U SCIENCE TECH
FACULTAT DE CIÈNCIES
I TECNOLOGIA
UVIC-UCC

EMBL-EBI

FINAL DEGREE PROJECT

"IMPROVED BIOLOGICAL ANNOTATION OF EMDB DATA"

Irene Solanes Valero

Degree in Biotechnology

Supervisor: Josep Mª Serrat Jurado

External supervisor: Ardan Pathawardan

Vic, January of 2016

TABLE OF CONTENTS

ABBREVIATIONS

3DEM	*three-dimensional electron microscopy*
CET	Cryo-electron tomography
cryo-EM	*Cryo-electron microscopy*
EBI	*European Bioinformatics Institute*
EF-G	*Elongation factor G*
EF-Tu	*Elongation factor Tu*
EM	*Electron Microscope*
EMBL	*European Molecular Biology Laboratory*
EMDB	*Electron Microscope Data Base*
EMPIAR	*Electron Microscopy Pilot Image Archive*
MCM	*Minichromosome maintenance*
NMR	*Nuclear magnetic resonance*
PDBe	*Protein Data Bank Europe*
SFF	*Segmentation File Format*
SPEM	*Single-particle electron microscopy*
wwPDB	*Worldwide Protein Data Bank*

Summary

Title: Improved biological annotation of EMDB data

Key words: Electron Microscope, structural biology, segmentation, database, tomogram, atomic model, biological annotation, Chimera.

Author: Irene Solanes Valero

Supervisor: Josep Mª Serrat Jurado

External supervisor: Ardan Pathawardan

Date: January 2016

In this Final degree Project are explained all processes required to create the segmentation and biological annotation parts of Segmentation File Format (SFF) project. The segmentation file format is an Electron Microscope Data Base (EMDB) project, which is a combination of two projects, the "OMERO volume slicer" and the "Segmentation annotation tool". This project was developed in order to build the segmentation and annotation sets for the "Segmentation annotation tool" step.

The project was divided into two steps, Segmentation using Chimera, that permitted the downloading of the EMDB and PDB structure and proceed to do the segmentation, then this segmentation was saved as an Segger or HDF5 file. The segmentation of a structure is the decomposition by the different components that forms it.
The second part consisted to create a biological annotation of the components previously segmented and building a link with other databases in order to identify them.

The objective of this project was to create a SFF, this application would permit to the user to visualize segmentations on a EM tomogram, in order to identify and distinguish all the components that composed it. The annotation and segmentation that is explained in this project was made as a data set to build this new EMDB application.

On the end of the project a collection of 100 segmentation and annotation sets were obtained, three types of structures can be distinguished; Helicases, Chaperones and Ribosomes.

1. INTRODUCTION

Structural biology is a branch of molecular biology, biochemistry, and biophysics concerned with the molecular structure of biological macromolecules, especially proteins and nucleic acids, on how these molecules acquire their structures, and on how alterations in their structures affect their functions.
Nowadays the electron microscopy and tomography[1] are two techniques growing up in use. With these techniques we can obtain 3D structures, that can be analysed and annotated in order to improve the knowledge of their structural biology.

At high resolution atomic models can be fit in to 3D structures and the biological interpretation is done via sequence information. At low resolution, when we do not have such accurate information, the 3D structures are often divided into regions called segments[2] then biological meaning is assigned to these regions by means of other knowledge, for example labelling experiments.
There is currently a public archive for 3D EM data, the EMDB, but it currently does not capture segmentation data. In order to make the low resolution data in EMDB more useful from the biological perspective it is important that segmentations are captured and they are biologically annotated in a coherent and consistent way.

The aim of this project was to provide a pioneer, to develop example datasets and use cases that drive developments in the field. We have worked with several segmentation programs, such as Chimera[3] in order to segment and annotate the different subunits or components of a protein; we have worked with a total of 100 proteins including 14 helicases (MCM), 82 chaperones (GroEL) and 4 ribosomes. This has been done with the objective to improve the quality of the EMDB database, and to know more information about proteins and facilitate the identification of the components of the tomogram in order to help user interpretation and analysis.

---

1 Tomography: refers to imaging by sections or sectioning, through the use of any kind of penetrating wave (Eg. EM)

2 Segment: Region of a biological structure such us a protein a macromolecule. This can be a subunit or a group of regions of the structure, the definition of this segment will be determined by the resolution of the map.

3 Chimera: Program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, etc

1.1 Present situation and motivations

As a last year student of Biotechnology, my intention is to consolidate and interlink technological concepts of different subjects studied during the degree.

In my case, the project will include a part of Biotechnology concepts, an important part of Biology, and Bioinformatics.

This project was proposed to me during in a internship on Cambridge, UK, called EMBL-EBI. The experience that I had in Cambridge was really satisfactory. They gave me a fantastic opportunity to jump into the labour market, working directly in a team, and developing new projects for the company.

1.2 Goals and structure of EMBL-EBI

The **European Molecular Biology Laboratory** (**EMBL**) is a molecular biology research institution supported by 21 member states, three prospect and two associate member states. EMBL was created in 1974 and is an intergovernmental organisation funded by public research money from its member states. Research at EMBL is conducted by approximately 85 independent groups covering the spectrum of molecular biology. The laboratory operates from five sites: the main laboratory in Heidelberg, and outstations in Hinxton (the European Bioinformatics Institute (EBI), UK, Grenoble (France), Hamburg (Germany), and Monterotondo (near Rome). EMBL groups and laboratories perform basic research in molecular biology and molecular medicine as well as training for scientists, students and visitors. The organization aids in the development of services, new instruments and methods, and technology in its member states. [59]

In my case I was working on the European Bioinformatics Institute (EMBL-EBI) which is a research centre and services in bioinformatics, developing and maintaining a large number of scientific databases, which are free of charge.

This project is a collaboration with a EBI database called Protein Data Bank Europe (PDBe). The Electron Microscopy Data Bank (EMDB) is a database managed by PDBe and was also in this section where the project was developed.

1.3 PDBe and EMDB

PDBe is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures. In collaboration with the other Worldwide Protein Data Bank (wwPDB), *Figure 1.3.1*, and EMDataBank partners, they work to collate, maintain and provide access to the global repositories of macromolecular structure data (the Protein Data Bank (PDB) and Electron Microscopy Data Bank (EMDB)).

Objectives:

- To provide an integrated resource of high-quality macromolecular structures and related data and make it available to the biomedical community via intuitive user interfaces.
- To maintain in-house expertise in all the major structure-determination techniques (X-ray, NMR and EM) in order to stay abreast of technical and methodological developments in these fields, and to work with the community on issues of mutual interest (such as data representation, harvesting, formats and standards, or validation of structural data).
- To provide high-quality deposition and annotation facilities for structural data as a wwPDB and EMDataBank deposition site.

PDBe also works actively with X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and **cryo-Electron Microscopy (EM) communities which founded the Electron Microscopy Data Bank (EMDB),** that has been run jointly by PDBe and RCSB since 2008. PDBe's active involvement with the scientific community has resulted in improved tools for structure, data deposition and analysis. [1]

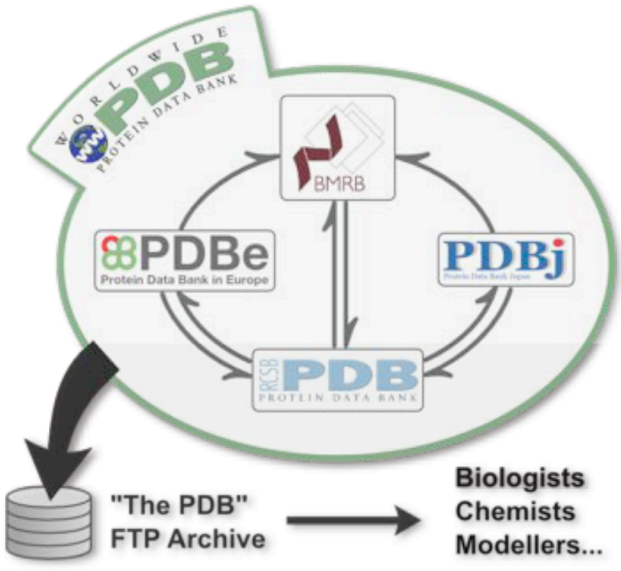


Figure 1.3.1 . Schema of the relationship between the different collaborators of PDB . wwPDB englobe all the PDB databases which forms the PDB archive.

The Electron Microscopy Data Bank (EMDB) is a public repository for electron microscopy density maps of macromolecular complexes and subcellular structures. It covers a variety of techniques, including single-particle analysis, electron tomography, and electron (2D) crystallography, *Figure 1.3.2*.

The EMDB was founded at EBI in 2002, under the leadership of Kim Henrick. Since 2007 it has been operated jointly by the PDBe, and the Research Collaboratory for Structural Bioinformatics (RCSB PDB) as a part of EMDataBank which is funded by a joint NIH grant to PDBe, the RCSB and the National Center for Macromolecular Imaging (NCMI). The actual leader of EMDB-PDBe is Ardan Patwardan and its department coordinated by Gerard Kleywergt, team leader of PDBe.

The map archive includes maps generated by a number of different electron microscopy reconstruction methods. The majority of entries (77%) are single particle reconstructions, which represens ensemble averages of thousands of individual imaged particles [2]

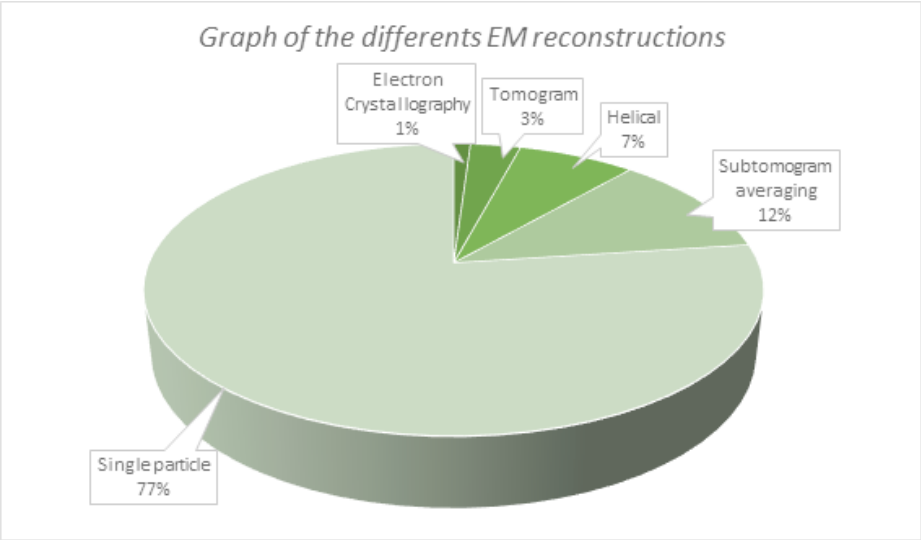


Figure 1.3.2 . Graph of the different EM reconstructions . Distribution of released maps, a total of 3326, in function of the technique used [40]

Related to the EMDB database we can find the **Electron Microscopy Pilot Image Archive (EMPIAR),** *Figure 1.3.3.* This portal is a public resource for raw, 2D electron microscopy images. Here, you can browse, upload, download and reprocess the thousands of raw, 2D images used to build a 3D structure.

The purpose of EMPIAR is to provide easy access to state-of-the-art raw data to facilitate methods of development and validation, which will lead to better 3D structures. It complements

the Electron Microscopy Data Bank (EMDB), where 3D images are stored, and uses the fault-tolerant Aspera platform for data transfers.

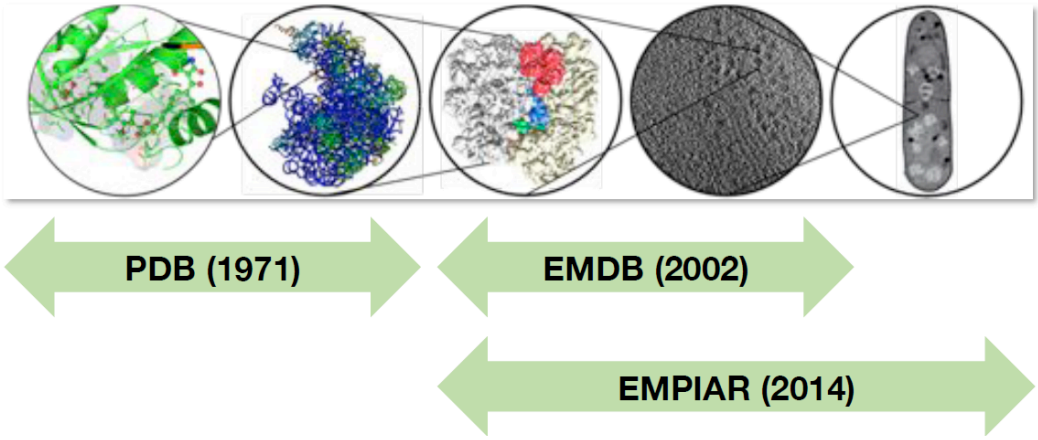


Figure 1.3.3 . Summary of the group organization. Relationship between the processed images of each of the PDB group of databases, with an associated image.

Another collaboration with PDBe is a database called **PDBShape**. This application contains structural biology volume data obtained from electron microscopy, subtomogram averaging, crystallography, and potentially other techniques. Structural alignments between the volumes have been pre-calculated, and these can be searched in PDBeShape to find volumes with similar morphology. This first release contains high quality volume data for prokaryotic and eukaryotic ribosomes, and class I and II chaperonins, taken from EMDB and PDB. The generated file is a Segger file - Chimera as the one generated on this project.

The results of this project have been published on this database, and they are now available for some of the segmented structures. Segger file can be downloaded and the segmentation can be viewed using a supported software as Chimera. You can find an example on the following link *(http://wwwdev.ebi.ac.uk/pdbe/emdb/pdbeshape_dev//volume_details/EMD-1046/)*, corresponding to EMD-1046 a GroEL Chaperone structure. [3]

1.4. 3D reconstruction of EM images

Three-dimensional electron microscopy (3DEM) is a technique used to obtain 3D reconstructions of macromolecular complexes and assemblies and cells. There are a number of sub-methods such as single-particle electron microscopy and electron tomography with their own strengths and weaknesses, but all techniques rely on combining 2D images taken on an electron microscope, combining this information using image-processing and producing one or more 3D reconstructions from the data.

In **single-particle electron microscopy (SPEM)** an amorphous ice layer containing a purified sample, for example many copies of a macromolecular assembly, is imaged. Individual 2D projection images of macromolecules are collected, but these will be of the macromolecule in random orientations. By making some assumptions, for example that all these views represent the same molecule but in different orientations, image processing can be used to obtain a 3D reconstruction (*Figure 1.4.1).*

The **electron tomography** is very similar to SPEM, the specimen is imaged several times while tilting the specimen. The orientations of the sample are known and standard tomographic techniques can be used to obtain the 3D reconstruction.

Single-particle EM now can go to very high resolution but the downside is that it assumes that all particles have nearly the same shape, so this technique is not well suited for imaging pleomorphic specimens. Electron tomography is however ideally suited for this as no assumptions of this nature are made.

The downside of electron tomography is that it is still limited to relatively low resolutions (~ 4nm or worse). A technique known as **sub-tomogram averaging** does the equivalent of single-particle averaging but on a 3D level, see Figure 1.4.1. This can reach a higher resolution that just tomography, but again at the expense of assuming that the specimen is homogeneous.

Cryo-electron microscopy (cryo-EM) is increasingly becoming a mainstream technology for studying the architecture of cells, viruses and protein assemblies at molecular resolution. However the resolution is directly influenced by the wavelength of the imaging radiation source: the shorter the wavelength, the higher the attainable resolution.

Additionally, established modalities for structure determination, such as X-ray crystallography and nuclear magnetic resonance spectroscopy, are being routinely integrated with cryo-EM density maps to achieve atomic-resolution models of complex, dynamic molecular assemblies.

Imaging biological objects in an electron microscope is, in principle, analogous in some respects to light-microscopic imaging of cell and tissue specimens mounted on glass slides. In light microscopy, visible photons serve as the source of radiation; once they pass through the specimen, they are refracted through glass optical lenses to form an image. The specimen is a **live sample**.

One approach to obtaining 3D structures of macromolecular assemblies using EM is tomography, in with a series of images are collected. With each image taken at a different tilt relative to the direction of the incident electron beam, see *Figure 10.5.2. in Appendix 6.*

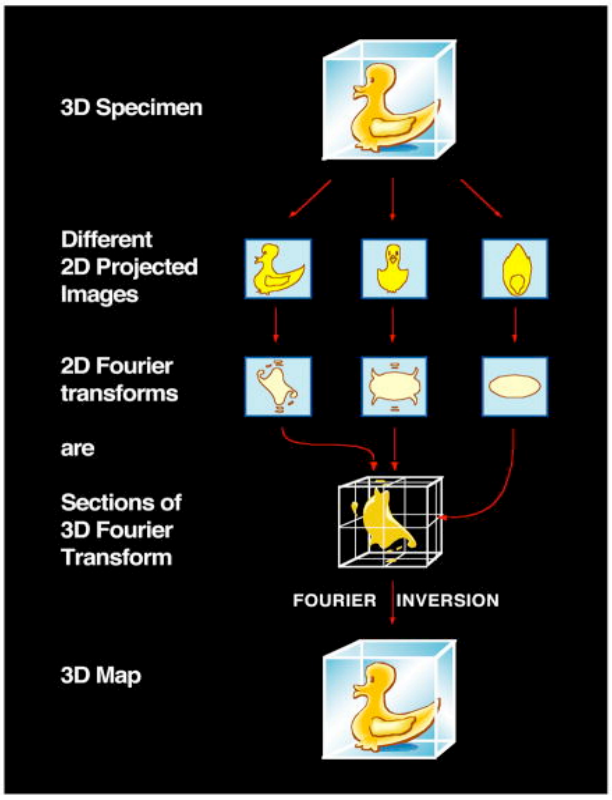


Figure 1.4.1. Principal of reconstruction of 3D structure by Fourier inversion. Using the "Fourier duck" as the prototypical biological specimen, the schematic illustrates that projection images of the object, each with a different orientation, have 2D Fourier transforms that correspond to sections (indicated by red arrows) through the 3D Fourier transform of the original object. Thus, once the 3D Fourier transform is built up from a collection of 2D images spanning a complete range of orientations, Fourier inversion enables recovery of the 3D structure [4].

EBI stores this data in two different databases, EMPIAR and EMDB. EMPIAR stores the scanned microscope tomograms (2D images) and EMDB stores the generated 3D model.

Now Cryo-Em is becoming an essential technique in structural biology, bridging the gap between cell biology, X-Ray crystallography and nuclear magnetic resonance (NMR) spectrometry. CryoEM reconstruction methods are being used to determine structures of large

macromolecules, macromolecular complexes and cell components involved in much key biological process. Find more information about the Cryo-electron tomography (CET) and sub-tomograms averaging technique in *Appendix 5*.

As is shown on the following Chart (*Figure. 1.4.2.*) we can see the increase in EM structures/maps in EMDB database. We can see an increase of 4 times the number of entries between 2002-2014.

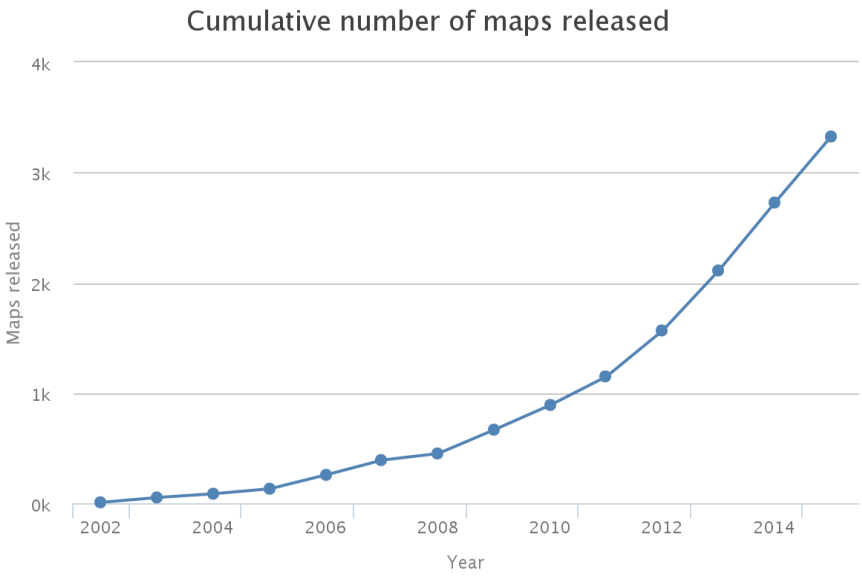


Figure 1.4.2. Chart of cumulative number of maps released of EMDB entries [41]

1.5. Present situation of segmentation

With the segmentation process, we focus images coming from single particles and tomographic reconstructions of cryoEM data. What we want to achieve processing this image is to distinguish the different components of the tomogram in order to obtain more information and facilitate the interpretation of the tomogram and its cellular components.

Nowadays segmentation can be done in three different ways:

- **Manually - *Segmentation on the tomogram*:** This type of segmentation requires a lot of patience and concentration as each generated EM tomogram is segmented separately (*Figure. 1.5.1.).* This process can lead up to 100 hours of work. Once all the slides are segmented, the images are joined by a supported software, resulting in a segmentation 3D model (*Figure. 1.5.2.)*. A recommended software in this case could be a program called Amira.

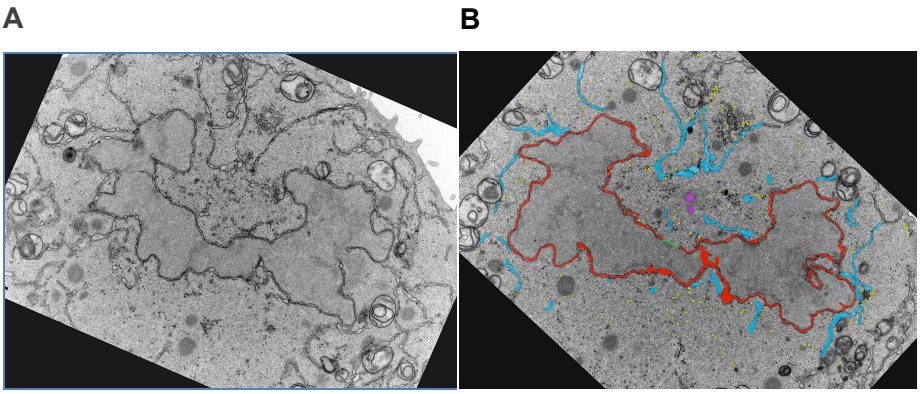


Figure 1.5.1.Original and segmented tomogram of a cell. This image was processed with Amira software by an EM group from Francis Crick Institute, (A) the original EM tomogram and (B) the segmented tomogram where we can distinguish the different components of the cell are labelled with different colours[4]

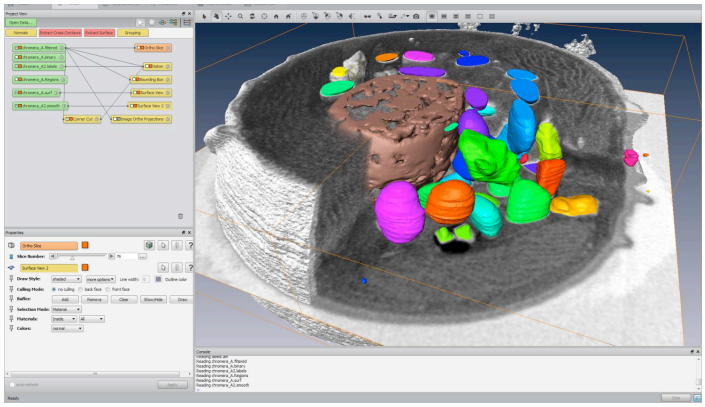


Figure 1.5.2.: Final result of the manual tomogram segmentation: This image is created by the Amira software by Dr. Robert Brandt FEI group. It consists of an EM image of a cell where the cellular components are distinguished from the tomogram.

[4] Lucy COLLINSON, Francis Crick Institute (2015), Images from - 3D segmentations and transformations – building bridges between cellular and molecular structural biology

- **Semiautomatic - Segmentation on a sub-tomogram:** The treated image is a digital 3D model, which was transformed from an EM tomogram by different procedures *section 1.4*. The isolated sub-tomogram average is processed with a program called Chimera. This type of segmentation is also seen in other cases, but never with the objective of storing and annotating the segments.

The methodology that was used in the framework of this project was divided in two principal steps: the EM tomograms extraction and the segmentation of the sub-tomogram average . In *Figure 1.5.3.*, an example of this technique is shown. First we obtain by the EM the ontological projections of a protein (Figure 1.5.3;A). Later on these tomograms are transformed to a digital 3D model. This 3D reconstructed model (EMD), is open then with Chimera software (Figure 1.5.3.;B;I), the contour level is modified to the interested volume (Figure 1.5.3;B;II) and finally is segmented with the different components. Explained in *section 3*, Methods.

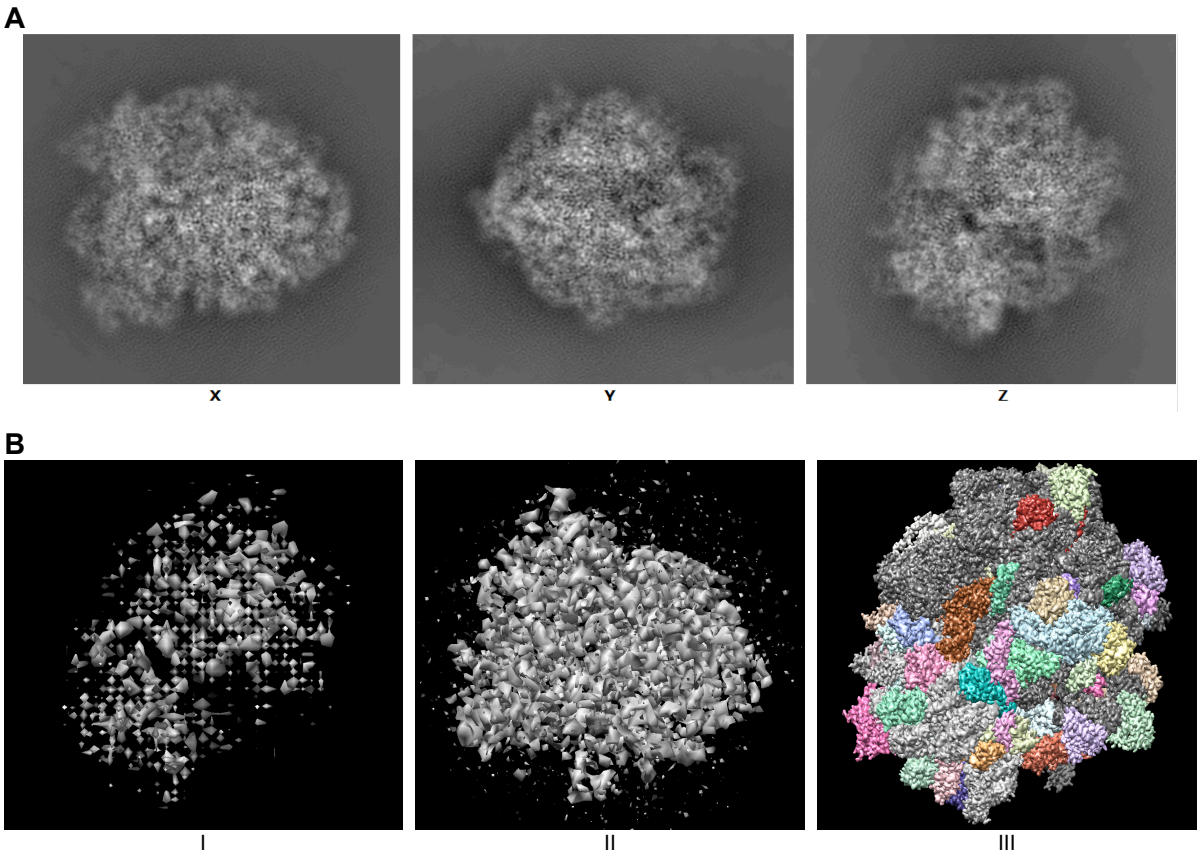


Figure 1.5.3 .Ribosome segmentation procedure (EMD-2847).

- **Automatic:** This methodology is not a very well developed technique, it is very ambiguous. The shape and structure of the sample is constantly changing, and depends on the moment that the photograph is taken. Some groups are working with some techniques to automatize this procedure to make easier and less subjective. A good example is The National cancer institute (USA) led by Sriram Subramaniam, where they automatize a procedure in which all the structures of the membrane proteins are scanned and the resulting average is analysed.

2. PROJECT GOALS

2.1 EMDB-SFF

The aim of this project is to combine the OMERO volume slicer and the segmentation annotation tool. See on "*Appendix 4"* a guide diagram of the life-flow of the project.

The **OMERO volume slicer** is the basis for the protein volume browser and segmentation annotation tool. This application allows the user to travel to all the tomogram using different tools integrated in it. There are provided three different slides views and orthogonal orientations from tomogram, *section 2.3.* This application is not yet available for the public, because is on an approbation system. But this application it is build for all the EMDB entries, we will found it as another EMDB tool.

On this application will provide of segmentations with structures and biological annotations. This data will be available to users that are already registered on the EMDB system. The registration system is already on process of development, but it is been discussed whether to make the registration procedure similar to EMPIAR, database where the 2D tomograms from the electron microscope are stored, explained on section 1.3. [5]

Also along this application will be provided some data sets, as a guide to do the users deposition. The web-service will help users with the format conversion and the biological annotations of segmentations. You can see a mock-up of how the application will look like when finished (*Figure 2.1.1*).

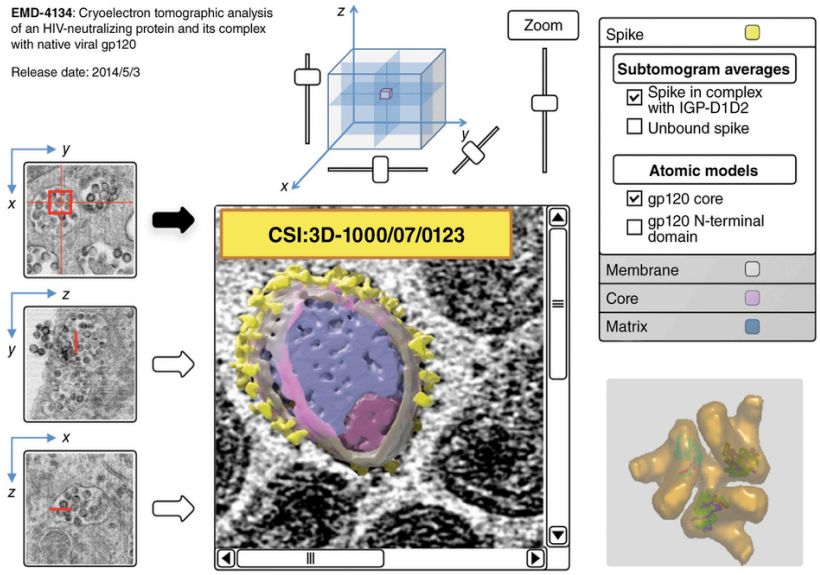


Figure 2.1.1. Mock-out of SFF. Organization of the structure of the SFF mock-out on EMDB application. It is divided on the segmentation tomogram window of a HIV and simian immunodeficiency virus (SIV), where the original tomogram is segmented with its components. Its components can be differentiated with the different labels. Also can be seen the related biological annotation, three orthogonal cross-sections and a cube that help users to orient themselves in the data [6].

SFF is funded by the Mol2Cell grant from MRC and BBSRC, and the EU BioMedBridges project. On December 2015 was organized an expert workshop called "3D segmentation and transformation", to discuss several items about the project, we end with a very productive contributions.

2.2. Segmentation data set

To set up this Web-service of EMDB, it was necessary to create sets of biological segmentation and annotation data in order to analyse the requirements of the service and the difficulties of the project. This is one of the objectives of this project.
We selected three different biological sets with different degrees of difficulties, these structures where: MCM helicases (14), GroEL Chaperone (82) and Ribosomes (4).

The segmentation was made with a Semi-automated segmentation technique. We call this procedure a semi-automated segmentation, because we worked with digital a 3D model from a sub-tomogram, not directly with a tomogram. The tomogram segmentation consists of segmenting each of the slides generated with the Electron Microscope, a very slow and difficult task. Our task was easer, as we worked with a 3D model sub-tomogram, where we could visualize the whole sub-tomogram on a single image, then we could edit and modify the image with the help of a program. In this project we used Chimera a molecule visualization program.

To develop this project, we had to learn the right way to segment an EMDB structure with Chimera. Chimera is a very complete program where you can edit the 3D molecule structure. Nevertheless, learning how to use all the tools included on the program consumes a lot of time to learn all the tools and its segmentation limits [7].
Simultaneously, we also annotated all the biological information of each of the segmented components on an Excel Sheet. This annotation was built thanks to the information provided on the related articles and databases. The information was stored and shared in a Google Drive with all the people involved on this project. The biological annotation was done for each of the EMDB entries, including: Species, **Fitted model, UniProt code,** Contour Level, Resolution, Source information, Chain ID, Length, Theoretical and Experimental MW, Copy number, Related Article and extra notes.

We annotated all this information for three reasons:

1. To check if this shared information was correct, in order to improve it
2. To build a first idea of all the structural components, to identify each of the segments of the structure before the segmentation step
3. To see the limits of the project

2.3 OMERO volume slicer

The volume slicer (protein volume) is one new application that was created in 2014 for the EMDB, on this moment is on validation process and unfortunately this service cannot yet be used, only demo links for some entries. See a demo on the following link: http://www.ebi.ac.uk/pdbe/emdb/3dslice/EMD-2187, and *Figure 2.3.2.*

OMERO is an scientific image translation software, to access and integrate image data. This gives power to combine images from any modern imaging modality — light and electron microscopy, digital pathology, high-content screening — in a single system. This software was used to create EMDB volume slicer and upload on the database, see *Figure 2.3.1*.

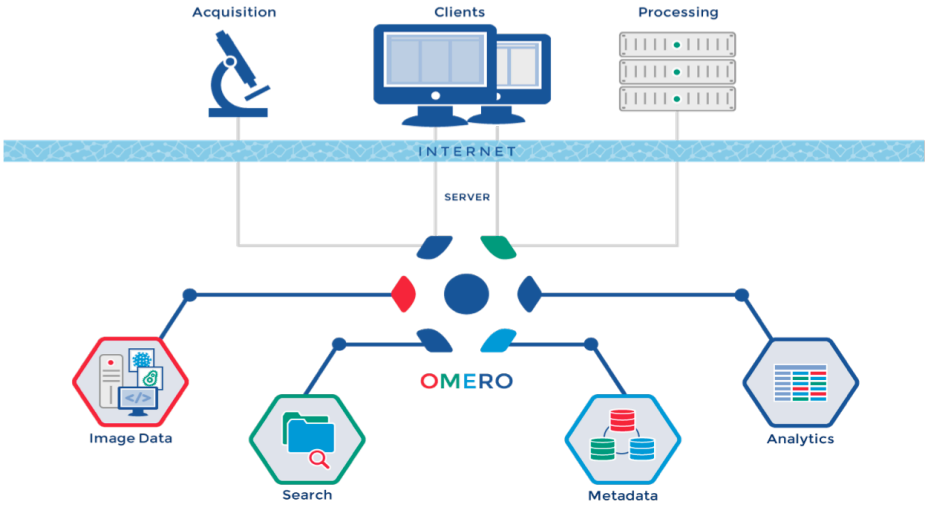


Figure 2.3.1. OMERO life flow diagram.

This EMDB tool one can be used for viewing slices in three orthogonal directions from the 3D EM reconstructions in EMDB. This application was created with the objective to view different direction, navigate in 3D, zoom and adjust the density range of one 2D EMDB image.

As you can see on the following image you have four different control options:

1. **Navigation panel:** Navigate to the region of interest by dragging the slides along each axis or by changing the centre coordinates of the slice being shown by changing the numbers in the input boxes
2. **Top, front and right views:** In order to facilitate navigation we show three orthogonal (in the top, front and right directions) thumbnail slices intersecting at the centre of the slice shown in the navigation panel.

3. **Density range limits:** The min and max sliders or the input boxes can be used to set the min and max of the linear grey scale mapping used to show the image.
4. **Main visualization panel:** The main visualisation panel displays the zoomed region from the active slice.

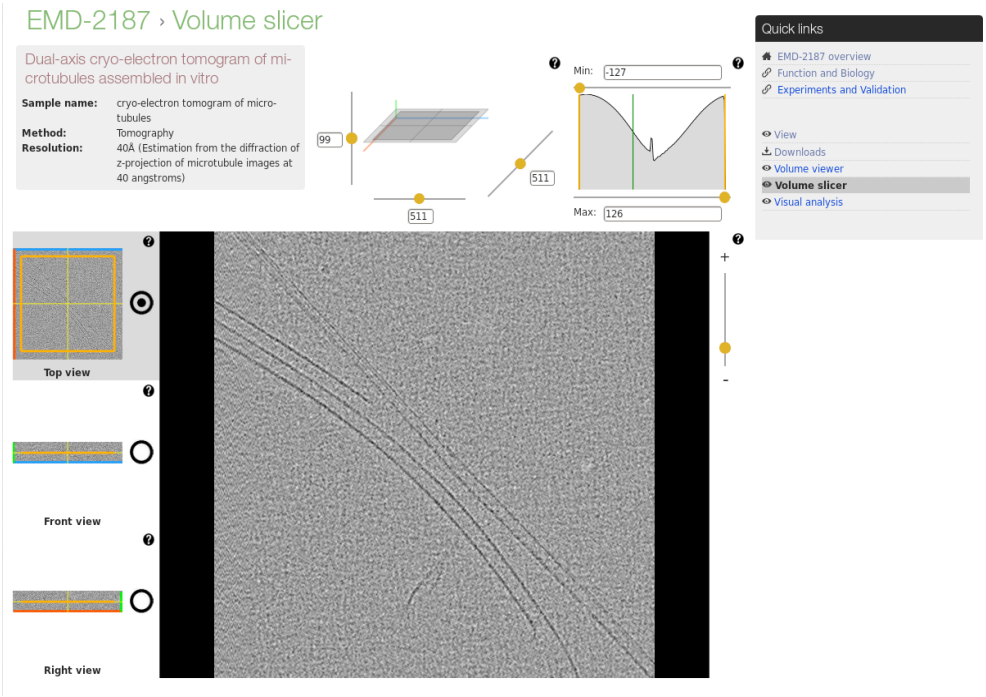


Figure 2.3.2. Volume Slicer application. Three orthogonal cross-sections and a cube that help users to orient themselves in the data. These tomogram belongs to microtubules sample.

3. METHODS

3.1. Proteins structures included in the work

We have worked on three different structures with different degrees of difficulty. First of all we started with a simple case to train the different possibilities of the project, and see then the difficulties that we could found with both; the segmentation and the biological annotation. We have worked three different structures; minichromosome maintenance complex (MCM), Chaperone (GroEL) and Ribosome.
All the EMDB entries are composed by; a *four digits code* which corresponds to a identifier of the structure, and the acronym "*emd"* on the begging that informs that corresponds to the EMDB.

3.1.1. Minichromosome maintenance (MCM)

The minichromosome maintenance (MCM) complex is an eukaryotic replicative helicase that is composed by six distinct, but related, subunits MCM (2–7). The relationship between the sequences of the subunits indicates that they are derived from a common ancestor and indeed, present-day Achaea have a homohexameric MCM [8].

There are cases where the helicase is conformed with a single subunit that forms a MCM homo-hexamer (Human; MCM10), but in other cases they are six different subunits conforming an hetero-hexamer.

We carried out the segmentation of some of the EMDB structures. As the MCM have very bad resolution (>15 Å), the segmentation was very difficult and sometimes impossible. The EMDB entries structures where: emd-6338, emd-5857, emd-1833, emd-1254, emd-2872, emd-5625, emd-1834, emd-1134, emd-2873, emd-5429, emd-1835, emd-2772, emd-1832 and emd-1526. [8]–[19]

The research of biological annotation was searched in the following databases:

- **Complex portal:** The problem that we have seen with this database is that there is not distinction between strains, and then the related sequence probably does not match 100% with our protein, as we could check it, sometimes using BLAST[5]. With these databases we have some advantages but also some disadvantages. On the one hand

[5] BLAST: Basic Local Alignment Search Tool is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences.

with a single search we obtain detailed information about the components that composes the complex, but on the other hand we did not find the interested specific strain. See an example in:
https://www.ebi.ac.uk/intact/complex/details/EBI-913722

- **Fitted PDB model:** PDB fitted structures information was very poor and there were not many links between other databases. This occasioned that the segmentation procedure was more complicated than in other cases, because of the absence of a template.
- **EMDB:** As the PDBe portal, we could see that the upload information was very poor. In future this could be improved, in order to generate more qualified information on the database.
- **Related Article:** Most of the related structures are not updated on the database. For this reason we had to search them on the linked articles in order to have more information to segment the structure.

3.1.2. Chaperones (e.g. GroEL)

Chaperones are proteins that provide favourable conditions for the correct folding of other proteins, thus preventing aggregation. Newly made proteins usually must fold from a linear chain of amino acids into a three-dimensional form. Chaperones belong to a large class of molecules that assist protein folding, called molecular chaperones. The energy to fold proteins is supplied by adenosine triphosphate (ATP). The following diagram characterizes the dynamic of these molecules *(Figure 3.1.2.1):*

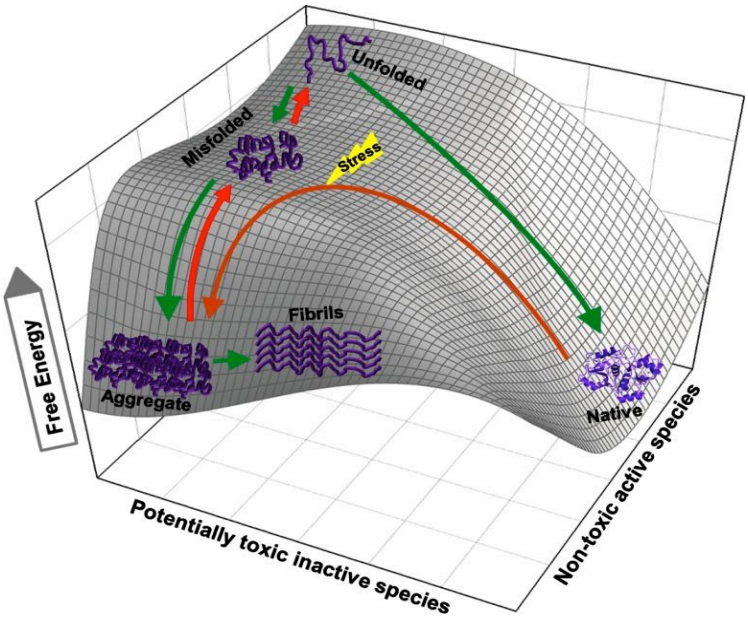


Figure 3.1.2.1. Lifeflow of proteins. Potentially toxic inactive species versus the non-toxic active species related with the energy released

Most molecular chaperones fall into five families of highly conserved proteins: the Hsp100s (ClpB), the Hsp90s (HtpG), the Hs70/Hsp110 (DnaK) and **Hsp60/CCTs (GroEL)**. On this project we have worked with the family of GroEL Chaperone.

GroEL is a chaperone complex, is also related with a smaller complex called GroES. The GroEL complex can be formed with different amount of subunits depending of the specie that we are looking for. They are composed with double-ring complexes with seven to nine 60 kDa subunits per ring and function by enclosing substrate protein for folding to occur unimpaired by aggregation. Two groups of chaperones complexes are distinguished; Group I and Group II.

- **Group I (Cpn60s or Hsp60s)**: found in the bacterial cytosol (GroEL in *Escherichia coli*) and in eukaryotic organelles of endosymbiotic origin, such as mitochondria and chloroplasts. It is a two-ring conformation which consist of eight paralogous subunits per ring.
- **Group II (TRiC or CCT):** present in Achaea and in the eukaryotic cytosol, and is also know as TCP1 ring complex (TRiC) or chaperone containing TCP1 (CCT). They are composed with double-ring structures with eight- or nine fold symmetry and in most cases consist of two to three homologous subunits per ring.

Also there is shown to contain three paralogous subunits, α, β and γ. This subunits produce on the structure a different assembly. [20]

We've worked with the following EMDB structures: emd-1042, emd-1046, emd-1047, emd-1080, emd-1081, emd-1180, emd-1181, emd-1200, emd-1202, emd-1203, emd-1286, emd-1289, emd-1295, emd-1296, emd-1297, emd-1298, emd-1396, emd-1397, emd-1398, emd-1457, emd-1458, emd-1531, emd-1544, emd-1545, emd-1546, emd-1547, emd-1548, emd-1587, emd-1588, emd-1960, emd-1961, emd-1962, emd-1963, emd-1997, emd-1998, emd-1999, emd-2000, emd-2001, emd-2002, emd-2003, emd-2221, emd-2325, emd-2326, emd-2327, emd-5001, emd-5002, emd-5043, emd-5137, emd-5138, emd-5139, emd-5140, emd-5143, emd-5145, emd-5148, emd-5154, emd-5157, emd-5159, emd-5244, emd-5245, emd-5246, emd-5247, emd-5248, emd-5249, emd-5250, emd-5258, emd-5336, emd-5337, emd-5338, emd-5339, emd-5340, emd-5391, emd-5392, emd-5395, emd-5396, emd-5640, emd-5645, emd-5646, emd-5767, emd-5768, emd-5769, emd-5770, emd-6422. [21]–[23], [23]–[50]

The search of the biological annotation was made in the following databases:

- **Complex portal:** Using this portal we only obtained a single result, "accession number: EBI-8769099", which belong to Escherichia coli (strain K12). Then we can not base the search only on this database, as there is not enough information to build a proper annotation.
- **Fitted PDB model:** we could found for almost all the EMDB structures a fitted PDB model, that helped to proceed with segmentation. This information was found on the

same PDB database and also EMDB and UniProt related fit model. Then was necessary to compare this obtained information in order to see if it was the correct one or not.

- **EMDB:** The information found on the EMDB portal was quite complet, as the links with other databases were properly annotated. In some cases there was not information about the fit model and other characteristics but they were completed with the information provided on the articles.
- **Related article:** We have found some information related with the PDB code, or segmentation organization procedure, that was useful as a guide to segment the structure.

3.1.3. Ribosomes

The ribosome is a complex molecular machine found within all living cells, that serves as the site of biological protein synthesis (translation). Ribosomes link amino acids together in the order specified by messenger RNA (mRNA) molecules.

The structure of the ribosome consists of two major components: the small ribosomal subunit, which reads the RNA, and the large subunit, which joins amino acids to form a polypeptide chain. Each subunit is composed of one or more ribosomal RNA (rRNA) molecules and a variety of proteins.

Depending on phylogenetic order this will have different structures:

- **Prokaryotic ribosome (bacteria and Achaea)**: is also known as 70S ribosome, is composed by a 30S subunit (Small) and the 50S subunit (Large)
- **Eukaryotic (cytoplasm, mitochondria and chloroplast),** depending on each location we can distinguish:
    - **Cytoplasm (80 S ribosome)** which is composed of:
        - 40S subunit (Small): 18S rRNA
        - 60S subunit (Large): 5S, 5.8S and 25S rRNA
    - **Mitochondrial (55S ribosome)** which composed of:
        - 28 subunit (Small): 12S rRNA
        - 39S subunit (Large): 16S rRNA
    - **Chloroplast (70S ribosome)** which is very similar to prokaryotic ribosomes.

Both subunits contain three binding sites for tRNA molecules. The A site binds the aminoacyl-tRNAs that is about to be incorporated into the growing polypeptide chain, the P site positions

the peptidyl-tRNA and the E-site is occupied all deacylated tRNAs before they dissociate from the ribosome.

Of central interest are mechanisms of peptide bond formation and mRNA decoding, which are crucial processes in the elongation phase of the protein synthesis by the ribosome. During this phase of the nascent polypeptides are elongated form N to the C terminus by the addition of one amino acid at a time. This process is facilitated by two proteins factors: elongation factor Tu (EF-Tu), which facilitates the delivery of aminoacyl-tRNA to the A site of the ribosome, and the elongation factor G (EF-G), which promotes the translocation of the tRNAs and associated mRNA from their positions in the A site and P site to the P site and E site, respectively, and dissociates and the previously bound E-site tRNA. [42]

The Ribosomes that we segmented on this project were the following: emd-3133 (Large subunit of a prokaryotic ribosome), emd-2847 (Prokaryotic ribosome), emd-2913 (Small subunit of a Mithocontridal Eucaryotic Ribosome) and emd-2914 (Mithocontridal Eucaryotic Ribosome).

It was impossible to proceed with the segmentation without a fitting model. All the ribosomes treated had a very precise atomic model structure, that was directly related with the resolution of the EM map then it was only possible to segment very high resolution structures < 3 Å. The segmentation of this type of structure is very slow and detailed; it is required time and patience to segment it (around 60 h each of them). The proteins that composes the ribosomes are not constants, they vary depending on its state and the interaction with other molecules/proteins, is not yet established a constant pattern with the proteins that composes its structure. This was also a problem to identifying the different ribosome proteins. [51]–[53]

- **Complex portal:** We did not find any information on the complex portal, as the ribosome is a complex that is composed with an unstable number of subunits. Depending on the state of the complex and the related proteins that interact with it on this moment, the ribosome has a different amount of molecules. Then that is why there is no information on this database, because nowadays they donnot know the constant proteins that compose this ribosome.
- **PDBe**: Most of the information was provided by Fitted PDB model, but we have seen that some of the annotation of the structure was annotated randomly. We had to compare this annotation with the related article to see if there was some coincidence.
- **EMDB:** Was used to see the components of the structure, but to see the subunits that compose the ribosome that we have found on the PDBe database, following the information on the atomic model chains.

- **Rfam:** One of the problems with the RNA annotation is that there is not a properly RNA database developed yet. That is a problem regarding the annotation of the RNA segments and its identification. The related database is RFam, in which there is a collection of RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models (CMs). Mostly we could see this problem on the identification of the different RNA that composes the ribosome. Like the P-site, A-site, E-site and other tRNA, we will expect that from each of these three t-RNA we will obtain 3 different RFam codes with its own personal information, but that is not the case, we only could found a single RFam code for the three t-RNA structures. The same happened with the RNA segments of each of the Ribosome. We could not find all the RFam RNA ribosome related, like the 12S and 16S mitochondrial RNA subunits, then we could not annotate as we would be expecting.
- **Related article:** Was very important to read properly these related articles, in order to understand the structure and see if there were some mistakes on the related databases

3.2. Annotated information

Annotation information was one of the main steps of the project. It is very important to make a very good annotation for the identification each of the segments. This information will be linked on the EMDB-SFF application. The biological annotation on the EMDB-SFF will be linked on each of the labelled segments by an interactive way, where the user, by pointing with the mouse to the interested segment, will be directed to a secondary window with the related biological annotation.

Sometimes it was difficult to search the related annotation, a table with all the description of the annotations and the advantages and disadvantages of this information (*Table 1*).

Table 1. Biological information annotated on SFF.

Annotation	Description	Advantages/ Disadvantages
Name of the segment	Biological name of the segmented component, this information was provided by the related article or databases like UniProt or PDBe.	Brief description of the segmented component but also a very general term
UniProt code	Corresponds to all the related information of a protein. In these database are annotated a high-quality and freely accessible resource of protein sequence and functional information. This code makes more accessible the related information of the segment.	The related information is very extended, but on the other hand, only with this code we can obtain a lot of information
Pfam code	Corresponds to the protein domain identification. This is specific and unique for its specie and protein subunits/domains.	It is a very concrete information, that can be very useful in order to identify the proteins domains. But on the other hand is difficult to obtain.
Rfam code	Is a RNA database. We will use it on cases that we want to identify a RNA segment, as the ribosomes	Is a very useful but poor database for RNA segments. Definitely there is a problem with the Rfam that needs to be improved, as sometimes there is no distinction with different RNA segments
GO code	The Ontology code is the identifier of the cellular components, these is specific and unique for each of the components on a cell	For the description of a tomogram is necessary to combine several GO codes. We will not use this code for the biological annotation of proteins
PDBe	As we explained previously, you can search information about high-quality macromolecular structures and related data, where you can found the fit model of the EM structures and chain structure information, as the different components that composes these one and related annotation.	Sometimes the stored information is annotated randomly or can be poor. But on the other hand you can obtain a lot of information in a single application.

Table 1, was build thanks the experience on this project. Where are annotated all the disadvantages and advantages of the best biological identification.

We concluded that a combination of all this biological annotation was needed to identify each of the segmentation regions, as some of this data are very general and other ones are more specific. Ideally we would use a single ID for the identification of a segment, but is complicated to decide which is the best way to do it.

3.3 Chimera – Proteins segmentation

The main work of this project was to create a segmentation data set for the SSF. This was possible with the help of Chimera. This program is an extensible program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, etc. You can upload a EMDB 3D model and with the help of some of the provided tools proceed with the segmentation. Once the segmentation is done the file is saved a Segger file (.seg) or an HDF5 file and the converted to an XML file.

It was used "HDFView 2.11", to see all the information in the saved file. It can be easly download from the official website. [54]

The segmentation process was structured in different steps:

1. Volume viewer adjustment
2. Segmentation step
3. Fitting model
4. Grouping and ungrouping
5. Attributes annotation
6. Saving

3.3.1. Tools

To work with a the EMDB structure and be able to segment it we had some preference on the Chimera tools. The Chimera tools most used:

- **Edit window:** Window where you can see and select the regions that you want to work with. [55]:
- **Volume Viewer:** Tool for visualizing volume data, 3D numerical data sets such as electron density maps. The data can be shown as solid or mesh isosurfaces (contour surfaces) or as partially transparent solids [56].
- **Model Panel:** Lists the models in Chimera and conveniently enables many operations upon them. Each file of atomic coordinates opened in Chimera becomes a model with an associated model ID number. Surfaces and other types of models also have ID numbers. Some tools in Chimera create models that are hidden from the Model Panel, however a molecular structure file may contain multiple sets of coordinates for the same set of atoms. From each crated mask that is made appears as new model on the table.

The model panel window shows *Figure 3.3.1.1.*:

- model ID number
- model colour (a colour well)
- whether Active (movable)
- whether Shown (display-enabled)
- Name (model name)

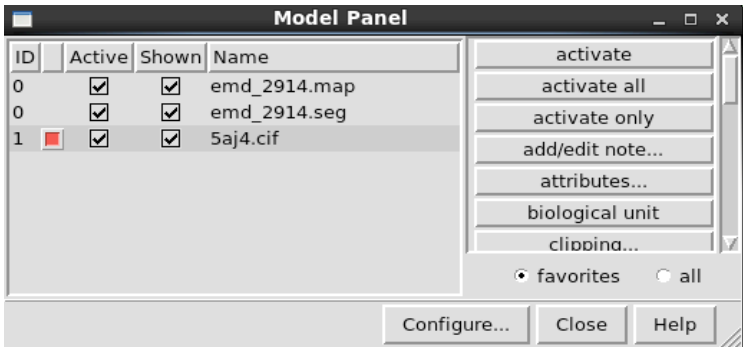


Figure 3.3.1.1.. Model Panel window. is an example of a model panel of a EMD-2914 and a PDB structure in this case 5aj4.

- **Fit to Segment:** rigidly fits atomic structures into segmentation regions from Segment Map. [55]
- **Segment map:** this tool allows us to do partitions volume data to create a surface model with one or more *segmentation regions* (specialized surface pieces) shown in different colours. Along with the Fit to Segments tool. [55]
- **Command line:** is one option that provides Chimera when you can execute some actions by a command line window, located at the bottom of the main window.
- **Segmentation region Attributes:** With this tool you can see all the names and information of the created segments; also you can add extra information as the annotation name of the segment. This extra information will be automatically saved on the same segmentation file, making it more accessible for the SFF process.

On *Figure 3.3.1.2* we can see one example of working window, where all the main tools are shown on the screen.

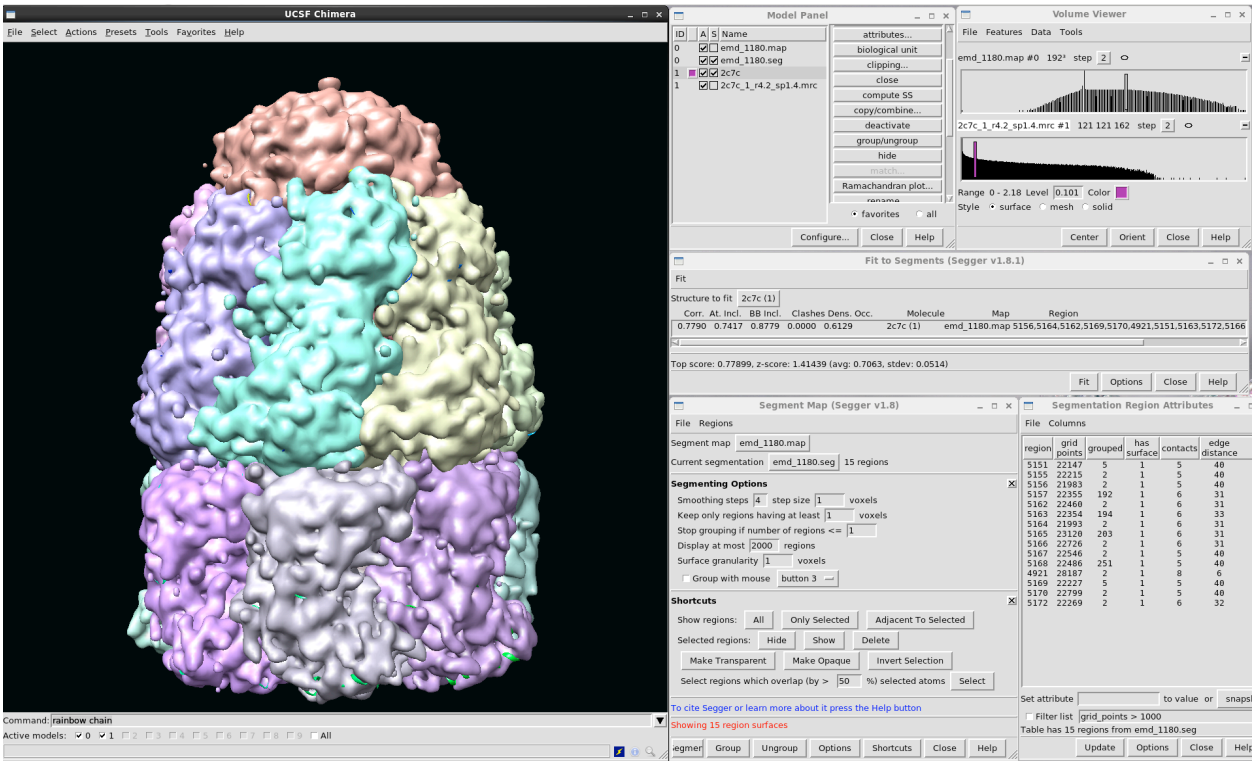


Figure 3.3.1.2. Chimera tools. Windows tools used to proceed with the structure segmentation. The image belongs to a GroEL and GroES complex with its components and subunits with an EMDB code EMD-1180.

3.3.2. Volume viewer adjustment

Volume Viewer is a tool for visualizing volume data, 3D numerical data sets such as electron density maps. The data can be shown as solid or mesh isosurfaces (contour surfaces) or as partially transparent solids.

With this tool you can edit the volume map on the amount of interest depending on your work, before preceding the segmentation. You can edit this working volume by two ways: manually by the volume graph *(Figure 3.3.2.1.; A)* or automatically adding the volume number on the bottom window *(Figure 3.3.2.1.; B).*

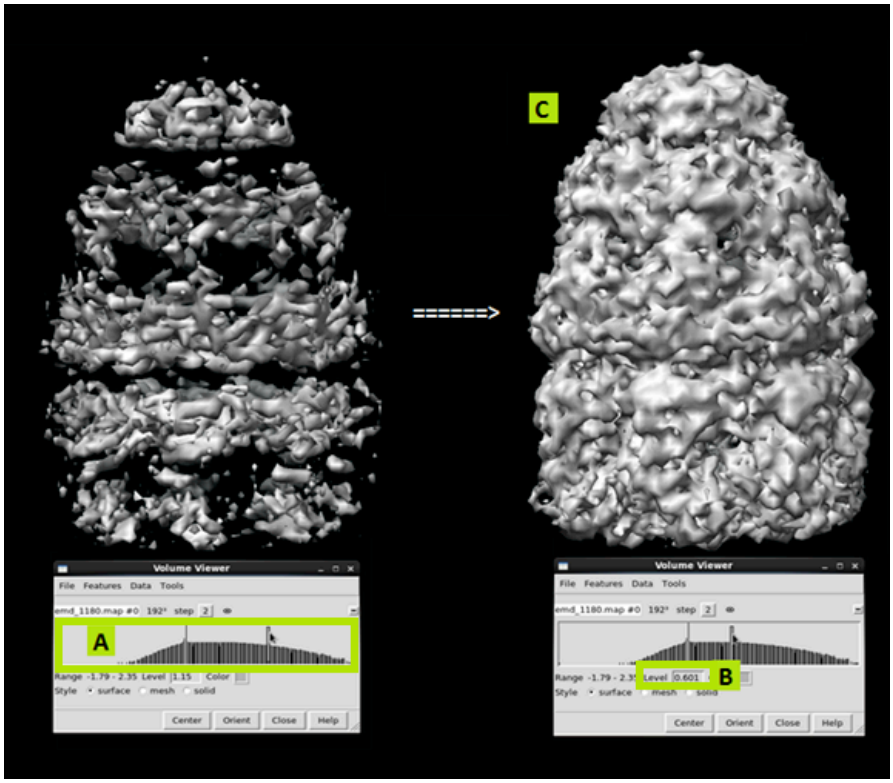


Figure 3.3.2.1.: Volume viewer results: (A) Manual way to edit the volume of the map, hover the mouse over and slide the bar to the right or left, (B) Automatic way to change the volume by adding the interested volume amount on the bottom window. (C) On the left image we can see the change of this contour level on the volume structure. This structure corresponds to EMD-1180

You have to options to establish a contour level, you can modify it by your own criteria (with a scientific basis), or you can use the EMDB service were is annotated the recommended contour level for each of the structures. Sometimes this recommended volume is not the correct one, as this one is calculated by human criteria, and we can't rule out the possibility that maybe is a wrong data (*Figure 3.3.2.2*). Then we always checked the result before the segmentation. We will see if these one escapes from the estimated limits when the volume is visualize by Chimera.

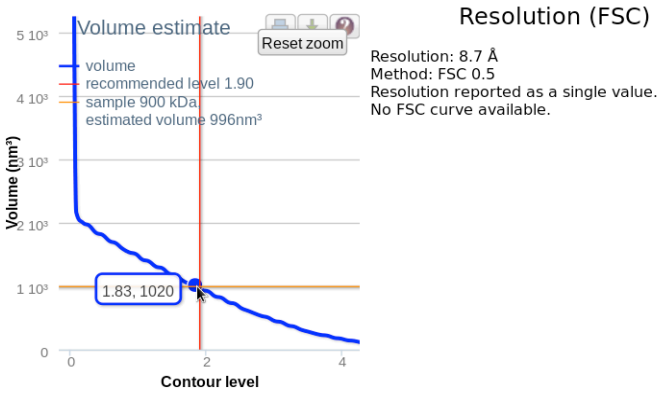


Figure 3.3.2.2.: Recommended contour level: Information found on the details of one EMDB entry, this case EMD-1180. To found the recommended contour level we have extrapolate the volume with the recommended sample volume line of the Volume estimate graph.

Many of its features are also implemented as the command volume, and several related tools can be also accessed from the Volume Viewer Tools menu.

3.3.3. Segmentation step

When contour level is already modified we start to segment the structure by the following route: Tool > Volume > Segment map > Segment

The segmentation task is easy but slow compared with the others. When you press the segmentation button the structure is divided by random voxels[6], next step is to try to reconstruct the order of the voxels by grouping and ungrouping the structure. Doing this step blindly is very difficult that is why we need information about the components that composes the structure and a fitted model to use like a template.

They are others situations where the worked structure has a very noisy map. Then the resulting segmentation is a "not real" segmentation, the subunits cannot be distinguished because they are hidden or are not well defined as the presence of artefacts. See some examples in *Appendix 2*. The procedure to solve this problem has to be applied once you have the interested volume and you want to denoise the structure, following the next route (*Appendix 2-Figure 10.2.1.)*:

Tool > Volume > Segment map > Options > Annotate the interested degree on "Keep only regions having at least *#border_level* voxels"

[6] Voxel: Represents a value on a regular grid in three-dimensional space.

The voxels degree is an estimation of the analyser, and this one can be accurate with its criteria. This function erases the segments/regions that are composed with a lower level from the border value that the analyser has established previously. Then the segments/regions that have an inferior number of voxels from the border value will be eliminated from the structure, ending with a denoised and clear map.

3.3.4. Fitting model

For this step, we used all the stored information and biological annotation stored in the Excel sheet (Google Drive). After the research some of the information founded was useful for the segmentation, but the key point was to find a related fit model in order to use it as a template for the segmentation.

Using the fitted model, the structure map was fitted by the Chimera tool "Fit to segment" on the EMDB structure. By the next route:

Tools > Fit to segment > Select the interested PDB structure > Fit

Sometimes when the PDB model is loaded is automatically fitted on the structure, then in this cases is not necessary using this tool.

Also we can be on the hypothetical case that we have more of one PDB model for a EMDB code, then on these cases it is needed to fit the segment separately, by selecting the interested volume map and then fit the related chain.

A good example could be EMD-1547 that corresponds to a GroEL structure composed by a GroEL complex, a mutant GroES (gp31) and a peptide on its core. On this case was needed to fit each of the structures separately and modify the PDB structure (2cgt) in order to obtain a GroEL structure without the WT GroES subunit, and then replace it for the 1g31 code, *(Figure 3.3.4.1.).*

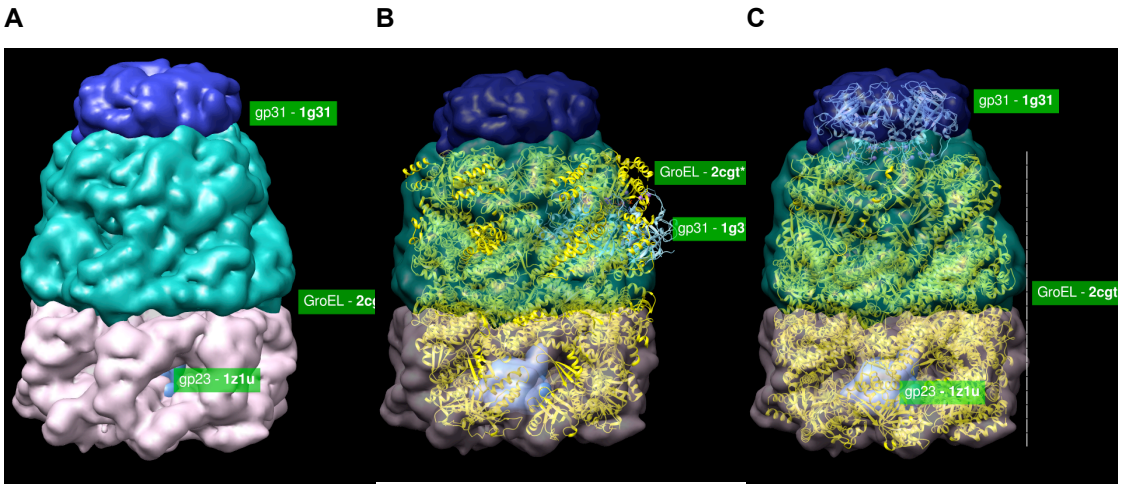


Figure 3.3.4.1. Fitting procedure with more of one PDB structure. (A) 3D structure of the Chaperone emd-1547, which is already segmented, these one is composed by 4 subunits: two that belongs to GroEL, one that belongs to the mutant GroES (gp31) and a last one that is a unfolded peptide (gp23). (B) The atomic model is added on the session file, as you can see these one have a random distribution and has to be fix. Next step (C) is to fit these structures on the EM map correctly.

3.3.5. Grouping and ungrouping

The Grouping and ungrouping is the most delicate and slow with stage. Depending on which structures we are working with, the grouping stage may lost from 1 hour to 80 hours. This range of time depends on several items, like the resolution of the EM model, the provided information and the existence or not of a fitted model. Also it is seen a singular difference on some cases that the resolution and the information provided was more or less the same but then the segmentation was quite difficult on one of them, that could be because the utilization of a different detector by the EM. A example of this phenomenon could been on EMD-1457 and EMD-1458,

On this grouping/ungrouping step was needed to reach a similar structure as the one shown on *Figure 3.3.5.1.*, a Chain with an EM structure around it.

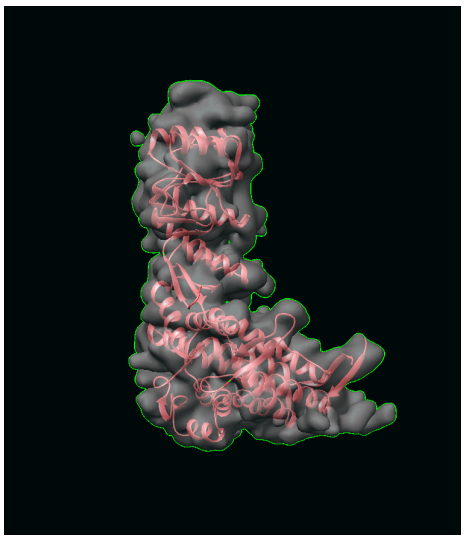
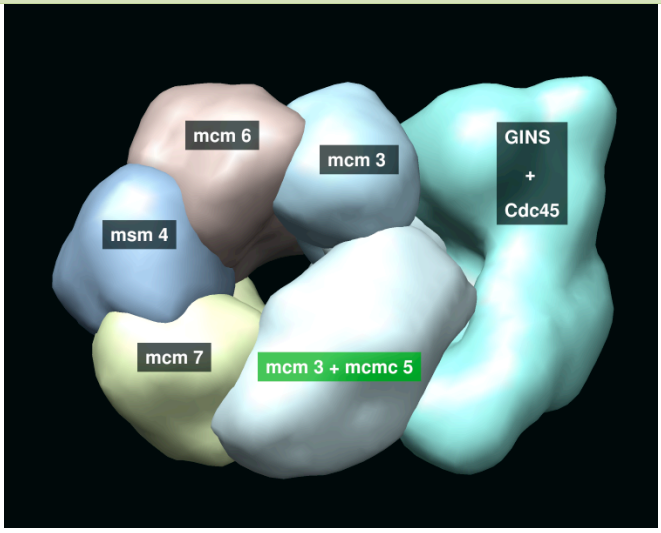
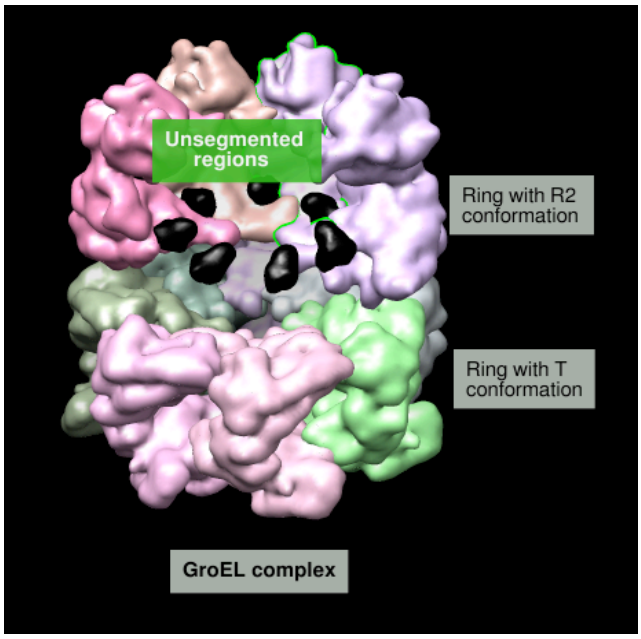


Figure 3.3.5.1. Expected fitting one to on (chain and segment). The segmentation structure surrounds all the PDB chain, as we would be expecting. Structure belongs to EMD-1181

Sometimes the EMDB structure does not fit that well. In these cases we will have two options; create a single segment with this unfitting part or joining the two problematic segments in a single volume, these differences are shown in *Table 2*.

Table 2. Images and examples of grouping and ungrouping procedure.

	Images and Examples of Grouping and Ungrouping problems	
Joining two problematic segments	 *Figure 3.3.5.2.: Segmentation of MCM complex: Image belongs to EMD-1832 structure.*	As you can see on this image (*Figure 3.3.5.2.*) we could not segment all the different components of the MCM complex, as we can appreciate on the green label there are two segments that are joined, because the map had a very bad resolution, 28 A, then the partitions of the structure are made according to this bad resolution. There is a close relationship between the resolution of the structure and the segmentation.
Create a single segment for the unfitting chain	 *Figure 3.3.5.3.: Segmentation of GroEL complex: Image belongs to EMD-1999*	In this case we could not segment properly the difference subunits that composes the GroEL complex. We have found some parts of the structure that the segmentation was very difficult because ofthese regions. On this example we decided to distinguish this regions on a separate segment labelled with black. On the other hand, sometimes we have a map with a high resolution but we can have some problems with the segmentation. In these cases, it can occurred that some parts of the fitted model does not fit properly with the 3D structure, then in order to separate these problematic regions, they are distinguished in a separate black segment, as is shown on *Figure 3.3.5.3.*

Using the tools that are provided by the "Segmentation map tool" we have played with different commands to perform the segmentation. On *Figure 3.3.5.4.* you can see an example of some of the options that you can use, hiding and making transparent some of the segments to facilitate the procedure and access to some components that are hidden by the surface structure.

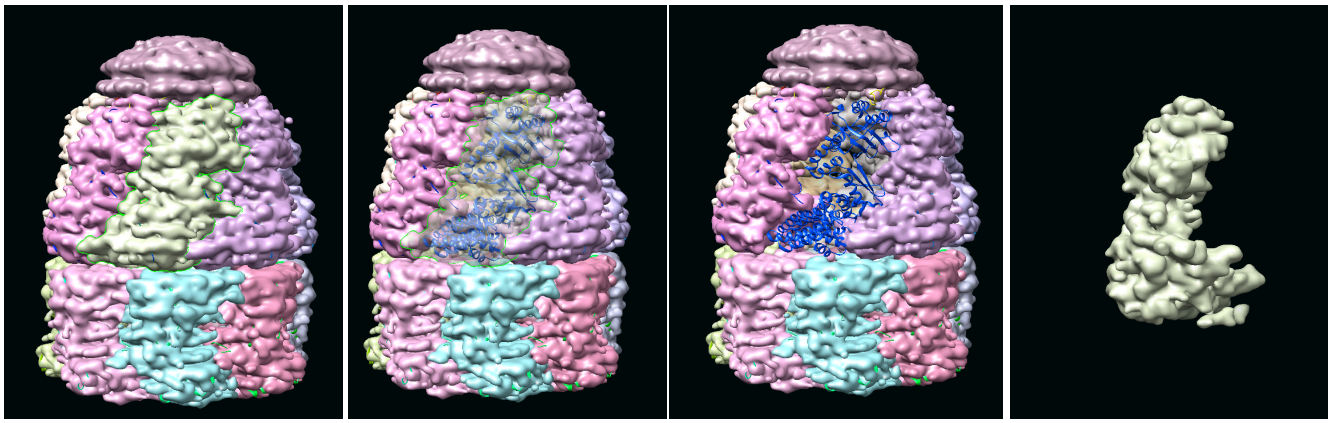


Figure 3.3.5.4.: Essential visualization segmentation tools options: Chaperone EMD-1181 on the segmentation procedure with Chimera program, where: (A) All the structure, (B) Make transparent a segment, (C) Hide a segment and (D) only visualize the selected segment

Also you can have others segmenting complications related with gaps on the structure, like the once observed on *Table 3* and *Table 4*.

Table 3. Images and examples with gaps and unfitting problems I. Unknown gaps on the structure (Low-resolution)

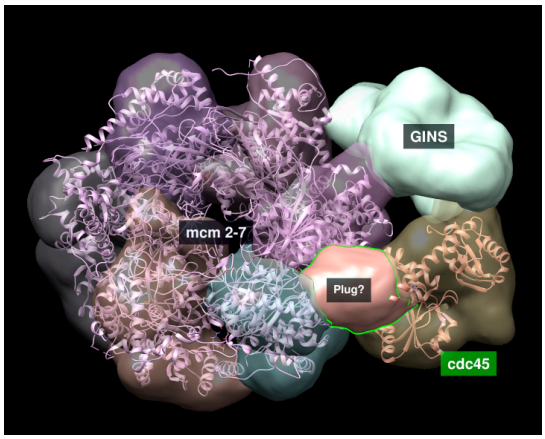
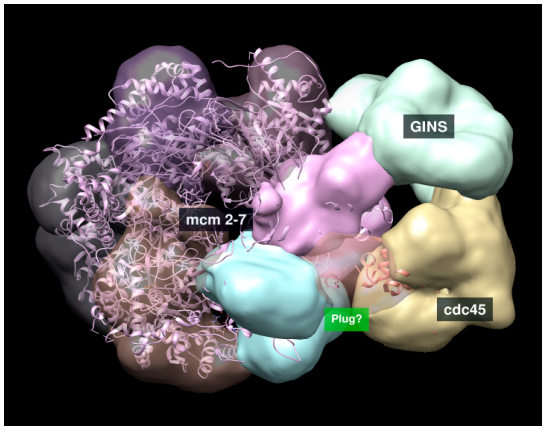
	Images and Examples with Gaps and unfitting problems I	
Unknown gaps on the structure (Low-resolution	You can find gaps/holes on the structure, where you do not have a fitted model and you do not know what could be inside of the structure. This problem can be found in low-resolution structures >15 Å. A representative case could be the EMD-2772 entry with a resolution of 17,8 Å , which corresponds to a MCM structure. First of all what we can see on this structure is that we did not found a PDB model of GINS, then the segmentation was more complicated as we did not have any template that indicates the limits of this one. Also we could see other problems like the following:	
	 *Figure 3.3.5.5. MCM with a problem that lies on the cdc45 subunit*	In this case we had the fitted model but a gap was generated inside the segmented structure. On this space a smaller structure could be fitted inside, then we cannot reject the possibility that a cdc45 is associated with another structure, in its hypothetical volume. We could hypothesize also that the problem remains on the fitted model, probably is not the correct one for this volume.
	 *Figure 3.3.5.6.: MCM with a problem that lies on the plug subunit*	We could observe that the inside part of the Plug 3D structure is empty; we did not find any related PDB structure. Reference [10] is mentioned a possible plug structure between the MCM2 subunit and the cdc45, but there is not any information about its function and components. Then in this case we decided to distinguish this component on a differentiated structure, as we know about its existence but we do not know a lot about it.

Table 4: Images and examples with gaps and unfitting problems II: Predictable Gaps on the structure (Large structures)

	Images and Examples with Gaps and unfitting problems
Predictable Gaps on the structure (Large structures):	Sometimes the fitted model is not completed, the structure finish with an open end because some of the atoms of the molecule where not found when the atomic structure was build. What we can see on these cases is that the EM structure generates an envelope of the structure but there is not an atomic model (skeleton) inside. We could see this phenomenon on the EMD-2914 that belongs to a ribosome. Probably this phenomenon occurs because of the big size of the structure. On this case we can see that the RNA that composes the 16S rRNA subunit of the ribosome is not completed, as we can see in *Figures 3.3.5.7. and 3.3.5.8.*, but there is an envelope around it that indicates us that perhaps there is a problem with the atomic structure. Also we can see that the end of the pink chain in *Figure 24* ends with a "*discontinuity line*", and does not finish with a loop or an other typical ending of atomic structures. This suggest us that there is a lack of information.

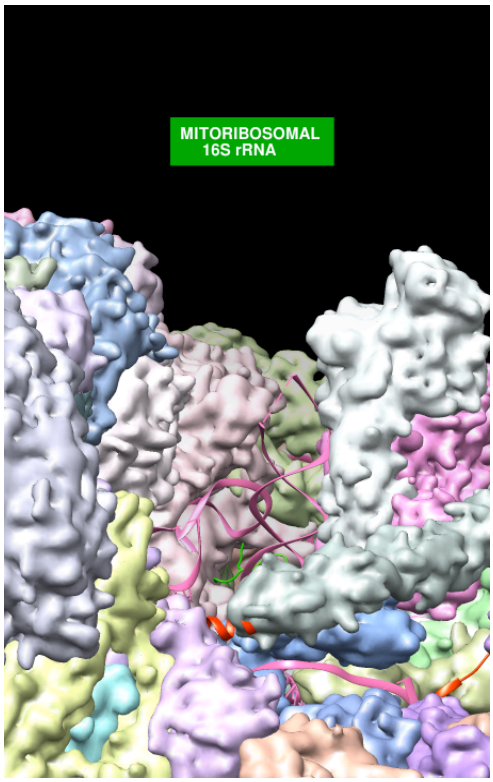


Figure 3.3.5.7.: Skeleton of EMD-2924 Ribosome, where a discontinuity is observed on the coordinates that composes the structure

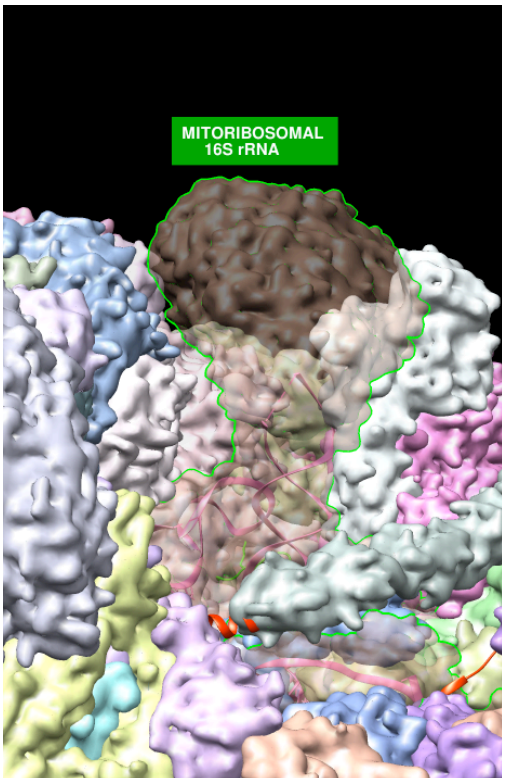


Figure 3.3.5.8: Envelope of the Skeleton of EMD-2924 Ribosome, where a gap is observed on its inside

3.3.6. Annotation attributes

When you download all the information from the Chimera - segger file without the biological annotation properly annotated on it, we have observed that these procedure involved a hardest future work, where a manually annotation had to be done by a programmer. On the beginning of this project all the annotation was added by this system by a programmer, but with structures such as ribosomes which have up to 100 components the task was more complicated, then we decided to find a better alternative to facilitate the task to the programmers.

Finally, it was found an attributes annotation on the "Segmentation map tool". On this one you could see all the segments with their respective codes and details, such as the region on the map, where is located or the grid points that composes the segment. By default on this table are shown some of the columns but you can add or edit these ones. On this project was added a column called "Segment name" on the attributes table *(Figure 3.3.6.1).* In a future, it could be added other columns as the related PFam or the UniProt codes.

To add this new column, the following procedure was performed:

- Select on the table the segment that you want to add more information
- Add on the "Set attribute" function the name of the new column
- Next step is to put the interested value or name with the extra information, on the "value" box

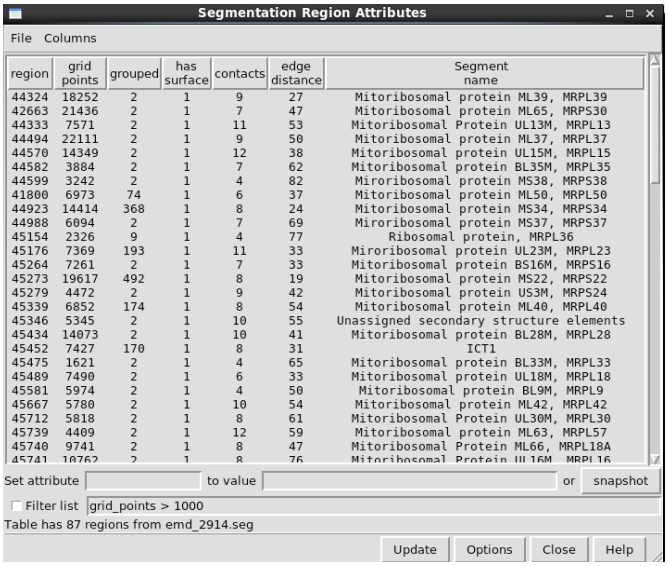


Figure 3.3.6.1. Segmentation region attributes. Example of the biological annotation added on the Segger-Chimera file, on a new column called "Segment name". On this one you can see the complete name of each of the segments that composes the ribosome. EMD-*2914*

4. RESULTS

At this moment, we cannot show a final result of the SFF project as it is already on the development process. In this project, we built several segmentation sets for the SFF application with the aim to guide users to do similar procedures related on the project, and share it with all the interested community.

The final result of the project was stored on a database (Google Drive link on *Figure 4.1)*. There can be found all the worked segmented structures and its biological annotation, including the Helicases, Chaperones and Ribosomes, with some representatives notes necessaries to proceed with the segmentation.

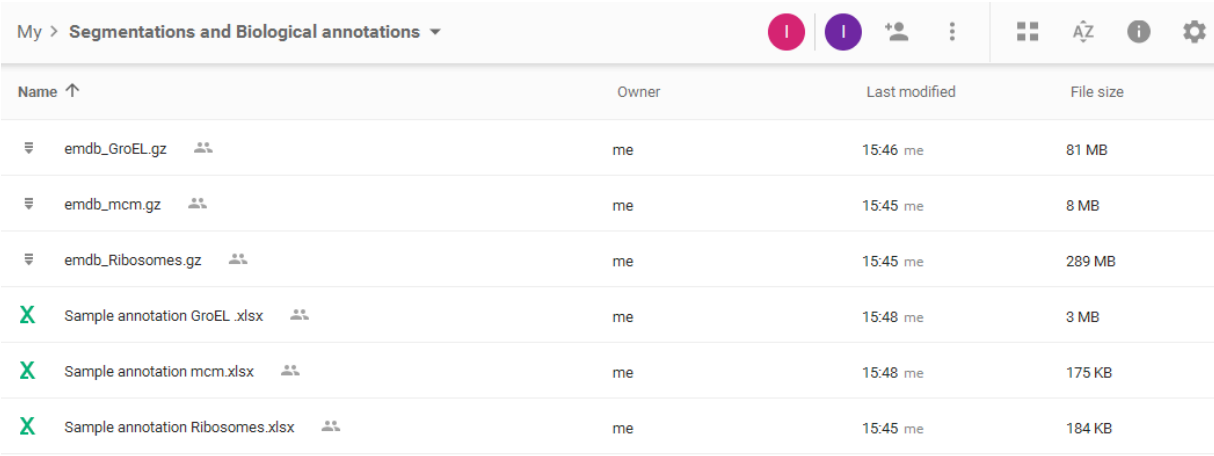


Figure 4.1. Google Drive Segmentation and biological annotation. Structure of the Google Drive, you can find first a zip folder with the segmentation structures and on the other hand an excel sheet with all the related biological annotation. You can find these files on the following link: *https://drive.google.com/open?id=0B_GxcVo8xvkuaGtfYkw4R1NhMEU*

Regarding the segmentation of EMDB structures with the Chimera program we could conclude that is very effective and useful software to segment 3D structures. We could use a lot of tools to segment and to edit the stored information. Also to save the file is easy, we didn't found any problem with it, and Google drive could support with any problem the Segger files, then that was very useful to share with the group components.

On the other hand, all this work is already available on the PDBeShape portal (*Figure 4.2.),* where you can download the structures of some of the structures segmented. On the following link you can find an example of a segmented GroEL, and see how this application works.
http://wwwdev.ebi.ac.uk/pdbe/emdb/pdbeshape_dev//volume_details/EMD-1046/

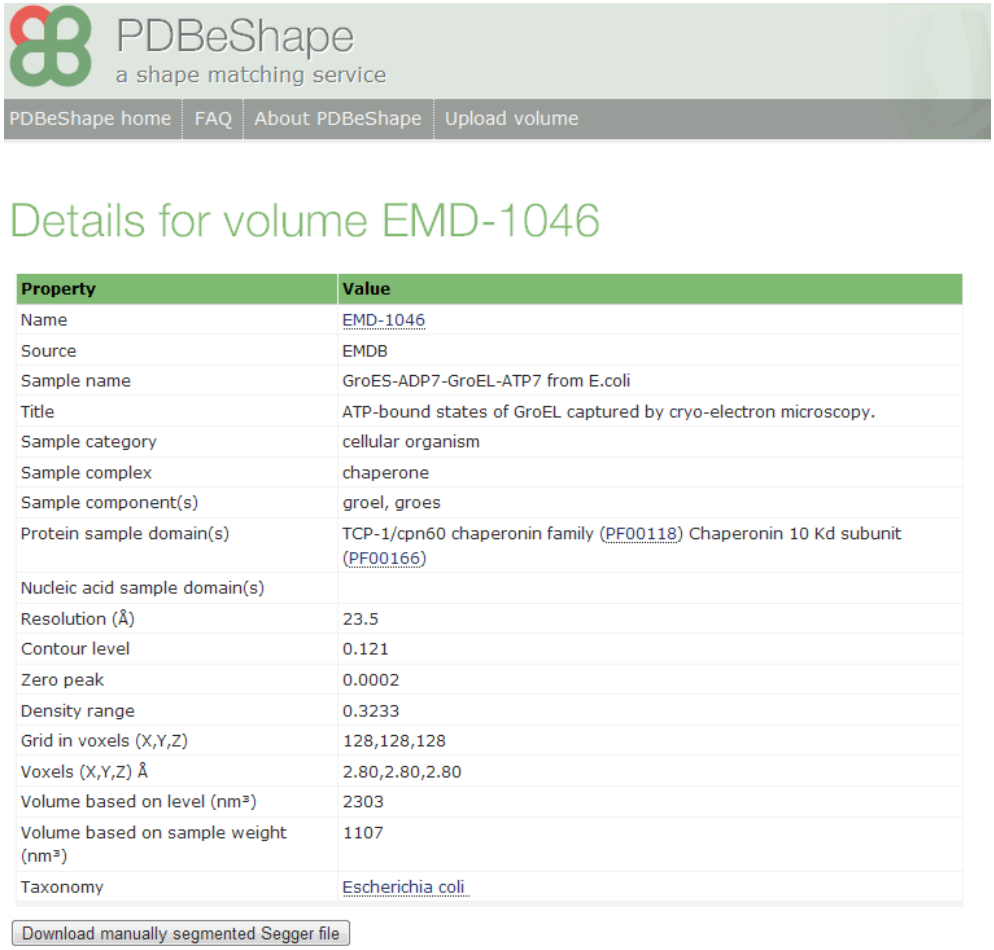


Figure 4.2. Segger file on PDBeShape. In this table it is shown the properties about this chaperone, and related links as the Pfam domains. On the bottom of this table download button is shown, where the user can obtain the segmentation Segger file. Once you have this file downloaded the user can visualize it by Chimera.

5. CONCLUSIONS

As can be seen on chapter 4 - "Final Result", the proposed work was a success, my company congratulated me for the job done, and they expressed their satisfaction with the segmentation sets done.

Regarding the biological annotation, we concluded that a combination of all the identifiers of "*Table 1- Section 3.2"* was needed to identify each of the segmentation regions. Some of this data is very general and other ones are more specific then a balance between the degrees of information has to take into account. Ideally we would use a single ID for the identification of a segment, but is complicated to decide which is the best way to do it.

They still wanted to introduce further improvements in the SFF application. That is why they organized an Experts Workshop on December of 2015 to discuss several points of the project. In this project the following improvements were proposed:

- To identify each of the cellular components of the segmentation with an Ontology code. This code is unique and specific for each of the segments and structures. Also you can annotate with this type of codes, as: experimental method, cell cycle, shape and phenotype of the component, etc. In a future, it could be useful to do the annotation by this system.
- It was also mentioned that, it would be very useful to visualize the segmentation and the annotation of the SFF-EMDB, additionally on an external program in order to edit and modify the structure, like IMOD or Amira. A collaboration with IMOD software with the purpose to visualize this type of segmentation and annotation is also in course.

In this project we didn't work with an automatic procedure, but we thought that in a near future this procedure could be implemented. With more segmentation sets, we could use them to automatize the system, or search an average structure. We were interested especially with ribosomes, as nowadays it is not known the constants regions on its structure.

Finally, with this automation will be save a lot of time and the opportunity to invest it in other sections of the project. Also we will obtain a highest amount of segmentations in a shorter period of time.

8.1. Acknowledgements

Firstly, I would like to express my sincere gratitude to my UVic-UCC advisor Josep Maria Serrat Jurado for the continuous support of my Final degree project, for his patience, motivation, and his guidance for the writing of this report.

My sincere thanks also goes to Dr. Ardan Patwardhan and Dr. Gerard Kleywegt, who provided me an opportunity to join their team (PDBe) as intern for 3 months, and who gave me access to bioinformatics applications and creation of new database tools. Especially without the precious support of Dr. Ardan Patwardhan, my advisor on the EBI-EMBL company, it would not be possible to conduct this project, also I'm extremely thankful to his guidance and help for my future career.

I thank my fellow colleagues: Paul Korir, Ingvar Lagerstedt, Andrii Iudin, Nurul Nadzirin and Mandar Deshpande, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives and for these relaxing but stimulating tea times where we could disconnect and talk a wide variety were I could learn a lot about life and with the kind of people that I was working with

Besides, I would like to thank the rest of my working PDBe team colleges who helped me to feel like home, this group is a gorgeous team where I could feel part of these big family. Also I would like to mention the input of Dr.Martyn Winn who collaborate on these segmentation project and he trust on it to integrate it on the PDBeShape application.

Last but not the least, I would like to thank my family: my parents; David and Teresa, and to my siblings; Berta and Joan for supporting me spiritually throughout writing this project and his support on my stay on UK and for the help given during the four years of my degree. I also wish to express my gratitude to my couple Pau, for the enormous patience he had during my stay abroad, and his complete support to keep doing my best and never give up.

9. BIBLIOGRAPHY AND WEBGRAPHY

[1] PDBe team, "About PDBe." [Online]. Available: http://www.ebi.ac.uk/pdbe/about. [Accessed: 15-Dec-2015].

[2] C. W. 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, "No CryoEM, EMDataBank.org: unified data resource forTitle," *Nucleic Acids Res.*, 2011.

[3] EMDB group, "Details for volume EMD-1046 - PDBeShape." [Online]. Available: http://wwwdev.ebi.ac.uk/pdbe/emdb/pdbeshape_dev//volume_details/EMD-1046/. [Accessed: 15-Dec-2015].

[4] S. S. JL, Milne, Borgnia MJ, Bartesaghi A, Tran EE, Earl LA, Schauder DM, Lengyel J, Pierson J, Patwardhan A, "Cryo-electron microscopy--a primer for the non-microscopist," *FEBS J.*, vol. 280, no. 1, pp. 28–45, 2013.

[5] E. Group, "Registration page of the EMPIAR deposition system." [Online]. Available: http://www.ebi.ac.uk/pdbe/emdb/empiar/deposition/register/. [Accessed: 15-Dec-2015].

[6] K. G. Patwardhan A, Ashton A, Brandt R, Butcher S, Carzaniga R, Chiu W, Collinson L, Doux P, Duke E, Ellisman MH7, Franken E, Grünewald K, Heriche JK, Koster A, Kühlbrandt W, Lagerstedt I, Larabell C, Lawson CL, Saibil HR, Sanz-García E, Subramaniam S, Verkade, "A 3D cellular context for the macromolecular world," *Nat. Struct. Mol. Biol.*, vol. 21, no. 10, pp. 841 – 845, 2014.

[7] F. T. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, "UCSF Chimera--a visualization system for exploratory research and analysis," *J. Comput. chimerstry*, vol. 25, no. 13, pp. 1605–1612, 2004.

[8] B. M. Bell SD, "The minichromosome maintenance replicative helicase," *Cold spring harboe Perspect. Biol.*, vol. 5, no. 11, 2011.

[9] B. J. Costa A, Ilves I, Tamberg N, Petojevic T, Nogales E, Botchan MR, "The structural basis for MCM2-7 helicase activation by GINS and Cdc45," *Nat. Struct. Mol. Biol.*, vol. 18, no. 4, pp. 471–477, 2011.

[10] B. J. Costa A, Renault L, Swuec P, Petojevic T, Pesavento JJ, Ilves I, MacLellan-Gibson K, Fleck RA, Botchan MR, "DNA binding polarity, dimerization, and ATPase ring remodeling in the CMG helicase of the eukaryotic replisome," *Elife*, 2014.

[11] B. J. Costa A, Renault L, Swuec P, Petojevic T, Pesavento JJ, Ilves I, MacLellan-Gibson K, Fleck RA, Botchan MR, "DNA binding polarity, dimerization, and ATPase ring remodeling in the CMG helicase of the eukaryotic replisome.," *Elife*, 2014.

[12] O. S. Costa A, van Duinen G, Medagli B, Chong J, Sakakibara N, Kelman Z, Nair SK, Patwardhan A, "Cryo-electron microscopy reveals a novel DNA-binding site on the MCM helicase," *EMBO J.*, vol. 27, no. 16, pp. 2250–2258, 2008.

[13] S. M. C. Gómez-Llorente Y, Fletcher RJ, Chen XS, Carazo JM, "Polymorphism and double hexamer structure in the archaeal minichromosome maintenance (MCM) helicase from Methanobacterium thermoautotrophicum," *J. Biol. Chem.*, vol. 280, no. 49, pp. 40909–40915, 2005.

[14] C. J. Hesketh EL, Parker-Manuel RP, Chaban Y, Satti R, Coverley D, Orlova EV, "DNA induces conformational changes in a recombinant human minichromosome maintenance complex," *J. Biol. Chem.*, vol. 290, no. 12, pp. 7973–7979, 2015.

[15] G. N. Li N, Zhai Y, Zhang Y, Li W, Yang M, Lei J, Tye BK, "Structure of the eukaryotic MCM complex at 3.8 Å.," *Nature*, vol. 524, no. 7564, pp. 186–191, 2015.

[16] B. J. Lyubimov AY, Costa A, Bleichert F, Botchan MR, "ATP-dependent conformational dynamics underlie the functional asymmetry of the replicative helicase from a minimalist eukaryote," *Proc. Natl. Acad. Sci. United Starares Am.*, vol. 109, no. 30, pp. 11999–12004, 2012.

[17] W. G. Okorokov AL, Waugh A, Hodgkinson J, Murthy A, Hong HK, Leo E, Sherman MB, Stoeber K, Orlova EV, "Hexameric ring structure of human MCM10 DNA replication factor," *EMBO Rep.*, vol. 8, no. 10, pp. 925–930, 2007.

[18] L. H. Sun J, Evrin C, Samel SA, Fernández-Cid A, Riera A, Kawakami H, Stillman B, Speck C, "Cryo-EM structure of a helicase loading intermediate containing ORC-Cdc6-Cdt1-MCM2-7 bound to DNA," *Nat. Struct. Mol. Biol.*, vol. 20, no. 8, pp. 944–951, 2013.

[19] L. H. Sun J, Fernandez-Cid A, Riera A, Tognetti S, Yuan Z, Stillman B, Speck C, "Structural and mechanistic insights into Mcm2-7 double-hexamer assembly and function," *Genes Dev.*, vol. 28, no. 20, pp. 2291–2303, 2014.

[20] A. M. Hirtreiter, G. Calloni, F. Forner, B. Scheibe, M. Puype, J. Vandekerckhove, M. Mann, F. U. Hartl, and M. Hayer-Hartl, "Differential substrate specificity of group I and group II chaperonins in the archaeon Methanosarcina mazei," vol. 74, no. 5, pp. 1152–1168, 2009.

[21] S. S. Bartesaghi A, Lecumberry F, Sapiro G, "Protein secondary structure determination by constrained single-particle cryo-electron tomography," *Structure*, vol. 20, no. 12, pp. 2003–2013, 2012.

[22] W.-S. P. Chen DH, Luke K, Zhang J, Chiu W, "Location and flexibility of the unique C-terminal tail of Aquifex aeolicus co-chaperonin protein 10 as derived by cryo-electron microscopy and biophysical techniques," *J. Mol. Biol.*, vol. 381, no. 3, pp. 707–717, 2008.

[23] R. H. Chen DH, Madan D, Weaver J, Lin Z, Schröder GF, Chiu W, "Visualizing GroEL/ES in the act of encapsulating a folding protein," *CellPress*, vol. 153, no. 6, pp. 1354–1365, 2013.

[24] S. H. Clare DK, Bakkes PJ, van Heerikhuizen H, van der Vies SM, "Chaperonin complex with a newly folded protein encapsulated in the folding chamber," *Nature*, vol. 457, no. 7225, pp. 107–10, 2009.

[25] S. H. Clare DK, Stagg S, Quispe J, Farr GW, Horwich AL, "Multiple states of a nucleotide-bound group 2 chaperonin," *Structure*, vol. 16, no. 4, pp. 528–34, 2008.

[26] S. H. Clare DK, Vasishtan D, Stagg S, Quispe J, Farr GW, Topf M, Horwich AL, "ATP-triggered conformational changes delineate substrate-binding and -folding mechanics of the GroEL chaperonin," *CellPress*, vol. 149, no. 1, pp. 113–123, 2012.

[27] C. W. Cong Y, Baker ML, Jakana J, Woolford D, Miller EJ, Reissmann S, Kumar RN, Redding-Johanson AM, Batth TS, Mukhopadhyay A, Ludtke SJ, Frydman J, "4.0-A resolution cryo-EM structure of the mammalian chaperonin TRiC/CCT reveals its unique subunit arrangement," *Proc. Natl. Acad. Sci. United Starares Am.*, vol. 107, no. 11, pp. 4967–4972, 2010.

[28] C. W. Cong Y, Schröder GF, Meyer AS, Jakana J, Ma B, Dougherty MT, Schmid MF, Reissmann S, Levitt M, Ludtke SL, Frydman J, "Symmetry-free cryo-EM structures of the chaperonin TRiC along its ATPase-driven conformational cycle," *EMBO J.*, vol. 31, no. 3, pp. 720–730, 2012.

[29] K. P. de Juanes S, Epp N, Latzko S, Neumann M, Fürstenberger G, Hausser I, Stark HJ, "Development of an ichthyosiform phenotype in Alox12b-deficient mouse skin transplants," *J. Mol. Biol.*, vol. 129, no. 6, pp. 1429–1436, 2009.

[30] L. S. DH, Chen, Song JL, Chuang DT, Chiu W, "An expanded conformation of single-ring GroEL-GroES complex encapsulates an 86 kDa substrate," *Structure*, vol. 14, no. 11, pp. 1711–1722, 2006.

[31] B. D. DiMaio F, Zhang J, Chiu W, "Cryo-EM model validation using independent map reconstructions.," *Protein Sci.*, vol. 22, no. 6, pp. 865–868, 2013.

[32] F. J. Douglas NR, Reissmann S, Zhang J, Chen B, Jakana J, Kumar R, Chiu W, "Dual action of ATP hydrolysis couples lid closure to substrate release into the group II chaperonin chamber," *CellPress*, vol. 144, no. 2, pp. 240–252, 2011.

[33] E. H. Elmlund D, "High-resolution single-particle orientation refinement based on spectrally self-adapting common lines," *J. Mol. Biol.*, vol. 167, no. 1, pp. 83–94, 2009.

[34] G. R. Han BG, Dong M, Liu H, Camp L, Geller J, Singer M, Hazen TC, Choi M, Witkowska HE, Ball DA, Typke D, Downing KH, Shatsky M, Brenner SE, Chandonia JM, Biggin MD, "Survey of large protein complexes in D. vulgaris reveals great structural diversity," *Proc. Natl. Acad. Sci. United Starares Am.*, vol. 106, no. 39, pp. 16580–16585, 2009.

[35] N. G. Harder J, "Functional expression of the intracellular pattern recognition receptor NOD1 in human keratinocytes," *J. Mol. Biol.*, vol. 129, no. 5, pp. 1299–1302, 2009.

[36] K. T. Hoersch D, Roh SH, Chiu W, "Reprogramming an ATP-driven protein machine into a light-gated nanocage," *Nat. Nanotechnol.*, vol. 8, no. 12, pp. 928–932, 2013.

[37] S. F. Huo Y, Hu Z, Zhang K, Wang L, Zhai Y, Zhou Q, Lander G, Zhu J, He Y, Pang X, Xu W, Bartlam M, Dong Z, "Crystal structure of group II chaperonin in the open state," *Structure*, vol. 18, no. 10, pp. 1270–1279, 2010.

[38] M. K. Kanno R, Koike-Takeshita A, Yokoyama K, Taguchi H, "Cryo-EM structure of the native GroEL-GroES complex from thermus thermophilus encapsulating substrate inside the cavity," *Structure*, vol. 17, no. 2, pp. 287–293, 2009.

[39] G. S. Kroeze KL, Jurgens WJ, Doulabi BZ, van Milligen FJ, Scheper RJ, "Chemokine-mediated migration of skin-derived stem cells: predominant role for CCL5/RANTES," *J. Invest. Dermatol.*, vol. 129, no. 6, pp. 1569–1581, 2009.

[40] C. W. Ludtke SJ, Chen DH, Song JL, Chuang DT, "Seeing GroEL at 6 A resolution by single particle electron cryomicroscopy.," *Structure*, vol. 12, no. 7, pp. 1129–1136, 2004.

[41] C. W. Ludtke SJ1, Jakana J, Song JL, Chuang DT, "A 11.5 A single particle reconstruction of GroEL using EMAN," *J. Mol. Biol.*, vol. 314, no. 2, pp. 253–262, 2001.

[42] P. C. Milazzo AC, Cheng A, Moeller A, Lyumkis D, Jacovetty E, Polukas J, Ellisman MH, Xuong NH, Carragher B, "Initial evaluation of a direct detection device detector for single particle cryo-electron microscopy," *J. Mol. Biol.*, vol. 176, no. 3, pp. 404–408, 2011.

[43] S. H. Ranson NA, Clare DK, Farr GW, Houldershaw D, Horwich AL, "Allosteric signaling of ATP hydrolysis in GroEL-GroES complexes," *Nat. Struct. Mol. Biol.*, vol. 13, no. 2, pp. 147–152, 2006.

[44] S. H. Ranson NA, Farr GW, Roseman AM, Gowen B, Fenton WA, Horwich AL, "ATP-bound states of GroEL captured by cryo-electron microscopy," *CellPress*, vol. 107, no. 7, pp. 869–879, 2001.

[45] B. D. Sanz-García E, Stewart AB, "The random-model method enables ab initio 3D reconstruction of asymmetric particles and determination of particle symmetry," *J. Mol. Biol.*, vol. 171, no. 2, pp. 216–222, 2010.

[46] K. J. Sergeeva OA, Chen B, Haase-Pettingell C, Ludtke SJ, Chiu W, "Human CCT4 and CCT5 chaperonin subunits expressed in Escherichia coli form biologically active homo-oligomers," *J. Biol. Chem.*, vol. 288, no. 24, pp. 177734–177744, 2013.

[47] P. C. Stagg SM, Lander GC, Quispe J, Voss NR, Cheng A, Bradlow H, Bradlow S, Carragher B, "A test-bed for optimizing high-resolution single particle reconstructions," *J. Mol. Biol.*, vol. 163, no. 1, pp. 29–39, 2008.

[48] C. W. Zhang J, Baker ML, Schröder GF, Douglas NR, Reissmann S, Jakana J, Dougherty M, Fu CJ, Levitt M, Ludtke SJ, Frydman J, "Mechanism of folding chamber closure in a group II chaperonin," *Nature*, vol. 463, no. 7279, pp. 379–383, 2010.

[49] C. W. Zhang J, Ma B, DiMaio F, Douglas NR, Joachimiak LA, Baker D, Frydman J, Levitt M, "Cryo-EM structure of a group II chaperonin in the prehydrolysis ATP-bound state leading to lid closure," *Structure*, vol. 19, no. 5, pp. 633–639, 2011.

[50] S. F. Zhang K, Wang L, Liu Y, Chan KY, Pang X, Schulten K, Dong Z, "Flexible interwoven termini determine the thermal stability of thermosomes," *Protein Cell*, vol. 4, no. 6, pp. 432–444, 2013.

[51] S. H. Fischer N, Neumann P, Konevega AL, Bock LV, Ficner R, Rodnina MV, "Structure of the E. coli ribosome-EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM," *Nature*, vol. 520, no. 7548, pp. 567–570, 2015.

[52] B. N. Greber BJ, Bieri P, Leibundgut M, Leitner A, Aebersold R, Boehringer D, "Ribosome. The complete structure of the 55S mammalian mitochondrial ribosome," *Science (80-.).*, vol. 348, no. 6232, pp. 303–308, 2015.

[53] G. N. Zhang Y, Mandava CS, Cao W, Li X, Zhang D, Li N, Zhang Y, Zhang X, Qin Y, Mi K, Lei J, Sanyal S, "HflX is a ribosome-splitting factor rescuing stalled ribosomes under stress conditions," *Nat. Struct. Mol. Biol.*, 2015.

[54] The HDF group, "OBTAIN HDF-JAVA PRODUCTS SOFTWARE." [Online]. Available: http://www.hdfgroup.org/products/java/release/download.html. [Accessed: 25-Dec-2015].

[55] Chimera, "Tools index of Chimera software." [Online]. Available: https://www.cgl.ucsf.edu/chimera/current/docs/UsersGuide/framecontrib.html. [Accessed: 30-Dec-2015].

[56] Chimera, "Volume Viewer of Chimera software." [Online]. Available: http://www.cgl.ucsf.edu/chimera/current/docs/ContributedSoftware/volumeviewer/framevolumeviewer.html. [Accessed: 05-Dec-2015].

[57] Research Group "Modeling of Protein complexes, "Sub-tomogram averaging." [Online]. Available: http://www.biochem.mpg.de/309018/PyTom. [Accessed: 10-Jan-2016].

[58] A. JI, Agulleiro and Martinez-Sanchez, "Computational Methods for Electron Tomography." [Online]. Available: http://sites.google.com/site/3demimageprocessing/research_projects/comet. [Accessed: 26-Dec-2015].

[59] The European Biology Laboratory [Online]. Availabe: https://en.wikipedia.org/wiki/European_Molecular_Biology_Laboratory [Accessed: January 2016]

[40] Stadistics of EMDB method used [Online]. Availabe: https://www.ebi.ac.uk/pdbe/emdb/statistics_emmethod.html/ [Accessed: January 2016]

[41] Stadistics of EMDB entries[Online]. Availabe: https://www.ebi.ac.uk/pdbe/emdb/statistics_main.html/ [Accessed: January 2016]

[42] Ribosome information [Online]. Availabe: https://en.wikipedia.org/wiki/Ribosome [Accessed: January 2016]

10. APPENDICES

APPENDIX 1: Saving a part of an atomic model

Sometimes the atomic model is hidden or is a subunit of the atomic complex. On those situations we have to try to extract this structure from the complex and save it in a separated file. To do this we have to open the complex structure on Chimera and select the interested atoms from the model.

We used two alternatives to select the interested atoms on the structure:

- Searching the coordinates of the interested part of the model on PDB database. Once coordinates are known, they will be introduced by "command line tool" from Chimera, with *Function 1*, then the interested atoms will be on the main window.

```
select :"AtomBeginning"-"AtomEnd"."ChainBeginning"-"ChainEnd"
```

*Function 1. Select atomic model residues. First you introduce the starting and the final atom that you want to select, and finally you introduce the interested chains that you want to select, if you want to select whole chains you will put a * instead the ID of the chain. Eg.* ***select :4.567-A.D***

- The other option is to select manually by selecting simultaneously the interested region with the mouse and pressing Ctrl.

Once the interested area is selected as the *Figure 10.1.1.*, we proceed to save this new structure as a PDB model, following the route below:

File > Save as a PDB file > Select the box "only selected atoms" > Save

Subsequently a new PDB file will be generated, then we reload this structure on Chimera, and finally we can work with the interested fragment, see *Figure 10.1.1*.

A B

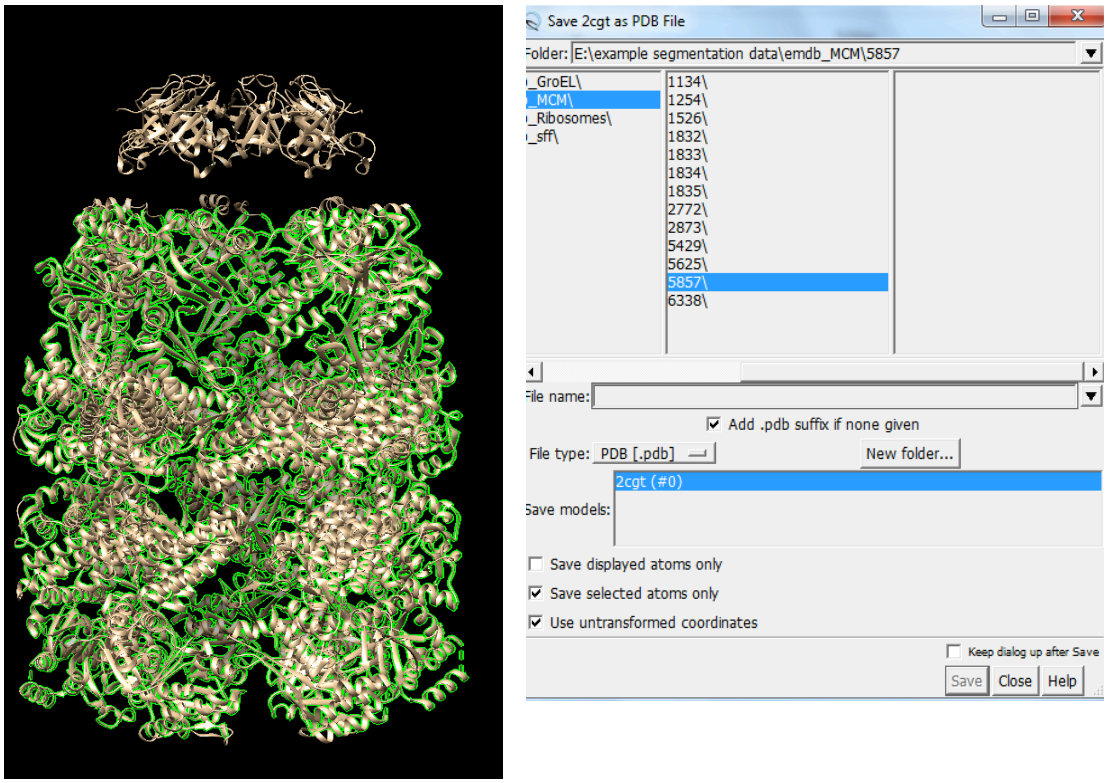


Figure 10.1.1. Selection of a fragment of an atomic model. (A) How looks the selected atoms on the PDB model on the main screen of Chimera. And on the right (B) the saving window to save the selected atoms as a separated PDB file

APPENDIX 2: Denoising procedure

In this project we have found some noisy samples that required to be clean. To perform it, we needed a Chimera tool that could be located on "Segmentation map tools" (*Figure. 10.2.1*).

To denoise the sample the analyser needed to stablish a "border value", this value belongs to the maximum group of voxels that want to be erased form the map. This procedure is explained also on section "3.3.3. *Segmentation step".* This procedure has to be done once the interested contour level is stablished and the map is ready to segment.

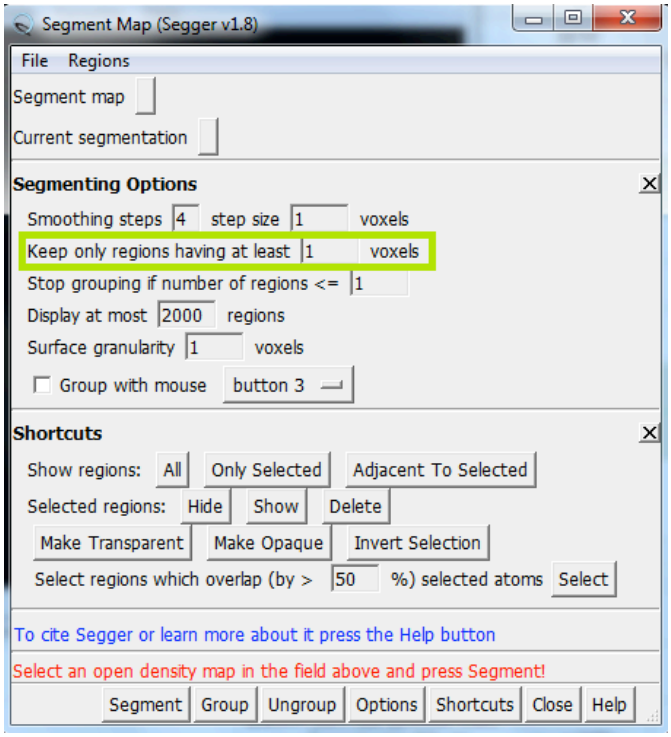


Figure 10.2.1. Denoising command on segmentation map window. Labelled on green is where the border number is entered, once the analyser is in conformity with it, needs to be executed by the "Segment" button

Along the creation of this segmentation sample sets, we have had to denoise several of the EM structures. Find some significant examples below, they belong to four different samples with different degrees of noise.

Low level of noise

As we can see on *Figure 10.2.2; A* the map was not very noisy, we could see some artefacts around it but not in excess. Sometimes this type of noise can be tread by the modification of the contour level before the segmentation. The stabilized contour level was 4 and the border value of voxels was **10**.

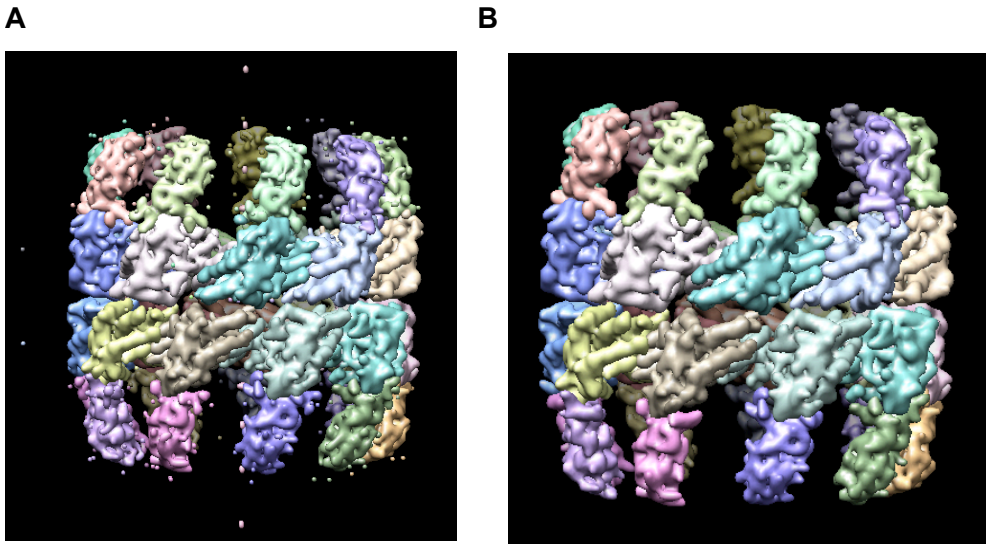


Figure 10.2.2. Segmentation map with low level of noise (EMD-5396). (A) Corresponds to a noisy segmented map that needs to be denoised, the result of this is shown on (B) image, without artefacts around it.

Medium level of noise

This type of noise was very common on the GroEL structures. Find an example of on *Figure 10.2.3*. The contour level of EMD-1200 was 1.84 and the "border value" was **20**.

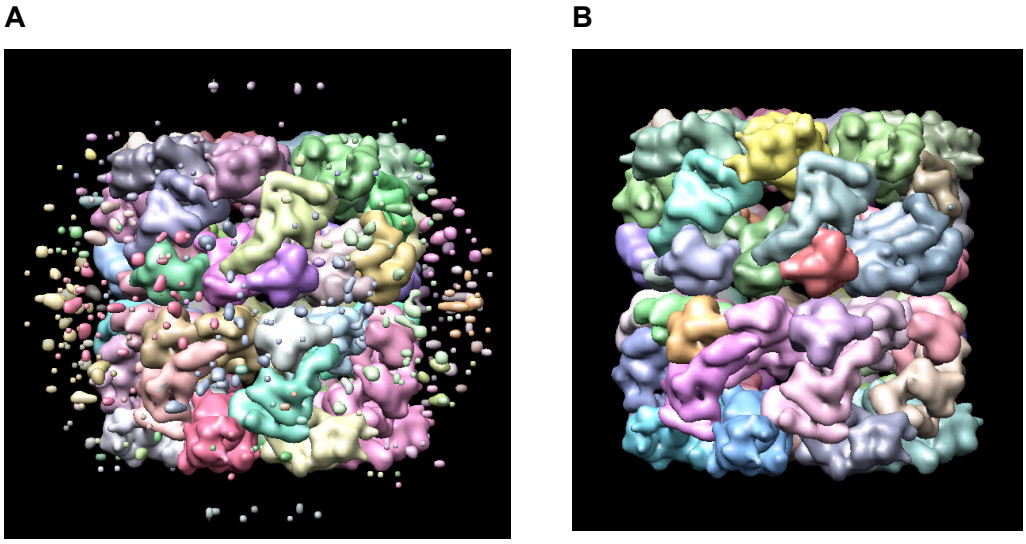


Figure 10.2.3. Segmentation map with medium level of noise (EMD-1200). (A) Corresponds to a noisy segmented map that needs. (B) Denoised map

High level of noise

On this case we could see a number slightly higher of artefacts around the structure, the structure is hidden by this noise and the segmentation would be impossible to realize without the denoising of the structure. The stabilized contour level on EMD-1458 was 2.08 and the border value of voxels was **20**. See *Figure 10.2.4.*

A **B**

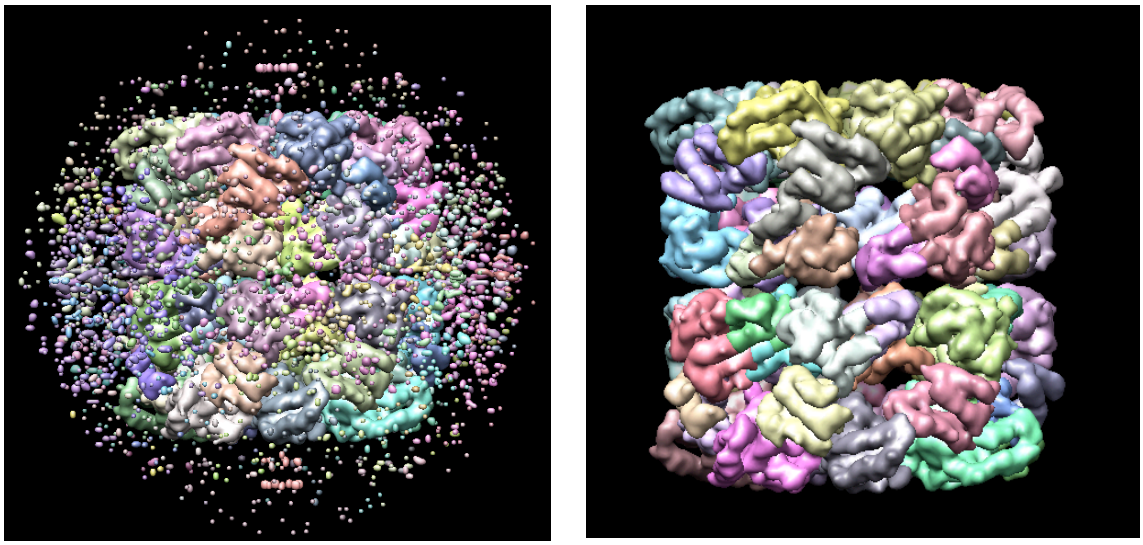


Figure 10.2.4 Segmentation map with low level of noise (EMD-1458): (A) Noisy and (B) denoised map

Very high level of noise

When you are on this extreme case, doesn't make sense to realize the segmentation of the EM map, as all the noise/artefacts hides the interested structure. To realize this type of denoising was required to put a very high "border value", on this case we needed a "border value" of 300, erasing then the artefacts composed with <300 voxels. On the *Figure 10.2.5*, we can see a denoising process passing with different amounts of border values. The stabilized contour level on EMD-5339 was 1.88 and the border value of voxels was **300**.

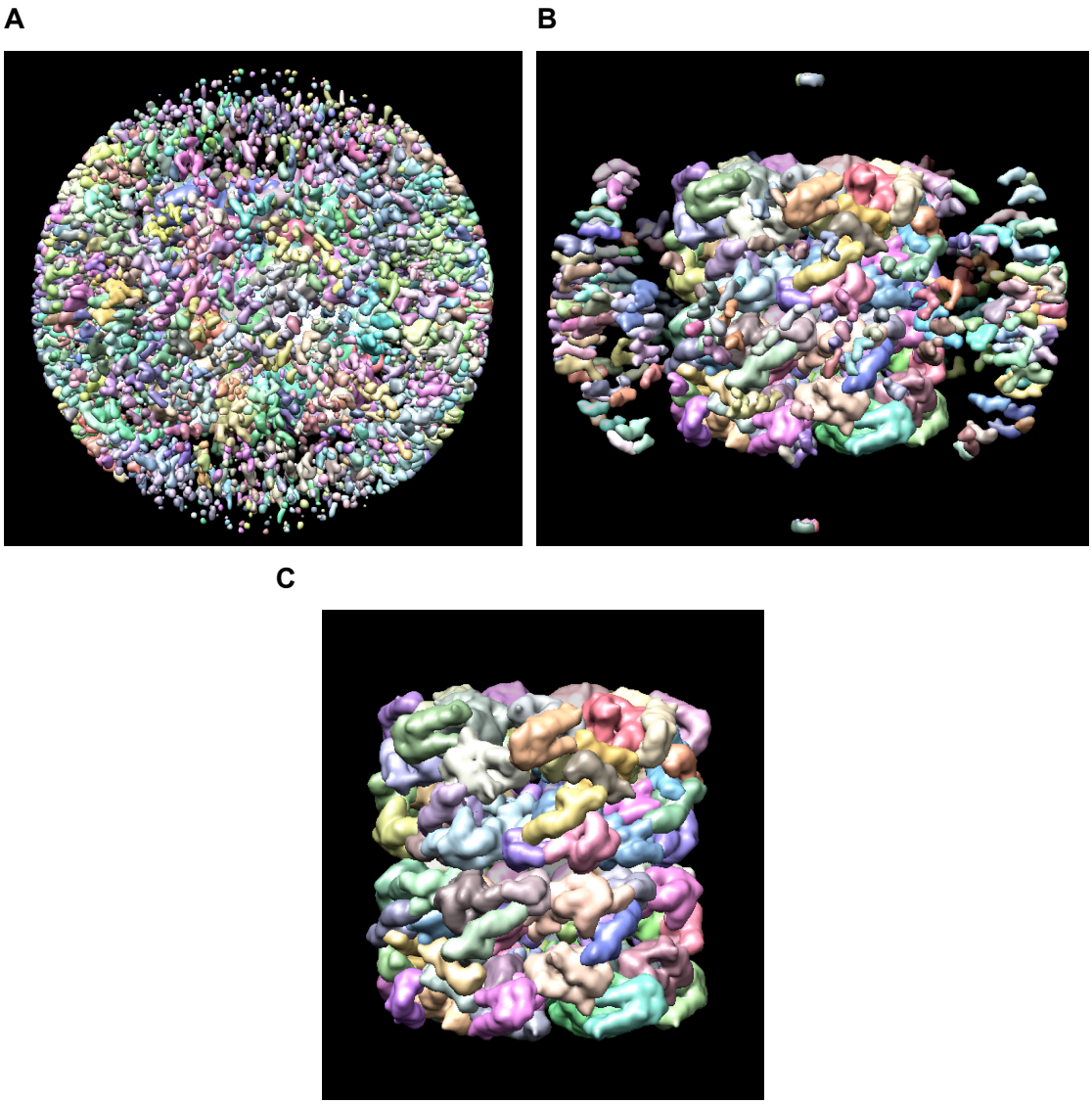


Figure 10.2.5. Segmentation map with low level of noise (EMD-5339). (A) Noisy with a borer value of 1, (B) intermedium denoised map with a border value of 200 and finally (C) corresponds to the denoised map with a border value of 300

APPENDIX 3: HDF5 information

This program was used to see all the stored information on the Segger file (.seg file format). Is a very useful alternative to visualize and access to this information.

When the program is already installed, you have to load the interested segmentation file:

File > Open > Select the interested file (.seg) > Open

Once the file is open, it will be generated a lateral column with several sections, see *Figure 10.3.1.* These sections store the information of the Segger file, as:

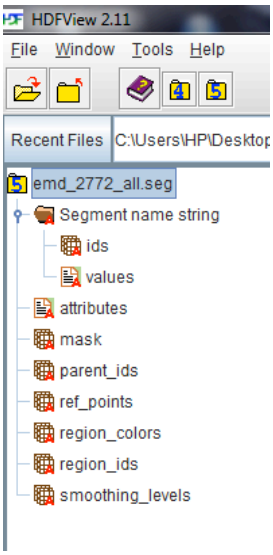


Figure10.3.1. HDFView 2.11 main options.

- **IDs:** Code that belongs to the identification of each of the segments. This code is generated automatically by the Chimera software. See *Figure 10.3.2.*
- **Values:** Name of the segmented regions. A biologist annotated the name manually on the "attributes tool" when the segmentation was done. See *Figure 10.3.2.*

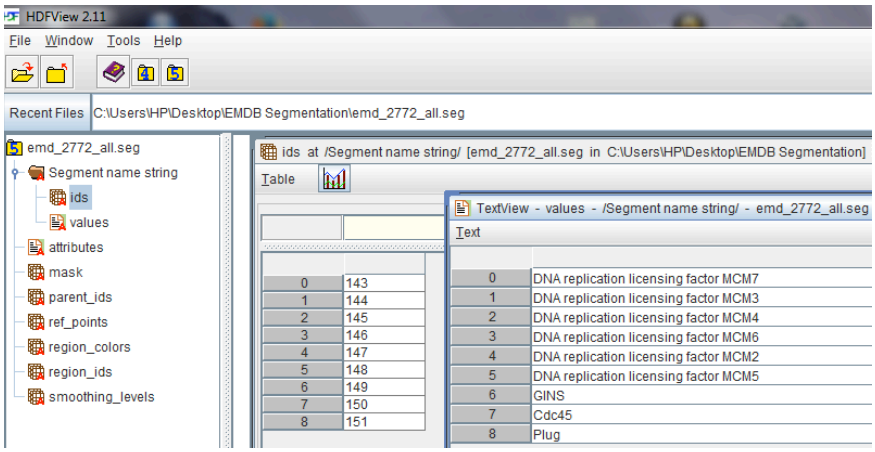


Figure 10.3.2. Data saved on the segmentation name attributes. (A) Codes IDs and (B) segments names of EMD-2772

- **Attributes:** Name of the new attributes added to the file. This section belongs to the "Segment name" value added as a new column in the "attributes tool" from Chimera.
- **Mask:** Belongs to the voxels information stored on the 3D document. To understand this tool you have to try to imagine a cubic matrix, where a segmented protein is localized on the centre of an imaginary cube (*Figure 10.3.3.*). The limits/edges of this cube corresponds to the borders of Chimera program, where the map is located. The size of the cube is related with the size of the EM map, a big map – big cube and small map – small cube.

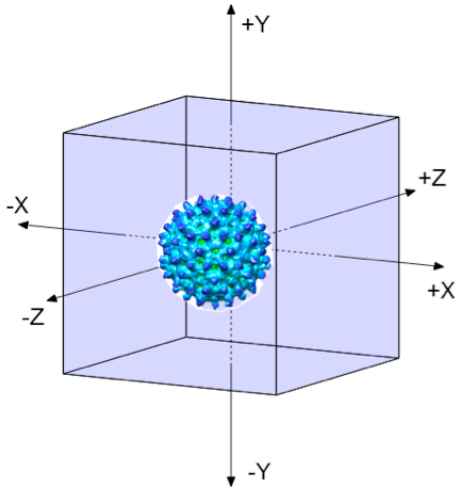
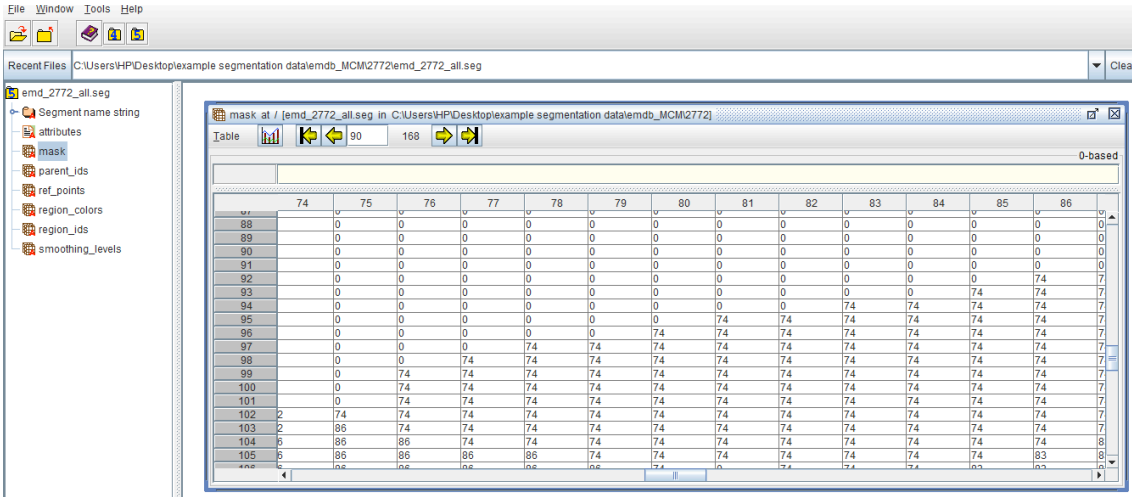


Figure 4.3.3 Cubic matrix. For the EMD-2772 the cube would be composed by 168 voxels for each side and a grid of 168^3

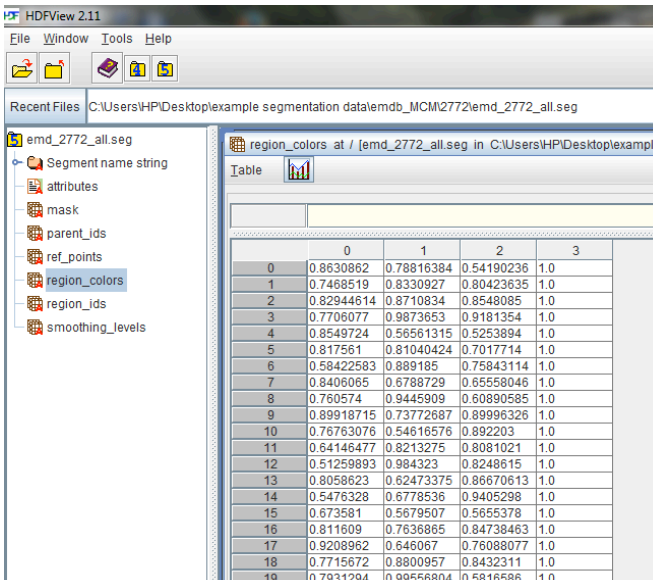
The information is stored on a matrix, if the voxel is empty will be represented on it like a 0 and on the other hand, if the voxel stores information will be annotated a number >0, depending on the type and information stored on each case. The final result would be a 3D matrix where the user would see the stored information written as numbers, on each of the boxes (*Figure 10.3.4.*)



o

Figure 10.3.4. Mask of the information saved on the file matrix.

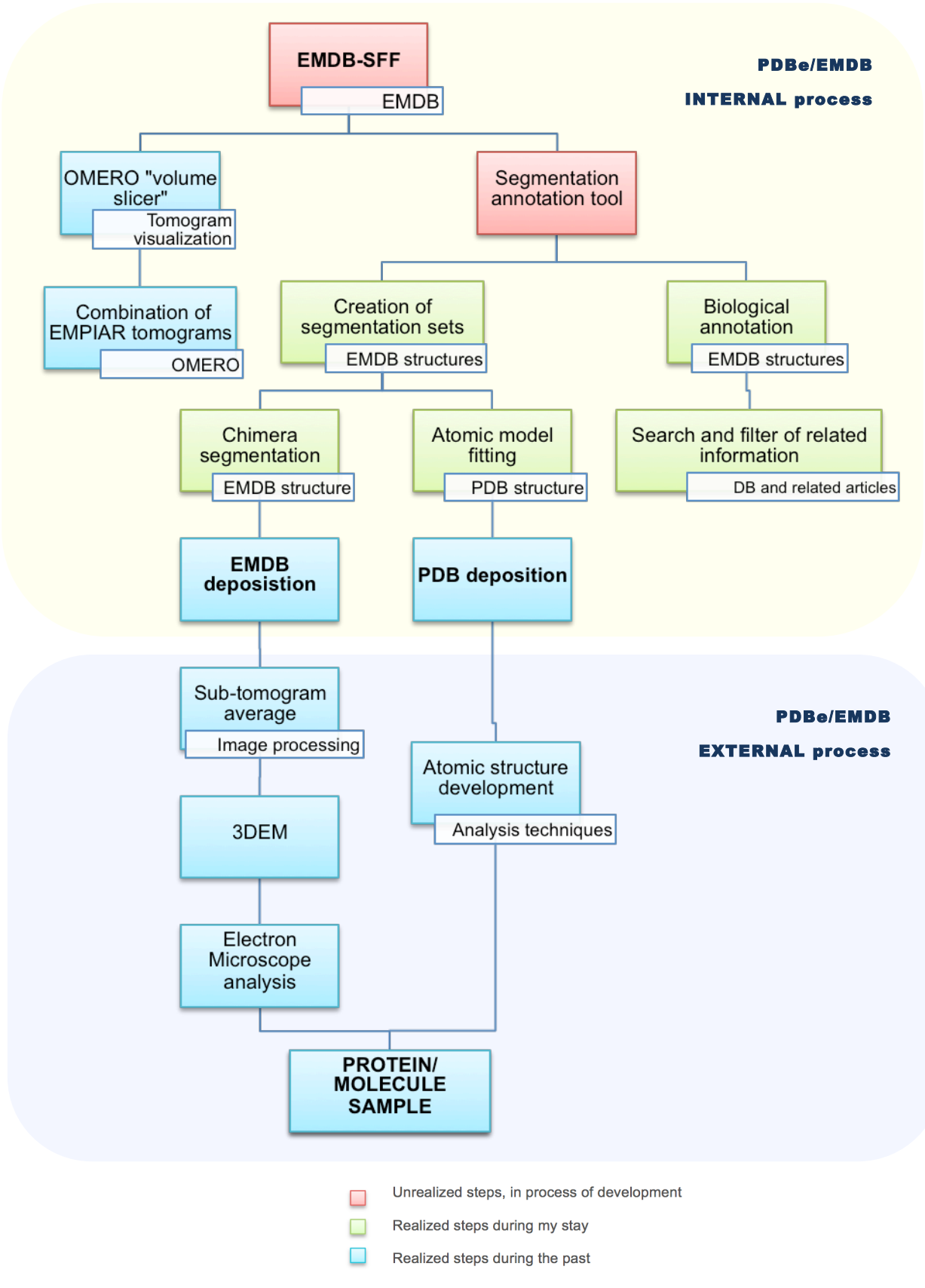
- **Region colours:** Corresponds to the colours of the regions that form the structure. The 0 corresponds to red, 1 green and 2 to blue. See Figure 10.3.5



	0	1	2	3
0	0.8630862	0.78816384	0.54190236	1.0
1	0.7468519	0.8330927	0.80423635	1.0
2	0.82944614	0.8710834	0.8548085	1.0
3	0.7706077	0.9873653	0.9181354	1.0
4	0.8549724	0.56561315	0.5253894	1.0
5	0.817561	0.81040424	0.7017714	1.0
6	0.58422583	0.889185	0.75843114	1.0
7	0.8406065	0.6788729	0.65558046	1.0
8	0.760574	0.9445909	0.60890585	1.0
9	0.89918715	0.73772687	0.89996326	1.0
10	0.76763076	0.54616576	0.892203	1.0
11	0.64146477	0.8213275	0.8081021	1.0
12	0.51259893	0.984323	0.8248615	1.0
13	0.8058623	0.62473375	0.86670613	1.0
14	0.5476328	0.6778536	0.9405298	1.0
15	0.673581	0.5679507	0.5655378	1.0
16	0.811609	0.7636865	0.84738463	1.0
17	0.9208962	0.646067	0.76088077	1.0
18	0.7715672	0.8800957	0.8432311	1.0
19	0.7931294	0.99556804	0.5816586	1.0

Figure 10.3.5. Colours of each of the file region.

APPENDIX 4: SFF project diagram



APPENDIX 5: Sub-tomogram averaging and Cryo-electron tomography

Is a three-dimensional imaging technique for structural studies of macromolecules under close-to-native conditions. A series of images are acquired from the sample at different tilt angles around a, typically, single axis. The images have to be aligned. In high-resolution structural studies, CTF has to be determined and its effects corrected for. Next, tomographic reconstruction combines the aligned images to yield the 3D volume or tomogram. Afterwards, the tomogram is typically subjected to denoising, to reduce noise with preservation of details (*Figure 10.5.1*).

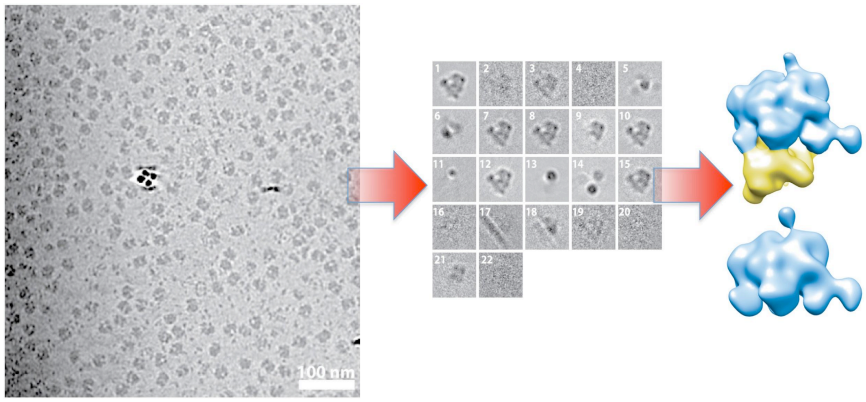


Figure 10.5.1. 3D volume tomographic reconstruction[57]

Segmentation then intends to decompose the tomogram into the structural components. Finally, if repetitive structures are present, they can be detected and extracted for further analysis. This analysis typically includes 3D alignment and averaging of the sub-tomograms to obtain a high resolution map. See *Figure 10.5.2*. [58]

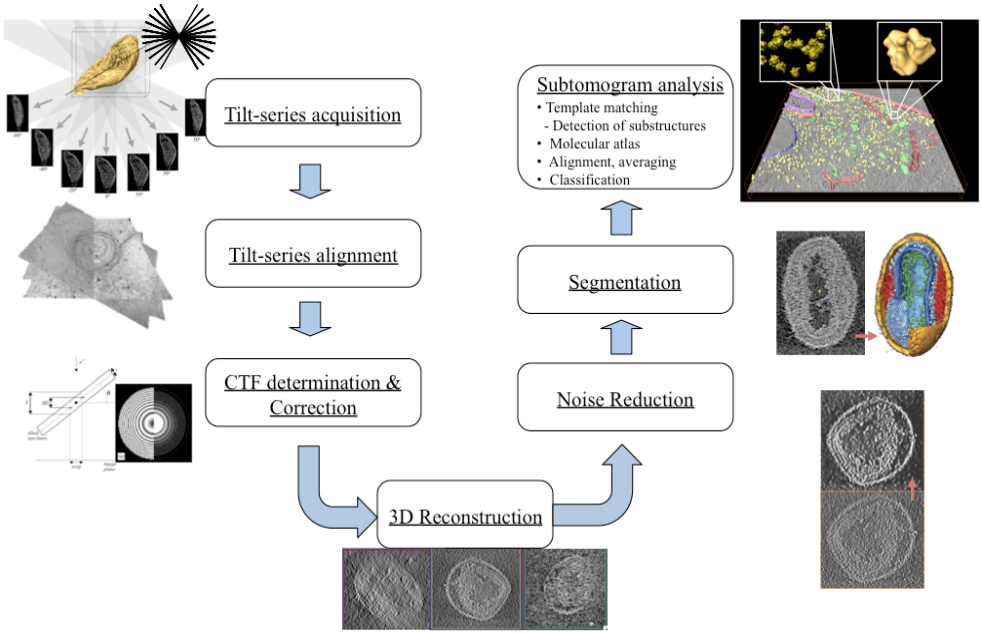


Figure 10.5.2. Image processing workflow in structural studies by electron tomography