Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the development of three new tools for organizing and sharing information in three-dimensional electron microscopy

Roberto Marabini,[a]* Jose Ramon Macias,[b] Javier Vargas,[b] Adrian Quintana,[b,c] Carlos Oscar S. Sorzano[b] and Jose María Carazo[b,c]

[a]Escuela Politécnica Superior, Universidad Autonoma de Madrid, 28049 Madrid, Spain, [b]Centro Nacional de Biotecnologia–CSIC, Campus Canto Blanco, 28049 Madrid, Spain, and [c]National Institute of Bioinformatics, Madrid, Spain

Correspondence e-mail: roberto@cnb.csic.es

Received 16 May 2012
Accepted 13 March 2013

Electron microscopy is a valuable tool for elucidating the three-dimensional structures of macromolecular complexes. As the field matures and the number of solved structures increases, the existence of infrastructures that keep this information organized and accessible is crucial. At the same time, standards and clearly described conventions facilitate software maintenance, benefit interoperability with other packages and allow data interchange. This work describes three developments promoting integrative biology, standardization and workflow processing, namely *PeppeR*, the EMX initiative and *Scipion*.

1. Introduction

Molecular biology is a discipline where traditional publishing has naturally merged with data sharing through electronic databases. This strong relationship is illustrated by the fact that most journals only accept manuscripts under the condition that data, such as gene sequences and model coordinates, are deposited in publicly available databases. Indeed, the accumulation of information in public repositories has been critical for the development of computational biology, which in turn generates new data for the scientific community.

Electron microscopy (EM) is a valuable tool for elucidating the three-dimensional structures of macromolecular complexes. Knowledge of the structure of a biological complex may provide important information about its function. As the number of solved structures increases, the existence of infrastructures that keep this information organized and accessible is crucial. Therefore, the establishment of the Electron Microscopy Data Base (EMDB), which provides access to the reconstructed maps as well as the conditions under which the sample was prepared and the data were recorded, has been an important milestone (Tagari *et al.*, 2002; Henrick *et al.*, 2003).

Building on the existence of the EMDB, the EM community is moving towards more ambitious goals and significant advances are expected in a number of areas, such as the following.

(i) Integrative biology. It is now possible to develop tools and services that allow the close integration of the structural information at medium–low resolution obtained by EM with atomic resolution data from X-ray diffraction and nuclear magnetic resonance spectroscopy (NMR), resulting in so-called 'hybrid models'. Furthermore, these hybrid models can be further integrated with the very rich body of different types of biological information contained in public databases of genes, proteins, motifs *etc.*

(ii) Standardization. Considering the maturity of the field, this is the right moment to advance in normalization and standardization, coming to an agreement on how to describe

the tasks associated with image processing. This standardization will in turn facilitate the use of software and increase the capabilities of collaborative work and data interchange between different platforms. Moreover, it will help users to change from one software suite to another, so that they can always use the optimal algorithm for the task in hand.

(ii) Workflow processing. The benefits of standardized image-processing workflows are multiple. On the one hand, users will have access to state-of-the-art algorithms, increasing the quality of their results. On the other hand, several software packages may cooperate in order to achieve the best results. Finally, results can be more easily shared and compared since the workflows will be well known to the community.

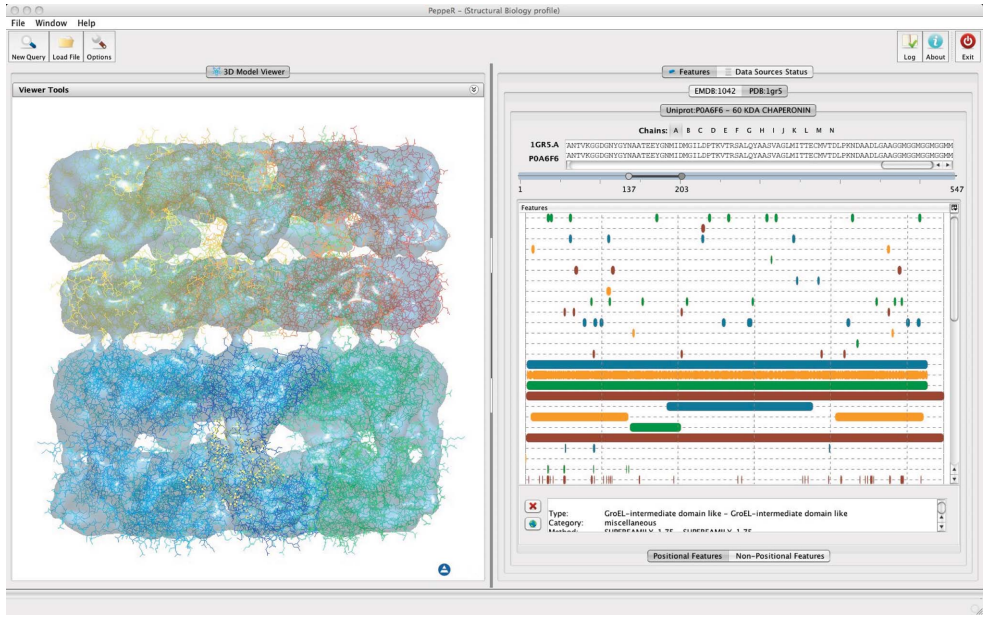

Figure 1
Starting from EMDB entry 1042 (*Escherichia coli* chaperonin GroEL), *PeppeR* is able to retrieve and display a variety of related information obtained from different DAS servers. This information includes fitted PDB models, protein sequences and genomic and secondary-structure annotations. For this particular example, 74 DAS servers were accessed and information was retrieved from 15 of them (those that contained relevant information for this chaperonin).

In the following sections, we present three developments by our laboratory intended to help the community to overcome some of the difficulties in sharing data and information.

2. *PeppeR*: an integrative tool for three-dimensional electron microscopy

PeppeR is an application that is intended for the visualization of hybrid models together with their biological annotation (Macías *et al.*, 2007; http://biocomp.cnb.csic.es/das/PeppeR/). These hybrid models combine information at medium resolution obtained from three-dimensional electron microscopy (3D-EM) with data provided by other techniques with higher (atomic) resolution: mainly X-ray diffraction and NMR. The establishment of this correspondence at different resolution levels allows the allocation of features and annotations related to the underlying protein and gene sequences into the three-dimensional maps.

PeppeR implements DASx3DEM (http://biocomp.cnb.csic.es/das/dasx3dem.jsp), an extension of the Distributed Annotation System (DAS; Dowell *et al.*, 2001) protocol which allows annotations of hybrid models to be shared. This annotation system has been built on the basis of three 'reference databases': the Electron Microscopy Data Bank (EMDB), which stores 3D-EM maps, the Protein Data Bank (PDB), which holds atomic coordinates, and the Universal Protein Resource (UniProt), which holds protein sequences. Through these reference databases, more than 200 further additional databases are accessed, providing new annotations that are collected and displayed using a single graphical visualization client. Thus, users can have an integrated view of all of the annotations in a large macromolecular complex (3D-EM map from the EMDB), the atomic resolution structures fitted into it (coordinates from the PDB) and the sequences corresponding to each of the structures (from UniProt). Currently, 74 DAS servers with almost 1500 DAS data sources are registered. These data sources originate from more than 20 projects served by over 44 institutions in 18 different countries.

PeppeR has been implemented as a set of web services and as a Java web start application. Web services are in charge of retrieving data from the repositories and databases, while the client provides graphical tools to the users for inspection of the information. The *PeppeR* graphical DAS client for 3D-EM basically consists of two panels (see Fig. 1). In the left-hand panel a three-dimensional viewer allows the user to display three-dimensional maps with their corresponding fitted atomic structures. In the right-hand panel the annotations provided by the different DAS servers are presented as coloured boxes overlaid on the protein sequence. Additional information about each annotation is provided when the user clicks on it. In this way, the user has a very powerful and intuitive integrated environment that links structural information to a plethora of other types of information, such as mutations, single-nucleotide polymorphisms (SNPs), links to pathology *etc*.

2.1. Current status of *PeppeR*

PeppeR is available for download *via* the *PeppeR* website (http://biocomp.cnb.csic.es/das/PeppeR/) or from the subversion repository (http://sourceforge.net/projects/dasx3dem).

Development of *PeppeR* is moving very rapidly; one of the main new features of the next version will be a new interface

for displaying annotations. In this new version, positional annotations will be shown in a table-like layout, allowing the user to reorder them based on different fields of the annotation (*i.e.* category, method, source *etc.*).

2.2. Electron-microscopy exchange (EMX) initiative

Standardization in 3D-EM is a long-pending issue. Until recently, it had only been achieved for those descriptors needed for public deposition of density maps in the EMDB. These descriptors provide a brief summary of how a particular three-dimensional map has been produced, but are not adequate to describe image-processing workflows or to allow metadata interchange between different image-processing packages.

Under the umbrella of the European Strategy Forum on Research Infrastructures (ESFRI; http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=home) and in particular the Integrated Structural Biology Infrastructure for Europe (INSTRUCT; http://www.structuralbiology.eu/), the Instruct Image Processing Center (I2PC; http://i2pc.cnb.csic.es/LoadHome.htm) was created. The first I2PC Developer Workshop took place in Madrid on 6 and 7 February 2012. Its main objectives were to standardize information exchange for single particles as well as to start a systematic comparison of the performance of the different algorithms used in 3D-EM. The list of participants, which includes representatives of most major single-particle software packages, is available at http://i2pc.cnb.csic.es/LoadNewsEvents.htm?type=DeveloperWorkshop. It should be noted at this point that EMX developments are made in close collaboration with the PDB/EMDB. We appreciate their strong support and, whenever possible, the same labels and units as used in the PDB/EMDB are also used in EMX.

At the Madrid workshop, an initial data model describing the minimum information that needs to be handled in order to understand the output of any EM software plus the relationships among pieces of this information was agreed on. The EMX initiative aims to help EM users to exchange data, so it is only concerned with the minimum amount of information that is needed to interpret the results of an image-processing step unambiguously. Therefore, it is not designed to replace the native formats, conventions and schemes that are used by the different image-processing packages to store metadata information or to guarantee the reproducibility of a particular experiment.

One of the more visible results of the First I2PC Developer Workshop (http://i2pc.cnb.csic.es/emx) was the creation of the EMX exchange website. The main objective of the EMX website is to standardize information interchange in single-particle analysis (SPA) among different 3D-EM software packages. The website is divided into three main parts: Dictionary and Format, Example and Tests, and Resources.

In the Dictionary and Format section, the proposed file format used for metadata and data exchange is described, as well as the list of labels agreed on at the I2PC workshop. In the Convention subsection, the different conventions followed in the standard for interchange are enumerated. These conventions, which are vital for data interchange in the field, include where the image centre is, a parametric description of the contrast transfer function and the transformation matrix used to define the alignment and projection directions.

In the Example and Tests section, examples as well as the material needed to test utilities for conversion to and from the interchange standard are available. For example, if the Coordinates section is selected, an example file and several tests will be displayed. In the first test, users should download one micrograph and a set of coordinates, convert them to a particular package format, extract the particles and upload the stack of images extracted from the micrograph at the given coordinates. The website will then compare the uploaded images with a reference gallery of images previously extracted from the micrographs at the same coordinates by the web maintainers. The test is successful if the galleries are identical (see Fig. 2). In addition to showing the image galleries, the EMX website makes

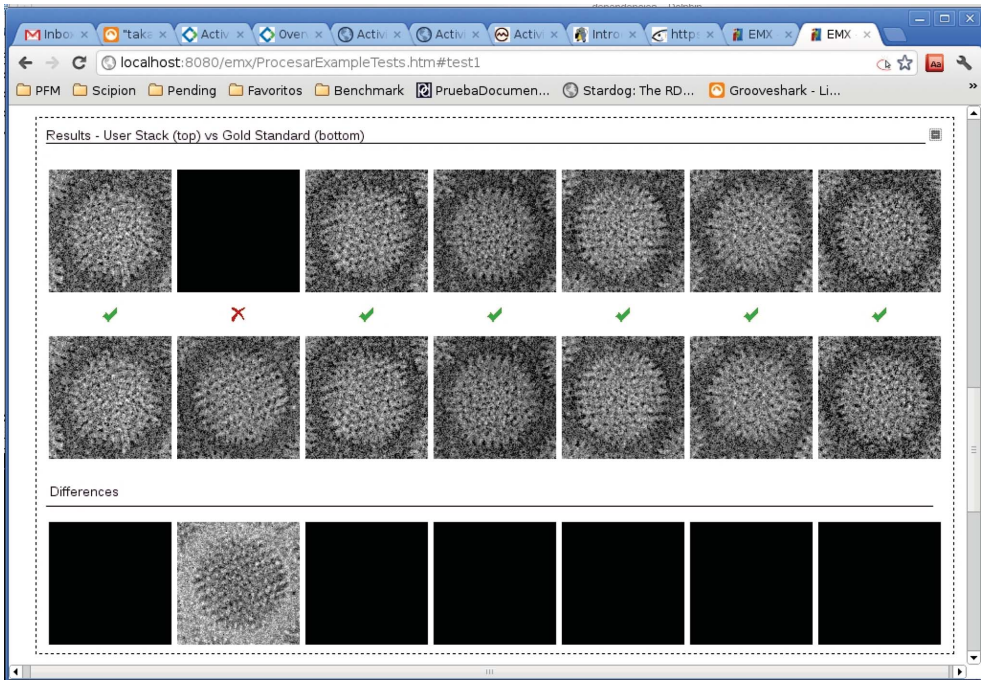

Figure 2
Test for particle selection showing the user input (first row), the reference gallery of images (second row) and the difference images (third row). In this case, the test has not been successful since the images in the first and second rows in the second column are not identical.

a pixel-by-pixel comparison between images belonging to both galleries, checking if any pair of pixels differ by more than 10^{-3}. All pairs of images which differ by more than this threshold are marked with a red cross since uploaded images should be identical to the reference gallery, apart from errors arising from numerical rounding. Finally, the last row shows the difference image between the reference gallery and the uploaded stack of images.

The last section is called Resources. It contains a table summarizing the convention utilities created by the developers of each package. If provided, there will be a link to online converters. In a second subsection, there will be a syntactic validator of the metadata files. Finally, in the third subsection the user can find libraries that have been found to be useful to create the conversion utilities.

2.3. Current status of EMX

The first version of EMX has been kept simple, following the consensus position among software developers of providing a simple standard to start with which will become more elaborate upon usage. This first version defines (i) micrographs (that is, images as obtained in an electron microscope) and (ii) particles (that is, each one 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. Of particular interest are 'classes' and 'three-dimensional maps' (volumes).

The EMX data model has been defined using a relational model. However, we envision that a data-acquisition and data-processing ontology will be necessary to allow a more thorough description of 3D-EM results, so that better integration of information with a variety of biological and biochemical data sources can become a reality.

As we write these words, we are updating the EMX web page in order to introduce the last modifications suggested by the community. The developers of several software packages, including *Appion* (Lander *et al.*, 2009), *Bsoft* (Heymann, 2001), *EMAN* (Ludtke *et al.*, 1999), *Relion* (Scheres, 2012) and *XMIPP* (Sorzano *et al.*, 2004), have shown interest in implementing export/import utilities to/from the EMX standard. An added benefit of EMX in the context of the EMDB is that it opens the way to meaningfully explore the pros and cons of archiving the images contributing to a given map on top of the map itself. Indeed, by archiving images and not only maps, new image-processing methods can be tested on existing image sets, potentially unlocking new information or resolving controversies. Obviously, EMX assures that sufficient and unambiguous information on the image sets has been stored so that they can be reprocessed using new algorithms.

2.4. *Scipion*

In transmission electron microscopy, software interoperability seeks to smooth the integration of different image-processing packages. The last development that we comment on in this work is *Scipion*, a project that is currently under development (http://scipion.cnb.csic.es). *Scipion* aims to integrate several software packages for the elucidation of three-dimensional structures through a workflow approach. The software will allow the execution of reusable, standardized, traceable and reproducible image-processing protocols. These protocols will integrate tools from the main 3D-EM software packages (providing full interoperability among them).

Scipion is being mainly developed in Java and Python. The system is formed by the following parts.

(i) *Scipion* client. A Java desktop application provides the primary user-interactive environment. This application is the main interface between the user and the *Scipion* system.

(ii) Web services. A set of Java web services that comprise the business logic. The *Scipion* client sends instructions to the different web services, which in turn execute the different tasks required for the proper execution of the workflows and related tasks. The web-service layer is the only part that has a connection to the database and the execution machine.

(ii) Execution machine (wrappers). A set of Python programs that are launched by the execution machine. These programs isolate *Scipion* from the input and output format. 3D-EM packages must be installed previously as a custom installation.

The initial *Scipion* design made use of a relational database (PostgreSQL), but soon we understood that a data model based on a relational approach was not an option since the number of tables needed was difficult to manage. Relational databases are great when the data model is well understood and it is known how the data will be used. However, for *Scipion* the developers can never quite know what requirements will be needed to accommodate the different algorithms of each software package, so high flexibility is mandatory. In this way, from an initial design in which database tables were strongly related to the different image-processing elements, we changed to a more flexible approach. In this new version the schema no longer reflects the basic concepts that programmers in image processing are used to. Instead, it is composed of rather abstract classes with properties assigned to them using a key/value approach. This new design, in which everything has been flattened to key/value pairs, allows the much needed flexibility.

At present, *Scipion* metadata storage is based on the Jena-TDB ontological database (http://incubator.apache.org/jena/documentation/tdb/index.html). We have developed a persistence layer using Empire which implements JPA (the Java Persistence API), which establishes specifications for accessing, persisting and managing data between Java objects/classes and a relational database. This layer queries the ontological databases using SPARQL, which is a query language for databases able to retrieve and manipulate data stored in Resource Description Framework (RDF) format.

The *Scipion* ontology is composed of almost 300 classes and a similar number of properties. The ontology is available at the URL http://scipion.svn.sourceforge.net/viewvc/scipion/DataBase/Ontologies/ and has been created using *Protégé* (http://protege.stanford.edu). Here, we present a simplified summary of the section related to workflows: in the *Scipion* ontology, a *workflow* is an entity that implements a logical

action and has a well defined minimum set of input and output *nodes*. Logical actions are any of the well defined steps in electron-microscopy image processing; examples are particle alignment, particle normalization *etc.* Each implementation of a workflow is called a *protocol*. In general, each image-processing package has one or many protocols for a given workflow. The type of input or output data needed or generated by each protocol are described using the node class. Finally, instances of the class ProtocolExecution store the execution of any protocol. Fig. 3 shows a diagram with all of the classes needed to implement workflows in *Scipion* plus their more relevant properties.

2.5. Current status of *Scipion*

A first prototype of *Scipion* will be available for download at the end of 2013 (at http://scipion.cnb.csic.es). This alpha release will integrate two of the main three-dimensional electron-microscopy packages: *XMIPP* and *EMAN*. The software will be able to remotely execute any 3D-EM task as well as to monitor the whole process, storing the current status and the input and output parameter relationship between the different steps in the ontological database. The application assures full traceability and, through a graphical interface, allows analysis of the 'pipeline' execution as well as providing reproducibility, versions from a previous execution, deletion *etc.* Finally, it provides a set of functionalities for visualization: it allows connectivity to any visualization tool previously installed by the user (*Chimera*, *ImageJ etc.*) and offers a wrapper layer over the protocol-visualization tool designed by each proprietary software.

As *Scipion* provides full traceability, once the user has obtained the final volume reconstruction *Scipion* will be able to provide a detailed description of the procedure and the parameters, allowing the automated deposition of data in the EMDB.

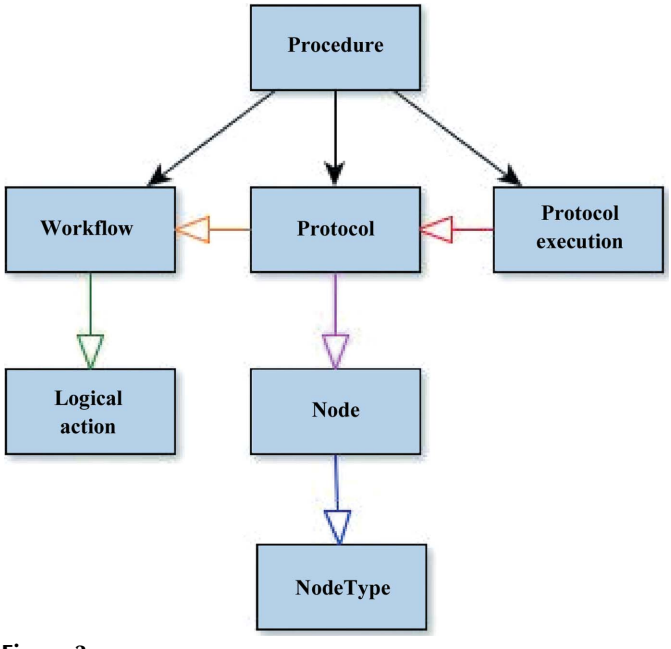

Figure 3
Snapshot showing a small set of terms from the *Scipion* ontology. Relationships between terms are represented by arrows. These arrows have solid arrowheads when the relationship is a specialization.

3. Conclusions

This article describes three developments that involve integration and data sharing: *PeppeR*, EMX and *Scipion*. The ultimate goal of these contributions is to provide integrative, high-resolution, high-throughput imaging to structural biology using electron microscopy. A summary is provided below.

(i) *PeppeR*. Twenty-first century biology is a data-intensive enterprise. An immense challenge is that of managing the variety and complexity of data types and the need to acquire data using a wide variety of methods. *PeppeR* allows the integration of sequence annotations in a three-dimensional model. Information such as secondary-structure prediction, domain prediction, binding sites, active sites, alternative splicing sites, SNPs *etc.* can be easily displayed and located within the three-dimensional volume.

(ii) EMX. The workflows proposed by the different EM reconstruction packages are similar and therefore it should be easy to exploit the different strengths of each of them. However, in practice small differences between the exact meanings of the different parameters heavily penalize users who try to use more than one package. The second development (EMX) addresses this lack of common conventions by defining an interchange format.

(iii) *Scipion*. Current software packages are actually composed of hundreds of small programs each performing an 'atomic' task, which must be assembled into a script constituting the image-processing pipeline. These scripts are usually run several times, varying the parameters among the different executions until the best reconstruction is obtained. Most of the time, the traceability of this process relies purely on the laboratory notebooks of the user and his/her good practice. Sometimes, the reproducibility of the final reconstruction is seriously compromised by poor notebook annotation. As a result, we have noticed that a few months after completing a project the researcher is usually unable to reproduce some details of her own results without major effort. The third development introduced in this paper, *Scipion*, is a workflow platform for image processing that allows the integration of several imaging-software packages, guaranteeing the traceability and reproducibility of the image-processing steps that were followed to obtain a final three-dimensional map.

This work was funded by the Spanish Ministerio de Economía y Competividad through grants BFU2009-09331, BIO2010-16566, ACI2009-1022, ACI2010-1088 and AIC-A-2011-0638, by the Comunidad Autonoma de Madrid through grant S2010/BMD-2305, by NFS grant No. 1114901 and by the Spanish National Institute of Bioinformatics (a project funded by the Instituto de Salud Carlos III). This work was conducted using the *Protégé* resource, which is supported by grant LM007885 from the United States National Library of Medicine. COSS is a Ramón y Cajal researcher financed by

the European Social Fund and the Ministerio de Economía y Competitividad. JV is a Juan de la Cierva Postdoctoral Fellow (JCI-2011-10185). This work was funded by Instruct, which is part of the European Strategy Forum on Research Infrastructures (ESFRI) and is supported by national member subscriptions.

References

Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R. & Stein, L. (2001). *BMC Bioinformatics*, **2**, 7.

Henrick, K., Newman, R., Tagari, M. & Chagoyen, M. (2003). *J. Struct. Biol.* **144**, 228–237.

Heymann, J. B. (2001). *J. Struct. Biol.* **133**, 156–169.

Lander, G. C., Stagg, S. M., Voss, N. R., Cheng, A., Fellmann, D., Pulokas, J., Yoshioka, C., Irving, C., Mulder, A., Lau, P.-W., Lyumkis, D., Potter, C. S. & Carragher, B. (2009). *J. Struct. Biol.* **166**, 95–102.

Ludtke, S. J., Baldwin, P. R. & Chiu, W. (1999). *J. Struct. Biol.* **128**, 82–97.

Macías, J. R., Jiménez-Lozano, N. & Carazo, J. M. (2007). *J. Struct. Biol.* **158**, 205–213.

Scheres, S. H. W. (2012). *J. Struct. Biol.* **180**, 519–530.

Sorzano, C. O. S., Marabini, R., Velázquez-Muriel, J., Bilbao-Castro, J. R., Scheres, S. H. W., Carazo, J. M. & Pascual-Montano, A. (2004). *J. Struct. Biol.* **148**, 194–204.

Tagari, M., Newman, R., Chagoyen, M., Carazo, J. M. & Henrick, K. (2002). *Trends Biochem. Sci.* **27**, 589.