Published in final edited form as:

J Struct Biol. 2016 May ; 194(2): 156–163. doi:10.1016/j.jsb.2016.02.008.

The Electron Microscopy eXchange (EMX) initiative

Roberto Marabini^{a,*}, Steven J. Ludtke^b, Stephen C. Murray^f, Wah Chiu^b, Jose M. de la Rosa-Trevín^c, Ardan Patwardhan^d, J. Bernard Heymann^e, and Jose M. Carazo^c

^aEscuela Politecnica Superior, Universidad Autonoma de Madrid, Campus Universidad Autonoma, 28049 Cantoblanco, Madrid, Spain

^bNational Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, USA

^cBiocomputing Unit, National Center for Biotechnology (CSIC), c/Darwin, 3, Campus Universidad Autónoma, 28049 Cantoblanco, Madrid, Spain

^dEuropean Molecular Biology Laboratory – European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK

^eLaboratory of Structural Biology Research, National Institute of Arthritis, Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD 20892, USA

^fGraduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030 USA

Abstract

Three-dimensional electron microscopy (3DEM) of ice-embedded samples allows the structural analysis of large biological macromolecules close to their native state. Different techniques have been developed during the last forty years to process cryo-electron microscopy (cryo-EM) data. Not surprisingly, success in analysis and interpretation is highly correlated with the continuous development of image processing packages. The field has matured to the point where further progress in data and methods sharing depends on an agreement between the packages on how to describe common image processing tasks. Such standardization will facilitate the use of software as well as seamless collaboration, allowing the sharing of rich information between different platforms. Our aim here is to describe the Electron Microscopy eXchange (EMX) initiative, launched at the 2012 Instruct Image Processing Center Developer Workshop, with the intention of developing a first set of standard conventions for the interchange of information for single-particle analysis (EMX version 1.0). These conventions cover the specification of the metadata for micrograph and particle images, including contrast transfer function (CTF) parameters and particle orientations. EMX v1.0 has already been implemented in the Bsoft, EMAN, Xmipp and Scipion image processing packages. It has been and will be used in the CTF and EMDatabank Validation Challenges respectively. It is also being used in EMPIAR, the Electron Microscopy Pilot Image Archive, which stores raw image data related to the 3DEM reconstructions in EMDb.

*Corresponding author. roberto@cnb.csic.es (R. Marabini).

Keywords

Electron microscopy; Data sharing; Conventions; Standardization

1. Introduction

Three-dimensional electron microscopy (3DEM) has become a very powerful technique for studying the structures of biomolecules without the need for 3D crystallization. Recent advances in detector technology and software algorithms now allow images of unprecedented quality to be recorded and structures to be determined at close to atomic resolution (Kuhlbrandt, Mar 2014; Wang et al., 2014; Bartesaghi et al., Jun 2015; Campbell et al., 2015; Cheng et al., Apr 2015b).

These advances are highly beneficial, but they create the need to generate effective pipelines to handle large amounts of data. An essential part of this effort is the definition of a common set of conventions for image processing that allows an easy exchange of information between different software packages.

The norm in the 3DEM field has been the development of software packages in many different laboratories. This lead to different sets of conventions, consistent within a package, but typically not with other packages. The growth of the Internet has simplified the problems of moving data between computers and acquiring new software packages but, unfortunately, this improved connectivity has had little impact on standardizing the conventions and formats used by software packages. One of the first attempts to address this issue was by Heymann et al. (Aug 2005), who proposed several standardized conventions for many of the parameters that are common to acquisition and processing in cryo-EM. The EMData-Bank project (<http://www.emdatabank.org/>) and in particular the Electron Microscopy Validation Task Force constitutes another ray of hope (Henderson et al., Feb 2012). Under this project several descriptors needed for the deposition of 3DEM reconstructions to EMDB have been introduced. These descriptors provide a brief summary on how a particular 3D map has been obtained, but they are not adequate to describe processing workflows or to allow rich metadata interchange between different software packages.

At present, the lack of a standardized set of conventions has forced several software packages (i.e. EMAN, Sparks, Scipion, etc.) to develop conversion libraries that allow users transparent access to algorithms implemented by other packages. This situation is clearly (a) unstable because packages (including their conventions) evolve with time and (b) only applies to a limited set of algorithms.

Building on top of the conventions already introduced by Heymann et al. (Aug 2005) and efforts by the EMDataBank, the EMX standard not only describes a set of missing descriptors, but provides tools and data to allow conformance testing, and directions and software for extracting data from metadata files, as well as the description of a particular design and implementation for metadata files.

Under the umbrella of the Integrated Structural Biology Infrastructure for Europe (INSTRUCT, <http://www.structuralbiology.eu/>), the first I²PC Developer Workshop took place in Madrid on February 2012. One of its main objectives was to standardize information exchange in single-particle analysis, as well as to start a systematic comparison of the performance of different algorithms used in 3D-EM. The list of participants, which includes a representative of most major single particle software packages, is available in the acknowledgments sections.

One of the main results obtained at the Madrid workshop was the EMX (Electron Microscopy eXchange) initiative, aimed at facilitating the interchange of EM data and metadata among software packages. EMX is only concerned with the minimum amount of information needed to interpret the results of an image processing step unambiguously. Therefore, it is not designed to replace the native formats/conventions/schemes used by the different image processing packages to store metadata information or to guarantee the reproducibility of a particular experiment.

The structure of this document is as follows: first, the need of standards is justified. Then, a formal introduction to EMX is presented, followed by implementation details. Finally, different practical cases in which the use of EMX is beneficial are presented. Two Appendices plus a Supplementary Material A section exhaustively describe the implementation details.

2. The need of standards for data sharing

Software packages need to manipulate information about real-world objects (or, in fact, the concepts and abstractions representing those objects). In most cases, the mapping between the real-world objects and their computer associated data structures is implicit and hidden in the data structure itself. Indeed, as long as these packages do not have to communicate with each other, the mapping information does not have to be explicit. However, if some sort of sharing is envisioned, each package has to be fully understandable by the others. In this case, an explicit description of the mapping of the metadata is required. In this way, and in order to ensure data sharing and information communication, researchers in different scientific disciplines have traditionally developed formal nomenclatures and agreed on conventions for the objects they study. In addition to the development of naming conventions, a great deal of effort has been devoted to the definition of data and metadata standard models, as well as corresponding file formats for data exchange among different laboratories. Some examples include Minimum Information About a Microarray Experiment – MIAME (gene expression data) (Brazma et al., 2001) and its corresponding format MAGE-ML (Smith et al., 2002); HUPO Proteomics Standards initiative (Hucka et al., 2003) and the suggested PSI Molecular Interaction Standard (protein–protein interactions) (Hermjakob et al., 2004) and PSI-MS Format (mass spectrometry); BioPAX (Demir et al., Sep 2010) for pathway information; and Systems Biology markup language (Hucka et al., 2003) for systems biology. These pioneer works were followed by several efforts to coordinate the development of standards, such as ISA-Tab (Rocca-Serra et al., Sep 2010) (describes the workflow of processing biological material with tab delimited files), FuGe (Jones et al., Oct 2007) (models the components of an experimental activity that are common across different-genomic related-technologies),

OBO (Smith et al., Nov 2007) (provides a framework for the developers of life-science ontologies), MIBBI (Taylor et al., Aug 2008) (repository and “harmonizer” of minimum information checklists (MIAME)), etc. Drager and Palsson (2014) (systems biology) and Deutsch et al. (2015) (human proteome) have published recent reviews covering the current state of the growing number of protocols, mark-up languages, graphical descriptions, etc.

In summary, in order to share data openly, it has become increasingly important to agree on what information should be shared and how this information is going to be described.

3. EMX initiative: a formal description

Work on the EMX (Electron Microscopy eXchange) format started during the Instruct Image Processing Center Developers Meeting in Madrid in February 2012. It evolved by means of a series of emails to the whole list of workshop participants, followed later on by focused technical work carried out by a reduced team.

The goal of the EMX format is to allow for an easy and unambiguous exchange of information about objects, such as particles, micrographs and others, among EM image processing packages. The consensus requirements for EMX were:

- **Simple** – The format should be as easy to process and conceptualize as possible.
- **Self-documented** – Understanding the exchange files should not depend on which software created them.
- **Flexible and backward compatible** – Conventions evolve with time, the design should guarantee that newer versions of the standard can inter-operate with programs designed for older versions.

Additionally, it was also agreed that, in most cases, a data exchange transaction would involve two types of files: (1) images (binary files containing pixel/voxel values) and (2) a separate text file containing parametric and descriptive information (metadata). Therefore, an instance of data exchange would consist of a set of homogeneous data, e.g. a set of particle images and a metadata text file referencing the images and additional parameters.

3.1. EMX data model

In order to implement exchange standards, it is vital to provide a framework in which different kinds of information can be represented and exchanged between researchers, that is, to define a data model.

Data models describe a small portion of the real world in terms of objects that can be easily represented by a computer, along with the properties and relationships between these objects. This first version of EMX has been kept simple, following the consensus position among software developers of providing a limited standard to start with, which will become more elaborate upon usage. This first version defines: (1) micrographs (images as obtained in a electron microscope) and (2) particles (each of the views of a specimen that may be identified in a micrograph). Future versions will address additional kinds of objects and

introduce the necessary relationships between them. “Classes” and “3D maps” (volumes) are of particular interest.

The next two subsections give an overview of the micrograph and particle objects. A detailed description of the conventions followed to define geometrical transformations as well as CTFs is deferred to Appendixes A and B, so as not to distract the reader with its mathematical complexities. The full set of available properties/attributes used to describe each object as well as which attributes are optional and which are mandatory is documented as Supplementary Material A.

3.1.1. Description of the object micrograph—A micrograph object contains all the information associated with a specific micrograph (defined as an image from either film, CCD, DDD movie average or DDD frame). This includes an image unique identifier, CTF information, sampling, etc. The image unique identifier is specified by a *fileName* and an *index*. *FileName* refers to the name of the stored file that may contain one or more micrograph images, and *index* refers to the image number within the file. The pair (*fileName*, *index*) must be unique, but only within a single EMX metadata file. There is no uniqueness requirement among multiple files, although it is strongly advised to have descriptive micrograph file names. For stacks of micrographs, each individual image in the file must have the same dimensions. If omitted, the value of index is assumed to be 1.

3.1.2. Description of the object particle—Particle objects contain the information associated with individual sub-images (boxes) extracted from a micrograph. This may include information such as the projection orientation and shift of the particle, given as a transformation matrix. How this transformation is understood is described in detail in Appendix A.

Note that in this first version of the standard, particle refers only to 2D images extracted from micrographs. and not to 3D images (volumes) as would be the case in tomography. The standard will be extended later to incorporate 3D volumes. Particles are specified by a *fileName* and an *index*, which follow the same rules as the *fileName* and *index* for micrographs. An object Particle is related to a single object Micrograph, while an object Micrograph may be related to several Particle objects.

4. EMX initiative: implementation

This section is divided in two main parts. The first one focuses on format issues, while the second one describes the EMX web site designed to facilitate the testing of EMX compliant software and comment on libraries and schemas developed by the authors that may be reused to implement tools and utilities related with the EMX initiative.

4.1. Format

For the first version of the exchange format, the CCP4/MRC file format has been selected for interchanging images as floating point values (pixels stored as real numbers) (Cheng, 2015a). All metadata information should be in an XML file as outlined below. If conflicting data information is stored in the CCP4/MRC header, the information in the XML file takes

precedence. In future versions, a new design in HDF may be considered for both metadata and images, allowing the flexibility of using single or multiple files for exchange.

Although the metadata files described in this article could be implemented using several formats (tab delimited text files, STAR files, JSON files, etc.) XML was finally selected for two main reasons: (1) EMX will, in the future, require data structures more complex than those provided by a simple “flat” row structure file can. (2) XML documents can be easily validated, ensuring that they conform to the specified format.

4.1.1. XML—Extensible Markup Language or XML is a tag oriented language in which tags denote the beginning and end of each element (<http://www.w3.org/XML/>). Specifically, EMX metadata files contain objects at the top level, such as Micrograph or Particle, and each object contains a set of metadata elements (or attributes). Particle records are not to be embedded inside Micrograph records or vice versa, but are referenced using identifiers.

A formal XML schema for the proposed model can be found at URL <http://i2pc.cnb.csic.es/emx/LoadTools.htm?type=Library>. This schema may be used to validate EMX files.

Any future expansions of the original XML schema would entail incrementing the version number of the EMX format, so that no confusion about compatibility of specific EMX files with specific software packages arises. An EMX version 1.0 compliant reader must be able to interpret any tags defined for version 1.0 and ignore extra tags, rather than raising errors. Warning messages should be displayed when unknown tags are encountered. Although not required, formatting the documents to promote human readability is encouraged when writing (such as presenting transformation matrices in apparent matrix-form).

The best way to explain how an EMX document is composed is to present one. Listing 1 shows a XML file containing two particle objects which have been extracted from a single micrograph object. The first line of the document is known as the XML declaration. This notifies a processing application about which version of XML is being used. The second line begins the root element, which has been named EMX. The contents of that element include everything between the right angle bracket (>) in < EMX> and the left angle bracket (<) in < /EMX>. The actual syntactic constructs < EMX> and < /EMX> are often referred to as the element start tag and end tag, respectively. Tags mark the boundaries of elements. Elements may include other elements, as well as text. An XML document must contain exactly one root element, that will contain all other content within the document.

The sample EMX document uses elements named micrograph and particle to describe micrographs and particles. Each micrograph/particle element must have at least an attribute named *fileName* or an attribute named *index*. Unlike elements, attributes can only contain textual content. Their values must be surrounded by quotes. The detailed list of elements/attributes that may be contained by the element micrograph/particle is available in Supplementary Material Section 1A.

4.2. Resources

One of the more visible results of the EMX initiative is the development of the EMX web site (<http://i2pc.cnb.csic.es/emx>). The web site is divided into three main parts:

- The *Dictionary and Conventions* section describes the proposed file format used for metadata and data exchange as well as the list of valid XML tags.
- The *Example and Tests* section provides the material needed to test conversion utilities to and from the interchange standard.
- The *Resources* section contains a syntactic validator of the meta-data files and libraries that have been found useful to create the conversion utilities (<http://i2pc.cnb.csic.es/emx/LoadTools.htm?type=Library>).

A video introducing the EMX web site in detail is available at URL <http://i2pc.cnb.csic.es/emx>.

5. Use cases

This section is divided in two parts. The first one includes an analysis, given a typical electron microscopy workflow, of the steps at which metadata exchange is useful for an average user. The second part contains several examples in which the EMX initiative is being used. They range from data archival and development of challenges/competitions in the area of image processing in Structural Biology, to data interchange between software packages.

5.1. Workflow interoperability

Computational tools for 3D-EM, at least in certain areas, are sufficiently mature to allow integration into higher-level workflows composed of several simpler steps. This is, in fact, the path followed by the major packages in the field – Bsoft (Heymann and Belnap, Jan 2007), EMAN (Ludtke et al., 1999), Imagic (van Heel and Keegstra, 1981), Spider (Frank et al., 1996), Xmipp (de la Rosa-Trevin et al., Nov 2013), etc. – as well as by integration platforms such as Appion (Fisher et al., Dec 2011) or Scipion (Quintana et al., Dec 2011). Ideally, different steps may be coming from different packages but, unfortunately, small differences in the conventions followed by each package heavily penalize the users that try to use more than one. EMX has been designed to bridge this gap and bring full interoperability among the different software packages; that is, packages should be able to cooperate on the same data sets without burdening the user with arcane data transformations.

From the very beginning, we understood that for an exchange initiative such as EMX to be useful to the community, it should be supported by several major software packages. EMX is presented here together with an announcement that four packages support the EMX v1.0 standard: Bsoft 1.9, EMAN 2.2, Xmipp 3.2 (requires Scipion 1.0) and Scipion 1.0 (to be released).

In the following, we first define a typical EM workflow for single particle, and then state the points at which we want to guarantee metadata exchange between our packages. Note that we are not trying to define an exhaustive EM workflow in which every possible step is

addressed, but to start from a simple sequence of steps and determine at which points metadata transfer may be of interest.

Typical EM workflow for single particle—A typical workflow starts by importing micrographs and the imaging conditions. If data were obtained as multiple frames using a DDD, preprocessing of the frames may be required to obtain an averaged micrograph after motion and/or radiation compensation. Once imported, the micrographs are screened to eliminate those with excessive astigmatism or drift, both limiting attainable resolution. Micrographs can also be downsampled to accelerate further calculations and remove noise. Particle coordinates are then selected (picked) from micrographs and subsequently used to extract sub-images, also called “boxes”. After extraction, particle images are stored in a gallery (stack). Some operations can be carried out while extracting them, such as filters, contrast inversion and other image manipulations. The extracted images may be subjected to 2D classification to detect possible heterogeneity due to contamination, different conformations or different specimens. Finally, an initial volume (provided by the user or computed by common lines, random conical tilting, etc.) is then refined by a number of different algorithms (e.g. projection matching, maximum likelihood, etc.). This step may be coupled with 3D classification.

Exchanging points

1. **Import raw data (micrographs) to the processing system.** Once data is obtained in the electron microscope, metadata containing a description of the microscope (spherical aberration, voltage, sampling and a first estimation of the defocus) is needed, in addition to the images themselves.
2. **Alignment of direct detector device frames.** Currently, the number of packages offering frame alignment possibilities is relatively small, so that transferring aligned images from one of these packages to another is appealing. In the long term, we may witness the appearance of a plethora of new alignment algorithms that would need to be compared.
3. **Particle picking information (coordinates).** A curated set of particle coordinates is one of the most cherished pieces of information since, in many cases, it requires extensive human intervention and allows raw data re-processing. For users who wish to try different processing workflows in different packages this may be a natural entry point.
4. **Particle alignment information (2D).** We envision few cases in which 2D alignment information obtained in one package needs to be transferred to another. The obvious exception are those cases in which the goal is to compare the final classification/alignment results produced by different algorithms.
5. **Particle projection direction information (3D).** A given set of projection directions obtained by an algorithm may be the seed for further refinements by others. Both 3D classification and alignment information belong to this category.

In summary, the import/export utilities are designed to be able to exchange information between the different software packages in the cases described above without requiring any detailed knowledge of the target or source package.

5.2. Examples in which the EMX has been or is being used

5.2.1. CTF challenge—The CTF estimation challenge goal was to estimate the key parameters of the CTF on a large micrograph data set produced by well-known laboratories under diverse experimental conditions (<http://i2pc.cnb.csic.es/3dembenchmark/LoadSubTasks.htm?subtaskId=2> (Marabini et al., Jun 2015)). EMX was selected as format for uploading a parametric description of the different microscopes used by the data providers to record the micrographs, and the estimates reported by the challenge participants. From the point of view of compliance with the EMX standard, results were very different for data providers, participants and organizers. Only one of the data providers created a file that was fully EMX compliant. In general, they sent the needed information in a collection of formats and conventions that the organizers standardized. On the other hand, files uploaded by the participants went through an automatic validation step and were all compliant. A few inquiries on how to format rejected files were received and indications on how to fix the formatting were provided to the participants. From the organizers' point of view, the use of the EMX conventions (and the libraries and XML schemas described in Section 4.2) greatly facilitated the comparison and analysis of the uploaded CTF estimations.

From EMX point of view, the lesson brought forward in this challenge was a quite obvious one: results uploaded by challengers (in any competition) must be standardized. Otherwise they are meaningless, since the analysis of the data provided by the participants should be blind and there is no opportunity for metadata disambiguation. This is not the case for the data contributing laboratories, whose data sets were carefully analyzed and discussed before launching the challenge.

5.2.2. EMDatabank validation challenges—Starting in 2015 the EM Data Bank (EMDB) is hosting *The 2015 Map Challenge* whose main objective is to establish a benchmark set of single-particle raw image datasets suitable for high resolution cryoEM (see http://challenges.emdatabank.org/?q=2015_map_challenge). In this challenge, metadata information provided to the participants will be encoded using EMX, while the use of EMX by the participant will be encouraged but not mandatory.

5.2.3. EMDB pilot image archive—As highlighted regularly by the community, public access to the raw 2D images used to obtain the 3D reconstructions deposited at EMDB would be extremely useful. In 2014, the Protein Data Bank in Europe (PDBe) launched EMPIAR – the Electron Microscopy Pilot Image Archive to explore issues surrounding the storage and transfer of large datasets and also to provide an immediate solution for the community's need for a public archive of image data related to EMDB reconstructions. PDBe will work with the other EMX partners to, when possible, put in place procedures for translating uploaded meta data files pertaining to particle locations etc. This will benefit the

archive by providing a standard means for distributing this information while simultaneously encouraging the use of the EMX format.

6. Conclusions

A key requirement for data sharing is standardization, that is, agreement on the vocabulary, data structures and formats used to access the shared data. EMDDataBank is a community-accepted portal for deposition and retrieval of 3DEM density maps, atomic models, and associated metadata. Nevertheless, the more ambitious goal of moving data between software packages in order to take advantage of their respective strengths is far from being solved.

Therefore, it is important to agree on standards and provide a semantic repository and test cases that can be used to systematically define the relevant concepts. Failure in doing so is already hampering the usefulness of our data and software tools and it threatens to restrict their future use.

As a final remark, we would like to comment on the strategy followed by EMX design, and on how we see its future in terms of development and maintenance.

In order to create EMX we established, first of all, the need for a standard followed by a specification of the standard requirements. Then, the list of requirements was prioritized and a subset was selected, providing an easier target for the first version. All these steps were carried out during the Madrid Developers Meeting (2012). The initial set of requirements and possible implementations was further discussed through a series of emails. Finally, a panel of experts volunteered to implement and verify a first prototype as well as polish the definitions. By “implement” we refer to: (1) create a set of numerical examples and tests to facilitate software testing, and (2) provide libraries that are able to read EMX. On the other hand, “verification” involves testing the project’s feasibility by implementing importing/exporting routines in their respective software packages and transferring data using them.

Now, we present EMX version 1.0. Future releases of this standard will need to go through similar development cycles involving requirement specification, implementation and validation. In the short term we offer support for using EMX in challenges, infrastructure projects as EMPIAR (<http://pdbe.org/empiar/>), centralized resources (institutions with public access to their microscopes that provide users with not only the raw data but preliminary processing, e.g. CTF estimation) or software projects (such as an automatic EMDB deposition routine). Any inquiries or suggestions regarding EMX may be sent to emx@cnb.csic.es.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors want to thank all the participants in the Instruct Image Processing Center Developers Meeting in Madrid (February 2012) for their valuable comments. These are (in alphabetic order and excluding the authors of

this paper): Vahid Abrishami, Christian Blaschke, Anchi Cheng, Yao Cong, Joachim Frank, Richard Henderson, Robert Langlois, Dean Leith, Stephan Nickell, Joaquin Otón, Pawel A. Penczek, Juergen Plitzko, Adrian Quintana, Carlos O. Sorzano, Michael Schatz, Sjors Scheres, Remco Schoenmakers, Henning Stahlberg, Javier Vargas, Marin van Heel and John Westbrook; as well as those who joined the email discussion list later: Bridget Carragher, Nikolaus Grigorieff, Clint Potter, Holger Stark and Airen Zaldívar.

The authors acknowledge financial support from the Spanish Ministry of Economy and Competitiveness through grants AIC-A-2011-0638, BIO2013-44647-R and BFU2013-41249-P; the Comunidad de Madrid through grant CAM (S2010/BMD-2305); the European Union (EU) and Horizon 2020 through grants: EINFRA-2014-2: EGI Engage (Proposal: 654142) and EINFRA-2015-1: West-Life (Proposal: 675858); and the USA National Institutes of Health through grants: R01GM079429, R01GM080139 and P41GM103832.

This work was partially funded by Instruct, a component of the European Strategy Forum on Research Infrastructures (ESFRI), and supported by national member subscriptions and by the Intramural Research Program of the National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health.

Appendix A. Geometric transformations

Particle objects may include information, such as the projection orientation and shift, encoded as a transformation matrix. There are multiple conventions to express orientations and shifts. Furthermore, there are at least two ways of interpreting them: either the object is transformed, or the object is kept fixed and the coordinate system is transformed. Because of this ambiguity, the exchange of angular information among different image processing suites is a major source of misunderstandings that leads to serious interoperability issues. In this appendix, in order to avoid misinterpretations, the geometrical conventions followed by EMX are described in detail. In addition, two numerical examples are included for testing purposes, so that potential users of EMX conventions can verify and compare their results.

Transformation matrices are normally used to describe the relative spatial relationship of two images (in this context images may be two or three dimensional objects). Before defining them, the coordinate system in which the transformations will be applied needs to be described. EMX follows the recommendations given by Heymann et al. (Aug 2005). In particular, it assumes a right-handed coordinate system and places the origin of coordinates at the *image center*. EMX defines image center as the pixel indexed by $[NdimX/2; NdimY/2; NdimZ/2]$, where $/$ is a rounded down division between integer numbers and $NdimX$, $NdimY$ and $NdimZ$ are the image dimensions. It follows therefore that the center of a 4×4 pixel image shown in Fig. 1(a) is at the center of the pixel labeled with $[2, 2]$. On the other hand, the center of the 3×3 pixel image shown in Fig. 1(b) is at the center of the pixel labeled With $[1, 1]$.

In EMX, transformation matrices are used to store the shifts and rotations needed for computing either 3D reconstructions or alignments. In the following we describe how these transformation matrices relate (1) projections and 3D maps and (2) images and average images. Before describing the geometrical transformation conventions used in EMX we will introduce the mathematical framework in which the transformations are described.

A.1. Mathematical framework

In this appendix we use two different 3D Cartesian systems of coordinates with coordinate axis X' , Y' , Z' and X , Y , Z respectively. The first coordinate system is attached to the imaging device, that is, the Z' axis is always parallel to the projection direction and X' and

Y' axis are parallel to the detector edges (CCD or equivalent device) used by the microscope. The second coordinate system is attached to the object being produced, either a reconstructed 3D map or an averaged image. We will use the notation $f(\mathbf{x})'$ and $f(\mathbf{x})$ to refer to the object f expressed in the systems of coordinates X', Y', Z' and X, Y, Z respectively.

A transformation matrix T is a 4×3 real matrix of the form:

$$T = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{pmatrix} = (R \mathbf{t}) \quad (1)$$

Matrix T comprises translations (given by the vector $\mathbf{t}$) as well as rotations (encoded in the rotation matrix R). In this document the matrix T will be used to transform the coordinates of a point given in the system X, Y, Z to the coordinates of the same point given in the

system X', Y', Z' and vice versa. In this way, if $\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix}$ and $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ are the coordinates of a point d in the coordinate systems X', Y', Z' and X, Y, Z respectively, then there is a transformation matrix T that satisfies:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} = T \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (2)$$

A.2. Projection

EMX follows the convention that a projection image p and a 3D map f are related by the transformation matrix T if they satisfy:

$$p \left(\begin{bmatrix} x' \\ y' \end{bmatrix} \right) = \int f \left(\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} \right) dz' = \int f \left(T \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \right) dz' \quad (3)$$

A.3. Image alignment

Let g and r (either 2 or 3D objects) be two identical images except for a geometrical transformation. Let us assume that we want to apply a transformation to g so the result matches r , that is, we want to align g using r as a reference image. EMX follows the convention that g and r are related by the matrix T if they satisfy:

$$r \left(\begin{bmatrix} x \\ y \\ z \end{bmatrix} \right) = g \left(T \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \right) = g \left(\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} \right) \quad (4)$$

A.4. Examples

The previous subsections of this appendix describe in detail the geometric conventions followed by EMX. In order to minimize the possibility of misunderstanding these conventions, two step-by-step examples with actual numbers are included.

A.4.1. Example 1: projection

Let us assume that we have: (a) a 3D map f which is zero valued in all the space but for a

small blob placed at coordinates $\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \\ 16 \end{bmatrix}$ and (b) the following transformation matrix T

$$T = \begin{pmatrix} 0.71461 & 0.63372 & -0.29620 & 4. \\ -0.61309 & 0.77128 & 0.17101 & 8. \\ 0.33682 & 0.05939 & 0.93969 & 12. \end{pmatrix} \quad (5)$$

The question is: for the given f and T , what projection image p satisfies Eq. (3)? The solution is a zero valued image for all points except for a small region close to the point that

satisfies $\begin{bmatrix} x' \\ y' \end{bmatrix} = HT \begin{bmatrix} 4 \\ 8 \\ 16 \\ 1 \end{bmatrix}$, where $H = \begin{pmatrix} 100 \\ 010 \end{pmatrix}$, that is $\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 7.2 \\ 14.4 \end{bmatrix}$

A.4.2. Example 2: alignment

Let us assume that we have: (a) a 2D image r which is zero valued in all the space but for a

small blob placed at coordinates $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 32 \\ 0 \end{bmatrix}$ and (b) the following transformation matrix T

$$T = \begin{pmatrix} 0.92388 & -0.38268 & 0. & 4. \\ 0.38268 & 0.92388 & 0. & 8. \\ 0. & 0. & 1. & 0. \end{pmatrix} \quad (6)$$

The question is: for the given r and T , what image g satisfies Eq. (4)? The solution is a zero valued image for all points except for a small region close to the point that satisfies

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = T \begin{bmatrix} 32 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{that is,} \quad \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 33.6 \\ 20.2 \\ 0 \end{bmatrix}.$$

Appendix B. CTF description

In the following we introduce the parametric form used to describe the CTF. We note that in EMX there is no CTF envelope term.

$$\begin{aligned} CTF(\mathbf{S}) &= w \cos(\gamma(\mathbf{S})) - \sqrt{1-w^2} \sin(\gamma(\mathbf{S})) \\ \gamma(\mathbf{S}) &= 180 \gamma(-\Delta Z(\angle \mathbf{S}) \|\mathbf{S}\|^2 + \frac{Cs 10^6 \|\mathbf{S}\|^4 \lambda^2}{2}) \text{ in degrees} \\ \Delta Z(\angle \mathbf{S}) &= Z_{avg} + Z_{diff} \cos(2(\angle \mathbf{S} - \Theta)) \text{ in nm.} \\ \Delta Z_{avg} &= \frac{Z_{max} + Z_{min}}{2} \\ \Delta Z_{diff} &= \frac{Z_{max} - Z_{min}}{2} \end{aligned} \quad (7)$$

where $\mathbf{S} = (S_u, S_v)$ is the 2D spatial frequency, Z denotes defocus, Z_{max} is the maximum

defocus value, Z_{min} is the minimum defocus value, $\angle \mathbf{S} = (\tan^{-1} \frac{S_v}{S_u})$ is the angle between the X axis and the vector defined by $\mathbf{S}$, Θ is the astigmatism angle (angle between the direction of maximum defocus and the X axis, see Fig. 2 for details), w is the fraction of amplitude contrast and Cs is the spherical aberration in mm. The factor 10^6 converts Cs from mm to nm and $\gamma(\mathbf{S})$ is termed in specialized literature as the wave aberration function (Frank,

1996). Finally, $\lambda = \frac{h}{\sqrt{2m_0 eV}} \frac{1}{\sqrt{1 + \frac{eV}{2m_0 c^2}}} \sim \frac{1.23}{\sqrt{10^3 V + 0.97845 V^2}} \text{ in nm}$. this last equation m_0 , e , c and V are the electron mass, electron charge, speed of light and microscope potential respectively.

The XML tags and units used to encode the different CTF parameters are:

Parameter	XML tag (units)
V	acceleratingVoltage (kVolt)
w	amplitudeContrast (dimensionless, range [0,1])
C_s	cs (mm)
Z_{max}	defocusU (nm)
Z_{min}	defocusV (nm)
Θ	defocusUAngle (degrees)

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jsb.2016.02.008>.

References

- Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Milne JL, Subramaniam S. Electron microscopy. 2.2 Å resolution cryo-EM structure of Beta-galactosidase in complex with a cell-permeant inhibitor. *Science*. Jun; 2015 348(6239):1147–1151. [PubMed: 25953817]
- Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*. 2001; 29:365–371. [PubMed: 11726920]
- Campbell MG, Veesler D, Cheng A, Potter CS, Carragher B. 2.8 Å resolution reconstruction of the *Thermoplasma acidophilum* 20S proteasome using cryo-electron microscopy. *Elife*. 2015; 4
- 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*. 2015a; 192(2):146–150. [PubMed: 25882513]
- Cheng Y, Grigorieff N, Penczek PA, Walz T. A primer to single-particle cryo-electron microscopy. *Cell*. Apr; 2015b 161(3):438–449. [PubMed: 25910204]
- de la Rosa-Trevin JM, Oton J, Marabini R, Zaldivar A, Vargas J, Carazo JM, Sorzano CO. Xmipp 3.0: an improved software suite for image processing in electron microscopy. *J Struct Biol*. Nov; 2013 184(2):321–328. [PubMed: 24075951]
- Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, Schacherer F, Martinez-Flores I, Hu Z, Jimenez-Jacinto V, Joshi-Tope G, Kandasamy K, Lopez-Fuentes AC, Mi H, Pichler E, Rodchenkov I, Splendiani A, Tkachev S, Zucker J, Gopinath G, Rajasimha H, Ramakrishnan R, Shah I, Syed M, Anwar N, Babur O, Blinov M, Brauner E, Corwin D, Donaldson S, Gibbons F, Goldberg R, Hornbeck P, Luna A, Murray-Rust P, Neumann E, Ruebenacker O, Reubenacker O, Samwald M, van Iersel M, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung KH, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Le Novère N, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD. The BioPAX community standard for pathway data sharing. *Nat Biotechnol*. Sep; 2010 28(9):935–942. [PubMed: 20829833]
- Deutsch EW, Albar JP, Binz PA, Eisenacher M, Jones AR, Mayer G, Omenn GS, Orchard S, Vizcaino JA, Hermjakob H. Development of data representation standards by the human proteome organization proteomics standards initiative. *J Am Med Inform Assoc*. 2015
- Drager A, Palsson B. Improving collaboration by standardization efforts in systems biology. *Front Bioeng Biotechnol*. 2014; 2:61. [PubMed: 25538939]
- Fisher LS, Ward A, Milligan RA, Unwin N, Potter CS, Carragher B. A helical processing pipeline for EM structure determination of membrane proteins. *Methods*. Dec; 2011 55(4):350–362. [PubMed: 21964395]
- Frank, J. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic Press; Burlington: 1996.
- Frank J, Radermacher M, Penczek P, Zhu J, Li Y, Ladjadj M, Leith A. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J Struct Biol*. 1996; 116:190–199. [PubMed: 8742743]
- 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*. Feb; 2012 20(2):205–214. [PubMed: 22325770]
- Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roehert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant S, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe

- B, Hogue C, Apweiler R. The hupo psi's molecular interaction format – a community standard for the representation of protein interaction data. *Nat Biotechnol.* 2004; 22:177–183. [PubMed: 14755292]
- Heymann JB, Belnap DM. Bsoft: image processing and molecular modeling for electron microscopy. *J Struct Biol.* Jan; 2007 157(1):3–18. [PubMed: 17011211]
- Heymann JB, Chagoyen M, Belnap DM. Common conventions for interchange and archiving of three-dimensional electron microscopy information in structural biology. *J Struct Biol.* Aug; 2005 151(2):196–207. [PubMed: 16043364]
- Hucka M, Finney A, Sauro H, Bolouri H, Doyle J, Kitano H, Arkin A, Bornstein B, Bray D, Cornish-Bowden A, Cuellar A, Dronov S, Gilles E, Ginkel M, Gor V, Goryanin I, Hedley W, Hodgman T, Hofmeyr J, Hunter P, Juty N, Kasberger J, Kremling A, Kummer U, Novere N, Loew L, Lucio D, Mendes P, Minch E, Mjolsness E, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff J, Shapiro B, Shimizu T, Spence H, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics.* 2003; 19:524–531. [PubMed: 12611808]
- Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, Degreiff J, Hardy N, Hermjakob H, Hubbard SJ, Hussey P, Igra M, Jenkins H, Julian RK, Laursen K, Oliver SG, Paton NW, Sansone SA, Sarkans U, Stoekert CJ, Taylor CF, Whetzel PL, White JA, Spellman P, Pizarro A. The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nat Biotechnol.* Oct; 2007 25(10):1127–1133. [PubMed: 17921998]
- Kuhlbrandt W. Biochemistry. The resolution revolution. *Science.* Mar; 2014 343(6178):1443–1444. [PubMed: 24675944]
- Ludtke SJ, Baldwin PR, Chiu W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J Struct Biol.* 1999; 128:82–97. [PubMed: 10600563]
- 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.* Jun; 2015 190(3):348–359. [PubMed: 25913484]
- Quintana A, Vargas J, Poza A, MacÃas J, Marabini R, Sorzano C, Carazo JM. Scipion: towards electron microscopy integration. *Methods.* Dec; 2011 55(4):350–362. [PubMed: 21964395]
- Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S, Hide W, Hofmann O, Neumann S, Sterk P, Tong W, Sansone SA. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics.* Sep; 2010 26(18):2354–2356. [PubMed: 20679334]
- Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Rutenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol.* Nov; 2007 25(11):1251–1255. [PubMed: 17989687]
- Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D Jr, S CJ, Brazma A. Design and implementation of microarray gene expression markup language (mage-ml). *Genome Biol.* 2002; 3:RESEARCH0046. [PubMed: 12225585]
- Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Michael Clark A, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Le Novere N, Leebens-Mack J, Lewis SE, Lord P, Mallon AM, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoekert CJ, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol.* Aug; 2008 26(8):889–896. [PubMed: 18688244]
- Thon, F. Phase contrast electron microscopy. In: Valdre, U., editor. *Electron Microscopy in Material Sciences.* Academic Press; New York: 1971. p. 571-625.

- van Heel M, Keegstra W. Imagic: a fast, exible and friendly image analysis software system. *Ultramicroscopy*. 1981; 7:113–130.
- Wang Z, Hryc CF, Bammes B, Afonine PV, Jakana J, Chen DH, Liu X, Baker ML, Kao C, Ludtke SJ, Schmid MF, Adams PD, Chiu W. An atomic model of brome mosaic virus using direct electron detection and real-space optimization. *Nat Commun*. 2014; 5:4808. [PubMed: 25185801]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1. Definition of the default origin in images with even and odd edge lengths. (a) The origin of a 4×4 pixels image is located at [2, 2], while (b) the origin of a 3×3 pixels image lies at [1, 1]. Each little green circle indicates the sampling point for that pixel and the numbers in parenthesis indicate the indices.

Fig. 2.

Definition of the defocus parameters for the contrast transfer function. The astigmatism angle is limited to the range $[0, 180)$, indicating the angle of the vector pointing to the direction of maximum defocus ($+45^\circ$ in this example). Note that the major axis of the ellipses formed by the Thon rings is perpendicular to the direction of maximum defocus (Thon and Valdre, 1971).

Listing 1

Example EMX file showing relationships between particles and micrographs.

```

<?xml version="1.0"?>
<EMX version="1.0">
  <!-- This is a comment.-->
  <micrograph fileName="mic001.mrc">
    <cs unit="mm">2.2</cs>
    <acceleratingVoltage unit="kV">300</acceleratingVol tage>
  </micrograph>
  <particle fileName="particleStack.mrc" index="1">
    <micrograph fileName="mic001.mrc"/>
    <centerCoord>
      <X unit="px">37.4</X>
      <Y unit="px">204</Y>
    </centerCoord>
    <boxSize>
      <X unit="px">128</X>
      <Y unit="px">128</Y>
    </boxSize>
    <transformationMatrix>
      <t11>0.1</t11><t12>0.21</t12><t13>0.3</t13><t14 unit=px>0.4</t14>
      <t21>0.2</t21><t22>0.02</t22><t23>0.0</t23><t24 unit=px>0.0</t24>
      <t31>0.3</t31><t32>0.22</t32><t33>0.1</t33><t34 unit=px>0.4</t34>
    </transformationMatrix>
  </particle>
  <particle fileName="particleStack.mrc" index="2">
    <micrograph fileName="mic001.mrc"/>
    <centerCoord>
      <X unit="px">502</X>
      <Y unit="px">194</Y>
    </centerCoord>
    <boxSize>
      <X unit="px">128</X>
      <Y unit="px">128</Y>
    </boxSize>
    <transformationMatrix>
      <t11>0.3</t11><t12>0.61</t12><t13>0.4</t13><t14 unit=px>0.1</t14>
      <t21>0.2</t21><t22>0.05</t22><t23>0.1</t23><t24 unit=px>0.2</t24>
      <t31>0.1</t31><t32>0.22</t32><t33>0.2</t33><t34 unit=px>0.3</t34>
    </transformationMatrix>
  </particle>
</EMX>

```