

Algorithm automation for tilt-series alignment and tomogram reconstruction in cryo-electron tomography

Autor:

de Isidro-Gómez, Federico Pedro¹

Tutors:

Sánchez-Sorzano, Carlos Oscar¹; Marabini Ruiz, Roberto²

- (1) Biocomputing Unit, National Center for Biotechnology (CSIC), c/Darwin, 3, Campus Universidad Autónoma, 28049 Cantoblanco, Madrid, Spain. Centro Nacional de Biotecnología (CNB-CSIC).
- (2) Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain.

ABSTRACT

Electron tomography is a technique that allows the elucidation of three-dimensional structures of biological samples at high resolution. However, the results obtained with this technique are limited to the performance of the algorithms used for the alignment of the tilted image series and the posterior tomogram reconstruction obtained from the sample. Although nowadays there exist many software solutions to this problem, results comparison and interoperability between different software packages is not obvious. Also, high-throughput data processing is not usually solved in many of these software packages. Thus, in this paper we present the integration of different tools for tilted image series and tomogram reconstruction into the Scipion framework. Scipion is a software tool that allows the interoperability of different packages for electron microscopy and tomography image processing into the same environment. All the developments presented in this paper are freely available in the Scipion software, as different plug-ins.

I. INTRODUCTION

A. *The electron tomography hardware*

Electron tomography is a specific technique within the electron microscopy field. The main particularity of this technique is the sample tilting while imaging process takes place. This sample tilting allows the imaging of different views of the same sample for its posterior three-dimensional reconstruction, as it can be seen in Figure 1 from the Supplementary material.

This technique arises structural information of samples from 20 Å up to 5 Å (Kovtun 2018) (Mattei 2018). This structural information helps scientists to determine the precise organization and relationship between the different components of samples. This technique has applications in

different fields, from structural biology, where it has helped to understand subcellular organelles becoming a cutting-edge tool in structural biology, to material science, where it has facilitated other different structural studies (Sorzano 2020).

The basis of this imaging technology relies on the radiation of the sample with an electron beam for the posterior detection of interactions between the sample and the radiated electrons. The projection images obtained from the sample with the usage of a transmission electron microscope (TEM) at different orientations allows the tomographic reconstruction of the sample.

The electron microscope is composed by several elements. A simplified structure of an electron microscope is presented in Figure 1.

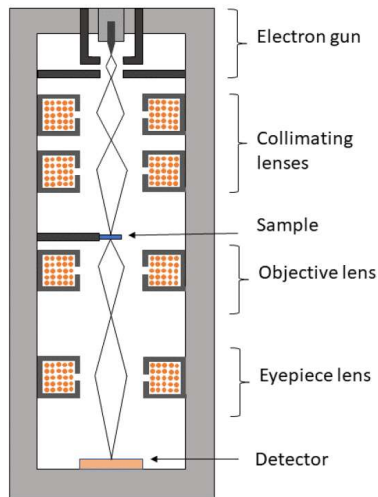


Fig. 1. Simplified schematic of an electron microscope indicating its main components and structures.

The first structural part in the electron microscope is composed by two elements: the source of radiation and the collimator condenser lens, performing the illumination of the sample.

The source of radiation of an electron microscope consist of an electron gun. This device is composed by a specific material, usually a metal, which under determined conditions can emit electrons. These conditions consist of an elevated temperature and the immersion of the gun into an electric field, which triggers the emission of the electrons from the surface of the material instead of generating an internal electrical current (de Isidro-Gómez 2019).

Once the radiation source expels the electrons, a set of rotationally symmetric magnetic lenses is used to collimate the electron beam to converge it towards a fixed point. This will ensure the correct radiation of the sample and the imaging formation process.

The next structural part is composed by the sample holder. This is the element where the main difference respect to the classical electron microscope appears. Also, the sample is embedded into an vitreous ice structure by a cryogenization process. This sample preparation technique allows the information acquisition of the inner structure of the specimen under study.

The third and last structural part of the microscope is related to the image formation process, composed by two elements: a new set of magnetic lenses and the detector.

This second group of lenses, the ocular system, produces a new magnification of the radiation to improve the image received by the detector.

Finally, the detector is a sensor responsible for the detection of the intensity pattern obtained from the electron beam after its interaction with the sample. A proper design of the detector is as important as the optical system to achieve the correct image formation.

Nowadays these detectors belong to the Direct Detector Devices (DDD) family (or Direct Electron Detectors, DED), whose main advantage is that the measurement of the radiation is done without the conversion of the electron beam into light. This is achieved by the usage of a phosphor layer; whose main drawback is that it will introduce some blurring in the detected image (de Isidro-Gómez 2019).

Detectors and its capabilities are characterized through several parameters as the picture element size, the resolution, and the dynamic range, between others.

Despite the complexity and accuracy of the exposed device, the raw images obtained by the microscope need further software processing to arise useful information, helpful in other fields. The developed project is immersed in the development of software tools that outperforms this process.

B. The tomogram reconstruction software

One set of images taken from the same sample at different views (different tilt angles) is commonly named as a tilt-series.

Before presenting the usual software pipeline to obtain a reconstructed tomogram from the input tilt-series, the concept of missing wedge must be introduced. When cryo-electron tomography is used for the study of biological samples, the tilting of the sample is not complete (the sample does not rotate 360°) as in other tomography techniques as CT (x-ray computed tomography).

For this reason, some of the possible views of the sample are avoided, introducing a lack of information in the reconstruction process in some directions. This missing information is known as the missing wedge of the acquisition.

There exist some other acquisition techniques implementing other geometries that try to avoid, or at least reduce, this missing information, as dual-axis (D. N. Mastrorarde 1997), multiple-axis (Messaoudi 2006) (Hata 2011), or conical tilt (Zampighi 2005).

One of the peculiarities that must be considered in the acquisition process is that it is a common practice to take several low-dose images with the same tilting. This is a common practice inherited from the classic electron microscopy and the single particle analysis (SPA) to increase the quality of the images while reducing the degradation of the sample due to the electron beam radiation. The obtained result is a stack of movies per each sample tilting.

As indicated in Figure 2, the first step consists in applying an averaging of these movies to obtain the tilt-series. In this step the obtained movies are wise-averaged, tracking possible sample movements or deformations from one frame to another. This will increase the signal-to-noise ratio (SNR), improving the quality of the input images to the pipeline.

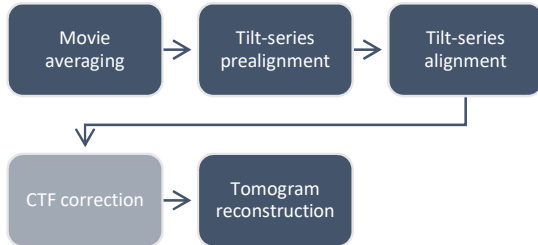


Fig. 2. Schematic presenting a typical workflow in electron tomography from the acquired images to the reconstructed tomogram. Decolored boxes indicate steps that may be unnecessary to obtain a three-dimensional reconstruction.

The next step in the pipeline is the tilt-series prealignment. In this step, the vertical and horizontal shifts introduced by the movement of the sample holder during the acquisition process are corrected.

To do so, contiguous images are iteratively compared with different relative shifts and the configuration presenting the minimum distance

(maximum similarity) is considered as valid. As a measure of similarity, the most widely used for this purpose (and the one in which the methods presented in this project are based on) is the cross correlation of two images:

$$(I * I')[x, y] \triangleq \sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} \overline{I[m, n]} I'[m + x, n + y]$$

The shifting introduced to reach this maximum similarity is characterized through a transformation matrix.

Once the tilt-series prealignment is concluded and the relative shift between images are corrected it is possible to finally align the tilt-series.

Prior to the final tomogram reconstruction as shown in Figure 2, it is necessary also to establish the relative projection direction of each image belonging to the tilt-series. This is due to the design of the reconstruction algorithms which are based on the existence of a single and well-defined axis.

Due to this, unlike the case of the tilt-series prealignment, the alignment performed is not relative to the position of each image respect to its contiguous ones but to the underlying movement described by the elements belonging to the tilt-series.

In the case of biological samples, the use of gold beads fiducial markers is widely spread. These gold beads act as fiducial markers, since due to its high reactivity with the electron beam, they present a high contrast in the acquired images. Also, it is important to remark that there exist alignment algorithms without fiducials, able to identify landmarks (patches of the image in which high contrast elements are present) and perform the same alignment based on its position. This process may be automatic, semi-automatic or manual.

Once these fiducials (or landmarks) are identified, it is possible to track its position through the whole tilt-series, generating fiducial (or landmark) chains. Being p_{ij} the projection coordinates of each landmark from its position in three-dimensional space r_j :

$$p_{ij} = A_i r_j + d_{ij}$$

It is possible to characterize the underlying movement described by the sample through the

projection matrix A_i and a vector accounting for the shifting of the projection d_{ij} .

Be aware that due to the geometry and the rotation of the acquisition it is possible that, depending on the position of the fiducial, it disappears in some images of the tilt-series, usually in high tilting ones.

Once the movement tracking and the chain generation is concluded, it is possible to identify the common tilt-axis for all images belonging to the tilt-series. Then, all the images will be interpolated, aligning their tilting axis to the image one, characterizing their shifts, rotations (and even deformations) with a transformation matrix of the form:

$$R = \begin{pmatrix} \cos \theta & -\sin \theta & d_x \\ \sin \theta & \cos \theta & d_y \\ 0 & 0 & 1 \end{pmatrix}$$

Where θ is the rotation angle, and d_x and d_y the shifts calculated for each direction.

Finally, the pipeline concludes with the tomogram reconstruction step. From the properly aligned tilt-series the reconstructed tomogram is obtained employing an application of the back-projection algorithm. This algorithm establishes an efficient reconstruction of a volume from its projections from different orientations in the Fourier domain. The algorithm consists in a stabilized and discretized version of the inverse Radon transform, known as the filtered back projection algorithm. With a sample discrete system, the inverse Radon transform is:

$$f(x, y) = \frac{1}{2\pi} \sum_{i=0}^{N-1} \Delta\theta_i p_\theta(t) k(t) (x \cos \theta_i + y \sin \theta_i)$$

Where N is the number of projections, $\Delta\theta$ is the angular spacing between the projections and $k(t)$ is the Radon kernel with frequency response $|\omega|$.

The back-projection algorithm is founded on the central slide theorem, which establishes that a volume in the real space $S(x, y, z)$ can be reconstructed from the Fourier transform of its oriented projections $P_\theta(S)$ along an θ angle by slicing them through the origin and performing the inverse transform.

$$D_\theta[\mathcal{F}(S)] = \mathcal{F}(P_\theta(S))$$

A schematic of this statement may be found in Figure 3.

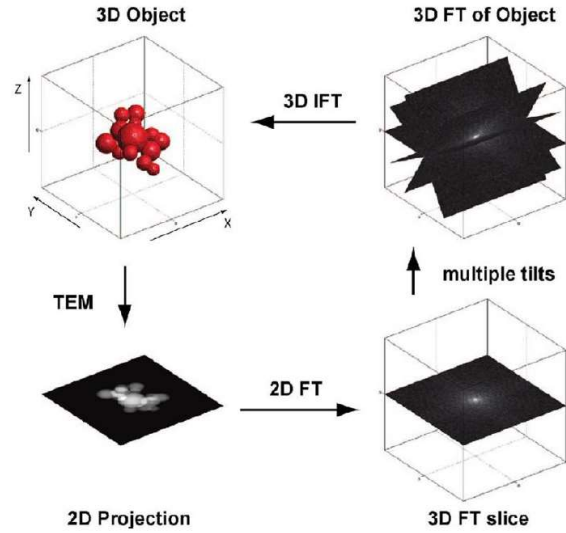


Fig. 3. Demonstrative scheme of the central slide theorem.

As can be seen in Figure 2, there is an unmentioned optional step before the reconstruction process, the contrast transfer function (CTF) correction.

The CTF mathematically describes the aberrations introduced by the microscope that affect the acquired images. Thus, this function quantifies the effect and variation of the obtained images and the ideal ones. Once the CTF is estimated it is possible to correct the obtained images from aberrations, increasing the SNR of the posterior reconstruction (Vilas 2019):

$$\mathcal{F}\{I_{obtained}\} = \mathcal{F}\{I_{real}\} \cdot CTF$$

The CTF estimation, and the posterior correction, is a process widely developed and applied in the SPA although due to the complexity of the geometry of electron tomography its definition and application present some peculiarities and complications.

One of the most important parameter to calculate in order to perform a correct estimation of the CTF is the defocus (O. A. Sorzano C.O.S. 2009). The correct estimation of this parameter will be fundamental for its posterior correction in the tilt-series and reconstructed tomograms.

II. METHODS

The developed project is based on two different toolboxes for the cryo-electron tomography image processing: IMOD (D. N. Mastronarde 1997) (D. N. Mastronarde 2017) and NovaCTF (Turoňová 2017). Both packages offer different tools presenting different underlying algorithmic strategies.

The presented project consists in the development of computational methods that allow the automation and high-throughput processing of cryo-electron tomography images using the presented packages. These methods are developed inside the Scipion framework (de la Rosa-Trevín 2016).

Scipion is a software framework that allow the agglomeration and interoperability of electron microscopy and electron tomography algorithms into a common platform.

In its last version Scipion integrates every software package as a plug-in, which allows not only the interoperability, but also the independence of every software package. Also, it offers a modular software in which each plug-in may be installed on user demand, fulfilling any possible requirement or dependency, but without including any undesired package by the user. This will keep the software to the maximum simplicity, making its installation lighter and more user-friendly, reducing its complexity.

All the different algorithms and functionalities inside Scipion are abstracted as the called protocols. Each plug-in holds its own protocols, offering a specific functionality in each of them. These protocols usually present a common structure in which they take an input and produces an output after some processing as a “black box” for the user, who only needs to properly specify the input settings (in most of the cases).

In this project, the integration and automation of both IMOD and NovaCTF packages will be achieved by the development of two different Scipion plug-ins. These packages will be integrated by the development of specific Scipion protocols that use these packages as libraries from which the different algorithms and functionalities will be employed.

As an important recall, Scipion relies in some functionalities of a special software package, Xmipp (de la Rosa JM 2013) (Sorzano C.O.S. 2004), to support some basic operations. Also, Xmipp is an independent plug-in which offers its own algorithms for image processing in electron microscopy and tomography.

To assure interoperability between the different protocols available, Scipion needs its own abstraction of every input and output for each protocol.

Also, Scipion presents a tree structure workflow, conformed by every protocol that form part of the designed pipeline. This tree structure is dynamically constructed as the user keeps aggregating new protocols to the workflow. The hierarchy of the workflow is constructed relating the output of some protocols as the input of others. An example of this can be observed in Figure 4.

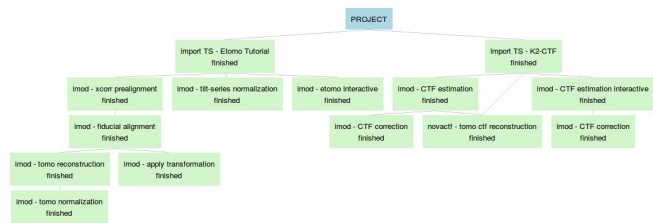


Fig. 4. Screenshot from an Scipion project containing all the available protocols in the IMOD and NovaCTF plug-ins.

To assure compatibility and to establish some design patterns, every Scipion plug-in follow a common structure. This structure includes every software need associated to the plug-in, both for the user (as the mentioned protocols or specific viewers) and the developer (as test or subjacent classes needed for the abstraction of some specific element of the plugin).

To make Scipion a user-friendlier application and improve the automation of the image processing workflows, some of the input parameters for the developed protocols have established some default values. These values usually retrieve acceptable results although its comprehension and personalization are crucial for more exhaustive studies from more dedicated users.

A. IMOD

The first software package to be automatized and integrated into Scipion is IMOD (D. N. Mastronarde 1997) (D. N. Mastronarde 2017).

IMOD is a widely used software package with many different functionalities and composed by several programs and algorithms. Its main pipeline allows the tomogram reconstruction from a non-aligned tilt-series. Several supplementary tools and features are included, as the CTF estimation and correction or the gold beads removal.

IMOD software integrates all these functionalities in its own graphical user interface (GUI), known as Etomo. From the Etomo GUI it is possible to process an input tilt-series selecting from the available programs those in which the user is interested in. The only previous step is the properly importing of the tilt-series and the associated acquisition information.

The first step in the project was to develop a protocol that allows the user to operate with the Etomo GUI inside Scipion. Although, from the automation point of view this suppose no advantages for the user, it will allow him to operate with the outputs obtained inside Scipion as plausible inputs for other protocols.

Also, those users who have already processed with IMOD outside the Scipion environment are used to this GUI and it could serve as a power testing processing tool before moving to a more high-throughput processing strategy.

Once this first wrapper was developed, it is possible to start the algorithm automation process. To do so, the first step is to define a basic pipeline that allows the reconstruction of a tomogram from an input unaligned tilt series and identify the logical steps that compose it. Each one of these steps will be encapsulated as one Scipion protocol.

It is important to mention that IMOD does not support as an input a tilt-series with its frames not averaged (as movies, see I Introduction. B). This is not a problem for the Scipion user since the framework expects this kind of data type, and it offers a specific protocol to perform the tilt-series movies averaging.

The first step and protocol developed is the tilt-series prealignment. In this protocol the first

automation effort is performed, allowing the user to automatically prealign a set of tilt-series without the interaction of any GUI.

This process requires the wise execution of two IMOD programs, *tiltxcorr* and *xftoxg*. These programs perform the relative shifts calculations to correct the relative movement of a tilt-image with the contiguous one to posteriorly homogenize the shift for the whole series. Thus, the obtained output is a list of transformation matrices associated with each one of the tilt-images from the series.

To execute this protocol the user only needs to indicate the input set of tilt-series to be processed. The output object generated is another set of tilt-series as the input one but with the transformation matrices associated to each tilt image.

Once the prealignment is performed, it is possible to move to the next step in the pipeline, the fiducial alignment. This the most complex protocol belonging to the IMOD plugin from the integration point of view.

This process requires the wise execution of three IMOD programs: *autofidseed*, to automatically detect the fiducial for its posterior tracking; *beadtrack*, to track the detected beads along the tilt series; and *tiltalign*, to calculate the transformation matrices associated to each image from the tilt-series in order to correct the tilt axis orientation.

To execute this protocol the user needs to indicate the input set of tilt-series to be processed, the approximated diameter of the fiducial to be detected and if the fiducials must be differentiated depending if they are located over or below the sample.

As in the previous step, one of the output objects is another set of tilt-series as the input one but with the transformation matrices associated to each tilt image.

Apart from this, this protocol generates other output classes. The first one is the associated landmark model to each tilt-series. This object is a Scipion abstraction that holds the information referred to the fiducial position in each tilt-image (a two-dimensional coordinate) used to perform the alignment. The landmark model is composed by a set of landmark chains, which holds the information of the position of the same fiducial through all the

images of the tilt-series in which it has been detected. In order to do so it is necessary to use the IMOD program *model2point*, which translates the model used for the alignment into a text file readable by Scipion.

Two landmark models are generated, with the difference that in one of them, the gaps coming from undetected fiducials in specific images from the tilt-series are filled by extrapolating its position after the alignment.

The last generated output is the set of three-dimensional coordinates associated to the position in the space of each fiducial used for the alignment.

In both prealignment and alignment protocols, there exists a user option to interpolate the input tilt series by the application to the obtained transformation matrices. This interpolation is performed with the IMOD program *newstack*, a very powerful tool that will be used in posterior protocols. Thus, the user may check the performance of the alignment by eye view.

There is a double reason for not to do this by default. First, for computational and disk space efficiency, and secondly, to avoid excessive interpolations, which is always a source of artifacts and errors in the output image.

This also means that, for every tilt-series with an associated transformation matrix that inputs an alignment protocol, it will be needed a transformation matrix composition (the input matrix and the obtained one) that will be assigned to the output tilt-series.

The last step in the first pipeline proposed for automation is the tomogram reconstruction.

This step requires the wise execution of the IMOD program *tilt*. This program generates the three-dimensional reconstruction from an aligned tilt-series.

To execute this protocol the user must indicate the input set of tilt-series to be processed, the usual shift and offset parameters, and the thickness of the output reconstruction. Apart from these, there exists the possibility to apply a Simultaneous Iterative Reconstruction Technique (SIRT) filter. Although the underlying algorithm is based in the back-projection algorithm, the application of this filter

emulates the results that would be obtained by the usage of this reconstruction technique.

The obtained results would be a new Scipion abstraction object, a set of tomograms containing all the reconstructions obtained.

Once the first pipeline has been completed, new protocols have been added to the plugin to offer new functionality to the user.

The first new developed protocol is able to calculate the interpolation described by the transformation matrix associated to each image from a tilt-series. This protocol also allows the application of some binning to the output interpolated tilt-series.

A binning operation consists in the reduction in size of some image (or set of images) by some factor, k :

$$I(r, c) = I'(r/k, c/k)$$

This is a common practice in the electron tomography field, and it is offered to the user in every algorithm in which some interpolation is calculated. There exist many benefits from the binning application to an image: computational efficiency, storage optimization, and an increment of the SNR (as expected for any averaging filter).

The second pair of protocols added to the IMOD plug-in are normalization protocols for both tilt-series and tomograms. These protocols allow the application of some image processing operations to these objects to obtain some modified version of the at the output.

These operations have different nature and applications. Apart from the already mentioned binning, there exist the possibility to change the storage data type (which may also have implications in the information contained in the images) or adjust the density values of some input images. This density adjustment allows the mapping of the values that compose every image to a determined distribution, as filling some data range or scaled to a common mean (μ) and standard deviation (σ):

$$J(r, c) = f\{I(r, c)\} \mid J(r, c) \in N(\mu, \sigma)$$

The last couple of protocols added to the IMOD plug-in are related to the CTF estimation and correction.

The CTF estimation protocol requires the wise execution of the *ctfplotter* IMOD program. This is one of the most complex programs from this software package which even has its own graphical user interface.

The first development of this protocol has the objective to avoid the user interaction with the *ctfplotter* GUI to fulfill the statements of automation and high-throughput processing.

In this first development the user dispose of several parameters to characterize different possible behaviors when performing the CTF estimation, although a few of them are mandatory (the input tilt-series, the expected defocus, and the defocus tolerance). Other optional parameters allow the angle or frequency auto-refinement, the search for astigmatism in the input tilt-series, the tile size, or the tilt-images grouping for the defocus calculation.

The output of the execution of this protocol is a text file that contains the defocus information of each tilt-image (or each range of images in case grouping parameters have been set) and astigmatism information (in case astigmatism parameters have been indicated).

After some user demands, an interactive version of this protocol has been developed too. In this version the user will execute the *ctfplotter* program and will operate directly with its GUI. Once the user finished the execution, the output data will be retrieved as in the automatic case without any extra operation needed from the user.

Due to incompatibilities with the IMOD CTF model foundations and other CTF models inside Scipion, this protocol will retrieve its output directly as the produced text information file, without any abstraction as a Scipion object. The main drawback is that the interoperability between protocols is reduced, although it presents the advantage that all the obtained information from the CTF estimation is preserved.

And the last protocol is the CTF correction. This protocol performs the correction of aberrations introduced by the microscope, calculated by the CTF estimation previous protocol.

To execute this protocol, the user only needs to input a previous IMOD CTF estimation protocol with an associated tilt-series to be corrected.

The output from this last protocol is another set of tilt-series which have their aberrations corrected with the CTF estimation previously performed.

There are three reasons why these two last protocols are not fused. First, this design will allow to input the CTF estimation to both CTF correction protocols and tomogram reconstruction protocols (as will be exposed in the next section). Second, this open the possibility for a consensus protocol able to compare different CTF estimations and keep only the ones that agree (as it is already developed for the single particle analysis in the Scipion framework). And finally, this design will allow the streaming execution (start correcting the first CTF estimation performed without the need of waiting until the estimations from the whole set of tilt-series is completed).

B. NovaCTF

Once the IMOD automation inside Scipion is completed, the second software package to be integrated is NovaCTF (Turoňová 2017).

NovaCTF is a less extensive software package (compared with IMOD) which has fewer tools but, due to its conception and design, it is highly appreciated and demanded by users.

NovaCTF package offers a tomogram reconstruction with the particularity of performing a three-dimensional CTF correction. This CTF correction is more powerful than the ones applied directly over the tilt-series (as it is the case of IMOD) since the geometry of the reconstructed tomograms will be considered and hence also its affectations over the estimated. This considerations about the obtained defocus values from the CTF estimation and its impact on the tomogram reconstruction geometry are fundamental since the defocus is directly related to the distance between the sample and the detector. This is illustrated in Figure 5.

The programs contained in this software package act as different steps of a single design, feeding one program with the outputs obtained from the previous one. Due to this and to the

specificity of the intermediate files generated, which does not seem compatible with any other software, the design decision was to generate a single Scipion protocol that agglomerates all the programs from this package.

The developed Scipion protocol requires from the usual inputs from a tomogram reconstruction program (the input tilt-series, parameters to control the tomogram size and positioning, and for some filtering), and other set of parameters related to the input CTF estimation. Also, NovaCTF allows different CTF correction strategies.

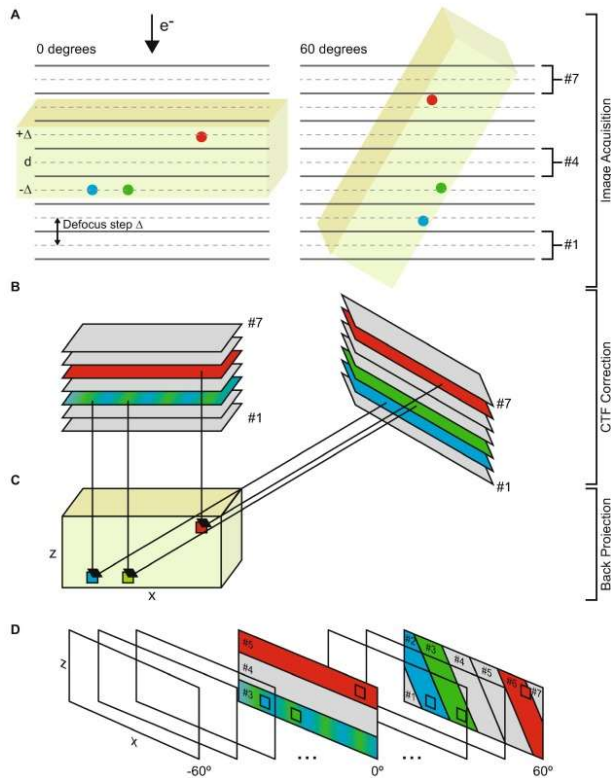


Fig. 5. Schematic of the reconstruction and three-dimensional CTF correction performed by NovaCTF. A) Acquisition process. B) Slice sectioning for the CTF correction. C) Image and defocus assignment to the reconstructed volume during the back-projection process. D) Final defocus calculation for the different regions of the tomogram (Turoňová 2017).

This software allows only two possible sources for the CTF estimation: the already mentioned IMOD and CTFFIND4 (Rohou A. 2015). This is the first limitation solved inside Scipion, in whose framework any CTF estimation protocol with a model design compatible with any of these two may be used to feed the protocol.

Once the input layer is completed the design of the protocols assures the proper coordination of the different programs inside Scipion.

Also, some of the operations performed inside this protocol are solved using IMOD programs. This has been indicated by the NovaCTF developers and it is not surprising due to the wide spread of the IMOD package. This is again easily solved inside Scipion since, due to its plug-in modular structure, it is straight-forward to use programs from other plug-ins.

The operations performed in each step inside Scipion protocol depends on the NovaCTF program to be executed. In the first step the *Defocus* program is executed, which will generate the intermediate defocus files from the input one.

Then, the *ctfCorrection* program is executed for each defocus file generated in the previous step. This program will perform the wise CTF correction over the input tilt-series considering the geometry of the output reconstruction.

Following, the IMOD program *clip* is used to flip the CTF corrected tilt-series to fit the geometry needed to perform the reconstruction.

Once the tilt-series are flipped, they are filtered with the usage of the *filterProjections* program.

And finally, all these filtered projections are used to reconstruct the final CTF corrected tomogram, using the *3dctf* program. For geometry convenience, this final reconstruction is again flipped before it is output.

Once all these steps are wisely coordinate and executed, the generated output will be a set of three-dimensional CTF corrected tomograms. This will increase the quality of the final reconstruction and improve the results obtained in later steps of the processing.

III. RESULTS

Several datasets have been used to measure the performance of the developed protocols.

In the performed tests, due to the specific nature of the developed methods, two main parameters are going to be under study. First, that the obtained results executing all the automatized protocols arise

the same result as if a manual execution with equivalent parameter would have been performed outside the Scipion environment. And second, that the execution time payout related to the execution of these methods inside the Scipion framework is negligible or at least it is compensated with the extra features that this platform offers.

A. Dataset 1: Etomo tutorial data

This first dataset is delivered directly from the IMOD developers team and it is the one used in the facilitated tutorial.

The dataset is composed by two tilt-series with identical dimension: 61 tilt-images of 512x512 pixels in size.

This dataset fits perfectly with the needs of this study due to two different arguments. First, this dataset has been widely processed due to the spread of this software package and its use as the dataset proposed in the IMOD tutorial. Also, the expected results from the processing are also delivered which makes even easier the comparison with the obtained results.

With this first dataset all the protocols from the IMOD plug-in have been tested, excluding those one related with the CTF estimation and correction. This data is not suitable for a proper CTF estimation due to it is from a plastic section (so it is dominated by amplitude contrast) and taken near focus. These protocols will be tested with the second introduced dataset in this paper.

Due to the small size of the test images, the execution time related to the Scipion operations will be proportionally increased, due to the faster execution of the IMOD programs. This will be useful to set a challenging test for the developed methods

Since not all the developed protocols have a direct correspondence with a tool from the Etomo GUI, proper performance tests for these protocols will be done by the execution through the command line. Individually, the performed tests raised positive results.

First, the interactive Etomo GUI integrated inside Scipion, presented the same execution time, and gave rise to the same results as equivalent executions outside the Scipion framework.

The rest automatized protocols will be exposed as follows. Concerning the protocols that has a direct correspondence in the Etomo GUI the first one is the tilt-series prealignment.

This protocol presented the same results (a detailed comparison can be performed though the transformation matrices). As for the execution time, no significant difference in the execution time has been observed, even considering the fastness of this step.

The next protocol tested in the pipeline is the fiducial alignment. As in the previous protocol, the obtained results are equivalent to the one obtained from the execution outside Scipion. Regarding the execution time, this protocol is the most demanding one in terms of operations related to Scipion from this pipeline. Also, performing accurate measurements of the execution time performance is complicated due to the number of intermediate operations that must be performed over the metadata obtained using IMOD programs. Anyway, the obtained results presented no significant difference in the execution time. To be confident with the results obtained and as a safety margin, the interpolation of the obtained tilt-series has been included in these measurements.

The last protocol that may be compared the directly from the execution with the Etomo GUI is the tomogram reconstruction step. In this case, as in the previous ones, the results obtained are identical, and there are no perceptible differences in the execution time.

Regarding those protocols that has not a direct correspondence with the Etomo GUI, they will be tested comparing though the command line. These tests are even more exhaustive than the previous one since the time related to the operations performed by the Etomo GUI are removed.

These protocols are the tilt-series interpolation (by the application of the transformation matrix) and the normalization protocols, both tilt-series and tomogram.

In all the cases the results obtained are equivalent to the ones obtained from the command line. Regarding the execution time, no perceptible differences have been observed comparing the execution in and outside Scipion.

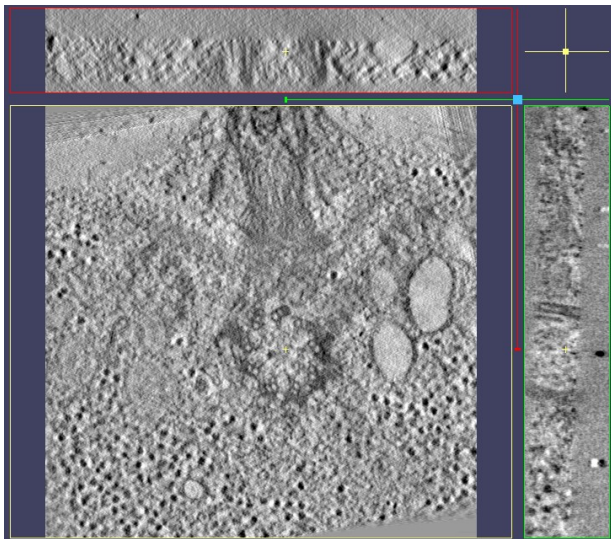


Fig. 6. Central slice of the reconstructed tomogram BBA.mrc from three different views.

In overall the obtained results are positive. The obtained interpolations and reconstructions are identical in all cases and the execution time difference is negligible (even more if we consider the pipeline as a whole).

In case there could be any discrepancy, these results are even more justified if we consider the benefits obtained from the execution of these protocols inside the Scipion framework (automation, interoperability, and easy results comparison), and if the size of the dataset is considered.

B. Dataset 2: K2-CTF tutorial data

This second dataset is also delivered directly from the IMOD developers team and it is oriented to the CTF estimation and correction with Etomo. This dataset has been chosen to properly test those protocols related to the CTF that have been integrated inside Scipion.

The dataset is composed by one tilt-series with 61 tilt-images with dimensions of 1,856x1,852 pixels in size.

This dataset has the same goodness as the one introduced before, it has been widely used for the same purposes and the expected results are also delivered as part of the dataset.

The two remaining protocols missing to be tested from the IMOD plug-in are the target of the

following analysis: the CTF estimation and correction.

The CTF estimation protocol has been tested twice, in its two operating modes: automatic and interactive.

The results obtained in both cases are identical to the obtained ones from the Etomo interactive GUI, once all the parameters are properly and equally set in both cases. These results have been checked through the defocus estimation output, since no output image is generated in this step

Regarding the execution time, measurements in case of the interactive mode are straightforward since the behavior of both GUI, in and outside Scipion are identical, and the time expended performing Scipion related operations is negligible.

In case of the automatic mode, execution time measurements are more complicated since no interaction with the GUI is needed in this case. In any case, the execution time in this case has been significantly shortened (principally due to the lack of user interaction).

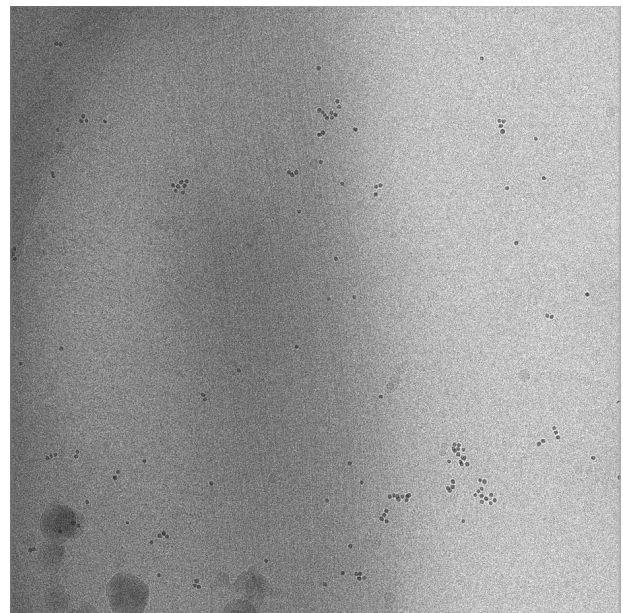


Fig. 7. Central slide at 0 degrees of the tilt-series WT1042413_1series4.st from the K2-CTF tutorial dataset after CTF correction.

This step is an example of the powerful benefits obtained from the protocol automation. Executing this protocol in automatic mode implies that the user does not requires to operate with the GUI,

accelerating the process, even more if several tilt-series are processed at one.

And finally, the CTF correction protocol. As in all the previous cases, the obtained results are equivalent as the one obtained from the Etomo GUI. In this case, the operation requires no operations through an extra GUI, so the test procedure will be closer to the ones presented with Dataset 1 (see section III Results. A).

Regarding the execution time, this process is significantly faster than the previous one and comparing the results obtained from the execution in and outside Scipion, no significant differences have been observed.

In overall, from this dataset the goodness of the performance of the CTF related protocols inside Scipion have been properly tested and both the obtained results and the execution time differences present acceptable results. As in the previous case, these results are even more justified due to the benefits obtained from the Scipion framework.

C. Purified breast cancer membranes dataset

This last dataset has been selected to test the performance of the integrated protocol from the NovaCTF package.

In this case the testing process is more complicated than in the previous case for the IMOD protocols since there is neither a user tutorial nor dataset with expected results. Due to this, a proper usage of the software to obtain quality results is more complicated and there is no reference to compare with.

Due to this, an experimental dataset has been selected for this purpose. This experimental data has been acquired from purified membranes from cancerous breast tissue where the SK-BR-3 tyrosine-protein kinase cellular receptor has been constitutively over expressed. This receptor is codified by the Erbb2 gene.

The dataset is composed by one tilt-series with 41 tilt-images with dimensions of 3,710x3,838 pixels in size.

The tilt-series from this dataset is already aligned, with the shifts and tilt axis corrected, so

only the previous CTF estimation is needed before executing the NovaCTF tomogram.

NovaCTF does not include any GUI or other environment, just the binaries needed for a posterior wise execution by the user. This has the advantage that the comparisons to measure execution time will not be affected by any intermediate interface, since it will be directly run in the command line through scripting.

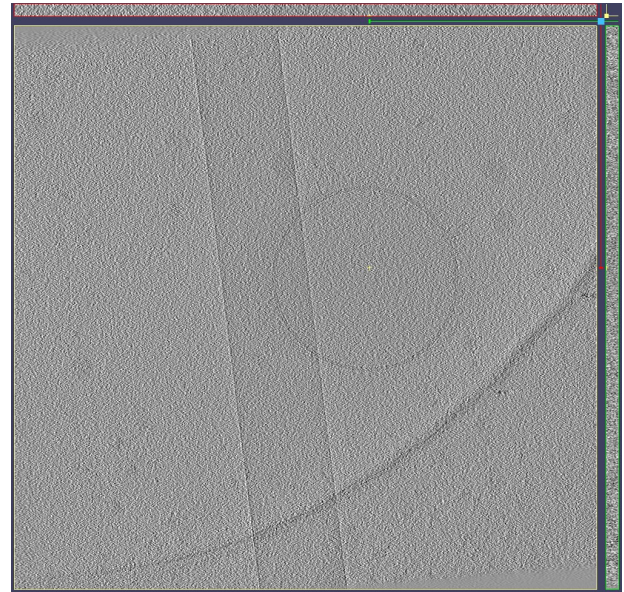


Fig. 8. Central slice of the reconstructed tomogram from three different views.

Compared with the IMOD software, NovaCTF present a more complex usability. So, for this software, although the integration process presents a higher complexity, the benefit received by the user is significantly enhanced.

The CTF estimation has been performed using the protocol from the IMOD package presented in the previous section. Once the CTF has been properly estimated all the information needed to perform the tomogram reconstruction is available.

Once a proper configuration and wise coordination of the different NovaCTF binaries have been achieved, obtaining acceptable results, it is possible to start the design of the protocol for its integration inside Scipion.

As in the previous cases, there is no difference in the results obtained from the tomogram reconstruction between in and outside Scipion.

Due to the long execution time that the NovaCTF tomogram reconstruction requires (almost 9 hours), the differences in execution time comparing both the execution in and outside Scipion are completely negligible.

In the case of this software, attending the positive overall results obtained, the benefits that the user receives by the execution of this software inside Scipion are even more significant than in the previous case. This is mainly because of the enhanced usability, no needing the development of any script to establish a workflow that properly coordinates the different binaries.

IV. DISCUSSION

The outcome of this project is the automation of tools that enhance the performance of image processing methods in cryo-electron tomography.

These developments for the automation are based on the possibility to operate with different tools into the same environment, with a high-throughput processing, and allowing a suitable comparison of the obtained results.

These image processing tools are based on specific underlying algorithms that will bias the results obtained, depending on the strategy followed to solve the problem and the input data provided to the algorithm.

Due to this biasing of the obtained results, the interoperability and combination of different software packages that offers different tools will increase the robustness of the obtained results. This robustness will be enhanced if these tools are based on underlying algorithms presenting different strategies to solve the same problem.

Thus, the combination of these algorithms will lead to the avoidance of systematic biasing errors in the obtained results, these being especially harmful if they are not noticeable at simple sight.

The next achievement in the developed project is the high-throughput processing. Through the integration of the mentioned software in this project, Scipion brings the opportunity to the user to automatize the processing of big batches of data without extra operations.

This is especially sensitive in the case of IMOD, whose interface Etomo is designed for the processing of single tilt-series at a time and through direct user operation. Also, the fact the Etomo GUI is also integrated into Scipion brings the opportunity to the user of combining both strategies, looking for a detailed processing and parameter selection with the graphical interface and then move to a posterior high-throughput batching processing.

In the case of NovaCTF, the benefits come from the easier implementation and user operation of this software through a simple Scipion protocol, instead with the directly delivered software which in any realistic use case will require some scripting implementation for a reasonable use.

Also, the Scipion project-oriented structure allows the user to design detailed workflows combining different tools that, once the proper design and parameters are found, are easily replicable with other data batches.

V. CONCLUSIONS

The previous work establishes the importance of the development of new tools for automation and high-throughput processing in cryo-electron tomography. Also, the software integration into a common framework allows the interoperability of tools from different software packages into the same environment and the result comparison.

These new features enhance the scope of the electron tomography image processing and improves the user experience of the integrated software tools into the Scipion framework.

A. Future work

One of the key future developments are software tools that allow the automatic detection of erroneous misaligned tilt-series and, in the best case, also calculate the associated correction.

This work is planned to be developed as part of a future PhD thesis that the student will develop once the master studies are concluded.

VI. REFERENCES

de Isidro-Gómez, F.P. 2019. "Image processing algorithms for the determination of the optical aberrations of an electron microscope." *Universidad CEU San Pablo Bachelor Thesis*.

de la Rosa JM, Oton J, Marbini R, Zaldivar A, Vargas J, Carazo JM, Sorzano COS. 2013. "Xmipp 3.0: An improved software suite for image processing in electron microscopy." *Journal of Structural Biology* 184 321-328.

de la Rosa-Trevín, Quintana, del Cano, Zaldivar, Foche, Gutiérrez, Gómez-Blanco, Burguet-Castell, Cuenca-Alba, Abrishami, Vargas, Otón, Sharov, Vilas, Navas, Conesa, Kazemi, Marabini, Sorzano, Carazo. 2016. "Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy." *Journal of Structural Biology* 195 93-99.

Hata, S., Miyazaki, H., Miyazaki, S., Mitsuahara, M., Tanaka, M., Kaneko, K., Higashida, K., Ikeda, K., Nakashima, H., Matsumura, S., Barnard, J. S., Sharp, J. H., Midgley, P. A. 2011. "High-angle triple-axis specimen holder for three-dimensional diffraction contrast imaging in transmission electron microscopy." *Ultramicroscopy* 111 1168-1175.

Kovtun, O., Leneva, N., Bykov, Y. S., Ariotti, N., Teasdale, R. D., Schaffer, M., Engel, B. D. 2018. "Structure of the membrane-assembled." *Nature* 561 561-564.

Mastronarde, D. N. 1997. "Dual-axis tomography: an approach with alignment methods that preserve resolution." *Journal of Structural Biology* 120 343-352.

Mastronarde, D. N., Held, S. R. 2017. "Automated tilt series alignment and tomographic reconstruction in IMOD." *Journal of Structural Biology* 197 102-113.

Mattei, S., Tan, A., Glass, B., Müller, B., Kräusslich, H.-G., Briggs, J. A. G. 2018. "High-resolution structures of HIV-1 gag cleavage mutants determine structural switch for virus maturation." *Proc. Natl. Acad. Sci. USA* 115 E9401-E9410.

Messaoudi, C., De Loubresse, N. G., Boudier, T., Dupuis-Williams, P., Marco, S. 2006. "Multiple-axis tomography: applications to basal bodies from paramecium tetraurelia." *Biology of the Cell* 98 415-425.

Rohou A., Grigorieff N. 2015. "CTFFIND4: Fast and accurate defocus estimation from electron micrographs." *Journal of Structural Biology* 192 216-221.

Sorzano C.O.S., Marabini R., Velazquez-Muriel J., Bilbao-Castro J.R., Scheres S.H.W., Pascual-Montano A., Carazo J.M. 2004. "XMIPP: a new generation of an open-source image processing package for electron microscopy." *Journal of Structural Biology* 148 194-204.

Sorzano C.O.S., Otero A., Olmos E.M., Carazo J.M. 2009. "Error analysis in the determination of the electron microscopical contrast transfer function parameters from experimental power Spectra." *BMC Structural Biology* 9.

Sorzano, C.O.S., de Isidro-Gómez, F.P., Fernández-Giménez, E., Herreros, D., Marco, S., Carazo, J.M., Messaoudi, C. 2020. "Improvements on marker-free images

alignment for electron tomography." *Journal of Structural Biology*.

Turoňová, B., Schur, F.K.M, Wan, W. and Briggs, J.A.G. 2017. "Efficient 3D-CTF correction for cryo-electron tomography using NovaCTF improves subtomogram averaging resolution to 3.4 Å." *Journal of Structural Biology* 187-195.

Vilas, J. L. 2019. "Local Quality Assessment of CryoEM Reconstructions and Its Applications." *CSIC Thesis* Vol 1.

Zampighi, G. A., Zampighi, L., Fain, N., Wright, E. M., Cantele, F., Lanzavecchia, S. 2005. "Conical tomography ii: A method for the study of cellular organelles in thin sections." *Journal of Structural Biology* 151 263-274.

VII. SUPPLEMENTARY MATERIAL

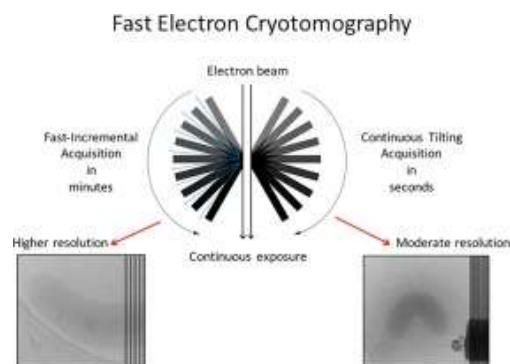


Fig. 1. Schematic of the acquisition process of a cryo-electron microscopy tomogram.