the experimental results (Glaeser, 2013; Henderson, 2013). The raw data is often orders of magnitude larger in size than the final 3D reconstructions and the EMDB infrastructure is not able to cope with the storage or transfer of these large datasets. In 2014, PDBe created EMPIAR (Iudin et al., 2016), a dedicated archive for raw EM image data designed to handle large data set transfers from the outset. EMPIAR now comprises over 45 datasets averaging 700GB in size with 5 datasets over a TB in size. In its short existence, EMPIAR has already been cited over 16 times and EMPIAR data is downloaded at an average rate of 10TB per month, underlining the important role it is playing for the EM community. EMPIAR data is used for a range of purposes including validation, methods development, testing and training. Based on input from the community, PDBe is also

working on extending EMPIAR to support related imaging modalities including 3D scanning electron microscopy, soft X-ray tomography and correlative light and electron microscopy.

2. Resources

The EMDataBank website (www.emdatabank.org) provides a unifying portal to resources relating to 3DEM map and model data deposited to EMDB and PDB. The EMDataBank project is a joint effort among PDBe, RCSB PDB, and the National Center for Macromolecular Imaging (NCMI) at Baylor College of Medicine (Lawson et al., 2016; Lawson et al., 2011). Resources for EMDB and EMPIAR from PDBe may also be accessed via the links http://pdbe.org/emdb and http://pdbe.org/empiar respectively.

2.1. Searching, Browsing, and Visualizing data

Two search services are currently available through EMDataBank. EMSEARCH (Lawson et al., 2011), hosted at RCSB, facilitates simple searches of EMDB based on author name, title, entry ID, sample name, citation abstract content, aggregation type, resolution, and/or release date. Advanced EMDB search (http://pdbe.org/emsearch; (Gutmanas et al., 2014)), hosted at PDBe, provides additional capabilities such as searches by sample molecular weight, taxonomy, reconstruction software package, microscope model and parameters, and has the ability to filter and further refine initial search results. Both search sites provide results listings with links to individual entry pages, where one can view overview information and access links for visualization tools and file downloads. The EMDB archive can also be investigated using the EMStats statistics service (http://pdbe.org/emstats; (Gutmanas et al., 2014)).

Three types of web-based visualization are available for 3DEM structures. First, a Java-applet-based volume viewer permits 3D visualization of maps and their associated PDB coordinate models (Lagerstedt et al., 2013; Lawson et al., 2011). Second, a volume slicer is available for inspecting 2D slices of EMDB entries from three orthogonal orientations (Salavert-Torres et al., 2016). Third, visual analysis pages (Gutmanas et al., 2014; Lagerstedt et al., 2013; Patwardhan et al., 2012) facilitate analysis and validation of maps, tomograms and models by providing static images of map orthogonal projections and central slices as well as graphs of FSC curves, map density distribution, rotationally averaged power spectrum, volume estimation vs. contour level, and model atom inclusion at the recommended contour level. These visualization tools have been designed to help non-experts and experts alike to gain insight into the content and assess the quality of 3DEM structures in EMDB and PDB without the need to install specialized software or to download large amounts of data from the structural data archives.

Maps and associated model files may also be downloaded for local analysis via links on individual entry pages. EMDB maps can be viewed along with associated models using locally installed software such as UCSF Chimera (https://www.cgl.ucsf.edu/chimera/; (Pettersen et al., 2004)), Pymol (www.pymol.org), VMD (http://www.ks.uiuc.edu/Research/vmd/; (Humphrey et al., 1996)), and Coot (http://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/; (Emsley et al., 2010)), enabling investigation with an extensive set of tools.

2.2. Validation

Visual analysis pages—The functionality of the visual analysis pages has been described above and they provide a basic simple first check of the entry based on the information available. They are available from links of the form: http://pdbe.org/emd-####/analysis, e.g., http://pdbe.org/emd-2852/analysis for EMDB entry EMD-2852.

Stand-alone validation servers—Two validation servers have been developed at PDBe for eventual integration into the 3DEM validation pipeline.

Fourier-Shell correlation (FSC; (Harauz and van Heel, 1986)) is the most commonly reported method for estimating the resolution of single-particle maps. However, the estimated resolution depends critically on the threshold criteria used, and several different conventions are followed. In order to simplify deposition of FSC curve data to EMDB, a web service for calculating FSC curves has been developed, community-tested, and placed into production (http://pdbe.org/fsc). A user can upload two independent maps, receive the calculated FSC curve in a standardized format for deposition into EMDB, and view and download a plot of the curve. Several reconstruction packages also produce FSC files suitable for direct upload to EMDB including EMAN2 (http://blake.bcm.edu/emanwiki/EMAN2; (Tang et al., 2007)), RELION (http://www2.mrc-lmb.cam.ac.uk/relion/index.php/Main_Page; (Scheres, 2012)) and Bsoft (http://lsbr.niams.nih.gov/bsoft/; (Heymann and Belnap, 2007)). More than 120 map entries in EMDB now include deposited FSC curves.

Tilt-pair analysis (Henderson et al., 2011; Rosenthal and Henderson, 2003) is a useful method for validating the hand and overall shape of a map, particularly for lower-resolution maps in which secondary structure features are absent. A tilt-pair validation server developed by the Rosenthal group (Wasilewski and Rosenthal, 2014) has made the method generally accessible. This server has now been migrated to PDBe and is available for public use (http://pdbe.org/tiltpair).

Validation pipeline—An initial EM validation report has been developed for use in the wwPDB Deposition & Annotation system (see section 3.1). The format closely follows the validation reports produced by wwPDB for structures from X-ray crystallography (Read et al., 2011) and NMR (Montelione et al., 2013), and is based on the same underlying validation software pipeline (Gore et al., 2012). We have initially focused on providing map-independent assessments of model quality. Model assessments include standard geometry (bonds, angles, and torsion angles), close contacts, protein and nucleic acid backbone geometry, and ligand geometry. A slider graphic compares the quality of the given structure, for key indicators, to all EM structures in the PDB archive, as well as all structures in the PDB archive.

Recognizing that nearly one quarter of all 3DEM models in the PDB contain polymers represented as atom traces (Cα-atom only for protein chains; P-atom only for nucleic acid chains), we are actively investigating new assessments for trace atom model geometry. Consecutive Cα-Cα distances are reported as outliers if they fall outside of ±3σ limits for *cis* and *trans* peptide distributions; consecutive P-P distances are reported as outliers if they are shorter than 4.4 Å or longer than 8.0 Å.

The EM validation report also provides a table of basic information about the map, e.g., the reconstruction method, reported resolution, resolution method, imposed symmetry, number of images used, microscope, imaging parameters and detector. Future report versions will, with guidance from the EM Validation Task Force (Henderson et al., 2012 and see below), add validation components for the map as well as for the fit of the model to the map, as the relevant methods and metrics evolve and become accepted community standards. We will encourage depositors to include the validation report when submitting manuscripts for review, and encourage journal editors and reviewers to request these reports.

2.3. Programmatic Access

A REST API providing easy access to EMDB meta-data and PDB data in JSON format is available from http://pdbe.org/api. A web-service based on the SOAP protocol for accessing EMDB meta-data is available from http://emdatabank.org/webservice.html.

3. Deposition and Annotation

3.1. EMDB and PDB

The wwPDB partners and the EMDataBank project recently launched a new Deposition & Annotation system that supports structures determined using 3DEM, NMR, and X-ray, neutron and electron crystallography. New entries can be submitted at http://deposit.wwpdb.org/deposition/. Depositors will be able to complete map-only (EMDB) and combined map+model (EMDB+PDB) submissions, providing information tailored to the particular 3DEM method selected (single particle, helical, subtomogram average, tomography, or electron crystallography). The new system produces an EM-specific validation report and features a revised and expanded metadata dictionary for 3DEM experiments (Patwardhan et al., 2012). For example, hierarchical sample description is implemented in way that can be tied to map segmentations (Patwardhan et al., 2014), and extensions for each 3DEM method have been added, following community-based recommendations (Henderson et al., 2012; Patwardhan et al., 2014). The new system also supports a rich set of auxiliary data including half maps used for validation, masks and FSC curves. Legacy systems for EM deposition (EMDEP for EMDB deposition and EM-ADIT (RCSB) or AUTODEP (PDBe)) will be kept running for a transitional phase, particularly to allow on-going depositions to complete. The shutdown of these systems will be announced well ahead of time on relevant forums in order to give the EM community time to prepare for the changes.

3.2. EMPIAR

A web-based deposition system for EMPIAR can be accessed via the “Deposition” tab at the top of the EMPIAR home page (pdbe.org/empiar). This is a user-based system that allows users to manage multiple depositions and to share access to depositions among multiple users. For EM experiments that require mandatory deposition of the final 3D reconstruction to EMDB, we require that an EMDB deposition is associated with the EMPIAR deposition. It is however also possible to deposit data relating to 3D SEM, and soft X-ray tomography. For these imaging modalities the requirement is that the data is associated with a journal publication. As with the EMDB depositions, once a deposition is submitted, the entry is

curated and released as per the instructions of the depositor. In contrast to EMDB, EMPIAR entries do not follow a weekly release process, but are released when they are ready. It should be noted though that this process is not instantaneous and multi terabyte datasets may take days to release from when the instruction is received.

4. Recent community wide initiatives

The structural archives serve a greater role than as mere data repositories. Reuse of data makes apparent issues related to data and meta-data formats, data storage and transfer, integration of data with other forms of structural data and other bioinformatics data, and data validation. The organizations, and partners involved in the running of the EMDB, PDB and EMPIAR archives play key roles as facilitators in helping bring about consensus and agreement on a range of issues to the wider benefit of the structural community. Here we provide an overview of some of the key initiatives and workshops that have helped move the field forward in recent years.

4.1. EM Validation task force

Assessment of structural data crucially requires community-accepted validation criteria (Montelione et al., 2013; Read et al., 2011). However, methods for validation of 3DEM structures are still in early development, and are applied inconsistently (Glaeser, 2013; Henderson, 2013; Subramaniam, 2013; van Heel, 2013). EMDataBank has been working with the 3DEM community to establish data validation methods that can be used in the structure determination process, to define key indicators of a well-determined structure that should accompany every structure deposition, and to implement appropriate validation procedures into a 3DEM validation pipeline.

In 2010 the EM Validation Task Force (EM VTF) was established. The international group of ~30 3DEM experts explored how to assess maps, models, and other data that are deposited into the EM Data Bank (EMDB) and Protein Data Bank (PDB) public data archives. Overall the need for more research and development of validation criteria for maps and map-derived models was strongly articulated (Henderson et al., 2012). ***For deposited maps***, the EM VTF recognized a critical need to develop standards for assessing map resolution and accuracy, and recommended reporting of map resolution in accordance with visible features, deposition of annotations specific to each map type, and validation of map symmetry. ***For deposited map-derived models***, EM VTF recommendations included establishment of criteria for assessing models both with and without regard to the fit to the map, creation of community-wide benchmarks for modeling methods, sequence annotation of all map components, and capability to archive coarse-grained representations of models. ***Additional recommendations*** included establishing deposition guidelines for publication of 3DEM structures in journals, and expanding the role of EMDataBank to work together with the 3DEM community to provide unified access to 3DEM structures and to facilitate development of validation and data standards. The EM VTF also recommended providing full FSC curves with deposited maps, indicating whether or not maps used for FSC calculation are fully independent, establishing benchmark datasets for maps and models, and providing multiple types of assessments for models. Participants at the 2011 “Data

management challenges in 3D electron microscopy" workshop (Patwardhan et al., 2012) reiterated the EM VTF's advice, and provided further recommendations regarding development of data models and validation-related services.

EMDataBank is following up on these recommendations with research efforts, community-wide challenges, and validation pipeline development.

4.2. "Data management challenges in 3D electron microscopy" workshop

The aim of this expert workshop held in December 2011 was to discuss the growing challenges of storing, sharing, transferring, analyzing, viewing, validating and annotating 3DEM data. The outcomes of the meeting included an acknowledgement for the need to set up an archive for raw image data relating to EMDB entries as vital for validation, laying the foundations for the initial ideas that would eventually result in the establishment of EMPIAR (Patwardhan et al., 2012). Overall the participants endorsed the vital role that EMDataBank and its partners could play in improving reporting standards for validation in 3DEM, and providing validation tools for 3DEM data. The participants also felt that it would be desirable to have community wide efforts on the subject of format standardization and harmonization. Subsequent to this meeting there has been an initiative to try and clarify and standardize the definition of the MRC map format (Cheng et al., 2015) and to develop an EM exchange format that could be used to represent EM related metadata such as particle coordinates and coordinate transformations (Marabini et al., 2016). Meeting participants also recognized a problem with the deposition rates for electron tomography lagging behind that of single-particle EM. It was agreed that there were several reasons for this, partly having to do with the cellular community being more closely related to the cellular imaging community than to the structural community where the concept of deposition had been long since established. Also the EMDB data model did not adequately capture meta-data for cellular EM, nor was integration adequately promoted as the support for segmentation data was rudimentary. In order to improve deposition of electron tomography EMDataBank highlighted the issue on the 3DEM bulletin board and at meetings and managed to garner agreement on the mandatory deposition of sub-tomogram averages, and a "strong recommendation" for the deposition of at least one representative tomogram from every electron tomography publication.

4.3. "A 3D cellular context for the macromolecular world" workshop

This expert workshop (Patwardhan et al., 2014) expanded on the discussions on archiving in cellular EM initiated at the "Data management challenges in 3D electron microscopy" workshop. A key question discussed was whether archiving needed to be expanded to support other imaging modalities to adequately provide a cellular context, and how to integrate structural data at different scales of imaging. There was a further endorsement for the development of an archive for raw image data with an eye to extend this to be able to easily accommodate new modalities of cellular imaging data. The participants recognized that the scope for this could be very wide and that it would make sense to focus on a few but important modalities, and the ones suggested were 3D SEM, soft X-ray tomography and correlative light and electron microscopy. In terms of data integration the participants recognized the need to capture segmentation data in EMDB and to ensure that segmentations

were annotated with biological terms from relevant biological ontologies and bioinformatics databases, and that tools and formats were created that facilitated this annotation and subsequent deposition. In 2014 PDBe received funding from the MRC/BBSRC for the very purpose of working on these outcomes from the meeting – the development of EMPIAR, the development of integrated web-based visualisation of molecular and cellular structural data, and the development of a web-based tool to facilitate the biological annotation of segmentations. A slight digression but nevertheless of importance for the field was a discussion on the need for archive movies relating to EMDB and PDB entries usually included in the supplementary materials for journal publications. The motivation was the lack of consistency in annotation and presentation between journals and questions about long-term sustainability – there was anecdotal evidence that many movies included as supplementary information were no longer available after some period of time. More recently PDBe (work by Vladislav Lysenkov) has developed a prototype movie archive, which subject support from the EM community and publishers and additional funding will be further developed into a full fledged movie archive for EMDB and PDB entries.

4.4. wwPDB Hybrid/Integrative Methods Task Force workshop

This workshop was convened to bring together experts in the field to discuss steps to enable better support for the archiving of hybrid methods experiments in the public domain (Sali et al., 2015). There are several structural biological problems that cannot be tackled by a single structural technique alone and require the close integration of information from both established structural techniques and other sources such small-angle X-ray scattering (SAXS), fluorescence microscopy and mass spectrometry. The information that can currently be deposited to public databases is often insufficient to provide a holistic view of hybrid methods experiments.

The participants agreed that all relevant experimental data should be archived, but it would be up to the expert communities to drive decisions on what should be archived and how. A flexible model representation needed to be developed to accommodate models at different scales – as integrative modelling would often yield coarse grained non-atomistic models (showing for example the positioning of domains rather than domain detail) and to accommodate multi-state models and model ensembles. Additionally methods for estimating the uncertainty needed to be established. Finally, due to the wide array of methods, user communities and funding sources involved a single archive model was deemed unrealistic and instead the approach would be to create a federated system of model and data archives.

4.5. EMDataBank map and model validation challenges

EMDataBank is sponsoring two new community challenges in 2015–2016 to raise awareness of the need for structure validation as a routine part of 3DEM studies and publications. Additional goals are to develop suitable sets of benchmark data, establish best practices, evolve criteria for validation, and compare and contrast different 3DEM methodologies. The new challenges have been formulated by committees composed of 3DEM community members, with benchmark targets of varying size and complexity selected from recently deposited 3DEM structures based on current state-of-the-art detectors and processing methods, in the resolution range 2.2–4.5 Å. The new challenges follow in the

positive spirit of previous community-based challenge activities for particle picking (Zhu et al., 2004), modeling (Ludtke et al., 2012), and CTF correction (Ludtke et al., 2012; Marabini et al., 2015; Zhu et al., 2004). We anticipate that the community-developed benchmarks will prove useful for methods evaluation, even beyond these challenge activities (Editorial, 2015).

For the map challenge, participants have been asked to create and submit reconstructions using supplied image data. The map challenge data, which totals 12 TB for seven benchmark targets and includes raw movie frame images, have been provided by the original authors of the selected targets, and are stored in EMPIAR. For the model challenge, participants have been asked to create and submit atomic coordinate models using supplied maps. Following a key recommendation of cryoEM specialists and modelling software developers at a planning workshop organized in June 2015, each benchmark target is an unfiltered, unsharpened map. Half-maps used for FSC curve calculation are also available for participants to try out various refinement and validation strategies. Maps for the eight targets have been provided by the original authors of the target structures, and are stored as supplemental files associated with EMDB entries.

Each challenge has challenger submission and results assessment phases. Follow-up discussions via participant workshops are planned, as well as dissemination of results in journal special issues. Anyone from the scientific community is welcome to participate as a challenger and/or assessor. Both challenges are hosted at http://challenges.emdatabank.org/.

5. Challenges and opportunities

5.1. Rise of multi-user facilities, CCP-EM and prospects for data harvesting

The rising costs of purchasing state-of–the-art cryoEM microscopy systems and maintaining the supporting infrastructure are putting them beyond the reach of many individual institutions. A growing trend is therefore for a more coordinated approach often involving regional or national collaborations between multiple institutions to set up multi-user facilities similar to the beam-lines at synchrotron facilities. Examples include Necen in the Netherlands (http://www.necen.nl/), the electron Bio-Imaging Centre (eBIC; (Saibil et al., 2015)) at the Diamond Light Source in the UK, the National Resource for Automated Molecular Microscopy in New York (NRAMM; http://nramm.nysbc.org/) and the National Center for Macromolecular Imaging in Houston (NCMI; http://ncmi.bcm.edu/ncmi/). A number of issues need to be considered to maximize the efficiency and throughput of these centres. For instance, the availability of lower-end microscopes that can be used for screening and fine-tuning to maximise the chance of getting high quality datasets when time is finally allocated. Automation of the imaging session using software such as EPU (https://www.fei.com/software/epu/), Leginon (Suloway et al., 2005) and SerialEM (Mastronarde, 2005) is essential to maximise the amount of data that can be collected. Data management and image processing can also pose major challenges. However the development of coherent software pipelines also gives rise to the opportunity for harvesting data directly to EMDB and EMPIAR, which could greatly facilitate the deposition process. In the context of the EMDataBank project we have previously demonstrated the feasibility of harvesting a partially populated meta-data XML file following the EMDB XML schema and populating

relevant form fields. The Collaborative Computational Project for Electron cryo-Microscopy (CCP-EM; (Wood et al., 2015)) was established following the model in the UK for Collaborative Computational Projects (e.g., CCP4 for macromolecular crystallography) for providing long-term support to scientific areas that require significant computation. The aims are to build a coherent cryoEM community, supporting users of cryoEM software, and supporting developers in producing and distributing software. Funding for the CCP-EM project was recently renewed with an additional emphasis on helping develop software pipelines and infrastructure for eBIC. In this context the prospect of direct harvesting of data from eBIC to EMPIAR and EMDB will be considered with PDBe, and any developments are likely to benefit other multi-user facilities as well.

5.2. New imaging modalities

With EMPIAR now accepting data from 3D SEM and soft X-ray tomography experiments it may be asked whether public archiving can and should be extended to other imaging modalities, including fluorescence microscopy, to accommodate for the rapidly changing landscape of cellular structural biology? The approach to archiving followed by EMDB and EMPIAR remains fairly traditional with centralised archiving, well-structured data, disciplined practices and transparent processes. The strengths of such a system are its robustness, high availability of data and ease of access. The high coherency and consistency in describing the data simplifies reuse and integration. However with centralised archiving costs can increase prohibitively for the archive provider if data volumes expand too quickly. Furthermore in fields such as super-resolution microscopy where there are a multitude of variations in how experiments are conducted, it may be difficult to abstract a coherent set of meta-data beyond a very basal level. On the opposite end of the spectrum is the prospect of distributed archiving, and unstructured data. From our perspective there are no “right” solutions and the relative merits of these solutions need to be considered on a case-by-case basis and may change over time. On an even more fundamental level a question that needs to be posed is what purpose public archiving of this data will serve? On the molecular end of structural biology it has largely been the community itself that has driven the need for public access of data for the purposes of reproducibility, validation, re-use of data and data integration. Even here some compromises have had to be struck, for example, it has been deemed impractical to archive raw X-ray imaging data in PDB and instead where possible links are maintained to the local sites where this data is stored. Similarly these considerations need to be made very carefully for cellular imaging data in order to arrive at solutions that are viable and sustainable.

Acknowledgments

We thank the many current and past colleagues who have made significant contributions to the development of data archiving for 3DEM methods. EMDataBank Unified Data Resource is funded by National Institutes of Health GM079429 to Baylor College of Medicine (Wah Chiu, PI), Rutgers University (Helen Berman, co-PI), and EMBL-EBI (Gerard Kleywegt, co-PI). Work on EMDB and EMPIAR at EMBL-EBI is also supported by the UK Medical Research Council with co-funding from the UK Biotechnology and Biological Sciences Research Council (BBSRC; grant MR/L007835), the BBSRC (grant BB/M018423/1), the Wellcome Trust (grants 088944 and 104948), and the European Commission Framework 7 Programme (grant 284209).

References

Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, Sali A, Rout MP. The molecular architecture of the nuclear pore complex. Nature. 2007; 450:695–701. [PubMed: 18046406]

Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol. 2003; 10:980. [PubMed: 14634627]

Berman HM, Kleywegt GJ, Nakamura H, Markley JL. The Protein Data Bank at 40: reflecting on the past to prepare for the future. Structure. 2012; 20:391–396. [PubMed: 22404998]

Brandt F, Carlson LA, Hartl FU, Baumeister W, Grunewald K. The three-dimensional organization of polyribosomes in intact human cells. Mol Cell. 2010; 39:560–569. [PubMed: 20797628]

Cheng A, Henderson R, Mastronarde D, Ludtke SJ, Schoenmakers RH, Short J, Marabini R, Dallakyan S, Agard D, Winn M. MRC2014: Extensions to the MRC format header for electron cryo-microscopy and tomography. J Struct Biol. 2015; 192:146–150. [PubMed: 25882513]

Danev R, Baumeister W. Cryo-EM single particle analysis with the Volta phase plate. Elife. 2016:5.

The difficulty of a fair comparison. Nat Methods. 2015; 12:273. Editorial. [PubMed: 26005724]

Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010; 66:486–501. [PubMed: 20383002]

Fitzpatrick AW, Debelouchina GT, Bayro MJ, Clare DK, Caporini MA, Bajaj VS, Jaroniec CP, Wang L, Ladizhansky V, Muller SA, MacPhee CE, Waudby CA, Mott HR, De Simone A, Knowles TP, Saibil HR, Vendruscolo M, Orlova EV, Griffin RG, Dobson CM. Atomic structure and hierarchical assembly of a cross-beta amyloid fibril. Proc Natl Acad Sci U S A. 2013; 110:5468–5473. [PubMed: 23513222]

Glaeser RM. Replication and validation of cryo-EM structures. J Struct Biol. 2013; 184:379–380. [PubMed: 24036314]

Gonen T, Cheng Y, Sliz P, Hiroaki Y, Fujiyoshi Y, Harrison SC, Walz T. Lipid-protein interactions in double-layered two-dimensional AQP0 crystals. Nature. 2005; 438:633–638. [PubMed: 16319884]

Gore S, Velankar S, Kleywegt GJ. Implementing an X-ray validation pipeline for the Protein Data Bank. Acta Crystallogr D Biol Crystallogr. 2012; 68:478–483. [PubMed: 22505268]

Gutmanas A, Alhroub Y, Battle GM, Berrisford JM, Bochet E, Conroy MJ, Dana JM, Fernandez Montecelo MA, van Ginkel G, Gore SP, Haslam P, Hatherley R, Hendrickx PM, Hirshberg M, Lagerstedt I, Mir S, Mukhopadhyay A, Oldfield TJ, Patwardhan A, Rinaldi L, Sahni G, Sanz-Garcia E, Sen S, Slowley RA, Velankar S, Wainwright ME, Kleywegt GJ. PDBe: Protein Data Bank in Europe. Nucleic Acids Res. 2014; 42:D285–291. [PubMed: 24288376]

Harauz G, van Heel M. Exact filters for general geometry three dimensional reconstruction. Optik. 1986; 73:146–156.

Henderson R. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. Proc Natl Acad Sci U S A. 2013; 110:18037–18041. [PubMed: 24106306]

Henderson R, Chen S, Chen JZ, Grigorieff N, Passmore LA, Ciccarelli L, Rubinstein JL, Crowther RA, Stewart PL, Rosenthal PB. Tilt-pair analysis of images from a range of different specimens in single-particle electron cryomicroscopy. J Mol Biol. 2011; 413:1028–1046. [PubMed: 21939668]

Henderson R, Sali A, Baker ML, Carragher B, Devkota B, Downing KH, Egelman EH, Feng Z, Frank J, Grigorieff N, Jiang W, Ludtke SJ, Medalia O, Penczek PA, Rosenthal PB, Rossmann MG, Schmid MF, Schroder GF, Steven AC, Stokes DL, Westbrook JD, Wriggers W, Yang H, Young J, Berman HM, Chiu W, Kleywegt GJ, Lawson CL. Outcome of the first electron microscopy validation task force meeting. Structure. 2012; 20:205–214. [PubMed: 22325770]

Heymann JB, Belnap DM. Bsoft: image processing and molecular modeling for electron microscopy. J Struct Biol. 2007; 157:3–18. [PubMed: 17011211]

Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph. 1996; 14:33–38. 27–38. [PubMed: 8744570]

Iudin A, Korir PK, Salavert-Torres J, Kleywegt GJ, Patwardhan A. EMPIAR: A public archive for raw electron microscopy image data. Nat Methods. 2016 Advanced online publication.

Kukulski W, Schorb M, Kaksonen M, Briggs JA. Plasma membrane reshaping during endocytosis is revealed by time-resolved electron tomography. Cell. 2012; 150:508–520. [PubMed: 22863005]

Lagerstedt I, Moore WJ, Patwardhan A, Sanz-Garcia E, Best C, Swedlow JR, Kleywegt GJ. Web-based visualisation and analysis of 3D electron-microscopy data from EMDB and PDB. J Struct Biol. 2013; 184:173–181. [PubMed: 24113529]

Lawson CL, Patwardhan A, Baker ML, Hryc C, Garcia ES, Hudson BP, Lagerstedt I, Ludtke SJ, Pintilie G, Sala R, Westbrook JD, Berman HM, Kleywegt GJ, Chiu W. EMDataBank unified data resource for 3DEM. Nucleic Acids Res. 2016; 44:D396–403. [PubMed: 26578576]

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 Res. 2011; 39:D456–464. [PubMed: 20935055]

Ludtke SJ, Lawson CL, Kleywegt GJ, Berman H, Chiu W. The 2010 cryo-EM modeling challenge. Biopolymers. 2012; 97:651–654. [PubMed: 22696402]

Marabini R, Ludtke SJ, Murray SC, Chiu W, de la Rosa-Trevin JM, Patwardhan A, Heymann JB, Carazo JM. The Electron Microscopy eXchange (EMX) initiative. J Struct Biol. 2016; 194:156–163. [PubMed: 26873784]

Marabini R, Carragher B, Chen S, Chen J, Cheng A, Downing KH, Frank J, Grassucci RA, Bernard Heymann J, Jiang W, Jonic S, Liao HY, Ludtke SJ, Patwari S, Piotrowski AL, Quintana A, Sorzano CO, Stahlberg H, Vargas J, Voss NR, Chiu W, Carazo JM. CTF Challenge: Result summary. J Struct Biol. 2015; 190:348–359. [PubMed: 25913484]

Mastronarde DN. Automated electron microscope tomography using robust prediction of specimen movements. J Struct Biol. 2005; 152:36–51. [PubMed: 16182563]

Montelione GT, Nilges M, Bax A, Guntert P, Herrmann T, Richardson JS, Schwieters CD, Vranken WF, Vuister GW, Wishart DS, Berman HM, Kleywegt GJ, Markley JL. Recommendations of the wwPDB NMR Validation Task Force. Structure. 2013; 21:1563–1570. [PubMed: 24010715]

Nogales E, Wolf SG, Downing KH. Structure of the alpha beta tubulin dimer by electron crystallography. Nature. 1998; 391:199–203. [PubMed: 9428769]

Patwardhan A, Carazo JM, Carragher B, Henderson R, Heymann JB, Hill E, Jensen GJ, Lagerstedt I, Lawson CL, Ludtke SJ, Mastronarde D, Moore WJ, Roseman A, Rosenthal P, Sorzano CO, Sanz-Garcia E, Scheres SH, Subramaniam S, Westbrook J, Winn M, Swedlow JR, Kleywegt GJ. Data management challenges in three-dimensional EM. Nat Struct Mol Biol. 2012; 19:1203–1207. [PubMed: 23211764]

Patwardhan A, Ashton A, Brandt R, Butcher S, Carzaniga R, Chiu W, Collinson L, Doux P, Duke E, Ellisman MH, Franken E, Grunewald K, Heriche JK, Koster A, Kuhlbrandt W, Lagerstedt I, Larabell C, Lawson CL, Saibil HR, Sanz-Garcia E, Subramaniam S, Verkade P, Swedlow JR, Kleywegt GJ. A 3D cellular context for the macromolecular world. Nat Struct Mol Biol. 2014; 21:841–845. [PubMed: 25289590]

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004; 25:1605–1612. [PubMed: 15264254]

Read RJ, Adams PD, Arendall WB 3rd, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Lutteke T, Otwinowski Z, Perrakis A, Richardson JS, Sheffler WH, Smith JL, Tickle IJ, Vriend G, Zwart PH. A new generation of crystallographic validation tools for the protein data bank. Structure. 2011; 19:1395–1412. [PubMed: 22000512]

Rodriguez JA, Ivanova MI, Sawaya MR, Cascio D, Reyes FE, Shi D, Sangwan S, Guenther EL, Johnson LM, Zhang M, Jiang L, Arbing MA, Nannenga BL, Hattne J, Whitelegge J, Brewster AS, Messerschmidt M, Boutet S, Sauter NK, Gonen T, Eisenberg DS. Structure of the toxic core of alpha-synuclein from invisible crystals. Nature. 2015; 525:486–490. [PubMed: 26352473]

Rosenthal PB, Henderson R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J Mol Biol. 2003; 333:721–745. [PubMed: 14568533]

Saibil HR, Grunewald K, Stuart DI. A national facility for biological cryo-electron microscopy. Acta Crystallogr D Biol Crystallogr. 2015; 71:127–135. [PubMed: 25615867]

Salavert-Torres J, Iudin A, Lagerstedt I, Sanz-Garcia E, Kleywegt GJ, Patwardhan A. Web-based volume slicer for 3D electron-microscopy data from EMDB. J Struct Biol. 2016

Sali A, Berman HM, Schwede T, Trewhella J, Kleywegt G, Burley SK, Markley J, Nakamura H, Adams P, Bonvin AM, Chiu W, Peraro MD, Di Maio F, Ferrin TE, Grunewald K, Gutmanas A, Henderson R, Hummer G, Iwasaki K, Johnson G, Lawson CL, Meiler J, Marti-Renom MA, Montelione GT, Nilges M, Nussinov R, Patwardhan A, Rappsilber J, Read RJ, Saibil H, Schroder GF, Schwieters CD, Seidel CA, Svergun D, Topf M, Ulrich EL, Velankar S, Westbrook JD. Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. Structure. 2015; 23:1156–1167. [PubMed: 26095030]

Scheres SH. RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol. 2012; 180:519–530. [PubMed: 23000701]

Subramaniam S. Structure of trimeric HIV-1 envelope glycoproteins. Proc Natl Acad Sci U S A. 2013; 110:E4172–4174. [PubMed: 24106302]

Suloway C, Pulokas J, Fellmann D, Cheng A, Guerra F, Quispe J, Stagg S, Potter CS, Carragher B. Automated molecular microscopy: the new Leginon system. J Struct Biol. 2005; 151:41–60. [PubMed: 15890530]

Tagari M, Newman R, Chagoyen M, Carazo JM, Henrick K. New electron microscopy database and deposition system. Trends Biochem Sci. 2002; 27:589. [PubMed: 12417136]

Tang G, Peng L, Baldwin PR, Mann DS, Jiang W, Rees I, Ludtke SJ. EMAN2: an extensible image processing suite for electron microscopy. J Struct Biol. 2007; 157:38–46. [PubMed: 16859925]

van Heel M. Finding trimeric HIV-1 envelope glycoproteins in random noise. Proc Natl Acad Sci U S A. 2013; 110:E4175–4177. [PubMed: 24106301]

Wasilewski S, Rosenthal PB. Web server for tilt-pair validation of single particle maps from electron cryomicroscopy. J Struct Biol. 2014; 186:122–131. [PubMed: 24582855]

Wood C, Burnley T, Patwardhan A, Scheres S, Topf M, Roseman A, Winn M. Collaborative computational project for electron cryo-microscopy. Acta Crystallogr D Biol Crystallogr. 2015; 71:123–126. [PubMed: 25615866]

Zeev-Ben-Mordehai T, Hagen C, Grunewald K. A cool hybrid approach to the herpesvirus ‘life’ cycle. Curr Opin Virol. 2014; 5C:42–49.

Zhu Y, Carragher B, Glaeser RM, Fellmann D, Bajaj C, Bern M, Mouche F, de Haas F, Hall RJ, Kriegman DJ, Ludtke SJ, Mallick SP, Penczek PA, Roseman AM, Sigworth FJ, Volkmann N, Potter CS. Automatic particle selection: results of a comparative study. J Struct Biol. 2004; 145:3–14. [PubMed: 15065668]

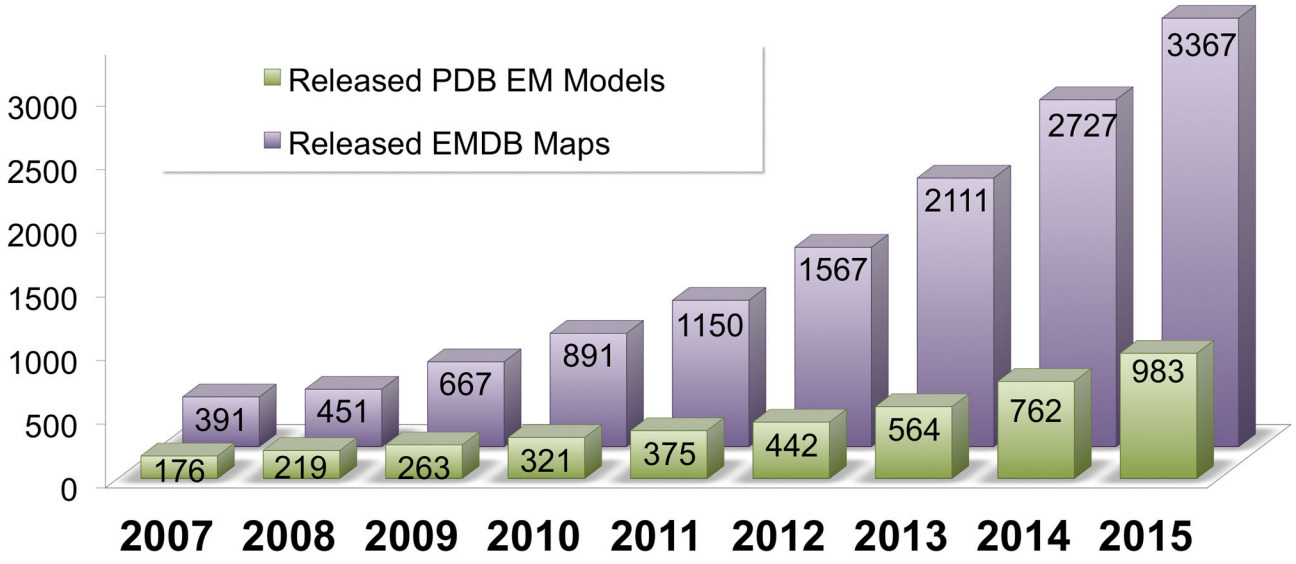


Figure 1.
Released 3DEM entries in EMDB and PDB, cumulative by year, since the start of the EMDataBank Project in 2007.

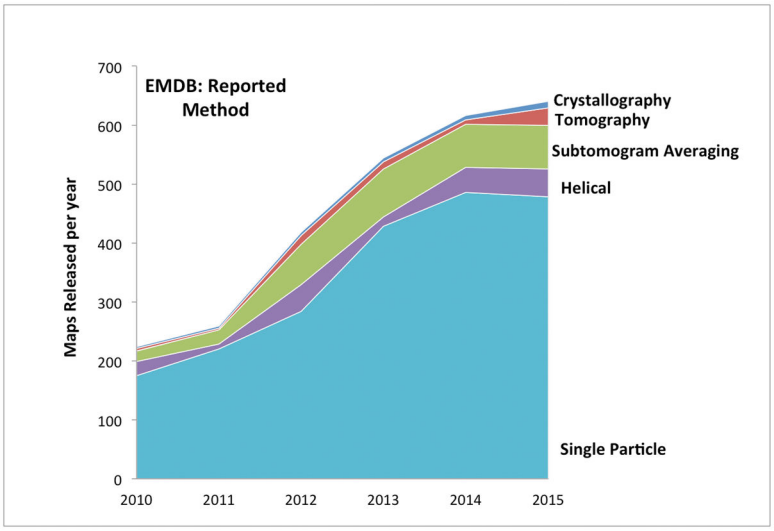


Figure 2.

Trend in reported 3DEM method for EMDB entries released between 2010 and 2015, showing annual releases.

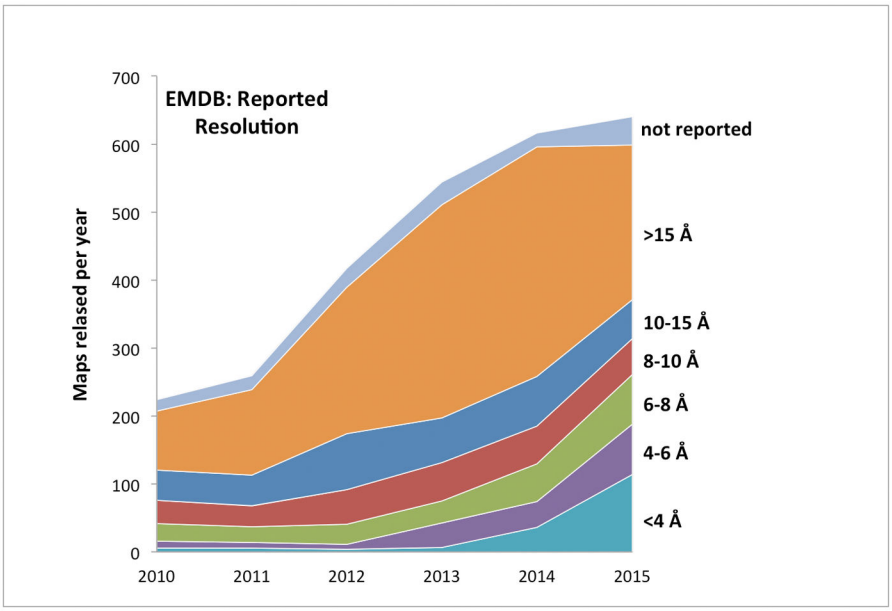
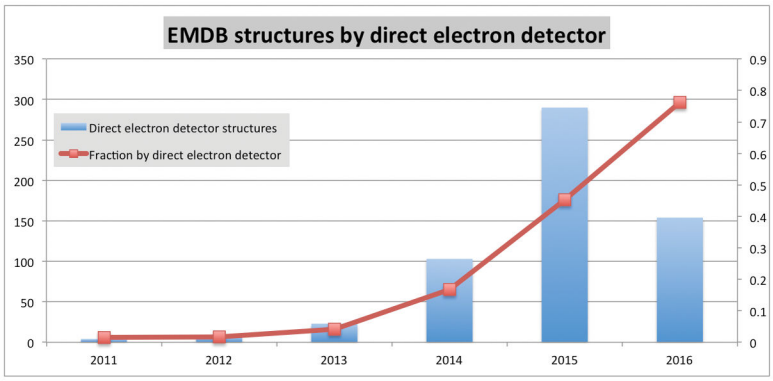


Figure 3.

Trend in reported resolution for EMDB entries released between 2010 and 2015, showing annual releases.



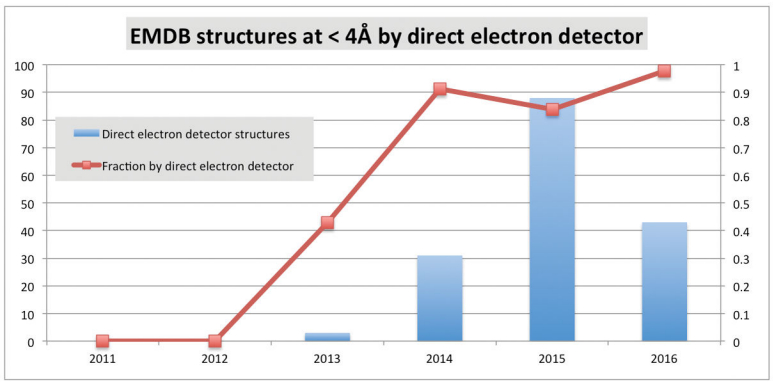


Figure 4.

Trends in released EMDB entries where a direct electron detector was used. The numbers for 2016 are for the period up to 23/3/2016. a) The column chart shows the total number of entries where a direct electron detector was used (axis to the left) and the line chart shows the fraction of all entries (axis to the right). b) These charts are similar to a) except that the resolution is restricted to 4Å or better.