Novel cryo-EM computational methods to process highly heterogenous samples

Swathi Adinarayanan

Anatomy and Cell Biology
McGill University, Montreal
August 2020

A thesis submitted to McGill in partial fulfillment of the requirements of the degree of Master's in Cell Biology

Table of Contents

The numbers refer to the order of appearance of the thesis sections

List of Abbreviations

3D: Three Dimensional

2D: Two-Dimensional

BIM: Beam Induced Motion

CCD: Charge-coupled Device

CL2D: Clustering 2D

CTF: Contrast Transfer Function

Cryo-EM: Cryogenic Electron Microscopy

Cryo-ET: Cryogenic Electron Tomography

DED: Direct Electron Detectors

DNA: Deoxyribonucleic Acid

DRV: Direction RANSAC volume

EM: Electron Microscopic

EMDB: Electron Microscopic Data Bank

EMPIAR: Electron Microscopy Public Image Archive

FSC: Fourier Shell Correlation

ML: Maximum likelihood

NMR: Nuclear Magnetic Resonance

OV: Original Volume

PCA: Principal Component Analysis

PDB: Protein Data Bank

RANSAC: Random Sample Consensus

RELION: Regularised Likelihood Optimisation

RNA: Ribonucleic Acid

SEM: Scanning Electron Microscopy

SNR: Signal to Noise Ration

SPA: Single Particle Analysis

TEM: Transmission Electron Microscopy

.

Abstrait en Francais

Généré une structure trois-dimensionnelle par « Single-Particle Cryo-Electron Microscopy » (SPA-CryoEM) d'une protéine ou d'un complexe de protéines permet l'élucidation de leur fonction. Une des difficultés les plus communes à rencontrer lors de l'analyse SPA-CryoEM est l'hétérogénéité des échantillons. Lors de mon projet, j'ai adressé deux problèmes liés à l'analyse des échantillons très hétérogènes. Mon premier but était d'obtenir la meilleure reconstruction possible de la conformation la plus abondante et mon deuxième but était d'obtenir des reconstructions de toutes les conformation présentes dans l'échantillon très hétérogène. Les techniques utilisées en ce moment ne sont pas capables de surmonter l'hétérogénéité des échantillons pour atteindre mes deux buts. Pour résoudre ce problème, j'ai développée deux méthodes : « Directional Pruning » et « Directional RANSAC ».

Directional Pruning est une méthode qui permet d'enlever des particules qui n'ont pas êtes filtrer hors de l'analyse lors des étapes avent. Cette méthode utilise des classes directionnelles (classifier les particules dans deux classes par direction) qui groupent les particules qui ont une orientation similaire et crée des classes deux-dimensionnelles. Les classes deux-dimensionnelles qui sont le moins abondantes sont élevées de chaque groupe directionnel. Les particules qui restent dans les classes deux-dimensionnelles sont considérées comme un échantillon homogène. Mon hypothèse était que cette méthode d'analyse augmentera la résolution de la structure finale de l'échantillon. Pour tester mon hypothèse, j'ai analysée diffèrent ensembles de données hétérogènes avec la méthode Directional Pruning. Mes résultats n'ont pas démontré une augmentation notable de la résolution de la structure finale.

Pour mieux comprendre mes résultats, j'ai conçu un nouveau ensemble d'expériences pour tester si la méthode de « maximum likelihood » (ML) (qui est la méthode standard en ce moment) est capable de générer une structure avec un haute résolution si l'échantillon de départ est très hétérogène. Mes résultats démontrent que si le volume initial est généré par un grand nombre de micrographes, l'algorithme ML ignore l'hétérogénéité qui reste dans l'échantillon. Donc, il n'est pas nécessaire d'inclure des filtres additionnels pour l'analyse.

Directional RANSAC est une méthode d'analyse qui génère de l'information structurale des échantillons hétérogènes. La moyenne des classes deux-dimensionnelles est génère pour chaque classe, puis ces moyennes sont sélectionnées de manière aléatoire pour générer plusieurs classes trois-dimensionnelles. Ensuite, un analyse « Principal Component Analysis » (PCA) est effectuée

pour générer des cartes trois-dimensionnelles pour identifier des patrons structuraux. Les groupes structuraux sont ensuite combines avec les autres groupes les plus similaires. Finalement, plusieurs cartes trois-dimensionnelles qui représentent plusieurs conformations de l'échantillon sont générées. J'ai testé ces approches avec plusieurs échantillons hétérogènes et mes résultats étaient satisfaisant compares aux approches qui sont utilisées en ce moment.

En plus, ces cartes trois-dimensionnelles peuvent servir de référence pour raffiner d'avantage les multiples structures finales trois-dimensionnelle d'un échantillon hétérogène.

English Abstract

The three-dimensional (3D) structures of biological complexes obtained from Single-particle Cryo-Electron Microscopy (SPA-CryoEM) provide crucial structural insights that help in the determination of their functions. One of the most challenging aspects of SPA-CryoEM is sample heterogeneity. In my research project, I address two different issues found in the processing of highly heterogeneous samples. My first aim is focused on obtaining the best possible reconstruction of the predominant conformation while my second pursues to obtain reconstructions of all different conformations in the heterogeneous sample. Current approaches fail to address these two problems in highly heterogeneous datasets. Hence, I developed two novel computational methods to address these two challenges separately: Directional Pruning and Directional RANSAC.

Directional Pruning is a method that enables us to remove remanent heterogeneities or the artifacts in the difficult datasets which are not removed by previous screening steps. This method uses directional classes, which groups particles with similar orientation together and performs 2D classification on these groups, classifying particles into two classes per direction. The particles belonging to the least populated 2D classes are removed from every directional group. The retained particles are considered homogenous particles. My hypothesis is that the retained homogenous particles will improve the quality of the final reconstruction. To test the hypothesis, different datasets of structurally heterogeneous samples were processed with directional pruning. The results showed only a slight improvement but did not really improve the quality (resolution) of the reconstruction.

In order to understand my results further, I developed a new set of experiments. These experiments were designed to test the theory of the maximum likelihood (ML) algorithm used in processing of the datasets. Our results show that when I provide a good initial volume, ML algorithm automatically disregards remanent heterogeneities/artifacts present in the dataset without requiring any screening or pruning process.

Directional RANSAC is a method that provides structural information of many different conformations in heterogeneous datasets. Directional classes are used to generate a number of class averages in all angular directions. Then, class averages are randomly selected along with their angular information to generate multiple 3D maps. Principal Component Analysis (PCA) is performed on generated 3D maps to identify the underlying structural components. These

structural components are clustered and averaged on the basis of their similarity. Then, multiple 3D maps, representing different conformations of the macromolecule in the heterogeneous dataset are produced. The proposed approaches were tested with different structural heterogeneous datasets and the results are found to be satisfactory in comparison with current approaches.

Moreover, these 3D maps can be used as reference volume in further 3D classification/3D refinement steps to obtain final 3D reconstructions of the macromolecule of interests in many different conformations.

Acknowledgement

First, I would like to thank my supervisor, Dr. Javier Vargas for his complete support, guidance, and encouragement throughout my master's degree.

I would also like you to thank my mentor, Dr. Khanh Huy Bui, and other committee members, Dr. Joaquin Ortega and Dr. Alba Guarne, for their guidance and concern about my project progression over past two years.

I want to acknowledge my lab members Dr. Josue Gomez Blanco, who helped me whenever I faced technical problems in the lab and Ms. Satinder Kaur, who provided professional and personal support during my master's degree.

Once again, I would like Dr. Joaquin Ortega, apart from being my committee member. His lab provided me with experimental data for my master's project. He also allowed me to participate in his lab's weekly meeting meetings.

I also want to thank Ms. Siobhan Schenk, who helped me with my thesis and translated my abstract into French. I would also like to thank, Ms. Muthu Lakshmi Muthu who helped me to adapt to the university and in the department.

I wish to avail myself of this opportunity to thank my beloved parents, sister, and friends for their continuous interest in my progress and constant support.

Lastly, I would like to Dr. Craig Mandato, Dr. Chantal Autexier, Ms. Joelle Denomy, and other Anatomy and Cell Biology department members. They provide support in many aspects, which helped student's life at McGill.

Introduction

Cells are building blocks of any living system [1]. There are different types of cells. These cells consist of a variety of structural components which signifies the function of the cells. To investigate the unknown features of the cells, structure components are studied in detail. Amino acids are the basic unit of the cells. The majority of all cell structures can be simplified to an amino acid sequence [2]. The series of different combinations of amino acids give rise to protein. Hence, the most identifiable constitute of all structural compartments of cells is the protein. Proteins are derived from genomic sequences. Cell's genetic material DNA or RNA contains information to synthesis protein [3]. Proteins help in cell shape and carry out a variety of cell functions. Hence, they are studied in detail to understand the different features of the cell. Proteins are classified based on their structure and function. These two parameters are highly correlated with each other in how the protein structure can determine the function and vice-versa. Studying different types of cellular proteins or their structural information will help in better understanding of their cellular function [4].

3D Structure Techniques

There are different techniques with different principles used to determine the 3D structure of macromolecular complexes, or cell compartments. The universal methods to acquire 3D structures are X-Ray crystallography, Nuclear Magnetic Resonance (NMR), and Transmission Electron Microscopy (TEM). Sometimes a single technique is insufficient to provide structural details of the biological sample. In such cases, the mentioned techniques are combined [5] to complement each other. Earlier, the most commonly employed complement technique is Electron Crystallography, where the sample is crystallized like X-Ray crystals, but the samples are imaged using electron in TEM [5].

X-Ray crystallography

X-Rays crystallography is a widely used technique to solve protein 3D structures at atomic resolution. The proteins need to be crystallized, so they are orderly packed inside the crystal. The diffraction patterns of the sample, obtained by X-rays at different orientations, are used to calculate spatial distances and identify the components present [6]. Thus, enabling X-Ray crystallography to solve atomic models of the sample protein. The Protein Database Bank (PDB) which is an archive containing 3D structure information of biological molecules,has more atomic model depositions from X-Ray crystallography than from any other structural methods [7].

Although there are many atomic models solved by X-Ray crystallography, the native structure of the protein may not be preserved. This is mainly due to the crystallization of the sample. Also, not all protein sample yield crystals. If the sample protein exhibits heterogeneity showing different confirmation, it is really tedious and challenging sometimes even impossible to obtain crystal for structure analysis [8].

NMR

There is another structural technique named Nuclear Magnetic Resonance (NMR), which was widely famous for solving small cellular protein structures and helping to investigate the different conformations of heterogeneous proteins structure as well. NMR is used extensively to identify molecular conformations in a solution and solves physical properties like solubility, phase changes, and diffusion at the molecular level. NMR uses an external intense magnetic field to identify structural elements based on their energy transfer [9].

The major limitation of NMR spectroscopy is the size restriction of the sample which is typically limited to 70-90 KDa.

3D Electron Microscopy

Initially, light microscopy was used to investigate the components of the cells. However, scientists were unable to see the structures beyond the provided resolution which was typically around $0.2\mu m$ in optical microscopy at that early times. The inability to study the cells lead to the discovery of the Electron Microscopy (EM) in 1931, which was developed by Ernest Ruska [10]. The invention of Electron Microscopy helped to visualize the cell components beyond the light microscope resolution.

Electron Microscopy comprises of several different of techniques such as TEM, Scanning Electron Microscopy (SEM), etc., which are actively used in the field of Cell Biology. TEM is often used to image and analyze of cell proteins along with their structure. TEM provides information about the samples internal structure where SEM focuses on morphology and chemical composition of the sample. Due to the high vacuum state inside the microscope, the living sample cannot be imaged directly like in light microscopy. The presence of any form of water components will degrade the high vacuum conditions of EM, causing contamination of the sample in the form of artifacts and noisy images [11].

Hence, there are unique sample preparation methods in EM. Namely resins, chemical and temperature treatments and, heavy metal staining are used to remove water content from the samples such that they can be imagined by Electron Microscopy [12]. Sometimes the sample preparation can disrupt the sample integrity. Using these methods will help visualize the outer structure of the sample. However, detailed information of inner components will remain unachieved, and the resolution of this structure will be far from atomic resolution [13].

Cryogenic Electron Microscopy (Cryo-EM)

These structural techniques outlined before are considered as universal techniques of structural biology to solve cellular structure or reconstruct macromolecules in three-dimensions. The performance of these techniques improved with the addition of new technology advances. One of the technology advances that made a breakthrough in TEM was introduction of Direct Electron Detectors (DED) in Cryo-electron microscopy (Cryo-EM).

Cryo-EM is a structural technique that uses TEM to study macromolecular assemblies in a thin layer of amorphous ice. Cryo-EM differs from classic EM in terms of its capacity to preserve the native state of protein or macromolecules [14]. To maintain the native state of the samples, the macromolecules in solution are plunged freeze at-198 ºC to obtain a thin layer of amorphous ice. As explained before, usually TEM cannot image biological samples due its water content, which will evaporate in the high vacuum conditions maintained inside TEM. Cryo-EM on freezing the biological sample at -198 ºC transforming water contents into vitreous ice [15]. The data collected from the EM may be used to reconstruct 3D structures of the macromolecules under study with the help of image processing computational methods. Cryo-EM overcomes the significant hurdles faced by X-Ray crystallography and NMR Spectroscopy, which requires crystals and smaller proteins for better results, respectively. Unlike X-Ray crystallography and NMR Spectroscopy, Cryo-EM doesn't require a highly concentrated sample for the imaging process. Cryo-EM widely is sub-divided into two main categories: Single Particle Analysis (SPA) and Cryo-Electron Tomography. These two methods are profoundly different from each other in terms of the samples used, imaging techniques, and software used for image processing.

In Cryo-Electron Tomography (Cryo-ET) (Figure 1.1), large macromolecular machines, an entire organelles or even cells scan be reconstructed. This image modality consists of tilting the sample at different angles during imaging processing inside the EM [16]. There is a physical restriction in the maximum tilting of the sample and usually the sample cannot be tilted more than ± 60 degrees. The tilt restriction results in the 'missing wedge problem' as structural information above this ± 60 degrees angle is unavailable. The missing wedge can be identified by an elongation or stretching of the reconstructed volume, misshapen or missing structural features. The missing wedge can be minimised by dual axis tomography where the sample is rotated to 90 degree and tilted.

Another major problem faced by Cryo-ET is radiation damage. In Cryo-ET the same sample area must be radiated multiple times during the collection of tilt series. The radiation damage occurs mainly due to prolonged exposure of sample to electrons while tilting. The radiation damage can be minimised by reducing the dose of electrons which in turn results in noisy projection images, which are challenging to be aligned.

The data collected from the electron microscope is computationally compiled in a way that the angular information during tilting is used to finally obtain a 3D reconstruction of the sample using typically using a weighted back projection [16]. The final reconstructed tomogram in Cryo-ET has a limited resolution of typically 12-40Å.

An extension of Cryo-ET, namely sub-tomogram averaging, overcomes the problems faced by classical Cryo-ET to achieve high resolutions reconstructions. As with Cryo-ET, sub-tomogram averaging analysis the complexes in-situ structure determination without any separate isolation or purification techniques allowing in-situ analysis. These complexes are found in multiple copies within the cell. These complexes are detected, aligned, averaged, and reconstructed to 3D structures [16]. The highest recorded resolution of sub-tomogram averaging achieved so far is 3.1 Å(EMDB-8986).

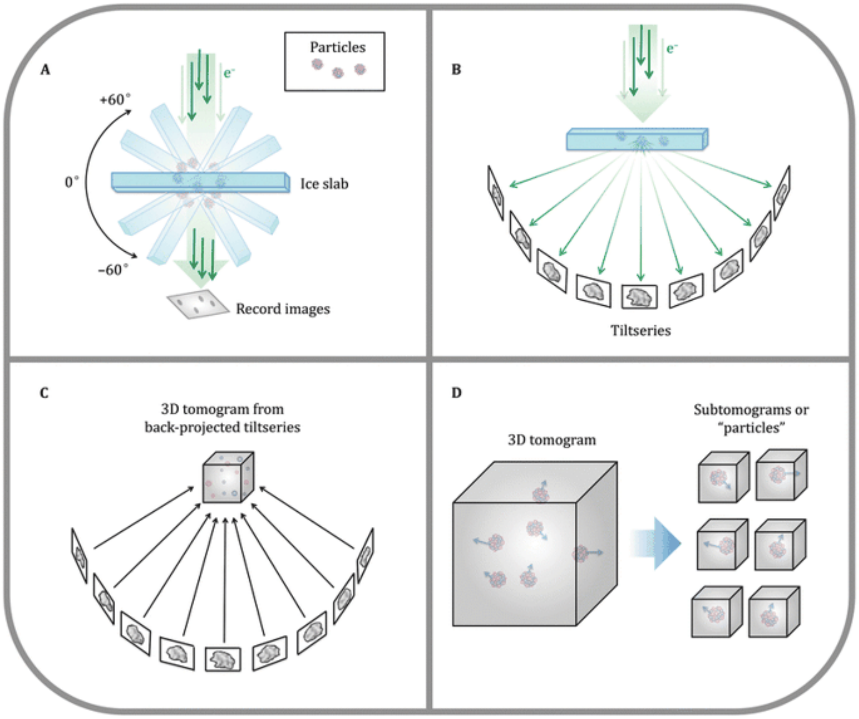

Figure 1.1: Principle of Cryo-Electron Tomography image processing [17] (A) Tilting of the sample during the imaging by electrons inside TEM. (B) Aligning images according to the tilt-series. (C) Back-projection of tilt-series images to form 3D tomogram. (D) Sub-tomogram averaging. Image is used with permission.

SPA

Among the Cryo-EM methods, Single Particle Analysis (SPA) is likely the most successful and famous because its capacity to provide near atomic reconstructions at the macromolecular native state. In this modality, macromolecular complexes to be analyzed are biochemically isolated and purified. After the purification process, the sample is flash-frozen by plunging it into the liquid ethane to obtain a thin layer of vitreous ice where the macromolecular complexes are embedded. In this technique, the vitreous ice layer has multiple copies of the identical sample found at different orientations (Figure 1.2) whose microscopic projection images are collected from the EM and used for 3D reconstruction. SPA does not require tilting of the sample for 3D reconstruction. Hence, the disadvantages of Cryo-ET are surpassed by SPA. Like sub-tomogram averaging the samples are averaged. There are some inconveniences during imaging that affects the 3D reconstruction, for example, the use of low electron dose conditions to avoid damage of the biological sample results in noisy images. Low contrast images are really difficult to interpret.

Despite these inconveniencies SPA is widely useful in solving biological structures at near-atomic resolution [18]. At that resolution, it is possible to build an atomic structure or model that plays a significant role in drug development or design sectors. In additions, SPA is able to solve 3D structures of small and large macromolecules at low sample concentration and at its native state. It is possible to obtain 3D structure of small and large macromolecules at low sample concentration and provides information about structural conformation as well. The number of structure depositions in the Electron Microscopy Data Bank (EMDB) is higher when compared to other Cryo-EM methods [19].

The conversion of the SPA microscopic data into 3D structures requires a series of computational steps. In general, they are referred to as a workflow. The typical SPA workflow includes motion correction, contrast transfer function (CTF) estimation, particle selection and extractions, 2D classification, initial volume, 3D classification, 3D refinement and finally, reconstruction.

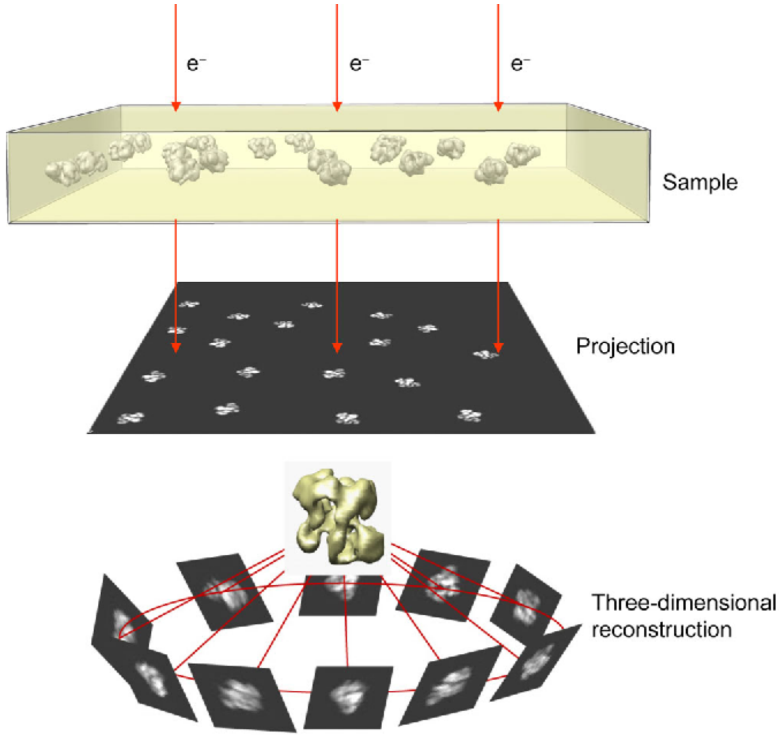

Figure 1.2: Principle of Single Particle Analysis Cryo-Electron Microscopy [20]. Image is used with permission.

Motion correction

This is the first step in the image processing workflow where the data collected from the electron microscope is aligned. The collected data is in the form of movies, not images, unlike traditional EM data. The frames comprising the movies are aligned to reduce the blurring in the resultant images caused by the beam-induced motion (BIM) [21-23]. The movement of the sample is due to the interaction of the sample with the electron beam and causes a blurring effect in the resultant images. Before the technology breakthrough experienced by Cryo-EM in 2012, it was not possible to compensate for the BIM, thus the obtained images and resultant final 3D reconstruction were inevitably blurred and of low resolution. However, with the advent of a new generation of camera, the DED that records electron impacts directly as opposed to previously used indirect electron detectors like CCDs (Figure 1.3) [23], it now is possible to achieve routinely near-atomic 3D reconstructions for appropriate samples. The correction of the BIM in the EM images, has become a routine pre-processing step that has allowed the so called resolution revolution in cryo-EM [24].After tracing the sample movement caused by the BIM and aligning the movies frames, the corrected frames are reduced to an aligned average micrograph that preservers the high resolution content.

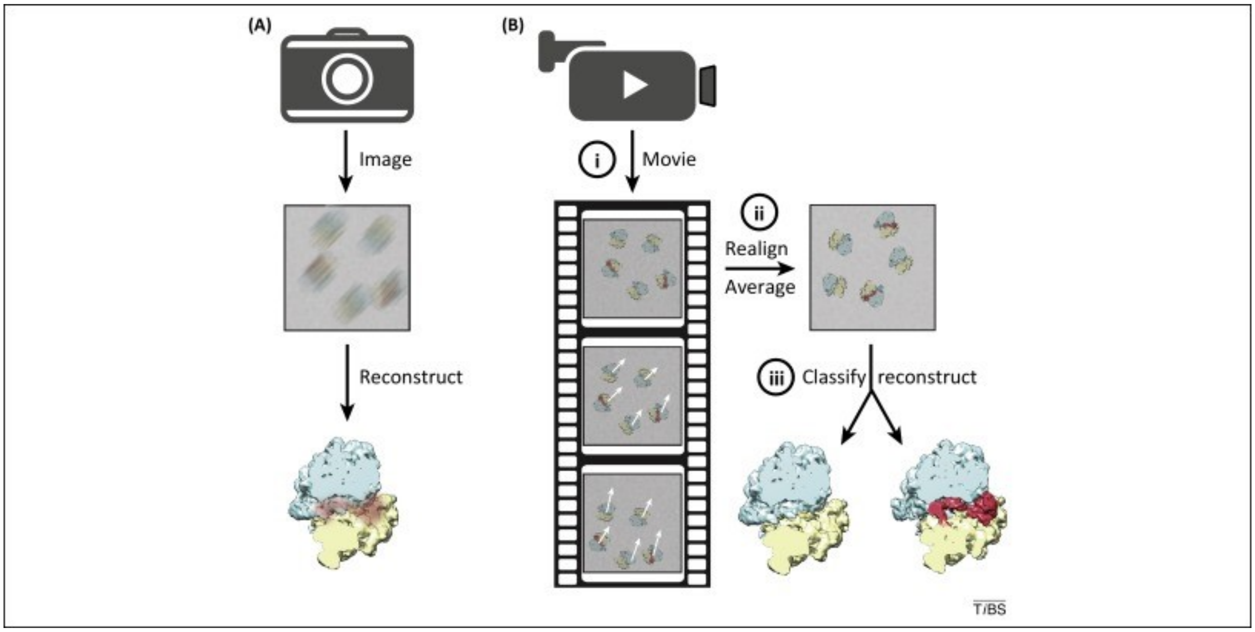

Figure 1.3: (A) Represents effect of Beam induced movement on CCD image. (B) DED movies which enable the motion correction induced by BIM [26]. Image is used with permission.

CTF estimation

CTF estimation step is one of the screening steps in the SPA workflow where the optical aberrations, mostly the defocus aberration affecting the micrographs are estimated. Based on this information, micrographs are screened.

CTF is estimated to identify the distortions introduced by the electron microscope to the EM images, correct them and screen the micrographs based on the results of the fitted CTF [27-30]. The CTF pattern of a good micrograph will have perfect concentric circles moving from the center towards the edge. The CTF pattern of a bad micrographs can be identified by asymmetric rings, corresponding to images affected by astigmatism aberration and/or faded rings showing the presence of sample drift in a particular direction during imaging (Figure 1.4) [28]. Micrographs showing a low quality CTF pattern (as shown in Figure 1.4) should be removed from the image processing workflow as they contain distorted information that will affect the quality of final 3D reconstruction.

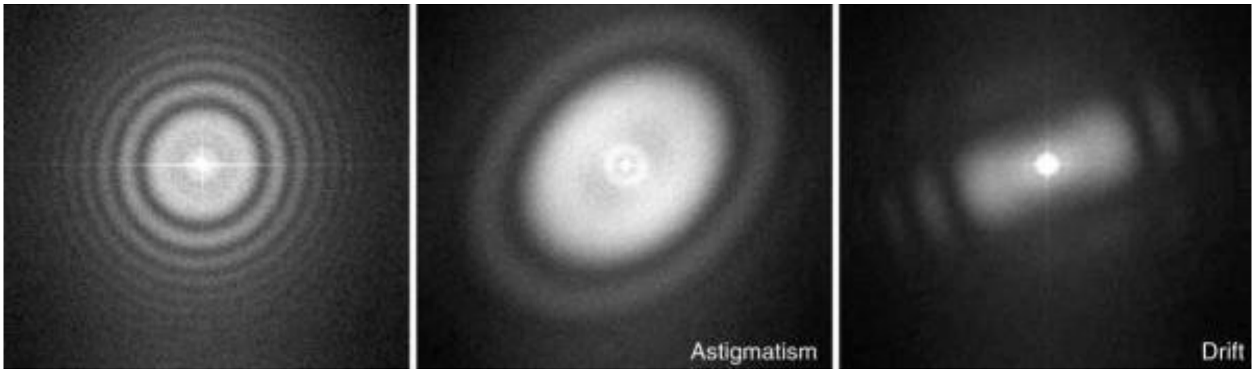

Figure 1.4: Screening of Micrograph based on CTF [28]. Image is used with permission.

Particle selection and extraction

The particle selection step is crucial as it directly contributes to the quality and resolution of the final 3D structure. Identifying and selecting each and every particle from the micrograph set is the aim of this step (Figure 1.5). Initially, this task was carried out manually to choose a few hundred of thousands of particles from the collected micrographs [34]. To facilitate this task, many automatic [31-33] and semi-automatic approaches have been proposed [34-35]. However, these approaches are prone to usually select also many incorrect picked particles corresponding to false positive.

After the particle selection step, the selected particles are extracted [32-35]. Then, a particle image set is created, where each image only contains one particle projection that is located at the centre of the image. Moreover, after the particle extraction step, it is highly recommended to perform a particle screening task on the extracted particles to detect and remove false-positive particles from the processing [36-37]

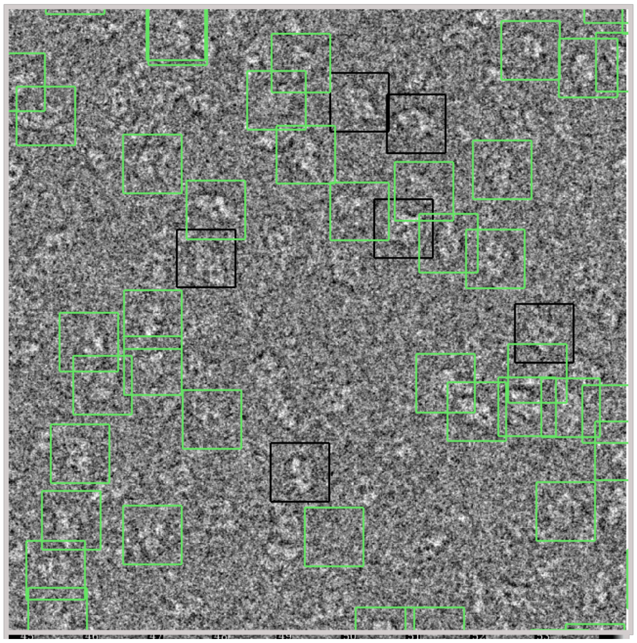

Figure 1.5: Particle Picking interface on a micrograph [35]. Image is used with permission.

2D classification

The aim of the 2D classification step is to group particle projection images with similar orientations to finally average them to improve the signal to noise ratio of the resultant average image [38]. Considering that the particle images provided by cryo-EM are always very noisy and affected by very low contrast, these averaged particles with improved the signal to noise ratio (SNR) allow to assess the quality of the dataset and preclude the existence of potential problems such as the presence of preferential orientations. The averaging of particles is efficient when the similar particles are grouped together. The particles are classified and aligned based on their similarity. Here particle similarity represents similarity in terms of particle orientation inside the vitreous ice layer. In other words, the particles with similar orientation are grouped together, aligned, and averaged (Figure 1.6). Different approaches have been proposed to perform the task of 2D classification [38-40]. The maximum likelihood approach is one of the most popular and effective methods to achieve the 2D classification. This approach assumes that each particle is in all possible orientations simultaneously but contributing with different statistical weights. Based on these weights, particles are averaged and assigned to the different classes. To improve the accuracy of the 2D classification task, sometimes reference images are considered [38]. In this case, the particles can be classified and aligned based on these references, which are provided by the user [40]. Along with the computed 2D class average and the orientation information, the 2D classification indicates the presence of symmetry or heterogeneity in the sample. Moreover, this process is usually employed as an additional particle screening task to remove particle images classified into groups providing yielding featureless, low resolution or inconsistent 2D class average, which are referred to as junk particles.

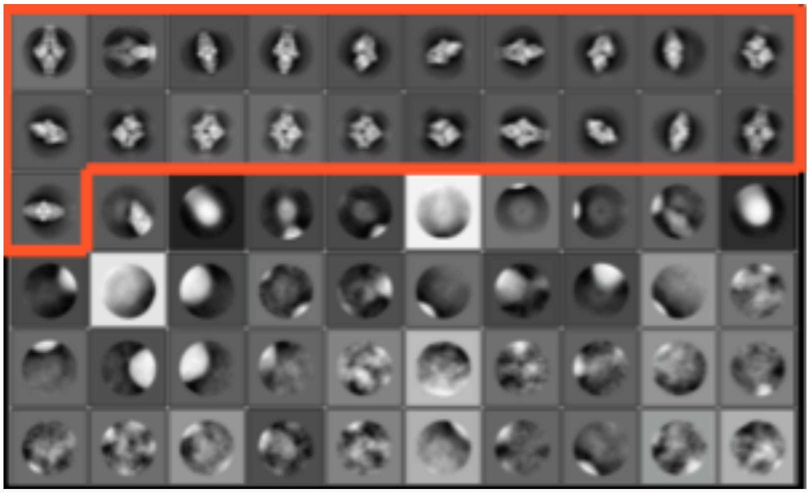

Figure 1.6: 2D classification of Beta-Galactosidase dataset. Highlighted 2D class averages are good class averages selected for next step in image processing while remaining class average are bad averages assumed to have junk particles [41]. Image is used with permission.

Initial Volume

A first guess of the final 3D structure that is required to further process the data. If the sample of interest has a predetermined model like a homology model that has been previously solved, this 3D reconstruction can be used as first guess or initial volume [42]. This initial volume will be low pass filtered to prevent bias during further processing steps. However, if the sample of interest doesn’t have a predetermined model, then this initial model should be obtained from the data computationally. Some approaches may provide an initial volume [42-44]. One of the commonly used approaches to generate an initial volume a RANSAC [44]. In the RANSAC approach, orientations are randomly assigned to a subset of 2D class averages, and multiple random volumes are generated. These volumes are also scored by RANSAC comparing their projections with the other class averages (Figure 1.7). The best-scored volume is selected as initial volume for further processing.

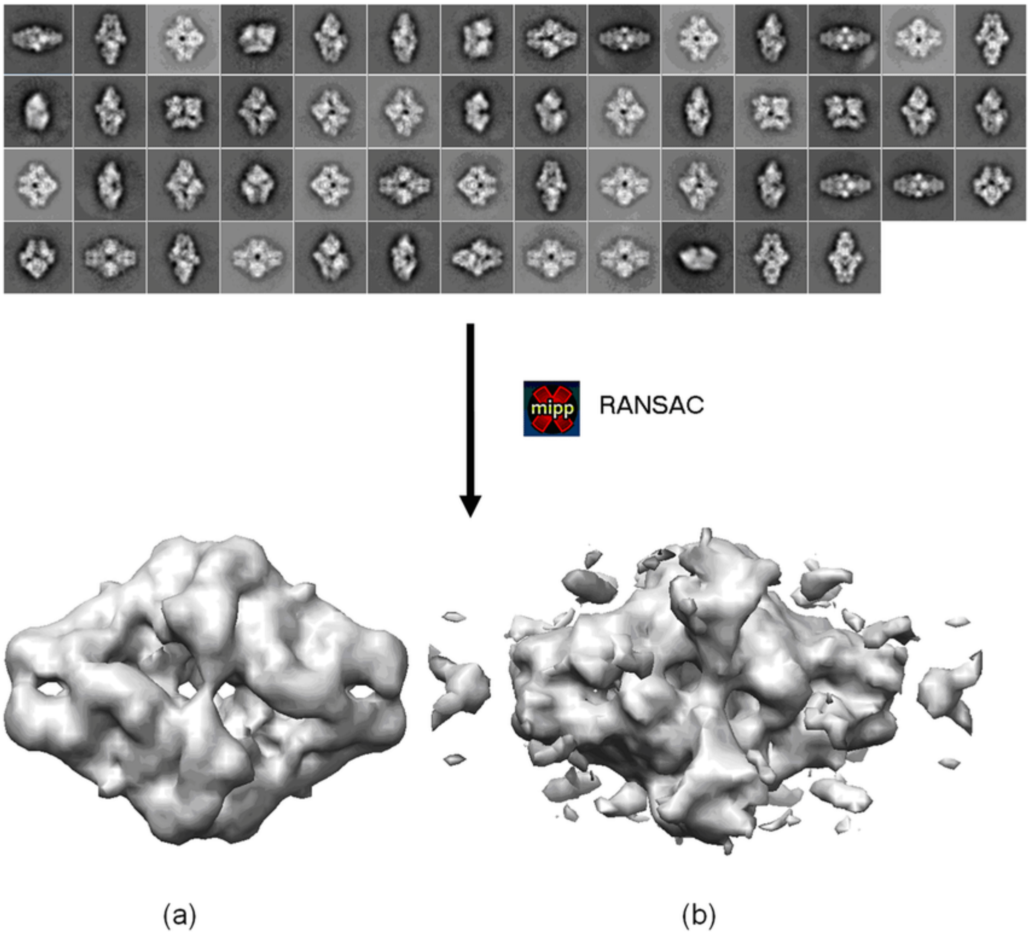

Figure 1.7: Initial volume generation employing the RANSAC approach (a) good and (b) bad (due to incomplete density information) initial volumes are generated from 2D averages of Beta-galactosidase dataset [45]. Image is used with permission.

3D classification

The aim of 3D classification approach is to group particle images based on their conformation. Usually, the 3D classification approach requires an initial volume to perform the classification [29,38-40]. If the dataset is homogenous showing an unique conformation the 3D classification will result in one 3D class comprising the majority of the dataset particles. Other minoritarian 3D classes can be formed but are assumed to contain only junk particles [42]. In the presence of a heterogeneous dataset i.e., a dataset collected from a macromolecular complex showing different structural conformations, each 3D class represents one different conformation. Along with this, the 3D classification step provides information of the prominent conformation identified in the sample inside the vitreous layer. The presence of an inappropriate initial volume in heterogeneous datasets sometimes causes bias in the classification, which causes that all particles migrate towards one class irrespective of their conformation [42]. Thus, it is essential to use accurate initial volume and low pass filtered initial volumes. During the low pass filtering, all high-resolution information is filtered out leaving only the low-resolution content.

3D refinement

In the refinement step, the angular information of particles bellowing to each 3D class is refined iteratively. Thus, the resulting 3D class is improved though different iterations to obtain finally a high-resolution 3D map. The 3D refinement has mainly two different approaches to refine the particles' angular information: maximum likelihood approach [39-40] and projection matching [46].

In projection matching, the 3D map is projected at different directions. Each experimental projection images are compared with all map projections so the orientation of the best-matched map projection is assigned to that particle image. This process is carried out iteratively such that angular information of all particles is obtained. The angular information provided by projection matching is not accurate due to the high level of noise in the images.

In the maximum likelihood approach, it is assumed that each particle is presented in all possible orientations but with different statistical weights. Based on these weights, particles are averaged and assigned to the different classes. The maximum likelihood approach has shown to be more robust to noise than projection matching.

The resolution of the 3D refined structure is evaluated by the Fourier shell correlation (FSC) [47-49] through the Gold Standard approach. In the Gold standard approach, the dataset is divided into two halves which are processed independently, giving rise finally to two independent 3D reconstructions. These two independent maps are band-pass filtered at different resolution, and at every resolution, the correlation is calculated between them. At low resolutions, the two maps are usually very similar due to the presence of low noise, so the correlation is found to be close to one. As the resolution increases, the two maps show differences between them, thus the correlation starts to drop off

The resolution of the 3D maps is typically evaluated by the 0.143 cut-off value. The 0.143 cut-off value means that when the correlation coefficient of the two independent values equals to 0.143, the resolution at the point is considered as the final resolution. Though 0.143 cut-off value is commonly used, some studies use 0.5 cut-off value and other criteria to evaluate resolution.

Reconstruction

The 3D reconstruction step is the final step of the SPA workflow, which generates the final 3D structure (a density map). All particles with the refined orientation information are re-projected back to obtain the final 3D structure [40]. After obtaining the 3D reconstruction, the map is usually low pass filtered at the resolution provided by the FSC during the 3D refinement process. This final map may be used, depending on its resolution to build an atomic model [50]. The constructed model can be deposited in the protein database bank whereas final 3D structure can be deposited in Electron Microscopy Data Bank irrespective of its resolution.

Objectives

Resolution could be intuitively understood as the ability to resolve minute details of the sample of interest [51]. Hence the smallest improvement in the resolution number sometimes would result in unlocking vital structural information of the sample of interest.

In Cryo-EM, there are different sub-categories of the resolution which are commonly used to categories the structures, high-resolution, medium-high resolution, medium-low resolution, and low resolution. At each of these sub-categories, different features of protein structures are unlocked.

The low-resolution range is typically defined at until 8-10 Å resolution, where alpha-helices look like tubes, beta sheets as tubes. The medium-low resolution range is between 5-7 Å, where the handedness of the secondary structures can be differentiated. Medium-high resolution ranges fall within 3-4 Å where the atomic backbone is traceable, deep grooves and precise pitches of the secondary structures are visible. This resolution is sufficient to build an atomic model. High resolution is also known as a near-atomic resolution which is the highest resolution one could achieve through Cryo-EM. High resolution ranges above 3 Å, where amino acid side chains are distinguishable [52]. The current highest recorded resolution in the Electron Microscopy Database Bank (EMDB) [19] is 1.25 Å (EMDB ID: 11103).

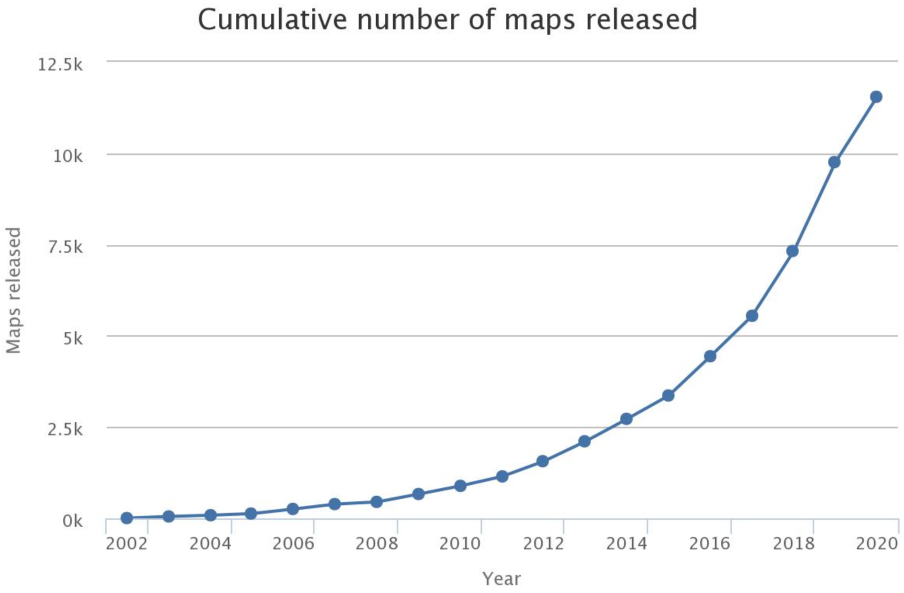

Figure 2.1: A chart representing maps released in EMDB over the last 18 years obtained from their official website

The invention of the Direct Electron Detectors (DED) [21] in 2012 enabled the transition of resolution from low resolution to the medium and high-resolution range. This transition phase is widely called a Resolution Revolution [25]. The number of Cryo-EM maps released in EMDB has increased geometrically (Figure 2.1)

It is a tedious effort to obtain high resolution or the medium-high resolution of any sample of interest. There are some challenges faced. In some cases, the challenges are due to improper sample preparation or purification techniques, but commonly identified challenges in Cryo-EM are the presence of preferred orientations and structural heterogeneity in the samples. These two significant challenges hinder the samples from achieving their best resolution.

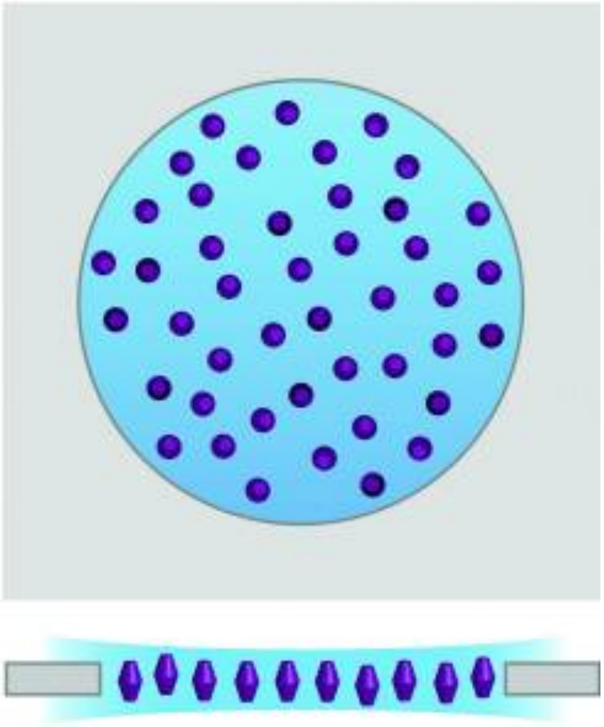

Figure 2.1: Representation of preferred orientation [55]. Image is used with permission.

Preferred orientations (Figure 2.2) are found to be present when the sample of interest interacts with the air-water interface during plunge freezing. This interaction causes the sample to orient at a particular view, making it difficult to obtain complete angle information. As SPA requires samples to be present in the random view to reconstruct 3D structures. On processing the samples of preferred orientation will results in elongated structure or misshapen structure wherever structural information is missing. The structures are similar to that of missing wedge problem in Cryo-ET. There are some specific types of samples, like membrane proteins, prone to be affected by preferred orientation. Accordingly, these samples require a particular type of sample preparation like pre-treating the grid with specific chemical compounds[53-55]

Structural heterogeneity refers to the presence of different structural conformations in the sample of interest. This is a common phenomenon that occurs even after the purification of the sample.

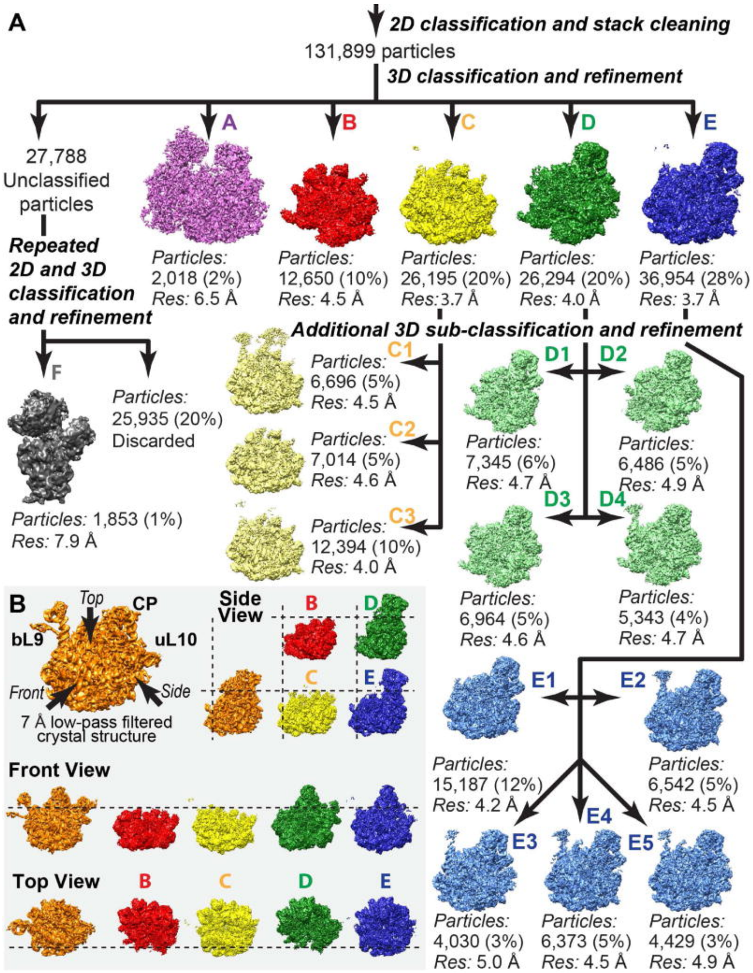

Figure 2.2: Cryo-EM dataset consisting different structural conformers[58]. Image is used with permission.

Thesis Challenges

There are two different problems faced while processing highly heterogeneous dataset. First, to obtain pure particles that belong to dominant structural conformation in the dataset. There are various cleaning and classification methods are available to overcome this challenge, but these methods are not as efficient for highly heterogeneous samples.

Second, to obtain structural information of all existing conformation in the dataset., the existing classification approaches can provide few different conformations from datasets showing moderate heterogeneity but they fall short when trying to obtain many different states from highly heterogeneous datasets.

Current available methods

Even after performing a thorough cleaning/screening process, there will be a presence of some junk particles or particles belonging to remanent heterogeneity present in the dataset. The current method of the standard image processing pipeline to distinguish different 3D structures from a heterogeneous structural dataset is the 3D classification by Bayesian inference [40]. Using an initial volume as reference for the classification of particles, the particles belonging to different 3D structures will be segregated, especially in cases showing massive heterogeneity, flexibility or cases where macromolecules shows a stable conformation but also some transient states, which are much less unlikely. These cases usually result in one dominant 3D class containing the majority of particles. This scenario is commonly termed as “attractor problem”. Though the heterogenous datasets consists of more conformations, the number of 3D classes are restricted. The restriction of number of 3D classes results in improper classification which “attracts” particles from other classes towards one class containing majority of particles in dataset [42].

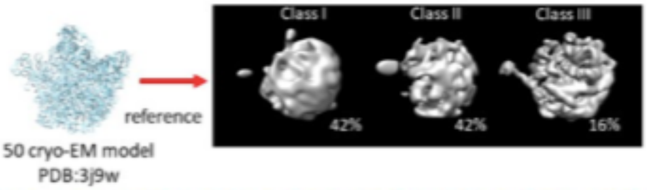

Figure 2.3: 3D classification methods fails to classify heterogeneous particles based on structure conformation [42]. Image is used with permission.

In such cases, the final 3D structure of that predominate conformation is highly influenced by remanent heterogeneity, which affects the overall resolution of the 3D structure (Figure 2.3). In addition, the other minor conformations are very challenging to obtain and require high expertise in cryo-EM methodology.

In Figure 2.3, it is observed that RELION 3D classification fails to perform an accurate 3D classification of a highly heterogenous sample. In this case, the sample is a mixture of immature 50S bacterial ribosome subunit particles are classified with initial volumes as mature 50S ribosomal subunit.

The classification can fail to provide (i) A single 3D class representing the predominant conformation in the heterogeneous dataset, which clearly is of low quality mainly because of the presence of particles from different conformation or states, and (ii) 3D classes representing other states that do not represent all existing conformation in the heterogeneous dataset.

Proposed methods

There are two approaches developed to overcome the two challenges indicated above during the image processing of a highly heterogeneous dataset.

The first method, called Direction Pruning, aims to achieve homogenous particles sets, free of remenant heterogeneities and artifacts so that the overall quality of the 3D map representing the predominant conformation will be improved.

The second method, called Directional RANSAC, aims to achieve to obtain individual maps representing all minority class/ structure present in the heterogeneous dataset so that conformation change among the same molecule can be studied in detail.

Methods

Scipion is a cryo-EM software platform mostly focused on Single Particle Analysis where many different image processing packages are integrated and can be used to create workflows. This software enables us to use different cryo-EM image processing packages at the same time without any compatibility issues.

The novel methods developed through this master thesis are included in Scipion platform and belong to the Cryomethods package included in Scipion. The methods are developed in Python programming language using Pycharm IDE software. The source code for these methods is publicly available from Cryomethods package public repository[56].

Input conditions

For both methods: Directional Pruning and Directional RANSAC, two forms of inputs are required.

1) Input Particles: Particles images with alignment information are required for the approaches. The methods fail to analyse the particles without alignment information.
2) Input Volume: An initial 3D reconstruction of the macromolecule is required for the algorithm to group particles with similar orientation and generate references for the 2D Classification step.

Steps involved

a. Grouping of particles

The first and foremost step in the proposed directional methods is the grouping of particles with similar orientation (directional classification). Then, a 2D Classification task is performed over the different directional classes.

Both methods use input parameters like angular distance and angular sampling, which are used to group the particles into the different directions. Angular sampling represents the number of different equi-spaced directions used to sample the projection sphere while angular distance represents the size (in degrees) of the solid angle used to group particles around each projection direction. The selection of appropriate values for these parameters depend on the number of particles available. Usually, the particle population for different orientations can be estimated during 3D classification through the angular distribution. The angular distribution is obtained based on the angular information of particles used in the reconstruction. Thus, from a 3D reconstruction process, it is easy to identify the number of particles found at a particular angle. After grouping the particles, the particles are ready for the 2D Classification step. Henceforth groups are referred to as blocks.

b. 2D Classification

To perform the 2D Classification, another set of input parameter are required.

(i) Classification type: This method can perform two different types of 2D Classification. The available methods are CL2D [38] and RELION 2D classification [40].

(ii) Number of classes: Total number of classes required per block. This parameter determines the number of 2D classes to be used in the 2D classification process run to sort out the particles of each directional class.

(iii) Number of iterations: Total number of iterations to be performed for the 2D classifications inside each block.

(iv) Number of particles: This parameter determines the minimum number of particles required for classification in each block. If the block has a lower number of particles than this value, the classification process is automatically skipped, and all the particles are present in that block are assumed grouped into a single class. Note that if there is not a sufficient number of particles present in a block, then the 2D classification process would fail. Hence it is essential to provide a correct number. A typical value for this parameter is 500-1000 particles.

Part 1: Directional Pruning

This method aims to remove particles corresponding to remanent heterogeneities and false-positive particles present in heterogeneous datasets.

A. Pruning

After classifying the particles of each block, low populated classes in every block will be removed. To perform this task, another important input parameter called the threshold value is required. The low populated class can be identified based on the class population. The class population refers to the number of particles that belong to a 2D class. The classes with the class population less than the threshold value will be automatically removed along with the particles belonging to those classes.

B. Output

Removing all the particles belonging to the low populated class in every direction, the homogeneity of the dataset is improved, then the output is generated containing only the good and homogeneous particles.

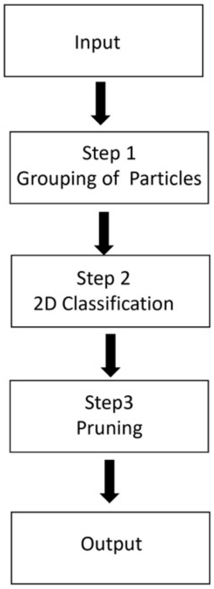

Figure 3.1: A schematic representation of steps involved in Directional Pruning method.

Part 2: Directional RANSAC

The aim of this method to generate multiple volumes representing all different conformation present in the heterogeneous dataset

A. Random selection

In this step, 2D classes averages from every direction are randomly selected. The selection occurs in a way that only one class average is selected per direction. This selection guarantees that one class average representing every direction is present to generate volumes. Randomly selected class averages are used for reconstruction.

B. Reconstruction

A large number of random volumes are generated by the randomly selected directional 2D class averages. The number of random volumes to be reconstructed is determined by a user defined parameter

C. Principal Component Analysis (PCA)

PCA is performed on the generated random volumes to identify the underlying conformations presented in the dataset. This is an essential and most crucial step in Directional RANSAC. The PCA estimates a new coordinate system where the variability of the dataset (random volumes in this case) is maximized. PCA reduces the data dimensionality to lower form where variability can be estimated with reduced data losses. When converting 3D volumes to a lower form where the structural variability is estimated, the estimated variability is called Principal components. Principal components refer to underlying structural components.

D. Cluster Analysis

The random volume generated previously can be represented as data points in the principal components space. Then, the coordinated representing the random volumes in the PCA space are clustered based on their structural similarity, and volume from each cluster is generated. Each cluster is different from each other, so the generated representative volumes are also structurally different from each other. The essential parameters required to perform this task are:

(i) Cluster methods: This method has two different cluster methods, namely K-means [59] and affinity propagation [60].

(ii) Cluster centers: total number of cluster centers required during clustering analysis.

The total number of final volumes generated during this method corresponds to cluster centers. As the one volume from each cluster is generated

E. Output

The output containing structurally different volumes from the heterogeneous datasets is generated.

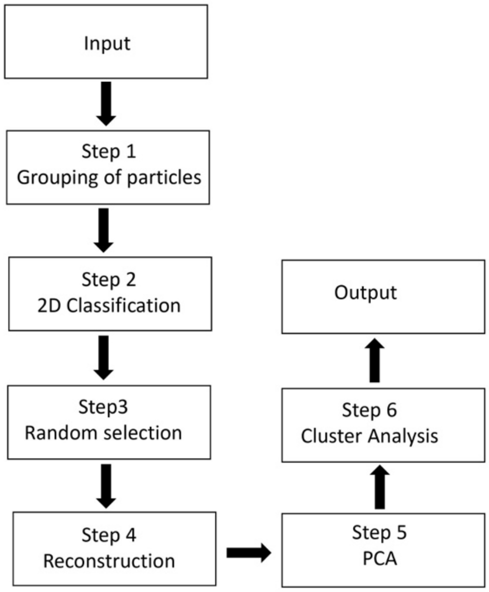

Figure 3.2: A schematic representation of steps involved in Directional RANSAC method.

Results

Three different massive heterogenous datasets were used for processing. Each method was developed and tested by using only two out of the three datasets. One of three datasets was unrefined and unpublished data whose structure information remains unknown. The other two datasets were refined and published. For Directional RANSAC, both refined and published datasets were used to compare the efficiency of the methods. For Directional Pruning, the unrefined and unpublished dataset and one of the refined and published dataset were used as it was available at the time of processing.

Preprocessing of the datasets

Ribosome I

Ribosome I or 45S Yphc depleted ribosome cryo-EM experimental data was collected by Dr. Ortega's laboratory that was shared. This dataset is composed by around 6303 movies which were converted into 6303 aligned micrographs generated after movie alignment using motioncorr [24]. These aligned micrographs were screened and corrected by CTF using CTFFIND [30]. Around 200 micrographs were discarded after manual CTF screening. Around 898603 particles were picked across 5949 CTF screened micrographs through Xmipp particle picking method, a semi-automatic approach particle picking [34]. RELION 2D classification was performed [40] to classify the dataset into 128 classes. Out of which, 22 good classes were selected for further processing. Then,10 Random initial volumes were generated from these 22 class averages by Xmipp RANSAC approach [44]. The best-ranked volume was selected as an initial volume. The particles belonging to the 22 selected classes and the initial volume were used as input for RELION 3D classification [40]. These particles were classified into three 3D classes. The particles belonging to the majoritarian 3D class were selected as input for Directional Pruning and Direction RANSAC

Spliceosome

Spliceosome data was found to have a highly flexible component in its structure which makes the processing of this sample difficult. The complete dataset of the spliceosome was downloaded from EMPIAR official website (EMPIAR ID code: 10180). This data has multiple EMDB entries (EMDB ID: 3682-3688), as its RNA structures and overall stability was studied in detail [57]. The downloaded data contains extracted particles with orientation parameters. CTF information and reconstructed volume at resolution 7.2 Å using 0.143 FSC criterion.

Ribosome II

This dataset is publicly available and corresponds to CryoEM particles images of L17-depleted 50S ribosomal intermediates. Ribosomal particles images and the corresponding reconstructed volume for processing were downloaded from the official EMPIAR website (EMPIAR ID code: 10076). The data available from EMPIAR consists of the particle stack along with CTF information but without alignment information. The processing was carried out exactly as mentioned on its corresponding published paper [58]. The paper has published several structural conformers (EMDB ID:8434,8440-8453, &8455-8457) obtained from the dataset. The paper mainly focuses on determining the possible ribosome assembly pathways using the obtained structural conformers. The published work contains five(A-E) main 3D classes. Out of the five classes, three(C-E) additional sub-classes were additionally sub-classified into a total of 12 3D classes. One of the conformers (Class D) was selected randomly for processing by our methods. The particles from class D containing around 28,000 particles were selected as input.

Directional Pruning

As mentioned earlier, only two of the three massive heterogenous datasets were used to test each of our methods. For Directional Pruning, Ribosome I (unpublished data) and Spliceosome data (published data) were used for processing. Ribosome II was not processed as the dataset wasn't pre-processed at the time of directional pruning testing.

Hypothesis: Directional Pruning will improve 3D reconstruction quality by removing remanent heterogeneities or artefacts after 3D classification/3D refinement.

Ribosome I Method Set-up

A 3D class containing 342,069 particles and an initial volume from Xmipp RANSAC were selected as input particles and volume. The parameters for the Directional Pruning method were as follows: RELION 2D was selected as the method for directional classification. Two 2D classes were generated over 30 iterations. The minimum number of particles required for classification was selected as 1000, and the threshold value to remove particles based on the class population, was 0.25. The Directional Pruning process was carried out, and around 20,000 particles were removed as remanent particles. The output particles are referred to as pruned particles.

Spliceosome Method Set-up

The downloaded particles and the reconstructed volume were selected as input particles and volume for the processing. The parameters of the Directional Pruning method were selected as follows, RELION 2D classification was selected as the directional classification method. The number of classes was selected as two and the number of classification iterations was 30. The minimum number of particles required for classifying the particles set at each direction in the 2D classification process was selected as 1000. To remove particles based on the obtained class distribution, 0.25 was selected as the threshold value. The Directional Pruning process was able to remove around 50,000 remanent particles from the dataset. Different types of tests/assessments are carried out to prove the efficiency of the pruned dataset:

(1) 3D Refinement test

(2) Random subset test

(3) Particle sorting test

3D refinement test

RELION 3D auto-refine method was performed before (complete) and after (pruned) the Directional Pruning method with the same parameters. This test was mainly done to check in the improvement in the map quality after the particle pruning method. The obtained Gold-standard FSC curve was the metric used for checking the improvement in the map quality. As the FSC curve provides the correlation at every frequency. The FSC curve of the complete and pruned particles were compared.

Figure 4.1 shows the FSC curves obtained by RELION auto-refine on Ribosome I dataset Though the correlation at every frequency is similar between both the input datasets(complete and pruned), the resolution of the pruned dataset was 5.88 Å which is slightly higher than the resolution obtained by the complete datasets that was 6.05 Å. Even though the number pruned particles were lower than the complete particles.

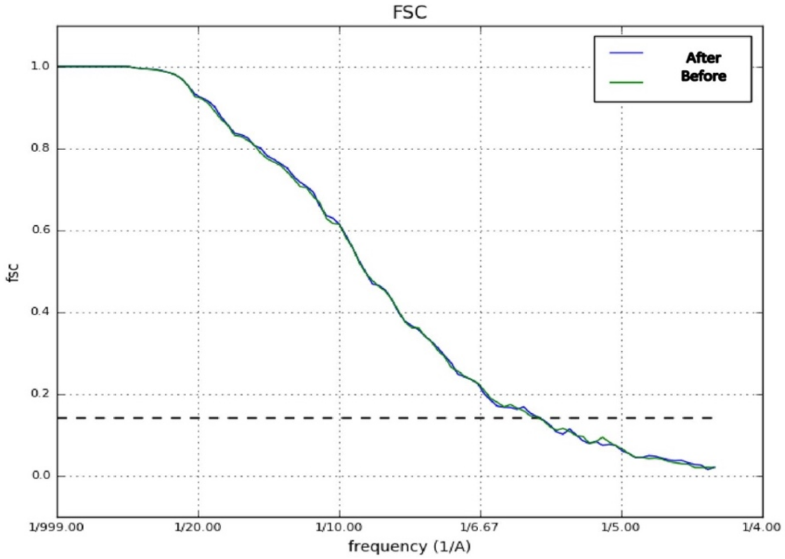

Figure 4.1 FSC curve comparison between after and before pruning for Ribosome I dataset.

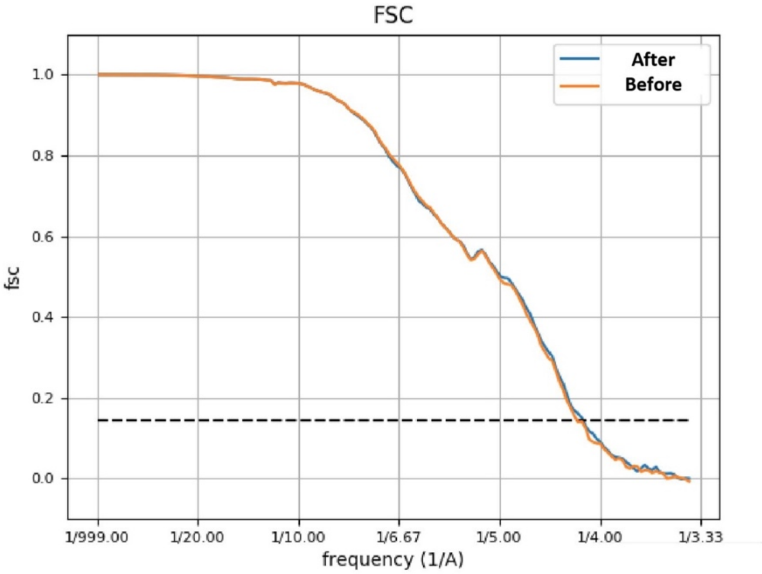

Figure 4.2 FSC curve comparison between before and after pruning for spliceosome dataset.

Figure 4.2 shows that FSC curves obtained when processing by RELION auto-refine on the Spliceosome dataset. The resolution of complete particle reconstruction was 5.91 Å whereas for pruned particle reconstruction, it was 5.97 Å. Though the resolution obtained from the complete particles set was higher than the pruned particles, the FSC curves of both datassts were similar at every frequency

Random Subset

This test was mainly done to show

(i) The adverse effect on the 3D map quality when a large number of particles are removed from the dataset.

(ii) Directional Pruning improves and retains the quality of 3D map even after removing a large number of particles.

A random subset contains of randomly selected particles from the complete dataset. The number of randomly selected particles was selected to match number of pruned particles in this test. For this test, ten different subsets were generated. The RELION auto-refine was performed on all these ten subsets to validate the aims of the random subset test.

Removing a large number of the particles from the dataset would generally affect the resolution. As the particles in these random subsets are randomly selected, sometimes, even good particles are removed. In the case of Directional Pruning only particles that where found different by the 2D classification approach for each direction are removed. In this scenario, the FSC curve of the Directional Pruning should be better when compared to the FSC of the random subset.

Figure 4.3 & 4.4, it shows the FSC curves from 3D auto-refinement test performed on pruned particles (blue curve) and ten different random subsets. In Figure 4.3, the FSC curve of the Ribosome I pruned dataset was founded to be slightly better when compared to the FSC curves of the random subsets whereas in Figure 4.4, all FSC curves of the spliceosome datasets remains the same at every resolution.

The differences obtained in the FSC analysis for the random subset test between the Ribosome I and the spliceosome datasets are mainly because of their origin. The Ribosome I dataset was a purely experimental dataset and its processing has been intermediately refined by expert users (unpublished datasets) whereas the Spliceosome dataset was already refined and published dataset.

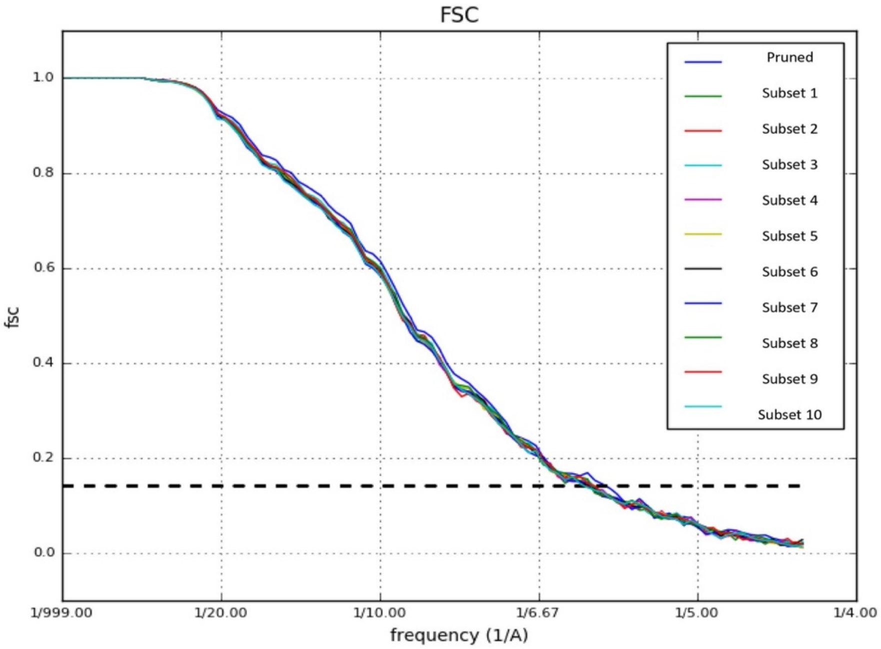

Figure 4.3: Random subset test for Ribosome I dataset.

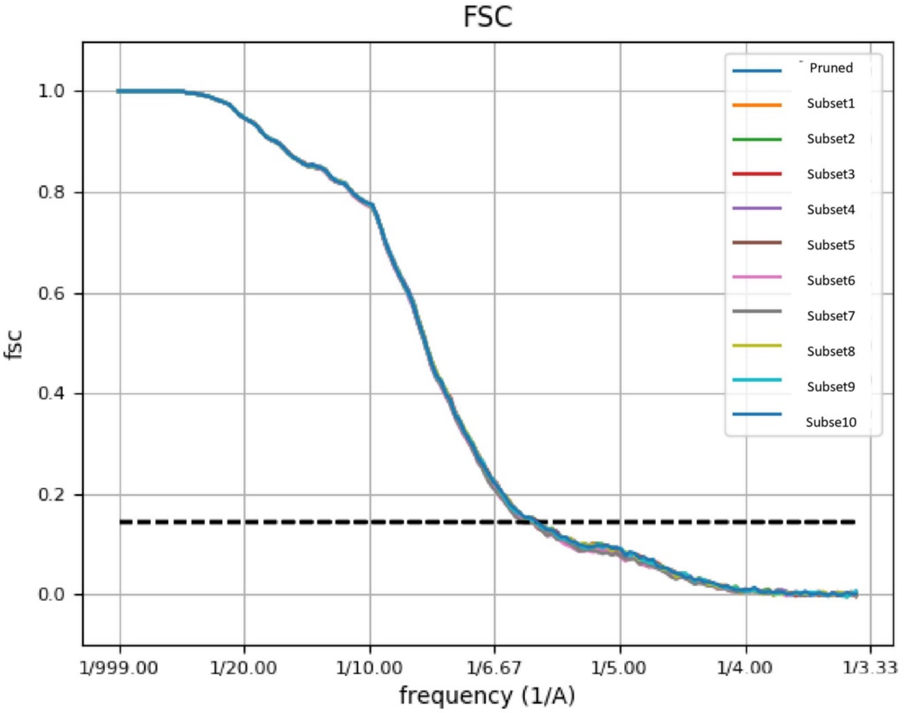

Figure 4.4 Random subset for spliceosome dataset

Particle sorting test

This test was mainly done to show that Directional Pruning method is as effective as other sorting and screening approaches in terms of removing the low-quality particles from datasets. Additionally, Directional Pruning removes heterogenous particles which makes it better when compared to other sorting and screening method.

RELION sorting approach [36] was performed over the complete datasets and a subset of particles containing the same number of pruned datasets were selected according to its Z-score. This test is mainly done to compare two cleaning methods: RELION sorting and Directional Pruning under the same conditions. The comparison between these sets was carried out by performing 3D refinement.

Figure 4.5 shows that the FSC curve of Directional Pruning is significantly better than the FSC curve of RELION sorting at all frequencies. This graph proves that the Directional Pruning approach was able to remove the artefacts and remanent heterogeneity particles and still manages to improve the FSC curve.

Figure 4.6 shows that the FSC curve of RELION sorting and Directional Pruning method is the same along with all different frequencies. This graph shows that Directional Pruning is as efficient as RELION sort in terms of removing artefacts and remanent heterogeneities.

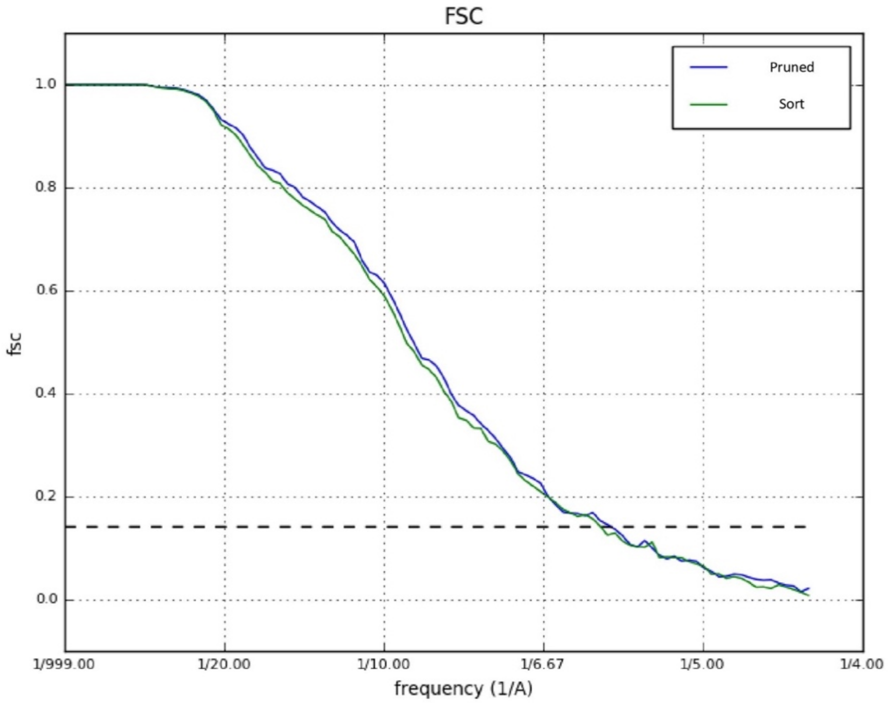

Figure 4.5: Ribosome I sort test

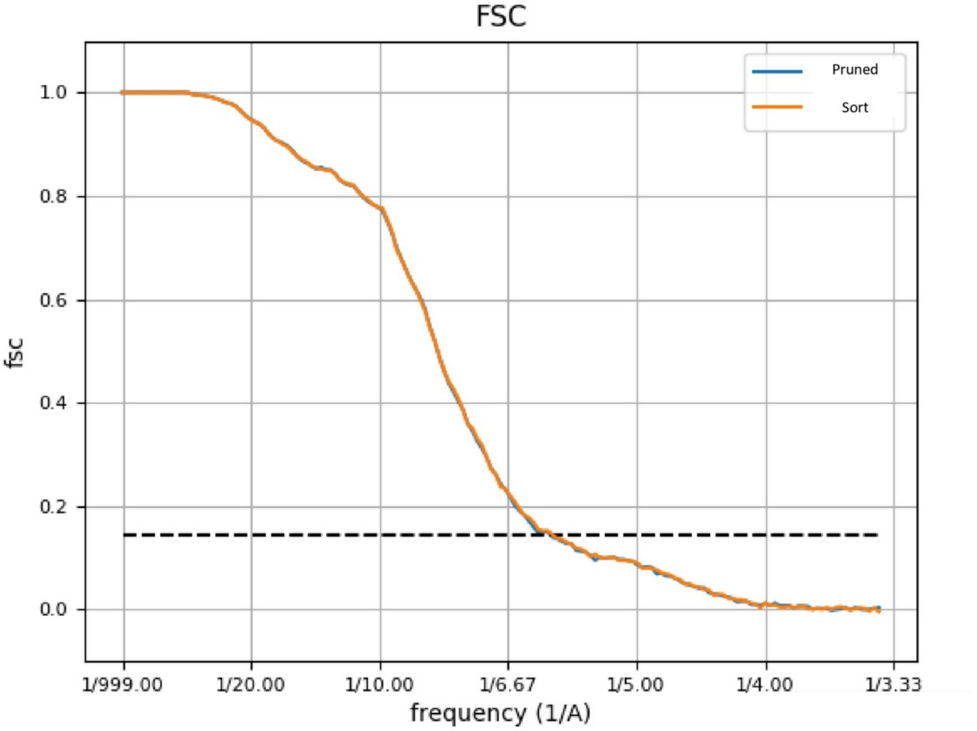

Figure 4.6: Spliceosome sort test

Alternate theory:

Even though performing different tests to evaluate the efficiency of Directional Pruning by using two different datasets, the results obtained and shown from figures 4.1 to 4.6 remain inconclusive whether the proposed hypothesis and efficiency of the Directional Pruning method is satisfactory.

The results are inconclusive because some figures show an improvement in the FSC curve, while for others we do not see this improvement. This difference could be caused by different origin of the dataset as both optimized and non-optimized datasets in terms of the image processing was used.

Initially, the similarity of the FSC curves from the 3D auto-refine test bodes well with the proposed hypothesis and the efficiency of Directional Pruning as the resolution and the FSC curve were able to maintain despite of removing a large number of particles from the dataset.

In terms of the results obtained from the particle sorting test, the figures showed similarity in the FSC curves between the sorting and pruning method. Though pruning method removed all insignificant particles whereas sorting methods removed only artefacts particles and small quantity of homogenous particles. In such scenario, the FSC curve of Directional Pruning should show improvement over FSC curve of sorting method.

This state of inconclusive results shaped us to design a new set of experiments whose aim was to understand the FSC curve similarity. Hence, another hypothesis was put forth to assess the efficiency of Directional Pruning and to understand in FSC curve similarity.

The hypothesis is based on an algorithm used in most commonly used refinement techniques. The speculation is that the maximum likelihood approach automatically disregards the presence of the remanent heterogeneity artefacts/noise particles automatically when provided with a good reference volume. The maximum likelihood approach is famously used in the 3D refinement technique. To validate this hypothesis different experiments were carried out

Control setup

To validate the newly proposed hypothesis, two different sets of the experiment was designed.

Set up 1: automatic pruning of a false-positive particle set by maximum likelihood approach

Ribosome I or 45S Yphc depleted ribosome microscopic experimental data was used to validate the second proposed hypothesis. Over 100,000 noise and artefacts were manually picked across 5000+ micrographs of Ribosome I dataset (Figure 4.7). Different percentage of artefacts or noise (insignificant) particles were added to the pruned dataset. The results are shown in Tables 4.1 & 4.2. As can be seen from these tables, adding around 100,000 false-positive particles to the pruned dataset up to 30% adulteration could be achieved. A 3D-auto refinement test was performed on these datasets affected by different number (percentages) of false-positive particles starting from 10%,15%, 20%,25% and 30%. For each percentage of adulteration, results from both noise and artefacts were compared with the 0% dataset containing only pruned particles without adding false-positives. This experiment was carried using reference volume the final map obtained previously from Xmipp RANSAC approach while pre-processing of the Ribosome I dataset and low pass filtered to 60 Å, where the resolution information above 60 Åis removed.

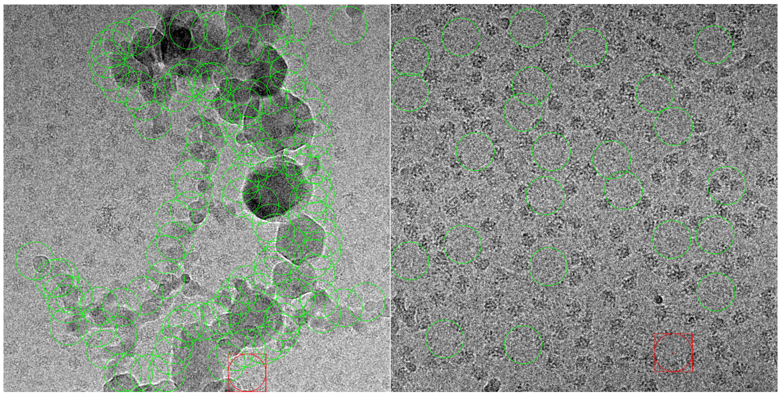

Figure 4.7: Manual particle picking of artefacts (left) and noise (right) particles

Table 4.1: Summarization of Noise particles result

% of noise	Noise particles	Total No.of Particles	3D auto-refine(A)	Final Resolution(A)
0	-	334,105	7.41	5.97
10	33,411	367,516	7.41	5.97
15	50,116	384,221	7.41	5.97
20	66,821	400,926	7.41	5.97
25	83,526	417,631	7.41	5.97
30	100232	434,337	7.41	5.97

Table 4.2: Summarization of artefacts particles results

% of artefacts	Artefacts particles	Total No.of Particles	3D- auto refine(A)	Final Resolution(A)
0	-	334,105	7.41	5.97
10	33,411	367,516	7.41	5.97
15	50,116	384,221	7.41	5.97
20	66,821	400,926	7.41	5.97
25	83,526	417,631	7.41	5.97

30	100232	434,337	7.41	5.97

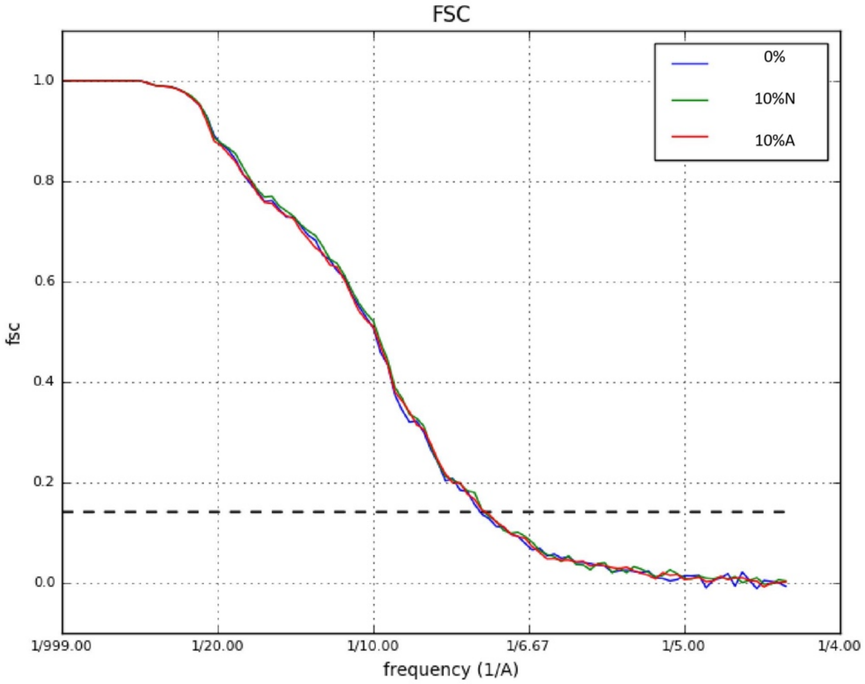

Figure 4.8: 0% vs 10% adulteration of particles

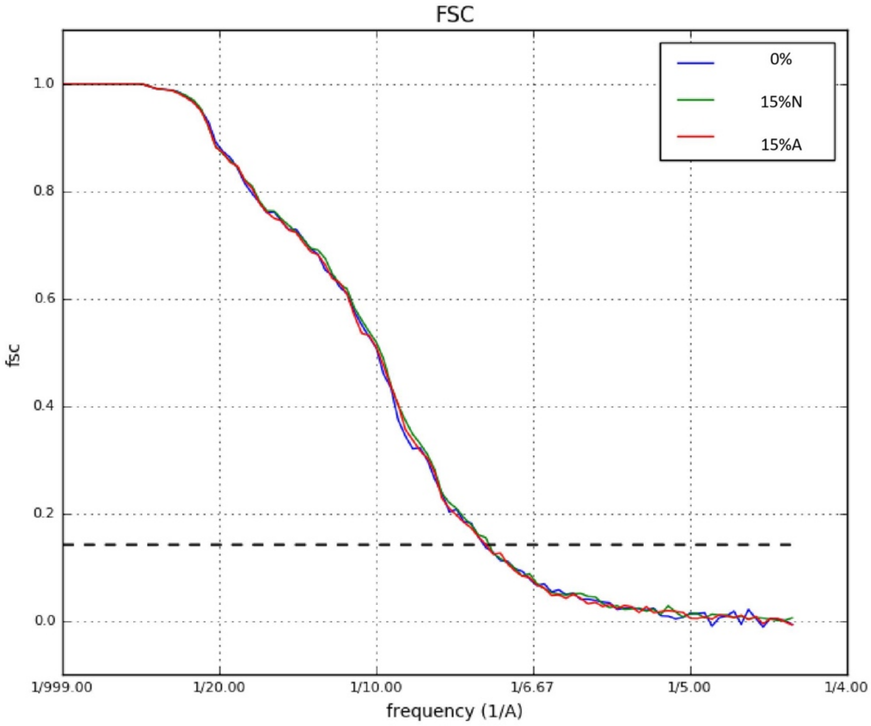

Figure 4.9: 0% vs 15% adulteration of particles

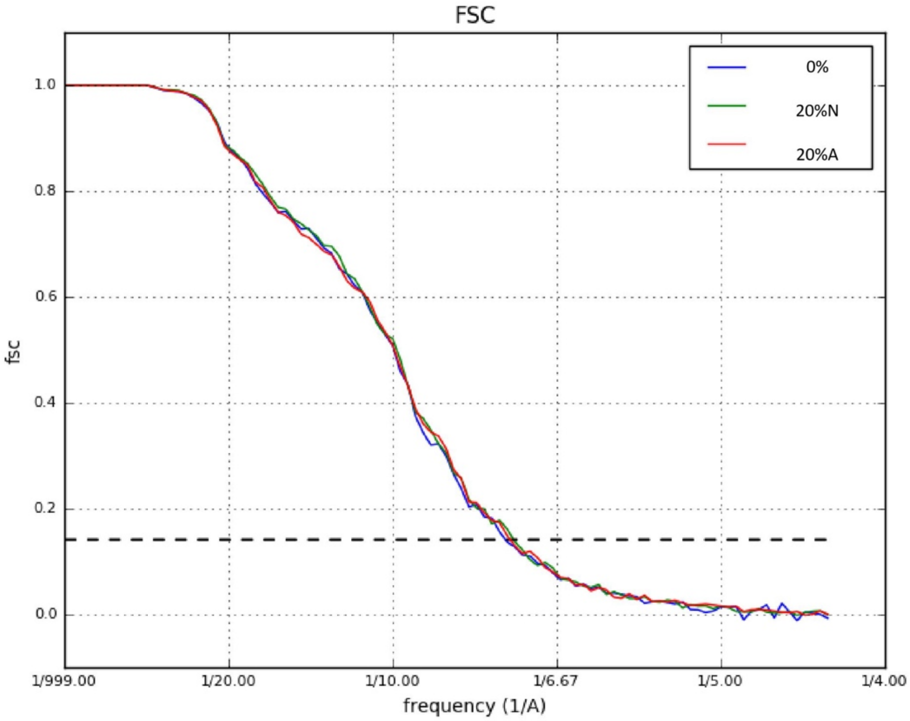

Figure 4.10: 0% vs 20% adulteration of particles

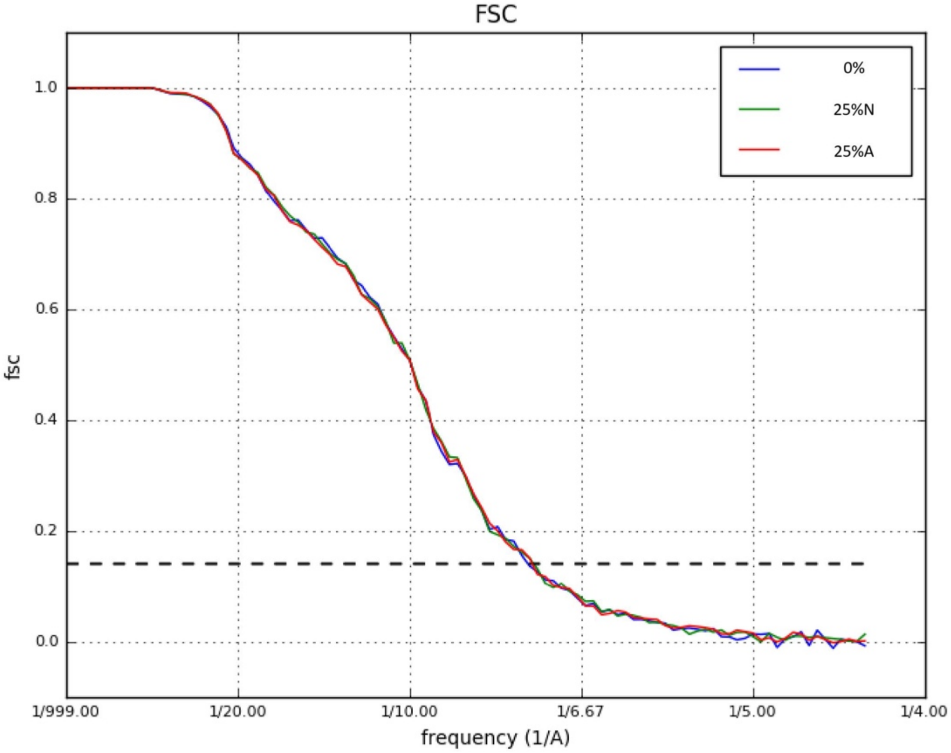

Figure 4.11: 0% vs 25% adulteration of particles

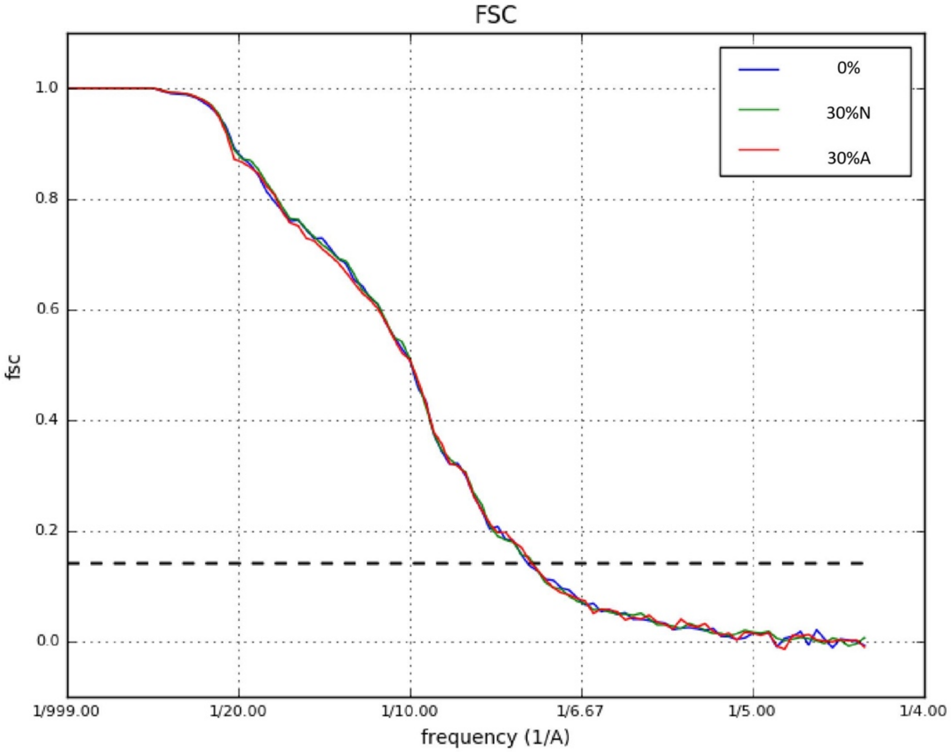

Figure 4.12: 0% vs 30% adulteration of particles

Figure 4.8-4.12 show that the similarity in the FSC curve remains the same irrespective of the percentage of particles adulteration.

Set up 2: automatic pruning fails in the absence of good reference volume

This experiment was mainly carried out to show the effect of the reference volume on automatic pruning by Maximum-Likelihood (ML) algorithm methods. According to the second proposed hypothesis, automatic pruning in ML works only when a good reference volume is provided. To test this part of the hypothesis, the initial volume obtained from Xmipp RANSAC approach while pre-processing of the Ribosome dataset was used. The initial volume was low pass filtered at two different resolution (60 Å &80 Å).

First, two different volumes were compared with each other using pruned particles as the input particles for the 3D auto-refine test which is shown in Figure 4.13. The FSC curve of the reference volume low pass filtered at 60 Å shows slight improvement over the FSC curve of the reference volume low pass filtered at 80 Å.

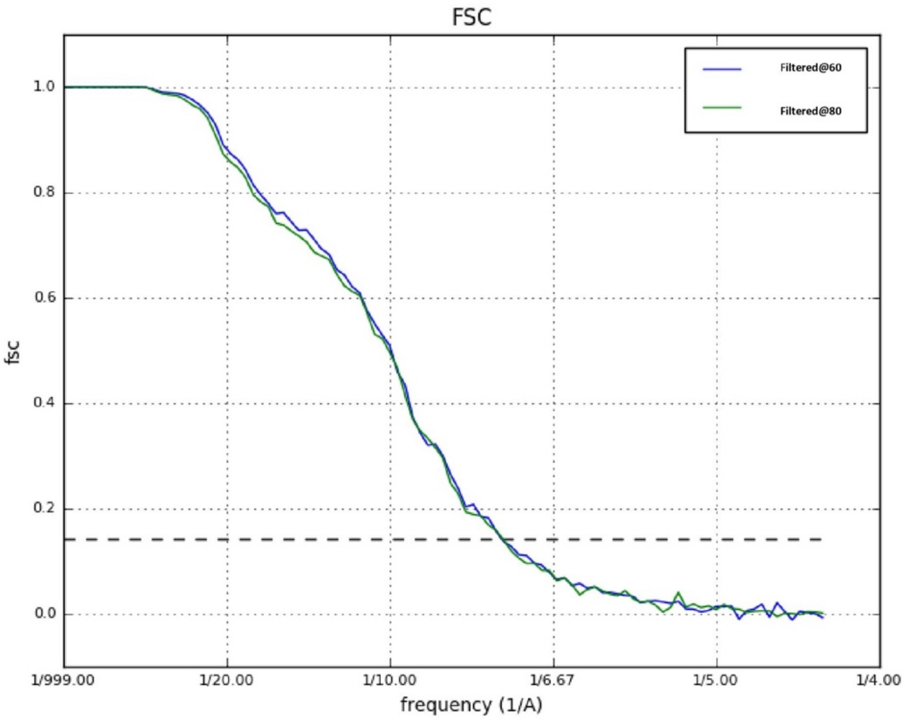

Figure 4.13: Initial volume low pass filtered at 60 Å Vs 80 Å using pruned particles as input

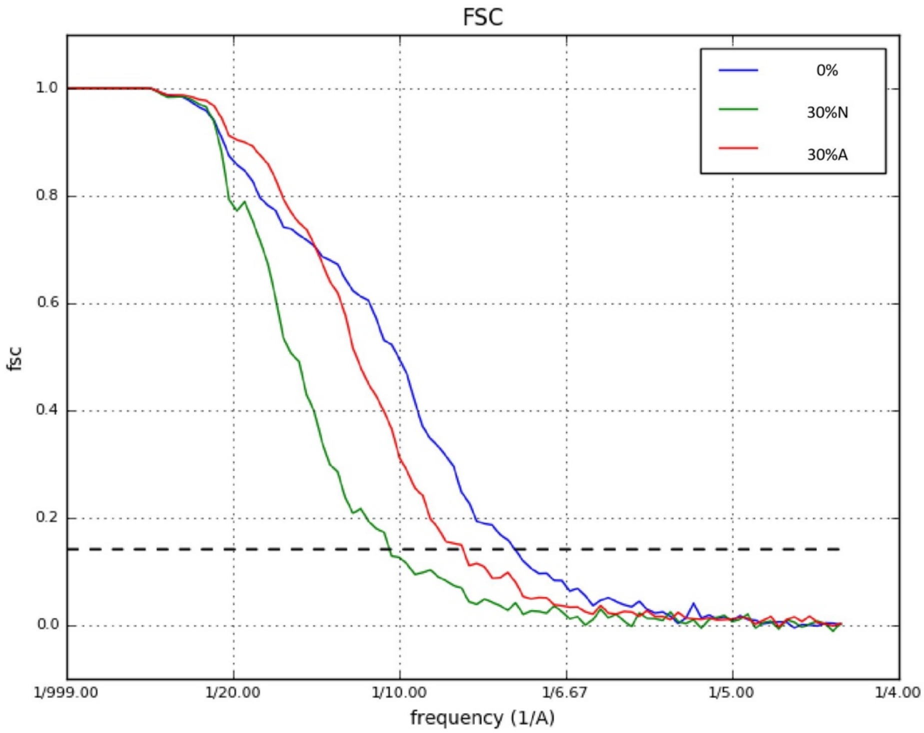

Figure 4.14: 0% vs 30% using Initial volume low pass filtered at 80 Å

Second, initial volume low pass filtered at 80 Å was used as input volume along with particle set composed by 30% noise, 30% artefacts, and 0% false positives added to the pruned particles were used as input particles. The 3D auto-refine test was performed on all these three different input sets. The obtained FSC curves shows the effectiveness of the initial volume as shown in Figure 4.14

The results of the control setup experiments were satisfactory and prove that the second proposed hypothesis that 'in presence of good reference volume disregards the presence of noise or artefact particles when the ML approach is used' was found to be true

Part 2: Directional RANSAC

As mentioned earlier, only two of three massive heterogenous dataset was used for each method. For Directional Pruning, Ribosome II and Spliceosome data (Published data) were used for processing. Ribosome I was not processed as the dataset the structural information of experimental dataset was remain unknown at the time of processing.

Hypothesis: Directional RANSAC will obtain individual maps representing all different minor classes present in a highly heterogeneous dataset.

Spliceosome method set up

Downloaded particles and volumes were selected as input particles and volumes. The parameters for directional RANSAC method were as follows, RELION 2D was selected as the method for directional classification. Eight 2D classes were generated over 30 iterations. The minimum number of particles required for classification was selected as 100 so 2D classification with these parameters will be performed for every direction. Five random volumes were generated and PCA was performed on the generated random volumes to obtain the principal components. The coordinates of the random volumes over the principal components were classified in 15 different clusters using K-means cluster analysis. From each cluster, one volume was generated, then, a total of 15 representative volumes were generated as output (Figure 4.15).

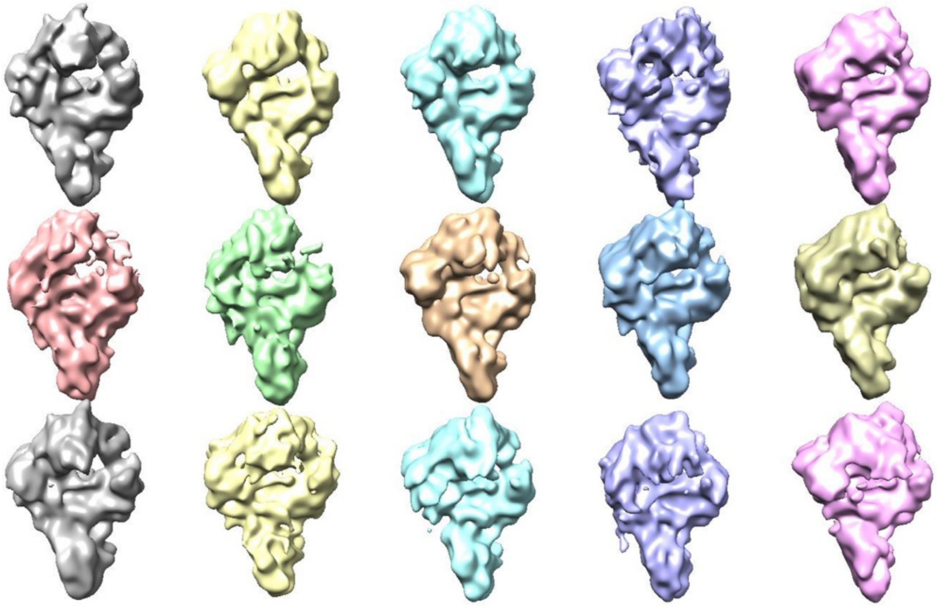

Figure 4.15: Spliceosome volume generated by Directional RANSAC

Ribosome II method set-up

A 3D class with 28,000 particles were selected. The parameters used were the same as in the spliceosome datasets except that in this case the number of selected classes per direction was two. The random volumes were projected and clustered into 5 clusters using the obtained principal components and K-means approach, giving rise to 5 different output volumes (Figure 4.16).

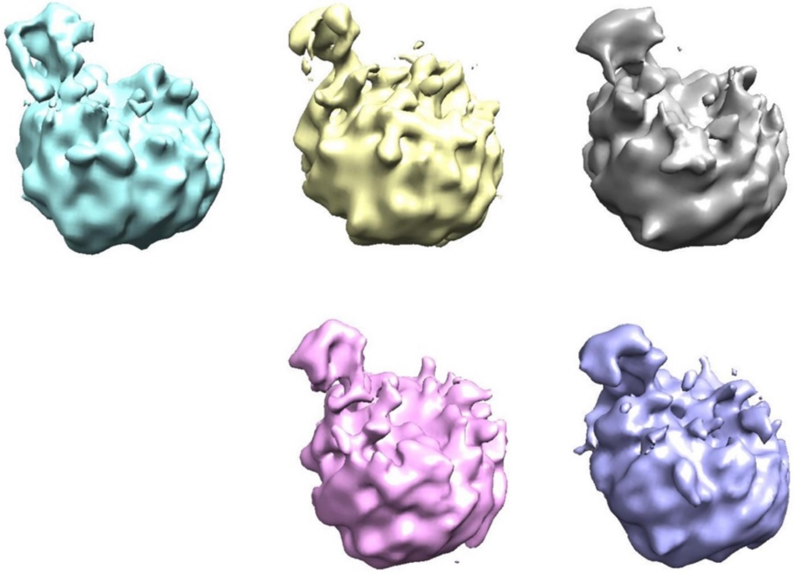

Figure 4.16: 50S Ribosome intermediate volumes obtained from Directional RANSAC

There are two different assessments were performed to evaluate the efficiency of the Directional RANSAC

(i) 3D classification test

(ii) 3D auto-refinement test

3D classification test

This test was performed to compare the existing approach to classify the data into different structural conformers (RELION) and the new developed approach, Directional RANSAC.

The 3D classification was performed using the same input particles and input volume as Directional RANSAC. The dataset was classified into ten 3D classes for the Spliceosome (Figure 4.17) and five 3D classes for Ribosome II (Figure 4.18) were obtained. Figures 4.17& 4.18 shows how 3D classifications fails to obtain the 3D classes when compared to the structures generated by Directional RANSAC (Figure 4.15&4.16).

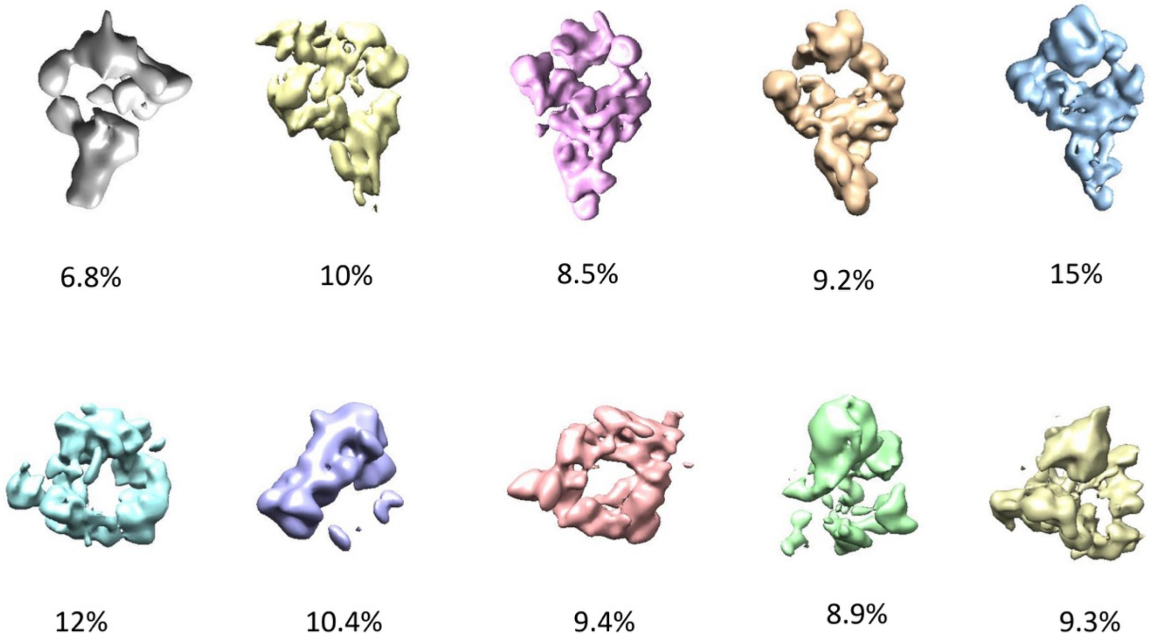

Figure 4.17: Ten 3D classes of Spliceosome dataset obtained by RELION and obtained particle distribution for each class

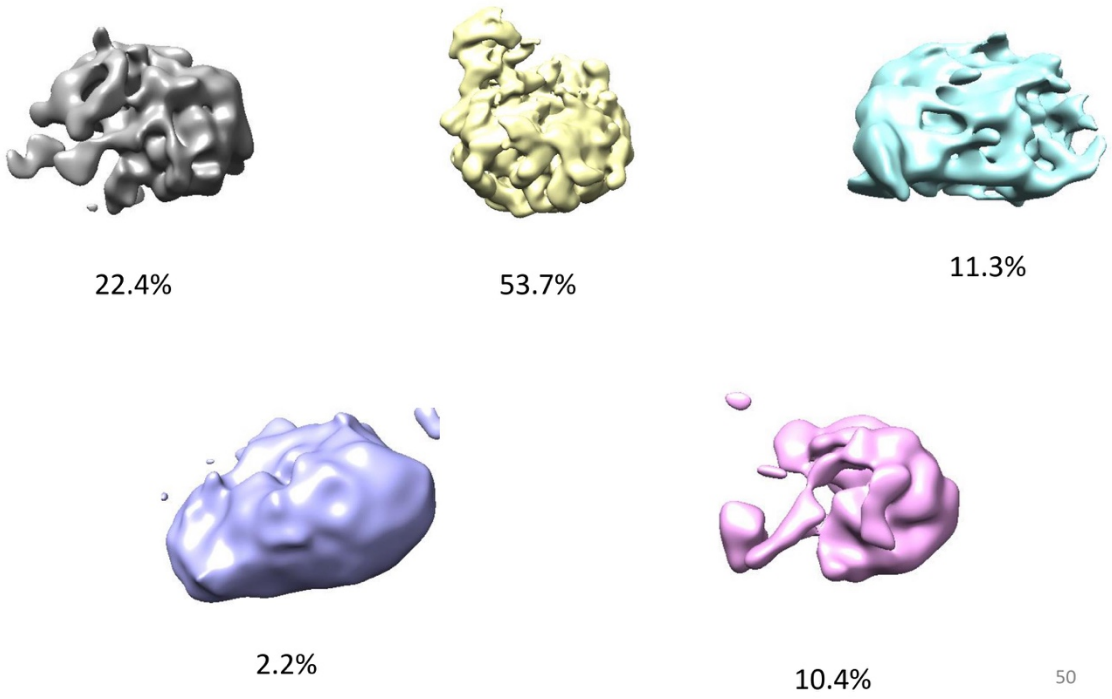

Figure 4.18: Five 3D classes of Ribosome II dataset obtained by RELION and obtained particle distribution for each class.

3D auto-refinement test

This test was mainly done to check whether the volume generated by directional RANSAC provides the same FSC curve when compared to initial volume which was used as input volume (original volume) for directional RANSAC approach.

The 3D auto-refine test was performed using 3 maps out of the fifteen and five output volumes generated by Directional RANSAC using the Spliceosome and Ribosome II dataset respectively. The volumes were selected randomly and used as input volumes for the 3D auto-refine test. Figure 4.19 shows that the generated volumes were able to deliver the same FSC curve and resolution that of the original volume (OV in the figure).

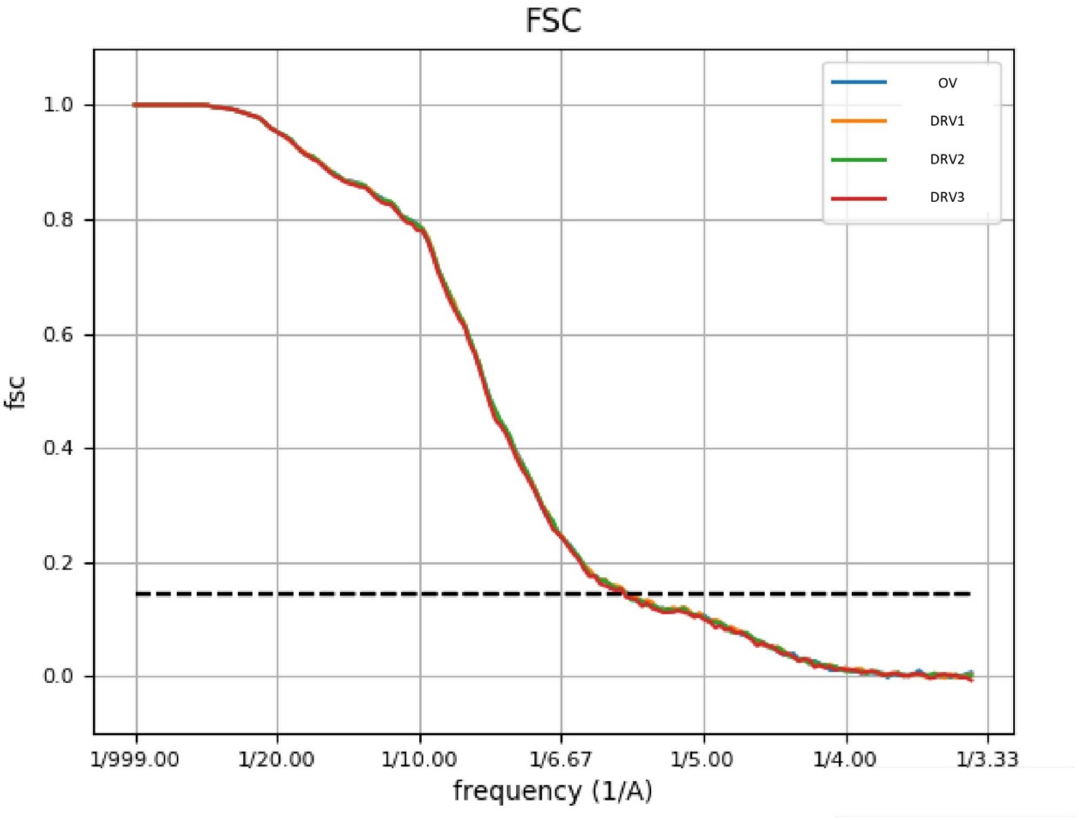

Figure 4.19: Comparison between Original initial volume (OV) vs Directional RANSAC volumes (DRV) for the Spliceosome dataset

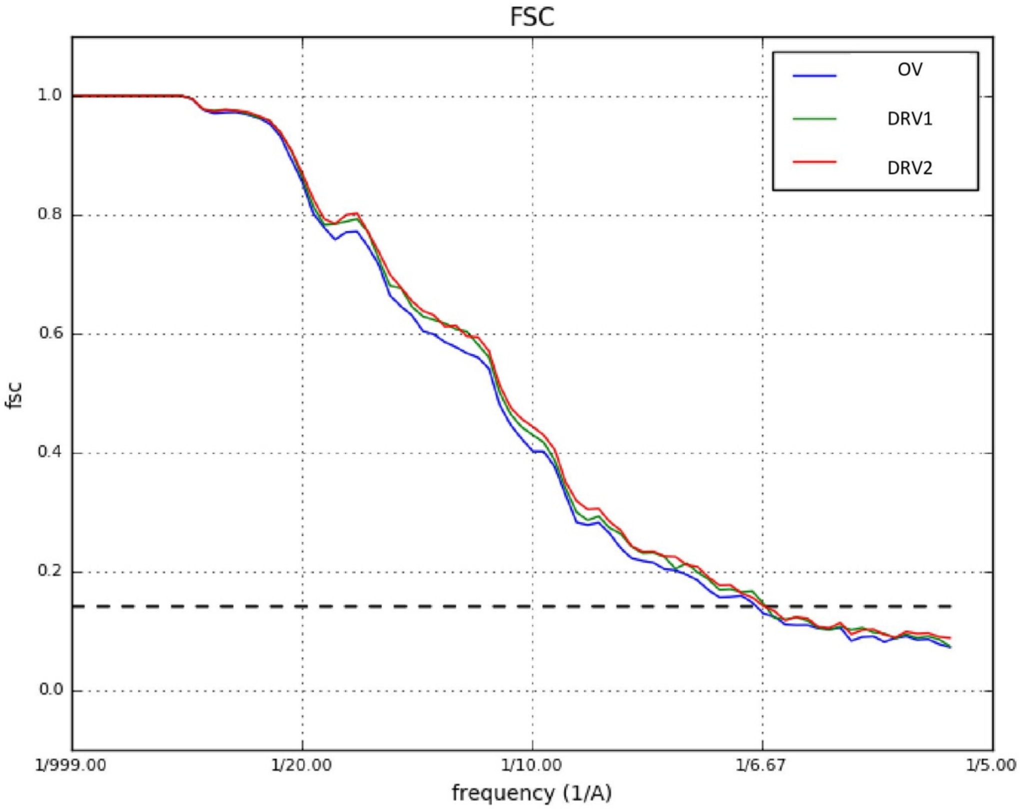

Figure 4.20: Comparison between Original initial volume (OV) vs Directional RANSAC volumes (DRV) for the Ribosome II dataset

On the other hand, Figure 4.20 shows that the FSC curve obtained when using the Directional RANSAC volumes as initial volumes significantly improves the FSC curve obtained by original volume.

Directional RANSAC was successfully able to reconstruct multiple conformations present in the heterogeneous dataset.

Discussion

Directional Pruning

We propose a novel method, Directional Pruning which aims to obtain homogenous particles images from datasets affected by massive heterogeneity. Initially, three different types of tests, namely: 3D auto-refine test, random subset test, and particle sorting test were performed on two different datasets affected by massive heterogeneous and obtained from two different sources (experimental/unpublished and published datasets).

These tests were performed to check the efficiency of the Directional Pruning method and validate our proposed hypothesis. From the results obtained, in some cases, the Directional Pruning method was shows improvement while for other cases, the results were neutral.

To clarify these results, another set of experiments called Control set-up were designed with a new hypothesis. The control set-up results were able to validate the second proposed hypothesis: “in the presence of a good reference volume Maximum-Likelihood refinement approaches disregards the presence of the noise or artefacts particles in the dataset”.

In conclusion, the Directional Pruning method was found to be ineffective when two closely associated conditions are met:

(i) The ML approach is used for 3D refinement

(ii) A good reference volume is provided

Directional RANSAC

We propose another novel method, Directional RANSAC, which aims to obtain all available structure conformations present in heterogeneous datasets, including minority classes. Two published datasets were used to test the method’s hypothesis.

The Ribosome II dataset (published) was used particularly because the several structural conformers were identified. This point enables us to check whether directional RANSAC would be able to obtain different structure conformers.

The Spliceosome dataset successfully generates 15 different structures, while its published paper was able to reconstruct only two different complete structures (EMDB ID: 3683 & 3688) [57].

According to the published data, the Ribosome II dataset contains five main (A-E), 3D classes. Out of which three (C-E) 3D classes were sub-classified into 12 different 3D classes. One of these 3D classes (Class D) was selected as input for Directional RANSAC. The published paper reported that Class D has five different structural conformers (EMDB ID: 8445-8449). Directional RANSAC was successfully able to reconstruct five different classes [58].

Two different tests, namely 3D classification and 3D auto-refinement tests were performed to check the efficiency of Directional RANSAC. The obtained results show that Directional RANSAC was able to reconstruct the complete structure of conformers present in these highly heterogenous datasets. Whereas 3D classification, the existing approach to obtain different structural conformers fails to provide more than one complete structure of the datasets. The 3D auto-refine test was performed using the generated RANSAC volumes as input volumes for the refinement process. In the case of the Spliceosome, the FSC curve of Directional RANSAC was the same as the one obtained when using original initial volume. In the case of the Ribosome II datasets, the FSC curve of the Directional RANSAC volume significantly improves over the one obtained by original volume.

Hence, Directional RANSAC was able to successfully reconstruct different structural volumes that can be used for refining the orientations of the single particles.

Concluding remarks

Both proposed methods, Directional Pruning, and Directional RANSAC leaves many unanswered questions. In the case of the Directional Pruning, will the method be productive when using other 3D refinement approaches like projection matching and Cryo-SPARC? In case of the Directional RANSAC, will the method provide the same results when purely experimental data is used? Were the generated volumes by Directional RANSAC similar to already published volumes? What are the limitations of directional RANSAC approach?

All of the mentioned questions can be achieved by carrying out different experiments. Future work should focus on answering the mentioned question, Wide range of heterogenous datasets should be used to attest further the efficiency of both the developed methods.

Bibliography

1. Marth JD. A unified vision of the building blocks of life. *Nat Cell Biol*. 2008;10(9):1015-1016.
2. Cooper GM. The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates; 2000. The Molecular Composition of Cells.
3. Breaker RR, Joyce GF. The expanding view of RNA and DNA function. *Chem Biol*. 2014;21(9):1059-1065.
4. Hvidsten TR, Laegreid A, Kryshtafovych A, Andersson G, Fidelis K, Komorowski J. A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PLoS One*. 2009;4(7):e6266. Published 2009 Jul 15.
5. Shi D, Nannenga BL, Iadanza MG, Gonen T. Three-dimensional electron crystallography of protein microcrystals. *Elife*. 2013;2:e01345. Published 2013 Nov 19. doi:10.7554/eLife.01345
6. Smyth MS, Martin JH. x ray crystallography. *Mol Pathol*. 2000;53(1):8-14.
7. Berman HM, Coimbatore Narayanan B, Di Costanzo L, et al. Trendspotting in the Protein Data Bank. *FEBS Lett*. 2013;587(8):1036-1045.
8. Zheng H, Handing KB, Zimmerman MD, Shabalin IG, Almo SC, Minor W. X-ray crystallography over the past decade for novel drug discovery - where are we heading next?. *Expert Opin Drug Discov*. 2015;10(9):975-989.
9. Mlynárik V. Introduction to nuclear magnetic resonance. *Anal Biochem*. 2017;529:4-9.
10. Müller SA, Aebi U, Engel A. What transmission electron microscopes can visualize now and in the future. *J Struct Biol*. 2008;163(3):235-245.
11. Tizro P, Choi C, Khanlou N. Sample Preparation for Transmission Electron Microscopy. *Methods Mol Biol*. 2019;1897:417-424.
12. Asadi J, Ferguson S, Raja H, et al. Enhanced imaging of lipid rich nanoparticles embedded in methylcellulose films for transmission electron microscopy using mixtures of heavy metals. *Micron*. 2017;99:40-48.
13. Winey M, Meehl JB, O'Toole ET, Giddings TH Jr. Conventional transmission electron microscopy. *Mol Biol Cell*. 2014;25(3):319-323.
14. Cheng Y. Single-particle cryo-EM-How did it get here and where will it go. *Science*. 2018;361(6405):876-880.

15. Carragher B, Cheng Y, Frost A, et al. Current outcomes when optimizing 'standard' sample preparation for single-particle cryo-EM. *J Microsc*. 2019;276(1):39-45.

16. Wan W, Briggs JA. Cryo-Electron Tomography and Subtomogram Averaging. *Methods Enzymol*. 2016;579:329-367.

17. Galaz-Montoya JG, Ludtke SJ. The advent of structural biology *in situ* by single particle cryo-electron tomography. *Biophys Rep*. 2017;3(1):17-35.

18. Sigworth FJ. Principles of cryo-EM single-particle image processing. *Microscopy (Oxf)*. 2016;65(1):57-67.

19. Lawson CL, Patwardhan A, Baker ML, et al. EMDataBank unified data resource for 3DEM. *Nucleic Acids Res*. 2016;44(D1):D396-D403.

20. Wang, Hongwei. "Cryo-electron microscopy for structural biology: current status and future perspectives." *Science China Life Sciences* 58 (2015): 750-756.

21. Bai XC, Fernandez IS, McMullan G, Scheres SH. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *Elife*. 2013;2:e00461. Published 2013 Feb 19.

22. Scheres SH. Beam-induced motion correction for sub-megadalton cryo-EM particles. *Elife*. 2014;3:e03665. Published 2014 Aug 13.

23. Brilot AF, Chen JZ, Cheng A, et al. Beam-induced motion of vitrified specimen on holey carbon film. *J Struct Biol*. 2012;177(3):630-637.

24. Zheng SQ, Palovcak E, Armache JP, Verba KA, Cheng Y, Agard DA. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods*. 2017;14(4):331-332.

25. Takizawa Y, Binshtein E, Erwin AL, Pyburn TM, Mittendorf KF, Ohi MD. While the revolution will not be crystallized, biochemistry reigns supreme. *Protein Sci*. 2017;26(1):69-81.

26. Bai XC, McMullan G, Scheres SH. How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci*. 2015;40(1):49-57.

27. Zhang K. Gctf: Real-time CTF determination and correction. *J Struct Biol*. 2016;193(1):1-12.

28. Scheres SH, Núñez-Ramírez R, Sorzano CO, Carazo JM, Marabini R. Image processing for electron microscopy single-particle analysis using XMIPP. *Nat Protoc*. 2008;3(6):977-990.
29. Zivanov J, Nakane T, Forsberg BO, et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife*. 2018;7:e42166. Published 2018 Nov 9.
30. Rohou A, Grigorieff N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J Struct Biol*. 2015;192(2):216-221.
31. Sanchez-Garcia R, Segura J, Maluenda D, Sorzano COS, Carazo JM. MicrographCleaner: A python package for cryo-EM micrograph cleaning using deep learning. *J Struct Biol*. 2020;210(3):107498.
32. Zhu Y, Carragher B, Glaeser RM, et al. Automatic particle selection: results of a comparative study. *J Struct Biol*. 2004;145(1-2):3-14.
33. Sanchez-Garcia, R., Segura, J., Maluenda, D., Carazo, J. M. & Sorzano, C. O. S. (2018). IUCrJ 5, 854-865.
34. Vargas J, Abrishami V, Marabini R, et al. Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. *J Struct Biol*. 2013;183(3):342-353.
35. Bell JM, Chen M, Baldwin PR, Ludtke SJ. High resolution single particle refinement in EMAN2.1. *Methods*. 2016;100:25-34.
36. Scheres SH. Semi-automated selection of cryo-EM particles in RELION-1.3. *J Struct Biol*. 2015;189(2):114-122.
37. Rosa-Trevín, Jose & Oton, Joaquin & Marabini, R & Zaldívar, Airen & Vargas, Javier & Carazo, J.M. & Sorzano, Carlos. (2013). Xmipp 3.0: An improved software suite for image processing in Electron Microscopy. *Journal of structural biology*. 184. 10.1016/j.jsb.2013.09.015.
38. Sorzano CO, Bilbao-Castro JR, Shkolnisky Y, et al. A clustering approach to multireference alignment of single-particle projections in electron microscopy. *J Struct Biol*. 2010;171(2):197-206.
39. Scheres SH. Classification of structural heterogeneity by maximum-likelihood methods. *Methods Enzymol*. 2010;482:295-320.

40. Scheres SH. Processing of Structurally Heterogeneous Cryo-EM Data in RELION. *Methods Enzymol*. 2016;579:125-157
41. Zhou, Ye et al. "Unsupervised Particle Sorting for High-Resolution Single-Particle Cryo-EM." *Inverse Problems* 36.4 (2020): 044002. Crossref. Web.
42. Gomez-Blanco J, Kaur S, Ortega J, Vargas J. A robust approach to ab initio cryo-electron microscopy initial volume determination. *J Struct Biol*. 2019;208(3):107397.
43. Sorzano CO, Vargas J, de la Rosa-Trevín JM, et al. A statistical approach to the initial volume problem in Single Particle Analysis by Electron Microscopy. *J Struct Biol*. 2015;189(3):213-219.
44. Vargas J, Álvarez-Cabrera AL, Marabini R, Carazo JM, Sorzano CO. Efficient initial volume determination from electron microscopy images of single particles. *Bioinformatics*. 2014;30(20):2891-2898.
45. Vargas, J., Melero, R., Gómez-Blanco, J. *et al.* Quantitative analysis of 3D alignment quality: its impact on soft-validation, particle pruning and homogeneity analysis. *Sci Rep* **7,** 6307 (2017).
46. Estrozi LF, Navaza J. Fast projection matching for cryo-electron microscopy image reconstruction. *J Struct Biol*. 2008;162(2):324-334.
47. van Heel, M.; Schatz, M. (2005). "Fourier shell correlation threshold criteria". *Journal of Structural Biology*. **151** (3): 250–262.
48. Böttcher, B.; Wynne, S.A.; Crowther, R.A. (1997). "Determination of the fold of the core protein of hepatitis B virus by electron microscopy". Nature. **386** (6620): 88–91.
49. Frank, J. (2006). Three-Dimensional Electron Microscopy of Macromolecular Assemblies. New York*: Oxford University Press*. ISBN 0-19-518218-9.
50. Burnley T, Palmer CM, Winn M. Recent developments in the CCP-EM software suite. *Acta Crystallogr D Struct Biol*. 2017;73(Pt 6):469-477.
51. Serna M. Hands on Methods for High Resolution Cryo-Electron Microscopy Structures of Heterogeneous Macromolecular Complexes. *Front Mol Biosci*. 2019;6:33. Published 2019 May 15. doi:10.3389/fmolb.2019.00033
52. Casañal A, Shakeel S, Passmore LA. Interpretation of medium resolution cryoEM maps of multi-protein complexes. *Curr Opin Struct Biol*. 2019;58:166-174.

53. Tan YZ, Baldwin PR, Davis JH, et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat Methods*. 2017;14(8):793-796.

54. Noble AJ, Wei H, Dandey VP, et al. Reducing effects of particle adsorption to the air-water interface in cryo-EM. *Nat Methods*. 2018;15(10):793-795.

55. Drulyte I, Johnson RM, Hesketh EL, et al. Approaches to altering particle distributions in cryo-electron microscopy sample preparation. *Acta Crystallogr D Struct Biol*. 2018;74(Pt 6):560-571.

56. Source code: https://github.com/mcgill-femr/scipion-em-cryomethods

57. Plaschka C, Lin PC, Nagai K. Structure of a pre-catalytic spliceosome. *Nature*. 2017;546(7660):617-621

58. Davis JH, Tan YZ, Carragher B, Potter CS, Lyumkis D, Williamson JR. Modular Assembly of the Bacterial Large Ribosomal Subunit. *Cell*. 2016;167(6):1610-1622.e15.

59. Li, Youguo & Wu, Haiyan. (2012). A Clustering Method Based on K-Means Algorithm. Physics Procedia. 25. 1104-1109.

60. L. Sun, R. Liu, J. Xu, S. Zhang and Y. Tian, "An Affinity Propagation Clustering Method Using Hybrid Kernel Function With LLE," in *IEEE Access*, vol. 6, pp. 68892-68909, 2018.